SM Journal of Infectious Diseases

Original Article

© Urazov S, et al. 2022

Classifier for Prediction of the Level of Disease COVID-19

Stanislav Urazov[1]*, Alina Cherkas[1,2], Alexey Boykov[2], Sergey Shcherbak[1,3], Anna Anisenkova[1], Sergey Mosenko[1], Oleg Glotov[1,4], Sergey Azarenko[1], Marina Bolsunovskaya[2], Oleg Popov[1] and Natalya Klenkova[1]

[1]City Hospital 40 of Saint Petersburg, Russia

[2]Laboratory of Industrial Systems for Streaming Data Processing of the SPbPU National Technology Initia-tive Center for Advanced Manufacturing Technologies, St. Petersburg, Russia

[3]Federal State Budgetary Educational Institution of Higher Professional Education, Saint-Petersburg State University, St. Petersburg, Russia

[4]Federal State-Financed Institution Pediatric Research and Clinical Center for Infectious Diseases under the Federal Medical Biological Agency, St. Petersburg, Russia

Abstract

The clinical spectrum of COVID-19 ranges from asymptomatic disease to pneumonia and life-threatening complications, including acute respiratory distress syndrome, organ failure and death [1-3]. In this regard, the development of models that allow medical workers to quickly assess the likelihood of an unfavorable development of COVID-19 seems to be an extremely urgent task. The aim of this study is to develop a model for predicting the severity of the course of COVID-19. For training and validation of 19 machine learning models, 117 clinical and laboratory parameters were used for 10487 patients with coronavirus infection who were treated at City Hospital No 40 of St. Petersburg, Russia from 01.09.2020 to 15.10.2021. As a result, 2 best models were obtained, including 21 and 10 features with AUC = 0.91 ± 0.01 and 0.86 ± 0.01, respectively. This paper provides an extensive overview of the available models for predicting the severity of COVID-19 disease and proposes 2 developed models. A mobile application has been created for the convenience of accumulating new data and using the model.

Keywords: COVID-19; Classifier; Risk Meter; Coronavirus Infection

Introduction

In medicine, decisions are made based on the perceived likelihood of a certain disease or condition (diagnostic setting) or a certain event that will occur in the future (predictive mindset) [4-6]. The overwhelming majority of models developed for COVID-19 are predictive. In a predictive model, several predictors (covariates, prognostic factors) are combined to estimate the likelihood of a particular outcome (event) at a particular period in the future [7-9]. This period can range from hours to weeks, months, or years. The main outcomes assessed for predicting the development of COVID-19 (endpoints) are admitted to the ICU/the need for non-invasive or mechanical ventilation/death in the 14-30-day period (time scales vary) [10-12]. Analysis of literature data over the past two years for queries: "COVID-19 (SARS-CoV-2, corona-virus)" in combination with "risk prediction model", "predictive model", "predictive index or rules (prognostic (or prediction) index or rule) ", "risk score ", "prognosis" identified over 70 studies, the ultimate goal of which was to create predictive models and build on them risk scales for patients with a confirmed diagnosis of coronavirus infection. Most of them were performed for a cohort of Chinese patients.

Submitted: 06 December, 2022 | **Accepted:** 13 December, 2022 | **Published:** 16 December, 2022

***Corresponding author(s):** Stanislav Urazov, City Hospital 40 of Saint Petersburg, Russia, Tel: +790-627-664-53; ORCiD: 0000-0002-5441-2911

Citation: Urazov S, Cherkas A, Boykov A, Shcherbak S, Anisenkova A, et al. (2022) Classifier for Prediction of the Level of Disease COVID-19. SM J Infect Dis 5: 12.

Predictive Assessment Models for Covid-19 Patients

An important aspect, which ultimately determines the scope of application of a particular model in practice, is the criterion for choosing predictors to be included in the model. Thus, the aim of a recent study in Spain (282 participants included in the shortlist) was to develop a simple predictive model focused on assessing the early symptoms of patients with COVID-19 [13]. The authors note the importance of such a model not only for the hospital but also for outpatients.

The predictors assessed were demographic characteristics (gender, age), existing comorbidities, and early symptoms (first 5 days). It is emphasized that the assessment of early symptoms of COVID-19 as possible predictors is rare in the literature. In the initial analysis, gender, age, several comorbidities (renal, respiratory, heart disease, diabetes, and hypertension), and two symptoms (shortness of breath and confusion) were statistically significantly associated with an increased risk of ICU hospitalization/death, while other symptoms (cluster symptom - rhinorrhea, myalgia, anosmia, dysgeusia) correlated with reduced risk. However, after multivariate adjustment, only age, confusion, dyspnea, and myalgia remained significantly associated with risk prediction. Based on the results obtained, the authors developed a simple predictive rule called CD65-M (abbreviation of the phrase "confusion, shortness of breath and age > 65 years", Dyspnoea, age > 65, Myalgias), having one positive score for the first three indicators, and negative for myalgia. In a more extended version in the predictive rule called CD65RD-WMA (granularity by age, gender) the score varies from 3 to 8. The score gives a negative score for each of the following factors: female sex, myalgia, and combined ageusia/anosmia/rhinorrhea. The authors note the unexpected appear acne in the risk scale of such an indicator as myalgia, which turned out to be independently associated with a decrease in the risk of critical outcomes after multivariate adjustments. There are reports on the relationship of myalgia with a favourable clinical course in other studies [14,15].

In the study mentioned above, Allenbach Y, et al. [14], aimed at identifying early prognostic factors for hospitalized patients, in addition to demographic data, comorbidities, and early symptoms, drugs taken before hospitalization, and the results of CT examinations were also taken into account. Clinical data (including respiratory) and laboratory parameters were recorded on the first day of admission to the hospital. Predictive factors for transfer to the Intensive Care Unit (ICU) or death on day 14 were assessed using multivariate logistic regression models. Based on the results obtained, age over 60 years, WHO scale score, CRP level (10-75, 75-150 or > 150 mg/L), and lymphocyte count below 800/mm3 were included in the scoring system. A score equal to or greater than 6 at baseline had a predicted more than 60% chance of a patient being ad-mitted to an intensive care unit or dying by day 14. The main limitation of this study, conducted in France, is the small sample size - 152 patients for model development and 132 patients for external validation.

A similar task, the development and testing of a model for assessing the risk of developing a critical form of COVID-19 upon admission to the hospital was set by the authors from China [16]. Epidemiological, clinical, laboratory, and imaging variables determined at admission to the hospital were examined (72 parameters in total). Data from 1590 patients were used to build a model using the Least Absolute Shrinkage and Selection Operator (LASSO). Logistic regression methods (Statistical software package R "glmnet" (R Foundation)) were used to build a predictive risk assessment model for COVID-GRAM. The accuracy of the estimate was measured by the Area under the Performance Curve (AUC). The model was validated on an external sample of 710 people. Critical COVID was defined as the cumulative rate of admission to an intensive care unit, invasive ventilation, or death. As a result of the statistical analysis, 10 variables were identified that were independent statistically significant predictors of the critical development of COVID-19. They were radiological abnormalities (yes *versus* no), age, hemoptysis (yes *versus* no), shortness of breath (yes or no), unconsciousness (yes *versus* no), number of concomitant diseases, history of cancer (yes and no), and also some laboratory parameters, the ratio of Neutrophils to Lymphocytes (NLR), the level of Lactate Dehydrogenase (U/L), Direct Bilirubin (DBIL) (μmol/L). The COVID-GRAM-based online calculator was designed to allow clinicians to enter values for 10 variables needed to assess risk, automatically calculating the likelihood (with a 95% confidence interval) that a hospitalized COVID-19 patient will develop critical illness. The accuracy of the COVID risk assessment in the validation cohort was similar to that in the derivation cohort with an AUC in the validation cohort of 0.88 (95% CI, 0.84-0.93).

An interesting study was carried out by the authors Gupta K, et al. [17]. As the researchers note, "despite the abundance of predictive models for assessing the risk of adverse outcomes of coronavirus infection, it remains unclear how well these proposed models work in practice and whether any of them are suitable for wide clinical use." In this regard, COVID-19 predictive models indexed in PubMed, Embase, Arxiv, medRxiv or bioRxiv until May 5, 2020 were considered. Also included were predictive models not specifically developed for COVID-19 patients, but which, according to the authors, could also be considered for use by clinicians in risk stratification for COVID-19 patients. For each identified candidate model, predictor variables, estimated outcome (including time horizons), modelling approaches, and final model parameters from the original publications were described. The final version summarized data on 22 predictive models (MEWS, REMS, qSOFA, CURB65, NEWS2, TACTIC, as well as a number of other studies, tentatively named after the first author - see source link).

Among the final predictors included in the models (risk scales) of these 22 studies, all possible combinations of the following can be distinguished: demographic characteristics (gender, age, ethnicity), the results of physical examinations at admission body temperature, respiratory rate, pulse rate, systolic and diastolic hypotension (systolic pressure), oxygen saturation, confusion. The severity of the patient's condition was assessed using various scales in accordance with the accepted recommendations in specific medical institutions the AVPU index, the Glasgow coma scale, the modified RALE scale, the NEWS2 scale. Among chronic diseases, possible predictors were diabetes (taking into account the age of onset), obesity, Chronic Obstructive Pulmonary Disease (COPD), chronic kidney disease, obstructive apnea, chronic respiratory diseases, cancer, immunosuppressive diseases and/or the drugs used. Among the laboratory parameters, the following indicators were included in various combinations - C-Reactive Protein (CRP), the content of neutrophils, platelets, lymphocytes, the ratio of neutrophils/lymphocytes, the level of albumin, the glomerular filtration rate, the level of Lactate Dehydrogenase (LDH), creatinine, D-dimer. More detailed information on specific studies is contained in the article.

A prospective Cahors study by Knight S, et al. [18] based on data from the International Severe Acute Respiratory and Emerging Infections Consortium (ISARIC) conducted in 260 hospitals in England, Scotland and Wales under the CCP-UK (Clinical Characterization Protocol UK). It is emphasized that the study fully complies with the requirements for the development of TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) predictive models, aimed at increasing the transparency of research reports using them, regardless of the methods used. 35,463 patients were included in the derivation dataset and 22,361 patients in the validation dataset. Comorbidities (chronic heart disease, chronic respiratory dis-ease (excluding asthma), chronic kidney disease (estimated glomerular filtration rate ≤ 30), mild and severe liver disease, dementia, chronic neurological conditions, connective tissue diseases, diabetes mellitus (diet, pills, or controlled by insulin), HIV or AIDS and malignant neoplasms) were determined by the modified Charlson comorbidity index. Clinically defined obesity has also been included as comorbidity due to its likely association with poor outcomes in COVID-19 patients. The clinical information used to calculate the prognostic scores was taken from the day of admission to the hospital. The primary outcome was hospital mortality. Potential predictor variables - 41 indicators - recorded at hospital admission represented patient demographic information, general clinical studies, and parameters consistently identified as clinically important in the COVID-19 cohorts. Using generalized additive modelling with multi-ply imputed datasets, eight key predictors of mortality were identified - age, sex, number of comorbidities, respiratory rate (breaths per minute), oxygen saturation, Glasgow coma score, blood urea nitrogen (mmol/L)) and C-reactive protein (mg/l). Next, continuous variables were transformed into factors with thresholds selected using smoothed component functions (on a linear predictor scale) from generalized additive modelling. The resulting scale is called the 4C Mortality Score. At the final stage, four risk groups were identified with corresponding mortality rates: low risk (0-3 points, mortality 1.2%), medium risk (4-8 points, 9.9%), high risk (9-14 points, 31.4%) and very high risk (≥ 15 points, 61.5%). Efficacy measures showed high sensitivity (99.7%) and negative predictive value (98.8%) for the low-risk group, covering 7.4% of the cohort and a corresponding mortality rate of 1.2%. Thus, the assessment of mortality on the 4C Mortality Score. Uses patient demographics, clinical observations and blood parameters that are usually available at the time of admission to the hospital, and allows you to accurately characterize the population of patients at high risk of hospital death. The authors emphasize that their 4C Mortality Score has advantages over other models.

In a study published a year later (in 2021) by Gupta RK, et al. [19], which is a continuation of the work of the authors Knight S, et al. [18] developed a multivariate logistic regression model for clinical deterioration in hospital, defined as a need for mechanical ventilation, transfer to ICU or death. The model was named 4C Deterioration and was also designed by TRIPOD standards. The work included data on 74,944 participants, also recruited within the ISARIC consortium under the CCP-UK protocol. Predictor scores were consistent with examinations on the first day of admission to hospital or the first day of clinical suspicion of COVID-19 for nosocomial cases. In the final version, the model already included 11 predictors, usually measured at admission to the hospital. They were age, gender, presence of nosocomial infection, Glasgow coma score, peripheral oxygen saturation on admission (SpO2), indoor air breathing or oxygen therapy (simultaneously with SpO2 measurement), respiratory rate, blood urea nitrogen (mmol/L), the concentration of C-reactive protein (mg/l), the number of lymphocytes (103 l/l), the presence of chest infiltrates on radiographs. The authors note that this model is intended to be used not only for hospitalization in case of out-of-hospital COVID-19 cases but also for evaluating COVID-19 of nosocomial origin.

It should be noted that the 4C Mortality Score and 4C Deterioration models have been successfully tested outside the UK (in other populations) and in more specific patient groups. For example, the 4C Mortality Score model is used to predict mortality in patients with COVID-19 with a history of cardiovascular disease [20]. In this case, the 4C Mortality Score model has been tested on the CLAVIS-COVID registry, developed in Japan to study the clinical features and outcomes of patients with COVID-19 with pre-existing or developing cardiovascular diseases or risk factors for coronary arteries. According to the CLAVIS-COVID criteria, cardiovascular disease was defined as heart failure, coronary heart disease, myocardial infarction, valvular heart disease, arrhythmia, stroke/transient ischemic attack, deep vein thrombosis, pulmonary embolism, peripheral arterial disease, aortic aneurysm, aortic dissection, cardiac arrest, heart transplant, left ventricular assistive device, electronic device implanted in the heart, pericarditis, myocarditis, congenital heart disease, and pulmonary hypertension. Risk factors included diabetes mellitus, hypertension, and dyslipidemia. The predictive power of the 4C Mortality Score was assessed for two different types of outcomes: in-hospital mortality and the combined outcome, defined as mechanical ventilation requirement and mortality. The discrimination of the 4C Mortality Score model showed a score of 0.84, an estimate of mortality an AUC of 0.78, and the calibration curve for both mortality and the combined score almost coincided with the ideal slope. According to Kuroda S, et al, these indicators indicate that the 4C Mortality Score model can be generalized to other clinically significant events (combined outcome of death and the need for IMV) and is useful for various ethnic groups and medical institutions.

The same group of authors - Matsumoto S, et al [21] set another task to analyze the effect of age on the profiles of cardiac biomarkers and on outcomes among hospitalized patients with COVID-19 with the presence of cardiovascular diseases (CVD) and/or factors their risk. CVDs were determined according to the CLAVIS-COVID criteria (see text above). This group included 693 patients. The control group included patients with COVID-19, but without CVD (825 people). The date of onset of COVID-19 was defined as the day the first symptoms appeared or, if patients were asymptomatic on admission, the day of the first positive SARS-CoV-2 PCR test result. The primary endpoint was hospital death. All patients were characterized by demo-graphic characteristics, results of physical and laboratory examinations upon admission, concomitant diseases, medications used at the hospital stage and were divided into 4 categories depending on their age (< 55, 55-64, 65-79 and ≥ 80 years). Older age (≥ 80 years) was shown to be closely associated with a worse hospital prognosis regardless of the patient's CVD history. Cardiac markers such as positive cTn troponin and elevated levels of B-type natriuretic peptide

BNP (or n-terminal pro b-type natriuretic peptide NT-proBNP) on admission were significantly associated with higher hospital mortality. In turn, cTn and BNP (or NT-proBNP) levels at the time of hospitalization showed a strong direct correlation with age in patients with COVID-19 and CVD. Specifically, approximately 80.0% of patients ≥ 80 years of age were cTn positive at admission, more than double the proportion of cTn positive patients < 55 years of age. At the same time, the authors emphasize that in elderly patients compared with younger patients, COVID-19 was asymptomatic with a higher frequency and/or patients had relatively less severe non cardiac biomarker profiles, which may lead to an underestimation of the severity of the condition of elderly patients, given the high level of their mortality.

Wu, et al. [22] set the goal of determining was the development of a multifactorial decision support system with different data sets to facilitate risk prediction and "tri-age" (quarantine in a home or mobile hospital, hospitalization or intensive care unit) of patients upon admission to the hospital. "Severe form of COVID-19" was defined if at least one of the following criteria was noted during hospitalization-respiratory failure requiring mechanical ventilation; shock state; admission to intensive care; organ failure or death. The model was built on the historical data. The indicators were determined on the first day of hospitalization. The patients were divided into two groups: 80% for machine learning of the model (239 patients) and 20% for internal validation (60 patients). Clinical characteristics included baseline information (five variables), comorbidities (11 variables), and symptoms (13 variables). All clinical characteristics were obtained at the first hospitalization of the patients. 42 laboratory results were also counted, including CBC, differential leukocyte count, d-dimer, C-reactive protein (CRP), cardiac markers, procalcitonin, liver function test, kidney function test, natriuretic B-type peptide and electrolyte test (for more details see the tables in the text of the article). It is interesting to note that the comparative analysis (Mann-Whitney test for continuous values and Fisher for discrete values) between the groups "Severe form of COVID-19" (see definition in the text above) and non-severe ones revealed highly statistically significant differences ($p < 0.001$) in the whole a number of evaluated variables. Among them are such characteristics as age, a feeling of tightness in the chest, a number of characteristics on CT (the semantics of CT was developed earlier), as well as the presence of diseases such as arterial hypertension, diabetes, cardiopathy, chronic obstructive pulmonary disease, cerebrovascular disease, and kidney disease.

In a multicentre retrospective cohort study, Zhang, et al. [23] also developed and validated a system for predicting the adverse outcome of COVID-19 (total number of participants - 828). The criterion "critically ill" was applied if the patients met at least one of the following criteria: shortness of breath with a respiratory rate of ≥ 30 breaths/min; oxygen saturation (at rest) ≤ 93%; PaO2/FiO2 ≤ 300 mm/Hg; x-ray showing more than 50% progression of the lesion within 24-48 hours; respiratory failure, shock, or other organ failure. Calculation based on guidance on sample size requirements for predictive models [12] showed that 6 variables could be included in this multivariate analysis. Variables were selected based on previous evidence base, clinical significance, correlations between predictors, and data availability. They turned out to be age, Neutrophil to Lymphocyte Ratio (NLR), Lactate Dehydrogenase (LDH), C-Reactive Protein (CRP) and Direct Bilirubin (DBIL) , respiratory rate, assessment of the severity of pneumonia CURB-65 (assessment CURB-65), rapid assessment of organ failure associated with sepsis qSOFA, reticular patterns and 15 laboratory parameters-all indicators reached statistical significance $p < 0.001$ - leukocyte count, abs., 109/l, lymphocytes, abs., 109/l, neutrophils, abs., 109/l, platelets, abs., 109/l, Neutrophil to Lymphocyte Ratio (NLR), albumin, g/L, total bilirubin, μmol/L, Direct Bilirubin (DBIL), μmol/L, urea nitrogen, mmol/L, d-dimer, mg/L, Prothrombin Time (PT), C-Reactive Protein (CRP, mg/L), the level of Lactate Dehydrogenase (LDH), U/l. In the training set, multivariate Cox regression showed that elderly age, the level of lactate dehydrogenase is more than 360 U/L, the ratio of neutrophils to lymphocytes is more than 8.0 and direct bilirubin. More than 5.0 μmol/L were independent predictors of 28-day mortality. These four independent predictors were used to build a predictive model, which was presented in the form of a nomogram scoring system. Nomogram scoring systems for predicting the likelihood of 14-day and 28-day survival in patients with COVID-19 showed good discrimination and calibration in two independent verification cohorts (external validity scores with a discriminatory C-index of 0.879 (95% CI, 0.856-0.900). the disadvantage of this study is its retrospective design.

The study by Berenguer J, et al. [24] was devoted to the analysis of death predictors among residents of Spain hospitalized with COVID-19. The work did not involve the creation of a risk calculator. However, it deserves attention within the framework of our review due to the large sample size - 4035 patients with COVID-19 and a wide range of assessed parameters. The study was of a retrospective nature. The primary endpoint was all-cause mortality. The starting point of reference is the date of admission to the hospital. Statistical analysis included univariate and multivariate Cox regression analysis. In order to obtain the most significant set of variables from a wide range of predictors, a block-by-block procedure was performed for direct distribution of predictors into five clusters: socio-demographic characteristics, comorbidities, grounds for hospitalization (symptoms), vital signs and laboratory parameters. For each block, multivariate regression analysis was performed using two criteria to achieve the best set of predictors: relevance to the clinical situation and statistical significance ($p < 0.10$). Statistical processing was performed using Stata 15.0 software (StataCorp, College Station, TX, USA).

In the work of Russian researchers Boytsov SA, et al. [25], the clinical picture and factors associated with adverse outcomes in hospitalized patients with COVID-19 were also studied. The work included 402 patients, whose demographic data, comorbid chronic diseases, clinical manifestations upon admission, and a number of laboratory parameters were taken into account. Disease severity at admission was assessed using the NEWS scale. Laboratory indicators of patients with COVID-19 were compared with the norm. Statistical processing was performed using Python v.3.8. Within the frame-work of univariate regression analysis, the age over 64 years with the level of statistical significance (*p* < 0.001) was associated with death during the hospitalization period; News score above 8 points; oxygen saturation < 92%, glucose level above 8.2 mmol/l, CRP above 133 mg/l, creatinine clearance less than 72 ml/min. According to multivariate regression analysis, the three most significant predictors of death from all causes during the period of hospitalization, according to multivariate regression analysis, are more than 5-fold increase in Aspartate Aminotransferase (AST) and/or Alanine Aminotransferase (ALT) levels in comparison with standard indicators (*p* < 0.001), changes in the lungs corresponding to the CT-4 pattern (*p* < 0.001), and MI/unstable angina during hospitalization (*p* = 0.023). COPD, decreased renal function (Cockcroft-Gault creatinine clearance < 60.0 ml/ min), type 2 diabetes, cancer and dementia also significantly increased the likelihood of death. It is noteworthy that in this study, the key laboratory parameters associated with the risk of death were the levels of transaminases AST and ALT.

The authors of Haimovich, et al. [26] set themselves the task of developing a predictive model for predicting early respiratory failure or death within 24 hours after hospitalization. Critical condition was defined as meeting one of the following criteria: oxygenation rate greater than or equal to 10 L/min, high flow oxygenation, non-invasive ventilation, invasive ventilation, or death. Predictive models were compared with the Elixhauser Comorbidity Index, Rapid Sepsis-Related Organ Failure Assessment (qSOFA), and the CURB-65 pneumonia severity score. The study was of a retrospective nature and was carried out on the basis of data on hospitalized 1,172 American patients with COVID-19.

The complete dataset for each patient included 713 variables that were determined within the first 4 hours after hospitalization. These included demographic data, bad habits, comorbidities, pre-hospitalization drugs, main complaints, vital signs, laboratory parameters, and radiographs. The patients were divided into 3 cohorts: for model creation, internal and external validation. The H-CUP statistical package (hcuppy package; version 0.0.7) was used to calculate the comorbidity indices for the Elixhauser, qSOFA and CURB-65 models. A variety of statistical procedures (see description in original source) have been applied to identify and rank potentially important predictive variables based on their occurrence in multiple selection methods. A "quick COVID-19 Severity Index" (qCSI) minimum score model and "COVID-19 Severity Index" (CSI) machine learning model were developed using the gradient boosting method. A logistic regression scoring system was used to quickly determine the COVID-19 severity index in the quick COVID-19 Severity Index model. The "COVID-19 Severity Index" model was developed using the statistical package XGBoost, and the "hyperparameters" were established using Bayesian optimization using the Parzen scoring tree structure. Clinical variables were divided into ranges of values according to clinical experience and logistic regression was used to derive the "weight" of a parameter for a rapid COVID-19 severity index scoring system. In the final version, the variables of the "Quick COVID-19 Severity Index" model included such indicators as respiratory rate, inhalation/min (≤ 22, 23-28; > 28), pulse oximetry (represents the lowest value recorded during the first 4 hours of observation of the patient), % > 92; 89-92); oxygen consumption, l/min. QCSI scores range from 0 to 12. For the CSI scale, in addition to the above variables, levels for the following indicators were included: aspartate transaminase, alanine transaminase, ferritin, procalcitonin, chloride, C-reactive protein, glucose, urea nitrogen, leukocyte count, and the patient's age. The SHAP (gradient-boosting Shapley additive explanation interaction values) method was used to analyze the effect of the range of values of individual variables on the model output. In particular, the authors noted interesting patterns in age - age showed an almost binary distribution of risk with an inflection point between 60 and 70 years, which suggests that younger patients had a higher risk of 24-hour critical ill-ness than older patients. In the independent validation cohort, the area under ROC for the qCSI and CSI models was 0.81 (0.73-0.89) and 0.76 (0.65-0.86), respectively, compared with the worst scores 0.61 (0.51-0.70) for Elixhauser, 0.59 (0.50-0.68) for qSOFA, and 0.50 (0.40-0.60) for and CURB-65. Thus, the "Quick COVID-19 Severity Index" model with three variables (respiratory rate, pulse oximetry and oxygen flow rate) and does not include any of the laboratory parameters surpassed all other models under consideration and was recommended by the authors as the main one.

The global interest in developing predictive models for COVID-19 is driven by the need to quickly and efficiently assess patients upon admission to hospital in order to facilitate adequate resource allocation and ensure appropriate treatment and follow-up of patients at increased risk of deterioration [27,28]. In addition, predictive models can be of added value in stratifying patients for new and/or expensive drugs. However, until now, there are no universal models recommended by the medical community for widespread use in all countries. This fact can be explained by a number of reasons. Patients of different races and nationalities may have differences in clinical and laboratory results. Hospital admission threshold and hospital admission management may vary from country to country. This is especially true during a pandemic with an acute shortage of hospital beds. It cannot be ruled out that patients of the older age category, despite the already proven fact of the negative influence of age on the clinical outcome of coronavirus infection, do not receive priority treatment in intensive care units. Treatment of patients in the pre-hospital period may affect clinical and laboratory results, however, none of the models we described included pre-hospital drugs as a predictor variable. Accounting for drugs at the hospitalization stage when building fore-casting models is technically extremely difficult. However, the fact that the difference in the applied therapies undoubtedly influences the assessed clinical outcome is clear. RNA viruses rapidly mutate with the emergence of more and more new strains, often more dangerous for humanity, which can affect the performance of models. The modification of the virus in the considered models was not taken into account. Most of the research on building forecasting models was of a retrospective nature. To date, there are studies that have detailed the differences in physical examination results, vital signs and laboratory parameters between COVID-19 survivors and non-survivors [29-31]. However, at the final stage of model building, a relatively small number of variables appear as predictor variables, the key role among which is played by vital signs (respiratory rate, pulse, saturation, etc.). And often, such simple models have indicators of discrimination and calibration that are comparable or superior to those of models built taking into account not only vital indicators, but also a large number of laboratory indicators, which undoubtedly testifies in favor of the imperfection of the latter. Therefore, there is a need to create and constantly update models for predicting the unfavorable development of COVID-19, in order to achieve their best statistical indicators, and, therefore, applicability in practice in a particular country, and especially in Russia with its multimillion population and not very favorable economic situation.

Materials and Methods

The development of the predictive model was carried out in five stages. All analysis was carried out using tools developed in Python v.3.9 and Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn. To assess the distribution of data by traits, at the first stage, descriptive statistics were carried out based on the study of clinical data and the results of laboratory and instrumental examination methods. To identify significant parameters for the prognostic model, at the second stage, the correlation analysis was transferred to data on 1272 patients. Were studied 93 clinical and 24 laboratory parameters at the time of hospitalization. Their correlation with three target parameters was assessed: desaturation, transfer to ICU, and death. At the third stage, quantitative characteristics were transferred to categorical ones based on the results of correlation analysis, literature review and expert opinion (Table 1).

At the fourth stage, 19 machine learning models were configured and applied to obtain an effective classifier, such as gradient boosting, logistic regression, decision trees (Figure 1), naive Bayes, k-nearest neighbors, etc. (Table 2). At this stage, information was used on 22 parameters of 9215 patients. The model was built on the basis of retro perspective data. The indicators of the predictors corresponded to the results of examinations on the first day of admission to the hospital. The patients were divided into 2 groups: 80% for machine learning of the model, 20% for internal validation. Due to the fact that initially we have a significantly larger number of records in one class than in another, the resampling method was applied to the training set. As a result of applying this method, we obtained 11266 balanced data in equal proportions. Since the parameters "Alanine aminotransferase" and "Aspartate aminotransferase" strongly correlated with each other (Figure 2), only "Aspartate aminotransferase" was retained. Also, the 11 most important features were selected (Figure 3) and the training of the models was repeated. The fifth stage included obtaining the best hyperparameters, as well as developing a mobile application.

Table 1: Normalizing data.

S.No	Feature	Categories
1	Sex	1 - m; 2 - f
2	Age	1 - up to 20, 2 - up to 30, etc.
3	Day of illness	1 - up to 7 days inclusive; 2- - 8 days or more
4	Body mass index (BMI)	1 - up to 18.5; 2 - 18.5-24.99; 3 - 25-29.99; 4 - 30-34.99; 5 - 35 - 39.99; 6 - 40 and higher
5	Dyspnea	0/1
6	Computed Tomography	0 - CT0, 1 - CT1, 2 - CT2, 3 - CT3, 4 - CT4
7	COVID-19 vaccination	0/1
8	Hypertension	0/1
9	Diabetes	0/1
10	Aspartate aminotransferase	1 - up to 40; 2 - 40-80; 3 - 80-120; 4 - 120-160; 5 - 160-200; 6 - more than 200
11	Activated partial thromboplastin time	1 - up to 35; 2 - 35-50; 3 - more than 50
12	D-dimer	1 - up to 0.5; 2 - 0.5-1; 3 - 1-1.5; 4 - 1.5-2; 5 - 2-2.5; 6 - 2.5 - 3; 7-more than 3
13	Interleukin-6	up to 33; 2 - 33-60; 3 - 60-93; 4 - 93-212; 5 - more than 212
14	CKD-EPI	1 - 90 and higher; 2 - 89-60; 3 - 59-30; 4 - 29-15, 5 - less than 15
15	Lactate dehydrogenase	1 - norm; 2 - 220-300; 3 - 300-400; 4 - 400-500; 5 - 500-600; 6 - higher
16	Lymphocytes abs	1 - higher than 3.18; 2 - 2-3.18; 3 - 1.5-2; 4 - 1.3-1.5; 5 - 1-1.3
17	PCR	0 - not found; 1 - higher than 25; 2 - lower than 25
18	C-reactive protein (quantitative)	1 - up to 30; 2 - 30-45; 3 - higher than 45
19	Thrombocytes	1 - above 150; 2 - 100-150; 3 - 50-100; 4 - lower than 50
20	Eosinophils	0 - norm; 1 - pathology (below 0.02)
21	Ferritin	1 - below 300; 2 - 300-600; 3 - 600-900; 4 - 900-1200; 5 - 1200-1500; 6 - more

Table 2: Comparison of models.

S. No	Model	Name	AUC (21 features)	AUC (11 features)
1	LDA	Linear Discriminant Analysis	0,794	0,819
2	QDA	Quadratic Discriminant Analysis	0,820	0,813
3	AdaBoost	AdaBoost Classifier	0,862	0,812
4	Bagging	Bagging Classifier	0,821	0,819
5	ETE	Extra Trees Classifier	0,819	0,809
6	GB	Gradient Boosting Classifier	0,869	0,819
7	RF	Random Forest Classifier	0,866	0,813
8	Ridge	Ridge Classifier	0,826	0,837
9	SGD	Stochastic Gradient Descent Classifier	0,871	0,816
10	BNB	Naive Bayes classifier (Bernoulli NB)	0,813	0,785
11	GNB	Gaussian Naive Bayes (Gaussian NB)	0,855	0,817
12	KNN	Classifier implementing the k-nearest neighbors vote	0,787	0,760
13	MLP	Multi-layer Perceptron classifier	0,863	0,843
14	LSVC	Linear Support Vector Classification	0,866	0,839
15	NuSVC	Nu-Support Vector Classification	0,742	0,667
16	SVC	C-Support Vector Classification.	0,803	0,825
17	LR	Logistic Regression	0,869	0,840
18	DTC	A decision tree classifier	0,777	0,747
19	ETC	An extremely randomized tree classifier	0,768	0,751

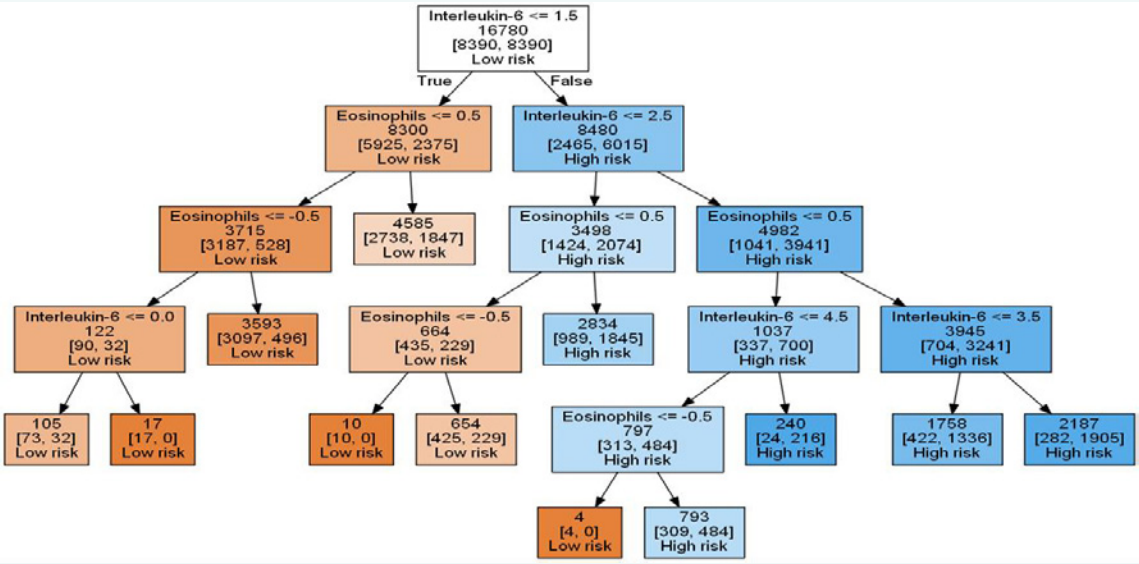

Figure 1 Part of the decision tree.

This study was carried out in the City Hospital No. 40 of St. Petersburg, Russia from 01.09.2020 to 15.10.2021. The study included only hospitalized patients aged 18 and over. The average observation time during the clinical course was 10 days. The study was approved by the City Hospital 40 of Saint Petersburg Expert Council on Ethics and was conducted in accordance with general principles of observational re-search.

Results

During the first stage, descriptive analysis was carried out on two datasets. The first dataset consists of 1272 people and includes: 518 people (40.7%) with de-saturation, 95 (7.5%) people who were transferred to the ICU and 34 (2.7%) people who died. The mean age of the cases was 56.4 years (standard deviation: 13.5), the mean body mass index was 28.8 (standard

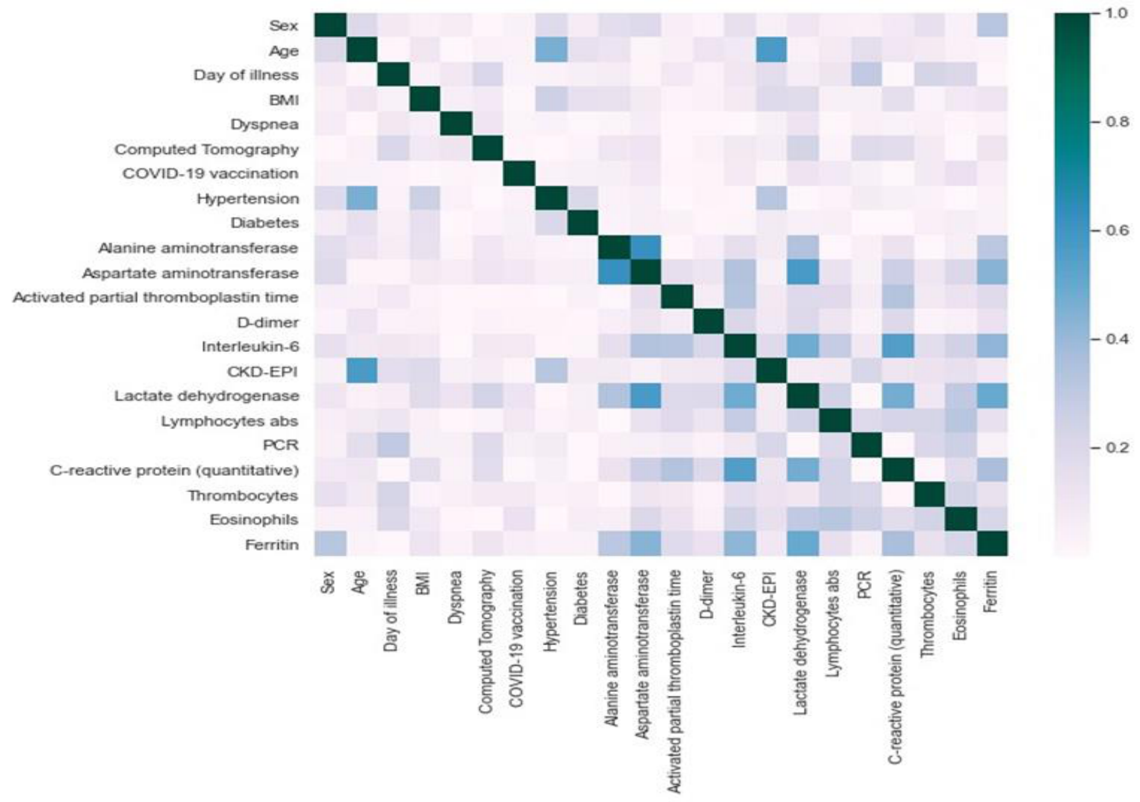

Figure 2 Correlation heat map.

deviation: 4.8) and 602 (47%) were men. The most common complaints were: fever (93.2%), weakness (78.3%), cough (73.7%), headache (45.7%) and muscle pain (34.7%). The most common comorbidities were: hypertension (46.7%), diabetes (12.6%) and coronary heart disease (10.0%).

The second dataset consists of 9215 people and includes: 825 (9.0%) people who were transferred to the ICU and 343 (3.7%) people who died. The mean age of the cases was 52.2 years (standard deviation: 16.3), the mean body mass index was 27.8 (standard deviation: 5.4) and 3989 (43.3%) were men. As a result of the performed correlation analysis, we received three ratings of signs that affect the target variables: "desaturation", "transfer to ICU" and "death". Having studied the literature, we noted that the signs that occupy a high place in the ratings are found in scientific articles. Several features with insignificant correlation coefficients were retained for the model based on the opinions of medical experts. Ultimately, 22 features were selected. After that, in order to normalize and increase the accuracy of the model, the numerical characteristics of these 22 features were divided into categories according to the principle: 1 - normal, 2 - bad, 3 - worse, etc. (Table 1) based on clinical guidelines. Then 19 models were tuned and trained for 21 and 11 features (simplified version) (Figure 4,5). The attribute "Transfer to ICU" was taken as a target variable as an important indicator of the severity of the disease. The experimental results are shown in (Table 2). An illustration of a part of the Decision Tree is shown in (Figure 1).

The simplified version of the model was created to be able to use the simplest, most affordable and cost-effective traits, such as age, Body Mass Index (BMI), day of illness, Chronic Kidney Disease-Epidemiology collaboration (CKD-EPI), lactate dehydrogenase, lymphocytes abs, C-reactive protein (quantitative), eosinophils, activated partial thromboplastin time, COVID-19 vaccination, and aspartate aminotransferase. The Stochastic Gradient Descent Classifier (SGB) method showed the highest results for the model trained on 21 features and was chosen as the main one (Table 2, Figure 4). For a model trained on 11 features, the Multi-Layer Perceptron classifier (MLP) method showed the highest accuracy (Table 2, Figure 5). The GridSearchCV method helped us find the optimal hyperparameters for each model. Using the Stratified K Fold method, we cross-validated and obtained the mean AUC (Figure 6,7).

Discussion

In this study, we developed two risk meter models to predict the development of critical illness in hospitalized patients infected with COVID-19. The main model includes 21 features and has an AUC score of more than 0.9 in the design and validation cohorts. However, we came to the need to develop a simpler model trained on 11 features to apply it to the main model, thereby saving resources. The classifier can be used by clinicians to assess the risk of developing a critical illness in an individual hospitalized patient. The metrics needed to calculate the risk of developing critical illness are usually available upon admission

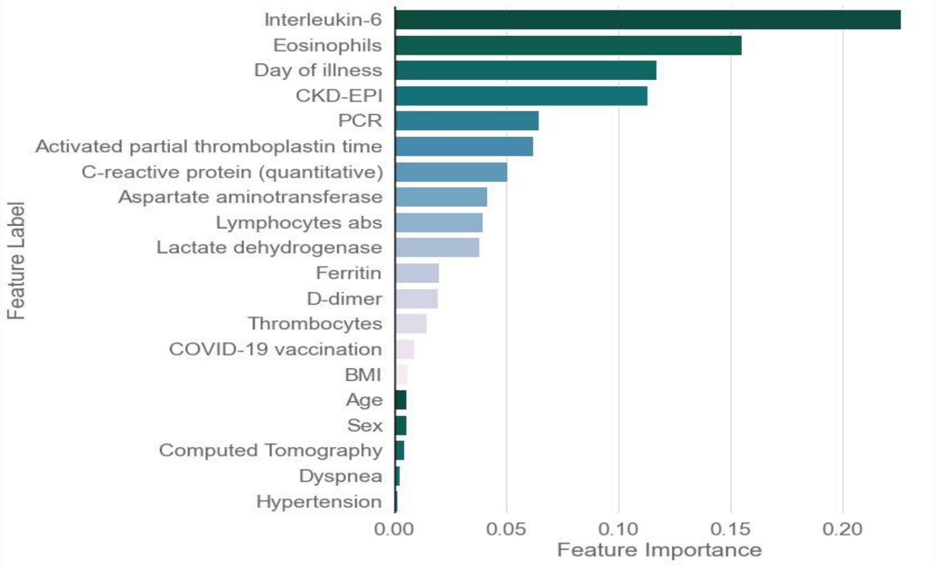

Figure 3 Feature importance: 20 most important features for model.

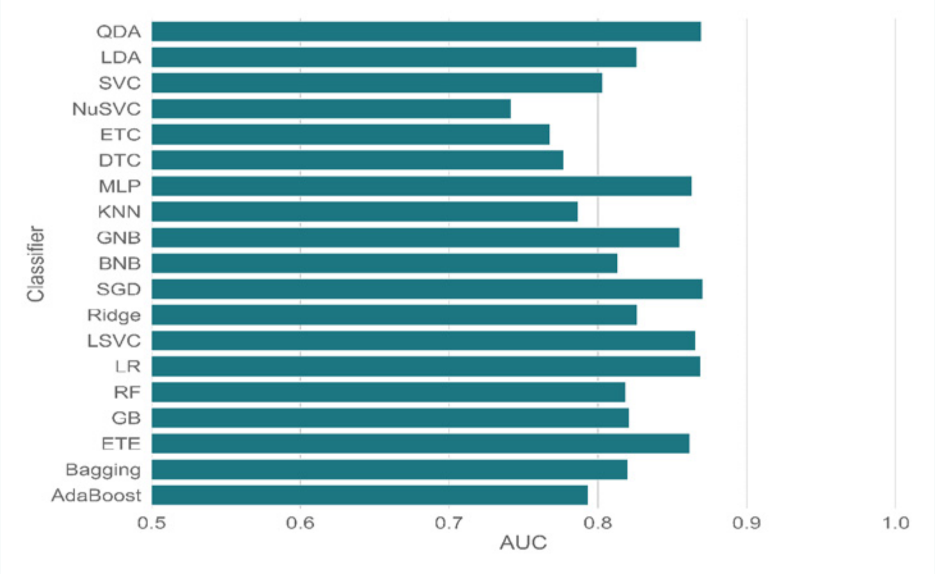

Figure 4 Comparison of models trained on 21 features.

to the hospital. If a patient's perceived risk of critical illness is low, the physician may choose to monitor, whereas high-risk assessments may support aggressive treatment or admission to an intensive care unit.

We deliberately did not divide risk into low, medium and high risk groups, as we believe that doctors are better informed, calculating a risk assessment for each individual patient and making decisions based on local or regional conditions. For example, in areas with good access to clinical and supportive care, patient outcomes can be optimized by choosing to provide more aggressive care to moderate risk patients. In contrast, in areas with high case counts and/or limited resources, the solution may be to provide less aggressive care to moderate risk patients to maximize the availability of beds and ventilators in intensive care units.

The first data set corresponds to the period of the 2nd wave of the pandemic in the Russian Federation, the second data set corresponds to the periods of the 3rd and 4th waves, which divides the periods into features of the etiological factor of the genotypes of the virus. It is necessary to conduct additional observational studies of the differences in the clinical course of COVID-19 in different variants of SARS-CoV-2 strains.

The main goal of this study is to show the possibilities of

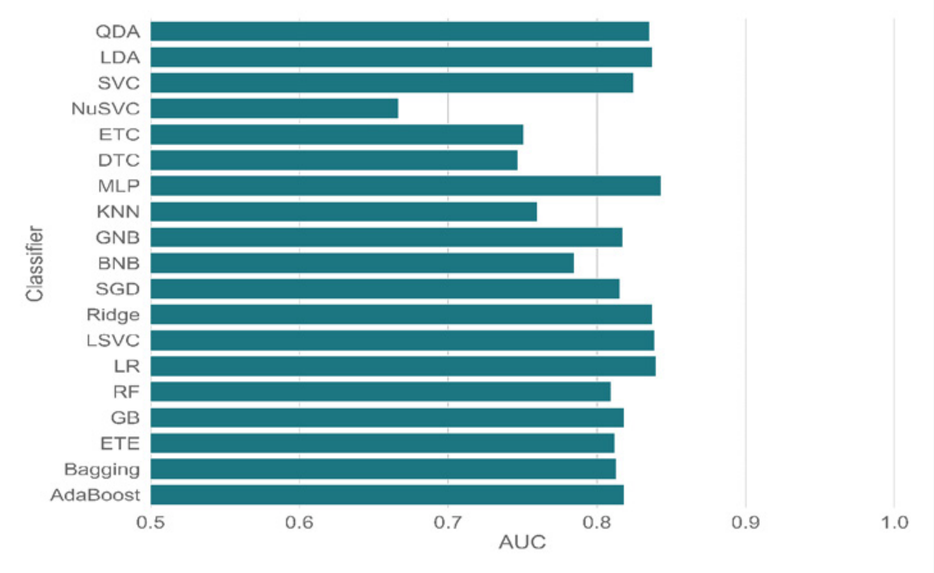

Figure 5 Comparison of models trained on 11 features.

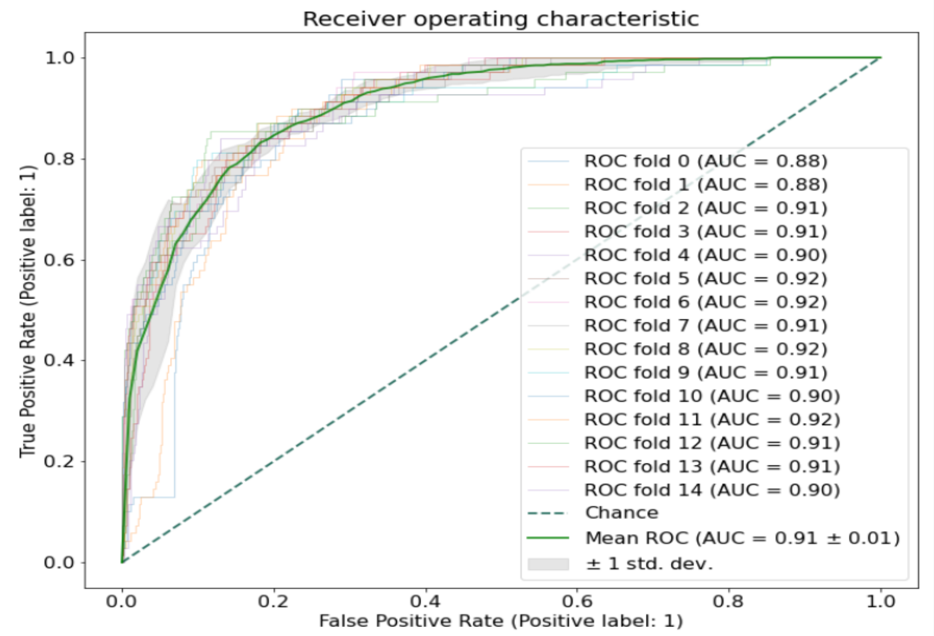

Figure 6 Receiver operating characteristic for model with 21 features.

developing a predictive model and draw attention to the need to collect more data and create a more efficient risk meter. The study was carried out on a cohort of patients with the one medical organization in St. Petersburg, and therefore there is a need to translate this experience of using the model in other institutions not only in St. Petersburg, but also in other regions of the Russian Federation and other countries. In pursuit of this goal, we simultaneously worked on the creation of a mobile application of a risk meter with possible widespread use among doctors working in red zones with patients with COVID-19. The accumulation of big data using the application will make it possible to draw more reliable conclusions and implement the use of the model in the practical health care of St. Petersburg and other regions of the Russian Federation. The clinical development of COVID-19 patients is initially unpredictable, but the use of predictive models such as the proposed “COVID-19 risk meter” are needed to support medical decision-making for COVID-19 patients data, it is possible not only to avoid a decrease in accuracy, but also to increase it.

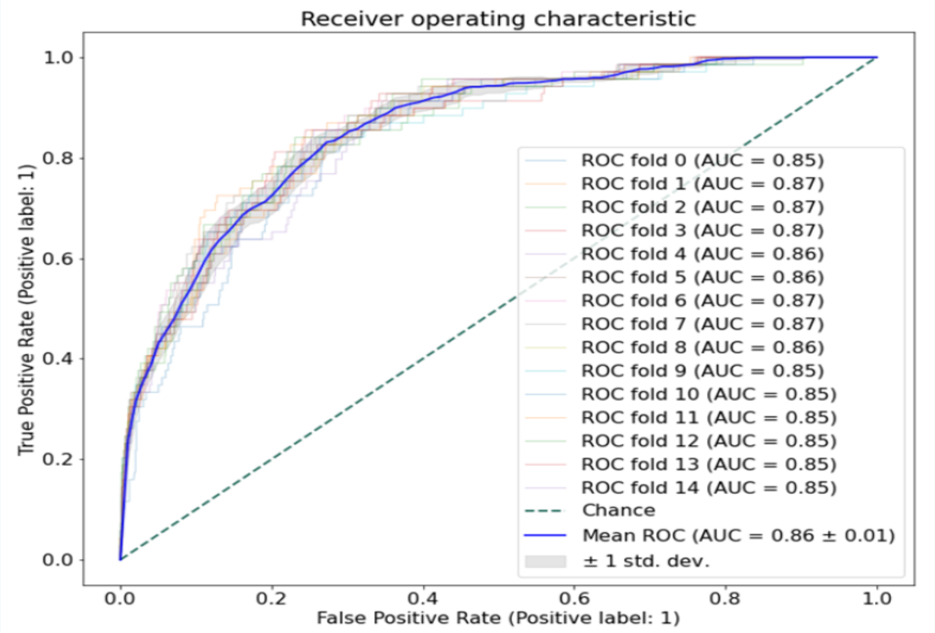

Figure 7 Receiver operating characteristic for model with 11 features.

Acknowledgments

The research is funded by the Ministry of Science and Higher Education of the Russian Federation as part of World-class Research Center program: Advanced Digital Technologies (contract No. 075-15-2020-934 dated 17.11.2020).

References

1. Shi Y, Wang G, Cai XP, Deng JW, Zheng L, Zhu HH, et al. An overview of COVID-19. J Zhejiang Univ Sci B. 2020; 21(5): 343-360. doi: 10.1631/jzus.B2000083. PMID: 32425000; PMCID: PMC7205601.

2. Eythorsson E, Helgason D, Ingvarsson RF, Bjornsson HK, Olafsdottir LB, Bjarnadottir V, et al. Clinical spectrum of coronavirus disease 2019 in Iceland: population based cohort study. BMJ. 2020; 371: m4529. doi: 10.1136/bmj.m4529. PMID: 33268329; PMCID: PMC7708618.

3. Zeng H, Ma Y, Zhou Z, Liu W, Huang P, Jiang M, et al. Spectrum and Clinical Characteristics of Symptomatic and Asymptomatic Coronavirus Disease 2019 (COVID-19) With and Without Pneumonia. Front Med (Lausanne). 2021; 8: 645651. doi: 10.3389/fmed.2021.645651. PMID: 33869253; PMCID: PMC8046922.

4. Martinez-Sanz J, Perez-Molina JA, Moreno S, Zamora J, Serrano-Villar S. Understanding clinical decision-making during the COVID-19 pandemic: A cross-sectional worldwide survey. EClinicalMedicine. 2020; 27: 100539. doi: 10.1016/j.eclinm.2020.100539. PMID: 32923995; PMCID: PMC7480231.

5. Metlay JP, Armstrong KA. Clinical Decision Making During the COVID-19 Pandemic. Ann Intern Med. 2021; 174(5): 691-693. doi: 10.7326/M20-8179. PMID: 33524281; PMCID: PMC7888345.

6. Anton N, Hornbeck T, Modlin S, Haque MM, Crites M, Yu D. Identifying factors that nurses consider in the decision-making process related to patient care during the COVID-19 pandemic. PLoS One. 2021; 16(7): e0254077. doi: 10.1371/journal.pone.0254077. PMID: 34214122; PMCID: PMC8253418.

7. Clarke BS, Clarke JL. In Predictive Statistics: Analysis and Inference beyond Models; Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 2018; 605-634.

8. Kelleher JD.; Namee BM, D'Arcy A. Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies. The MIT Press. 2015; ISBN 0-262-02944-8.

9. Max Kuhn, Kjell Johnson. Applied Predictive Modeling. Springer. 2013; ISBN 978-1-4614-6848-6.

10. Lorenzoni G, Sella N, Boscolo A, Azzolina D, Bartolotta P, Pasin L, et al. COVID-19 ICU Mortality Prediction: A Machine Learning Approach Using SuperLearner Algorithm. Journal of Anesthesia, Analgesia and Critical Care. 2021; 1: 3. doi:10.1186/s44158-021-00002-x.

11. Banoei MM, Dinparastisaleh R, Zadeh AV, Mirsaeidi M. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care. 2021; 25(1): 328. doi: 10.1186/s13054-021-03749-5. PMID: 34496940; PMCID: PMC8424411.

12. Icten Z, Chan C, Munsell M, Menzin J. ML4 Application of Machine Learning Models to Evaluate COVID-19 Related ICU Utilization in a US Population. Value Health. 2020; 23: S404. doi: 10.1016/j.jval.2020.08.044. PMCID: PMC7834221.

13. Vila-Corcoles A, Satue-Gracia E, Vila-Rovira A, de Diego-Cabanes C, Forcadell-Peris MJ, Ochoa-Gondar O. Development of a predictive prognostic rule for early assessment of COVID-19 patients in primary care settings. Aten Primaria. 2021; 53(9): 102118. doi: 10.1016/j.aprim.2021.102118. PMID: 34139400; PMCID: PMC8162822.

14. Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Development of a multivariate prediction model of intensive care unit transfer or death: A French prospective cohort study of hospitalized COVID-19 patients. PLoS One. 2020; 15(10): e0240711. doi: 10.1371/journal.pone.0240711. PMID: 33075088; PMCID: PMC7571674.

15. Sudre CH, Lee KA, Lochlainn MN, Varsavsky T, Murray B, Graham MS, et al. Symptom clusters in COVID-19: A potential clinical prediction

tool from the COVID Symptom Study app. Sci Adv. 2021; 7(12): eabd4177. doi: 10.1126/sciadv.abd4177. PMID: 33741586; PMCID: PMC7978420.

16. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern Med. 2020; 180(8): 1081-1089. doi: 10.1001/jamainternmed.2020.2033. PMID: 32396163; PMCID: PMC7218676.

17. Gupta RK, Marks M, Samuels THA, Luintel A, Rampling T, Chowdhury H, et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. Eur Respir J. 2020; 56(6): 2003498. doi: 10.1183/13993003.03498-2020. PMID: 32978307; PMCID: PMC7518075.

18. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ. 2020; 370: m3339. doi: 10.1136/bmj.m3339. PMID: 32907855; PMCID: PMC7116472.

19. Gupta RK, Harrison EM, Ho A, Docherty AB, Knight SR, van Smeden M, et al. Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: a prospective cohort study. Lancet Respir Med. 2021; 9(4): 349-359. doi: 10.1016/S2213-2600(20)30559-2. PMID: 33444539; PMCID: PMC7832571.

20. Kuroda S, Matsumoto S, Sano T, Kitai T, Yonetsu T, Kohsaka S, et al. External validation of the 4C Mortality Score for patients with COVID-19 and pre-existing cardiovascular diseases/risk factors. BMJ Open. 2021; 11(9): e052708. doi: 10.1136/bmjopen-2021-052708. PMID: 34497086; PMCID: PMC8438580.

21. Matsumoto S, Kuroda S, Sano T, Kitai T, Yonetsu T, Kohsaka S, et al. Clinical and Biomarker Profiles and Prognosis of Elderly Patients With Coronavirus Disease 2019 (COVID-19) With Cardiovascular Diseases and/or Risk Factors. Circ J. 2021; 85(6): 921-928. doi: 10.1253/circj.CJ-21-0160. PMID: 33952834.

22. Wu G, Yang P, Xie Y, Woodruff HC, Rao X, Guiot J, et al. Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. Eur Respir J. 2020; 56(2): 2001104. doi: 10.1183/13993003.01104-2020. PMID: 32616597; PMCID: PMC7331655.

23. Zhang S, Guo M, Duan L, Wu F, Hu G, Wang Z, et al. Development and validation of a risk factor-based system to predict short-term survival in adult hospitalized patients with COVID-19: a multicenter, retrospective, cohort study. Crit Care. 2020; 24(1): 438. doi: 10.1186/s13054-020-03123-x. PMID: 32678040; PMCID: PMC7364297.

24. Berenguer J, Ryan P, Rodríguez-Baño J, Jarrín I, Carratalà J, Pachón J, et al. Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. Clin Microbiol Infect. 2020; 26(11): 1525-1536. doi: 10.1016/j.cmi.2020.07.024. PMID: 32758659; PMCID: PMC7399713.

25. Boytsov SA, Pogosova NV, Paleev FN, Ezhov MV, Komarov AL, Pevsner DV, et al. Clinical Characteristics and Factors Associated with Poor Outcomes in Hospitalized Patients with Novel Coronavirus Infection COVID-19. Kardiologiia. 2021; 61(2): 4-14. Russian, English. doi: 10.18087/cardio.2021.2.n1532. PMID: 33734042.

26. Haimovich AD, Ravindra NG, Stoytchev S, Young HP, Wilson FP, van Dijk D, et al. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. Ann Emerg Med. 2020; 76(4): 442-453. doi: 10.1016/j.annemergmed.2020.07.022. PMID: 33012378; PMCID: PMC7373004.

27. Knottnerus JA, Tugwell P. Methodological challenges in studying the COVID-19 pandemic crisis. J Clin Epidemiol. 2020; 121: A5-A7. doi: 10.1016/j.jclinepi.2020.04.001. PMID: 32336471; PMCID: PMC7180157.

28. Why Predictive Modeling is Critical in the Fight against COVID-19? Department of evidence and intelligence for action in health.

29. Zhou X, Cheng Z, Shu D, Lin W, Ming Z, Chen W, et al. Characteristics of mortal COVID-19 cases compared to the survivors. Aging (Albany NY). 2020; 12(24): 24579-24595. doi: 10.18632/aging.202216. PMID: 33234724; PMCID: PMC7803528.

30. Bhaskaran K, Bacon S, Evans SJ, Bates CJ, Rentsch CT, MacKenna B, et al. Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. Lancet Reg Health Eur. 2021; 6: 100109. doi: 10.1016/j.lanepe.2021.100109. PMID: 33997835; PMCID: PMC8106239.

31. Chong WH, Saha BK, Medarov BI. Clinical Characteristics Between Survivors and Nonsurvivors of COVID-19 Patients Requiring Extracorporeal Membrane Oxygenation (ECMO) Support: A Systematic Review and Meta-Analysis. J Intensive Care Med. 2022; 37(3): 304-318. doi: 10.1177/08850666211045632. PMID: 34636697.