

Predictive model for determining the severe condition of patients with SARS-CoV-2 (COVID-19) based on a modified application of logistic regression*

Valentyna Yunchyk^{1,†}, Tetyana Mamchych^{*1,2†}, Ivan Mamchych^{1,2†}, Anna Yushchuk^{1,3†} and Olha Korzhyk^{1,†}

¹ Lesya Ukrainka Volyn National university, Voli 13 43000 Lutsk, Ukraine

² Tetyana Mamchych Consulting, Gruvvägen 29, 74333, Sala, Sweden

³ Volyn regional infectious disease hospital, Shevchenko Str. 30, Lutsk, Ukraine

Abstract

This study presents the development of a prognostic model designed to predict the likelihood of severe disease progression in patients newly diagnosed with SARS-CoV-2 (Covid-19). The aim is to enable early identification of high-risk individuals upon hospital admission, using rapid and accessible diagnostic technologies based solely on blood test indicators.

Timely recognition of a patient's risk of developing severe Covid-19 is critical for immediate referral to intensive care units (ICUs). A prognostic model capable of estimating this risk using readily available and fast diagnostic tools could significantly improve hospital workflow and patient outcomes.

Covid-19 presents a broad spectrum of clinical manifestations, and the course of the disease is often unpredictable. Diagnostic complexity is further compounded by the fact that severity is not always clinically apparent at the time of admission. Blood analysis and imaging procedures such as computed tomography (CT) are typically initiated immediately. Modern technologies allow for key blood parameters to be obtained within 20 minutes. A predictive model based exclusively on these parameters could be extremely valuable for preserving life and health.

To develop the model, we conducted a study evaluating the diagnostic potential of various blood parameters. In addition to empirical data, internationally recognized biomarkers with high diagnostic value were included. Our analysis confirmed the diagnostic relevance of these indicators; however, none demonstrated sufficient predictive power when used independently.

The study details the search for an optimal predictive algorithm, including the identification of predictor variables and the selection of a suitable computational model. Model performance was assessed using a bootstrap validation approach: 100 random subsamples of 140 cases (out of a total of 200) were used to train the model, while prediction errors were evaluated on the remaining 60 cases.

Results: We developed an algorithm based on the sequential application of logistic regression models to selected blood-based predictors. This model effectively estimates the risk of severe disease in SARS-CoV-2 patients. The predictors include the neutrophil-to-lymphocyte ratio (NLR), derived NLR (dNLR), platelet-to-lymphocyte ratio (PLR), absolute lymphocyte count (IGLR100), and the systemic inflammation index (SII). These are computed from five standard blood test indicators.

In our dataset, the algorithm achieved an accuracy rate of 92% in correctly predicting severe outcomes. The model has been implemented in a software application that was successfully validated on new clinical cases. Notably, it also demonstrated utility in selected pneumonia cases unrelated to Covid-19.

The proposed model has been accepted for practical use. The study was conducted as part of a collaborative project between the Volyn Regional Infectious Disease Hospital (Laboratory Center) and Lesya Ukrainka Volyn National University (Faculty of Medicine and The Center for Data Science).

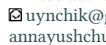
Keywords

prognostic model, Covid-19, logistic regression, classification, statistical distribution

Proceedings of the Workshop on Scientific and Practical Issues of Cybersecurity and Information Technology at the V international scientific and practical conference Information security and information technology (ISecIT 2025), June 09–11, 2025, Lutsk, Ukraine

[†] Corresponding author.

[†] These authors contributed equally.

 vynchik@gmail.com (V. Yunchyk); tetyana.mamchych@gmail.com (T. Mamchych); mamchych.i@gmail.com (I. Mamchych); annayushchuk2010@gmail.com (A. Yushchuk); olga.korgik@gmail.com (O. Korzhyk)
 0000-0003-3500-1508 (V. Yunchyk); 0000-0002-1934-1484 (T. Mamchych); 0009-0005-6399-6727 (I. Mamchych); 0000-0001-8108-8216 (A. Yushchuk); 0000-0002-2526-8877 (O. Korzhyk)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

COVID-19 is an acute infectious disease caused by SARS-CoV-2, which has spread widely across the globe, reaching more than 704 million cases and 7 million deaths by June 2024 according to officially registered data ([1]).

In practice, the severity of a patient's condition is not always obvious when they are admitted to the hospital. As a result, patients are sometimes mistakenly assigned to regular wards when they actually require intensive care –and vice versa.

The task of effectively identifying patients in critical condition is vital for saving lives and becomes especially important during epidemics. During the spread of the disease, hospitals were overwhelmed, and both hospital staff and resources were overburdened around the world.

Undoubtedly, computed tomography (CT) technology is a key tool in diagnosing infectious pneumonias. However, under conditions of a very high number of patients, the use of CT scans becomes time-consuming, while critically ill cases require immediate intensive care. Moreover, not all hospitals are equipped with CT scanners.

When a patient is hospitalized, a wide range of laboratory test results can be obtained within 20 minutes thanks to modern technologies. Therefore, the idea emerged to build a predictive model based solely on laboratory blood test indicators, without relying on additional data. From a data analysis perspective, efforts were made to find an appropriate method and identify the minimally necessary and relevant indicators.

Desirable characteristics of such a predictive model include:

- Speed of obtaining results;
- Technological accessibility (i.e., the ability to use it with typical diagnostic equipment found in hospitals, avoiding technologies that are unique, complex, time-consuming, or particularly expensive);
- Instrumental automation (for example, having a computer program to perform the calculations);
- Comprehensible interpretation of results for hospital staff.

The development of the predictive model was carried out as part of research at the Laboratory Center of the Volyn Regional Infectious Disease Hospital. The hospital's Laboratory center is equipped with modern analytical instruments.

2. Methods

During the years 2020–2023, empirical data were collected on patients diagnosed with SARS-CoV-2 (COVID-19) who were admitted to the Volyn Regional Infectious Disease Hospital.

The working sample of this study consisted of 200 records, divided into two groups: 100 recovered patients and 100 fatal cases. Among the 100 fatal cases, 38 were patients who were not immediately referred to the intensive care unit but were transferred there from the general ward after their condition worsened during treatment.

The empirical data are represented by a set of laboratory blood test indicators, determined at the time of hospital admission, along with a grouping variable Group, which takes the value 0 for recovered patients and 1 for deceased patients.

The analysis involved five integral indicators that have been recognized in international scientific publications as having high diagnostic potential.

The final predictors selected were:

- Neutrophil-to-lymphocyte ratio (NLR);
- Derived neutrophil-to-lymphocyte ratio (dNLR);
- Platelet-to-lymphocyte ratio (PLR);
- Absolute lymphocyte count (IGLR, IGLR100);
- Systemic inflammation index (SII).

All of these are calculated using components from the blood test results:

- Neutrophils (NEUT);
- Lymphocytes (Lymph);
- Leukocytes (WBC);
- Immunoglobulins (IG);
- Platelets (PLT).

Previous research conducted prior to building the predictive model had already established that these diagnostic indicators differ significantly between the two patient groups. However, they were still insufficient in accurately distinguishing cases that require intensive care. Specifically, the Mann-Whitney test demonstrated a significant difference in the values of each of the five predictors between the recovered and deceased groups at a significance level of 0.05. Nevertheless, using these indicators as classifiers resulted in an unacceptable number of errors on the available data.

During the search for an appropriate forecasting technique, the following methods were tested for applicability: discriminant analysis, support vector machines (SVM), and both binary and multiple logistic regression. For multiple regression, various combinations of predictors were used as factors. None of these approaches were found to be successful in correctly classifying the available data with at least 90% reliability.

The logistic regression method is a recognized approach for classification tasks, including in the field of medical research. For example, in [4]–[6], logistic regression was used to illustrate the predictive capabilities of certain inflammatory markers in patients diagnosed with COVID-19. That study presented cut-off points for distinguishing severe conditions based on the diagnostic indicators, calculated using their empirical data. The logistic regression was chosen as the primary method for the predictive model.

Data analysis was conducted using the R software [3]. Model parameters were calculated using a generalized linear model through the `glm()` function.

For each of the five diagnostic indicators, logistic regression coefficients were estimated using the maximum likelihood method. The cut-off point that maximizes the number of correct

classifications was determined, and the critical value of each indicator was identified at which the odds ratio (OR) equals 1.

The following logistic models were obtained:

$$mNLR = -2.71 + 0.36 * NLR, \quad (1)$$

$$mdNLR = -12.56 + 14.53 * dNLR, \quad (2)$$

$$mIGLR = -0.89 + 0.06 * IGLR, \quad (3)$$

$$mPLR = -1.35 + 0.005 * PLR, \quad (4)$$

$$mSII = -2.08 + 0.001 * SII. \quad (5)$$

The coefficients of the logistic regression models are statistically significant at the $p < 0.01$ level.

For each patient, the odds ratio (OR) value for the Group variable was calculated using the logistic models for each of the five integral indicators, and based on this, a predicted value for the Group variable was determined. To compute the predicted odds ratio (OR) value for a given patient, the following formula is used:

$$Y = \exp(model)/(1 + model), \quad (6)$$

where the argument "model" refers to one of the values (1) – (5), depending on which predictor indicator is being used.

The predicted value of the Group variable was set to 1 if the patient's NLR value exceeded the critical threshold at which $OR = 1$ (similarly for dNLR, IGLR100, PLR, and SII); otherwise, the value was set to 0. The predictor values at which $OR = 1$ were computed using the R software.

Table 1

Pivotal values (OR=1) and numbers of incorrectly predicted cases for the five predictors with using logistic regression model

	NLR	dNLR	IGLR100	PLR	SII
OR=1	7.61	0.86	14.44	299.8	1917.59
Group = 1	16	25	41	40	24
Group = 0	6	30	16	12	10
Total	22 (11%)	55 (28%)	57 (29%)	52 (26%)	34 (17%)

As shown in Table 1, none of the five predictors provides 90% correct classifications – either for the total sample of 200 cases or for the subgroup of 100 fatal cases. The lowest number of errors – 22 out of 200 (11%) – was achieved by the NLR indicator.

Specifically, the logistic model based on the NLR predictor resulted in 16 misclassifications for patients in Group 1 (deceased) and 6 misclassifications for patients in Group 0 (recovered) – and this was the best result among all models. Attempts to build effective multiple logistic regression models using all five predictors together, or different combinations of them, were unsuccessful. For example, out of 100 deceased patients, 8 were not identified as high-risk by any of the five diagnostic indicators, and 7 were detected by only one of the five. Moreover, applying multiple logistic regression with all five factors resulted in 35 out of 100 fatal cases being misclassified.

A crucial step in the search for an effective algorithm was a detailed analysis of the dataset to identify which specific cases were misclassified. It was found that different predictors made errors on different cases. At the same time, only 8 out of 100 fatal cases were missed by all predictors, meaning that 92% of fatal cases were correctly identified by at least one of the indicators.

Therefore, it is logical to classify a case into Group 1 (high risk of severe condition) if at least one of the five predictors classifies it as such.

3. Results

The algorithm consists of the sequential application of five logistic models for the specified indicators as factors. In the final algorithm, classification using simple logistic regression is performed five times in succession. If at any step a case is identified as belonging to the high-risk group for developing a severe condition, that result is accepted as final, and the verification process stops.

Next, the task is to determine the appropriate order for applying the predictors.

The fewest incorrect predictions were made by the NLR indicator (16 errors for Group = 1 and 6 errors for Group = 0, totaling 22 errors based on the OR = 1 criterion). Therefore, NLR was chosen as the initial predictor in the sequential application.

To identify the next predictor, the number of correct predictions following misclassifications by NLR was determined.

Table 2

Numbers of correctly predicted cases for the four predictors after 16 incorrect predictions with NLR. Here (*) means “predicted value”. Only cases with Group =1 were under consideration here.

	DNLR [*] =1	IGLR [*] =1	PLR [*] =1	SII [*] =1
NLR [*] =0	5	2	4	2

That is, out of the 16 severe cases that the mNLR misclassified, mNLR correctly identified 5 cases, the mPLR identified 4 cases, and so on (Table 2).

Since the highest number of correct classifications (5) corresponds to the dNLR predictor, it was chosen as the next in sequence. Subsequently, the sequence PLR, IGLR, and SII was selected. Ultimately, after sequentially applying all five predictors, only 8 severe cases remained misclassified.

To automate the calculations, a software product was developed, and its operation is illustrated in Figures 1 and 2. These are new cases that confirmed the effectiveness of this algorithm.

Fig. 1 shows data of a patient who was not considered severe upon admission to the hospital and was assigned to a general ward. However, over the course of several days, his condition worsened, and he was transferred to the intensive care unit, where his life was saved. This case occurred before the prognostic model was implemented for use. The scientific interest was in testing the model on new data. It was demonstrated that the model predicted the risk of severe condition for this patient based on his blood test results at the time of hospital admission.

NEUT	NLR	NLR (OR = 1)	NLR Result
8,94	13,545453	7,61	High
Lymph	dNLR	DNLR (OR = 1)	dNLR Result
0,66	0,97811806	0,86	High
WBC	IGLR100	GNLR100 (OR = 1)	IGLR100 Result
9,8	3,030303	14,44	Low
IG	PLR	PLR (OR = 1)	PLR Result
0,02	221,21211	299,88	Low
PLT	SII	SII (OR = 1)	SII Result
146	1977,6362	1917,59	High

Figure 1: A case of a severe patient condition determined by blood test indicators.

Fig. 1 shows data of a patient who was not considered severe upon admission to the hospital and was assigned to a general ward. However, over the course of several days, his condition worsened, and he was transferred to the intensive care unit, where his life was saved. This case occurred before the prognostic model was implemented for use. The scientific interest was in testing the model on new data. It was demonstrated that the model predicted the risk of severe condition for this patient based on his blood test results at the time of hospital admission.

The program included the blood test indicators, calculated the values of all predictors, and 3 out of 5 predicted a severe condition.

Fig. 2 illustrates a mild course of the disease, where none of the predictors forecast a severe condition.

NEUT	NLR	NLR (OR = 1)	NLR Result
4,94	2,3301888	7,61	Low
Lymph	dNLR	DNLR (OR = 1)	dNLR Result
2,12	0,82470787	0,86	Low
WBC	IGLR100	GNLR100 (OR = 1)	IGLR100 Result
8,11	4,2452836	14,44	Low
IG	PLR	PLR (OR = 1)	PLR Result
0,09	194,81133	299,88	Low
PLT	SII	SII (OR = 1)	SII Result
413	962,368	1917,59	Low

Figure 2: A case of a mild patient condition determined by blood test indicators.

Note 1. Attempts to exclude even one of the predictor indicators from the analysis led to an increase in the number of incorrect predictions in our data.

Note 2. This algorithm does not take into account errors for the group of recovered patients, since its primary goal is to minimize errors of failing to detect the risk of severe conditions (false negatives), unlike errors where a false prediction of severe condition is made (false positives). In

other words, the main quality criterion of the model was maximizing the preservation of human health and life, and errors related to excessive use of hospital resources were not considered here.

Note 3. The model parameters were obtained based on data from hospitalized patients, i.e., those considered sufficiently severe cases when the patient seeks or is referred to the hospital. The model's effectiveness has not been tested on other cases.

4. Discussions

After developing the model and creating the software product, it was implemented for use at the Volyn Regional Infectious Diseases Hospital (Lutsk) and the Lutsk Perinatal Center. A detailed analysis of fatal cases from the existing database confirmed the usefulness of this model. It was found that in those cases (from the available database) where a severe condition was not recognized at the time of patient admission, it could have been identified immediately if this predictive model had been applied. In practice, however, patients were admitted to general therapy wards, and even during treatment, they developed severe conditions.

Moreover, this model was tested on some pneumonia cases without a confirmed Covid-19 diagnosis, where the proposed algorithm indicated a risk of developing a severe condition, which was not obvious at admission.

The work was carried out as part of a joint study with medical specialists, but this publication is more accurately categorized as a work in the field of applied mathematics and computer science, as it reflects only the technical aspect – the algorithm for constructing a tool that generates a recommendation-based conclusion.

The medical aspect of the study, including the interpretation of the obtained parameters, we consider more appropriate to present in forums for medical professionals.

From a medical standpoint, the use of NLR, dNLR, PLR, IG and SII as important predictors of the development of a severe condition is not new. Proposals to use them for prognostic purposes through logistic regression have been published in previous years in the medical literature ([4] – [6]). The effectiveness of these predictors was also confirmed in our data, and the results of the research and interpretations have been published ([7] – [9]).

The main achievement of our work is the proposal to apply these regressions sequentially, one after the other in a predetermined order. This algorithm, on our data, led to a reduction in classification errors compared to using each predictor individually.

When building forecasts, errors may occur. These are of two types: false positives and false negatives. Let us consider the substantive interpretation of these concepts. A value of 1 for the grouping indicator meant that the patient died, while a value of 0 indicated recovery. In this context, a false positive forecast means that a severe course was predicted, but it did not occur. A false negative means that the model failed to classify the case as being at risk of a severe condition, but the severe condition still occurred.

Clearly, both types of errors are undesirable, as in both cases the patient may receive not optimal treatment. However, the implications of these errors differ. A false positive means that a patient is sent to the intensive care unit without an urgent need. This leads to excessive use of hospital resources (not just medications, but also additional strain on the medical staff and unjustified use of equipment), occupying a bed that could be used by a less critical patient, and potentially harming the patient through an inappropriate treatment protocol. This case can also be considered economically wasteful.

The second type of error means that a seriously ill (or potentially seriously ill) patient is sent to a general ward without recognizing their risks. The result of such an error could be a significant deterioration in the patient's condition or even death. In fact, the cost of this error could be a life.

In developing the prognostic model, we aimed to minimize the second type of error – that is, to minimize false negatives. In doing so, false positive cases were not taken into consideration.

Note: The completion of our study coincided with the end of the disease epidemic, when the overall number of patients significantly decreased. Additionally, the nature of the disease course changed, and the proportion of patients in severe condition drastically declined. At present, severe cases of pneumonia accompanied by a positive COVID-19 test are not common. A sufficient number of severe cases, from which a statistically adequate sample could be formed, has not yet been accumulated in this department. All instances of using the model on new data have been successful. The effectiveness of the model is still being tested.

Currently, in Ukraine, as well as globally, there is no recognized COVID-19 epidemic. By resolution of the Cabinet of Ministers of Ukraine, quarantine restrictions caused by the SARS-CoV-2 coronavirus were lifted as of June 30, 2023 ([10]). However, the possibility of the epidemic situation re-emerging still exists. Cases of illness continue to occur. Increased activity was recorded in some countries of East Asia, West Africa, Central America and the Caribbean, Western Asia, South Asia, and Southeast Asia ([11]). Therefore, the relevance of our predictive model remains valid.

5. Acknowledgements

This work was conducted as part of a joint project between the Volyn Regional Infectious Diseases Hospital (Laboratory Center) and Lesya Ukrainka Volyn National University (Medical Faculty and The Center for Data Science).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] An official website of the United States government. National Library of Medicine. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11787180/>
- [2] <https://index.minfin.com.ua/ua/reference/coronavirus/ukraine/>
- [3] <https://cran.r-project.org/doc/manuals/R-intro.pdf>
- [4] Role of leukocytes and systemic inflammation indexes (NLR, PLR, MLP, dNLR, NLPR, AISI, SIR-I, and SII) on admission predicts in-hospital mortality in non-elderly and elderly COVID-19 patients / H. Ghobadi [et al.]. Frontiers in medicine. 2022. No 9. DOI: <https://doi.org/10.3389/fmed.2022.916453>
- [5] Çakırca T. D., Torun A., Çakırca G., Portakal R. D. Role of NLR, PLR, ELR and CLR in differentiating COVID-19 patients with and without pneumonia. Int. J. Clin Pract. 2021. Vol. 75, iss. 11
- [6] Alisik M., Erdogan U.G., Ates M. and others. Predictive value of immature granulocyte count and other inflammatory parameters for disease severity in Covid-19 patients. International Journal of Medical Biochemistry. 2021, 4(3), 143-149. doi: 10.14744/ijmb.2021.03164

[7] Mamchych T., Yushchuk H., Mamchych I. One case of a prognostic model with using series of logistic regressions (in Ukrainian). I International Scientific and Practical Conference "Problems of Computer Science, Software Modeling and Security of Digital Systems", Abstracts. Lutsk-Svityaz, 13-16 June 2024. URL:<https://apcssm.vnu.edu.ua/index.php/conf/article/view/112/25>

[8] Yushchuk A., Pykaliuk V., Korzhyk O. Hematocytological, biochemical, and hemostasis parameters' role in predicting the possibility of the various forms of the COVID-19 course in hospitalized Ukrainian patients: A cross-sectional study. *Health Sci Rep.* 2023. 6(7). e1403. doi:10.1002/hsr2.1403 <https://onlinelibrary.wiley.com/doi/10.1002/hsr2.1403>

[9] Yushchuk H., Korzhyk O., Pykaliuk V. Prediction of the state of critical patients with SARS-CoV-2 infection (laboratory aspect). *Ukrainian Journal of Laboratory Medicine.* 2024. 2(3). P. 15-21. <https://doi.org/10.62151/2786-9288.2.3.2024.03> <https://ujlm.org.ua/index.php/journal/article/view/46>

[10] Resolution of the Cabinet of Ministers of Ukraine dated 27 June 2023 № 651 <https://zakon.rada.gov.ua/laws/show/651-2023-п#Text>

[11] Yushchuk H., Korzhyk O., Pykaliuk V. The role of complete blood count for predicting the course of coronavirus infection in patients with COVID-19: a systematic review. *Notes in Current Biology.* 2023. Vol. 6 No. 2. P. 73–76. doi.org/10.29038/NCBio.23.2-10 <https://journalbio.vnu.edu.ua/index.php/bio/article/view/620>

[12] WHO COVID-19 dashboard <https://data.who.int/-dashboards/covid19/summary?n=o>)