Revista Thema

V.24 | n.2 | 2025

p.1-17

DOI: http://doi.org/10.15536/revistathema.24.2025.4002

ISSN: 2177-2894 (online)

CIÊNCIAS DA SAÚDE

DEVELOPMENT OF MODELS FOR ASSESSING INSTANT RELATIVE RISK IN COVID-19 PATIENTS AT EARLY AND LATE STAGES FOR USE AS A SCORING SYSTEM

DESENVOLVIMENTO DE MODELOS PARA AVALIAR O RISCO RELATIVO INSTANTÂNEO EM PACIENTES DE COVID-19 EM ESTÁGIOS INICIAIS E TARDIOS PARA USO COMO UM SISTEMA DE ESCORE

Silvio Alencar Cândido-Sobrinho[1]

https://orcid.org/0000-0002-6055-5791 http://lattes.cnpq.br/1245397610888677

Beatriz Ximenes Braz[2]

https://orcid.org/0000-0002-2241-5800 http://lattes.cnpq.br/7442313976777182

Juvêncio Santos Nobre[3]

https://orcid.org/0000-0002-7321-3221 http://lattes.cnpq.br/4610025058115796

Danilo de Menezes Daloso[4]

https://orcid.org/0000-0003-1842-420X http://lattes.cnpq.br/0306680503261422

ABSTRACT

COVID-19, recognized as an outbreak in January 2020 and declared a pandemic shortly after, remains a global clinical concern. This study developed a predictive model to classify patients by Relative Risk, which can be used as a scoring system. It is a retrospective cohort study using data from patients admitted to the National Institute for Infectious Diseases "Lazzaro Spallanzani" (Italy) between January 29 and March 28, 2020. Cox Proportional Hazards models were created for patient admission (n = 218) and for the late stage of the disease (n = 385). The models were validated using Machine Learning and tested for Cox assumption violations and statistical power. Coefficients were extracted, and equations were derived to calculate Relative Risk. For Early-stage Relative Risk (ERR), the following variables were used: Age Hazard Ratio (HR): 1.05 [95% CI: 1.02 - 1.08], RDW-SD (HR): 1.08 [95% CI: 1.02 - 1.14], Lymphocytopenia (HR): 3.23 [95% CI: 1.32 - 7.95], and

[1] Universidade Federal do Ceará (UFC), Fortaleza/CE – Brasil. E-mail: silvio.sobrinho@prof.ce.gov.br

[2] University of Miami/Jackson Memorial Hospital Internal Medicine (UM/JMH), Miami, FL – Estados Unidos. E-mail: beatriz.x.braz@gmail.com

[3] Universidade Federal do Ceará (UFC), Fortaleza/CE – Brasil. E-mail: juvencio@ufc.br

[4] Universidade Federal do Ceará (UFC), Fortaleza/CE – Brasil. E-mail: daloso@ufc.br

Revista Thema
V.24 | n.2 | 2025
p.1-17

DOI: http://doi.org/10.15536/revistathema.24.2025.4002

ISSN: 2177-2894 (online)

Obesity (HR): 2.88 [95% CI: 1.08 - 7.69]. For Late-stage Relative Risk (LRR), the variables were: LDH (HR): 7.41 [95% CI: 3.00 - 18.28] and RDW-CV (HR): 5.74 [95% CI: 2.28 - 14.45]. This study provides models and equations that can be implemented as a scoring system to aid decision-making by calculating instant Relative Risk (RR).

***Keywords**: COVID-19; Score; Cox Proportional Hazards; Machine Learning.*

RESUMO

A COVID-19, reconhecida como um surto em janeiro de 2020 e declarada uma pandemia pouco depois, ainda é uma preocupação a nível clínico. Este estudo desenvolveu um modelo preditivo para classificar pacientes por Risco Relativo, que pode ser empregado como sistema escore. Trata-se de um estudo de coorte retrospectivo usando dados de pacientes admitidos no Instituto Nacional de Doenças Infecciosas "Lazzaro Spallanzani" (Itália) entre 29 de janeiro e 28 de março de 2020. Modelos de Riscos Proporcionais de Cox foram criados para a admissão do paciente (n = 218) e para o estágio tardio da doença (n = 385). Os modelos foram validados através de Machine Learning e testados quanto às violações das premissas de Cox e Poder estatístico. Coeficientes foram extraídos, e equações foram derivadas para calcular o Risco Relativo. Para o Risco Relativo em Estágio Inicial (ERR), as seguintes variáveis foram usadas: Razão de Risco (HR) da Idade: 1,05 [IC 95%: 1,02 - 1,08], RDW-SD (HR): 1,08 [IC 95%, 1,02 - 1,14], Linfocitopenia (HR): 3,23 [IC 95%, 1,32 - 7,95] e Obesidade (HR): 2,88 [IC 95%, 1,08 - 7,69]. Para o Risco Relativo em Estágio Tardio (LRR), as variáveis foram: LDH (HR): 7,41 [IC 95%, 3,00 - 18,28] e RDW-CV (HR): 5,74 [IC 95%, 2,28 - 14,45]. Este estudo fornece modelos e equações que podem ser implementados como um sistema de pontuação para auxiliar na tomada de decisões calculando o Risco Relativo instantâneo (RR).

Palavras-chave: COVID-19; Escore; Riscos Proporcionais de Cox; Aprendizado de Máquina.

1. INTRODUCTION

The COVID-19 pandemic, identified as an outbreak in January 2020, continues to pose significant challenges globally. Despite the production of highly effective vaccines, the situation remains concerning due to several factors. At the time, these included disparities in healthcare resources, such as the concentration of intensive care units (ICUs) in metropolitan areas, as well as shortages of essential supplies and personnel (Alves, 2021; French et al., 2021).

Additionally, vaccination hesitancy (Middleman; Klein; Quinn, 2022; Shakeel et al., 2022) and the emergence of new variants (WHO, 2022) contribute to the complexity of the situation. While vaccines offer a high level of efficacy, they do not guarantee complete immunization. Furthermore, the uneven distribution of vaccines globally means that a significant portion of the population remains unvaccinated or under-vaccinated. Consequently, individuals are still at risk of developing severe forms of the disease (Botton et al., 2021; Yek et al., 2022).

The pandemic has propelled modelling and forecasting to the forefront of public policy making. However, it has also highlighted significant challenges, particularly the lack of standardisation in data and the absence of unified data repositories for data-driven policy decisions (Bertozzi et al., 2020; Ros et al., 2021). There has been a noticeable lack of effort in data anonymisation and the sharing of raw data worldwide (Subbian et al., 2021; Wolkewitz; Puljak, 2020). Nevertheless, independent research groups within health facilities have swiftly investigated the use of routine data, including haematological variables, enzymatic biomarkers, cytokines, and other parameters (Velavan; Meyer,

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

2020; Yuan et al., 2020), to establish monitoring parameters for patients. Some groups have developed Machine Learning models that demonstrate good accuracy (Guan et al., 2021; Ma et al., 2020; Vaid et al., 2020).

At the clinical level, two noteworthy models for monitoring are the '4c Score' (Jones et al., 2021; Knight et al., 2020) and CIRC (Garibaldi et al., 2021). These models have become popular among physicians due to their ease of use, as they are implemented as online calculators readily available on MDCalc. The 4c Score is calculated using 9 variables, while the CIRC score uses 23 variables. However, using a large number of variables in a model can lead to undesirable outcomes due to model complexity. This complexity can cause the model to fit not only the relationship between variables but also the random error, potentially leading to bias, overfitting, and other unwanted effects (Lever; Krzywiski; Altman, 2016).

COVID-19 is still a matter of concern, and tools to assist in patient monitoring and resource allocation remain pertinent. The objective of this research was to develop two models for use at different stages of the disease, which can aid in patient follow-up at early and late stages of disease progression. Both models allow for ranking patients by Relative Risk (RR), which can be adopted as a score using easily acquired data such as complete blood count (CBC), comorbidities, and demographic variables obtained during routine.

2. STUDY DESIGN AND METHODS

2.1 MODEL DEVELOPMENT GUIDELINES

This study was conducted in compliance with the *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD statement)* guideline (Collins et al., 2015).

2.2 DATA SOURCE AND ETHICS

Data for this study were obtained from PCR-confirmed COVID-19 patients admitted to the Italian National Institute for Infectious Diseases "Lazzaro Spallanzani" (INMI) between 29th January and 28th March 2020 (Lanini et al., 2020). This data had been previously released under Ethics Committee approval. According to the Brazilian National Council for Ethics in Research (CONEP), data shared in public repositories containing anonymized data are waived of Ethics Committee evaluation (Brazil, 2016).

2.3 VARIABLE CHARACTERIZATION

The following variables were considered for analysis: White Blood Cell count (WBC), Lymphocytes' count (LYM), Monocytes' count (MONO), Neutrophils' count (NEU), Eosinophils' count (EOS), Basophils' count (BAS), Red Blood Cell count (RBC), RDW-SD, RDW-CV, Mean Corpuscular Volume (MCV), Haemoglobin concentration (HBC), Haematocrit (HCT), Mean Corpuscular Haemoglobin Concentration (MCHC), Mean Corpuscular Haemoglobin (MCH), Platelets' count (PLT), Mean Platelet Volume (MPV), Total Bilirubin (BILTOT), Direct Bilirubin (BILDIR), Lactate Dehydrogenase (LDH), Alanine Aminotransferase (ALT), Aspartate Aminotransferase (AST), and Glucose (Glycaemia).

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

The following comorbidities were also considered: Diabetes, Neoplasm, Cardiovascular Disease, Immune Disorder, Chronic Obstructive Pulmonary Disease, Chronic Renal Failure, Metabolic Disorder, and Obesity, along with Age and Sex. Right-censored patients were considered as discharged (not deceased).

Additional discrete variables were computed such as anaemia, neutrophils per lymphocyte ratio, lymphocytopenia, and thrombocytopenia, as they have been previously associated with increased severity in other studies (Terpos et al., 2020; Zhang et al., 2020).

2.4 STATISTICAL ANALYSES

All data processing and analyses were conducted using R (4.0.1) (R Core Team, 2025). The original dataset was retrieved (Lanini et al., 2020) and processed using the *tidyverse* (Wickham et al., 2021) library. Baseline tables were created using the *tableone* (Yoshida; Bartel, 2021) library.

Missing data were imputed using Bayesian Principal Component Analysis (BPCA) (Schmitt; Mandel; Guedj, 2015) from the *pcaMethods* (Stacklies et al., 2007) library. Continuous variables were transformed using natural logarithm, square root, and an optimal lambda for Box-Cox transform calculated using the *forecast* (Hyndman; Khandakar, 2008) library.

Cox Proportional Hazards models were created using the library *survival* (Therneau, 2021). Forward-backward step-wise regression was performed along with MASS library (Venables; Ripley, 2002).

Machine Learning models screening was conducted using Generalized Linear Model (GLM), Bayesian Generalized Linear Model (Bayes GLM), and Lasso/Elastic-Net Regularized GLM (GLMNet) models implemented by the caret Framework for Machine Learning in R (Kuhn, 2021). The training control for cross-validation method settings was Repeated 20-fold cross-validation, 20 times. Prediction results were assessed with AUC-ROC, PRG, PRC, and CC curves using the library *MLeval* (John, 2020).

Cox Proportional Hazards models of interest were tested for assumption violation using Schoenfeld's test with *survival* (Therneau, 2021) and *survminer* (Kassambara; Kosinski; Biecek, 2021) libraries. Forest plots were created using the *forestmodel* (Kennedy, 2020) library. Power effect analyses for proportional hazards were calculated using the library *powerSurvEpi* (Qiu et al., 2021).

3. RESULTS

There were 10,671 entries from up to 508 patients. Patient characteristics can be found in the baseline Tables 1 and 2, respectively. There were 4,840 and 8,778 entries for continuous variables, of which 237 (4.89%) and 626 (7.13%) were missing, leaving 4,603 (95.1%) and 8,152 (92.86%) available entries. There were no missing data for comorbidities and demographic data (Table 1 for more details per variable).

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

Table 1. Missing values from subsets used for Early-stage and Late-stage modelling. There were, respectively 4840 and 8778 total entries for continuous variables, of which, respectively 237 (4.89 %) and 626 (7.13 %) were missing, 4603 (95.1 %) and 8152 (92.86 %) available. There were no missing data for comorbidities and demographic data.

Variable	Early-stage subset		Late-stage subset	
	Missing values	Proportion of missing values(%)	Missing values	Proportion of missing values(%)
White Blood cell count	0	0	0	0
Lymphocytes' count	0	0	0	0
Monocytes' count	0	0	0	0
Neutrophils' count	0	0	0	0
Eosinophils' count	0	0	0	0
Basophils' count	0	0	0	0
Red Blood Cells' count	0	0	0	0
RDW-SD	0	0	0	0
RDW-CV	0	0	0	0
Mean Corpuscular Volume	0	0	0	0
Haemoglobin	0	0	0	0
Haematocrit	0	0	0	0
Mean Corpuscular Haemoglobin Concentration	0	0	0	0
Mean Corpuscular Haemoglobin	0	0	0	0
Platelets' count	0	0	0	0
Mean Platelet Volume	1	0	8	2
Total Bilirubin	50	23	135	34
Direct Bilirubin	51	23	136	34
Lactate Dehydrogenase	58	26	162	41
Alanine Aminotransferase	29	13	71	18
Aspartate Aminotransferase	29	13	72	18
Glicemia	19	9	42	11
TOTAL	237	4.89	626	7.13

Source: authors' own elaboration.

Missing data were estimated using Bayesian PCA (BPCA) for each subset. After missing value imputation, continuous variables were also subjected to transformation using natural logarithm, square root, and optimal lambda for Box-Cox transform.

The data were split into two subsets using unique entries per patient. The subset used for early-stage modelling included entries from the 8th day, while the subset for late-stage modelling included the last lab data entry, up to 21 days of follow-up.

An initial screening was performed by univariate Cox Proportional Hazards regression. Only significant variables ($p < 0.05$) with an increase on Hazard Ratio ($HR > 1.1$) were listed. This list would further assist on manual variable removal on step-wise regression. There was a total of 26

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

variables of interest on subset for Early-stage and 37 variables for Late-stage modelling (Tables 2 & 3).

Table 2. Summary of univariate Cox Proportional-Hazards regression for each variable transformed in the Early-stage subset, including β coefficients, Hazard Ratios (HR), 95% confidence intervals (CI) for HR, Wald test scores, and Wald p-values. Results displayed were filtered by p-value < 0.05 and HR > 1.1.

Variable	Data transform	β	HR	95%CI for HR	Wald test	p-value
RDW-CV	-	0.21	1.2	[1.1 - 1.4]	9.3	0.0023
Diabetes	-	1.1	3	[1.4 - 6.4]	7.7	0.0054
Cancer	-	1.2	3.2	[1.1 - 9]	4.7	0.03
Cardiovascular disease	-	1.4	4.2	[2 - 8.8]	14	0.00015
Chronic kidney disease	-	0.98	2.7	[1 - 6.9]	4.1	0.043
Metabolic disorder	-	1.4	4.2	[1.6 - 11]	8.7	0.0032
Anemia	-	1.1	3.1	[1.5 - 6.4]	10	0.0016
Neutrophilia	-	0.93	2.5	[1.1 - 5.6]	5.2	0.022
Lymphocytopenia	-	1.5	4.3	[1.8 - 10]	10	0.0013
Neutrophils	Square-root	0.54	1.7	[1.1 - 2.7]	5.4	0.02
RDW-SD	Square-root	2.1	8	[3.5 - 18]	24	1.10E-06
RDW-CV	Square-root	2.3	10	[2.4 - 43]	9.9	0.0016
MCV	Square-root	1.7	5.5	[1.9 - 16]	9.9	0.0017
Glicemia	Square-root	0.25	1.3	[1.2 - 1.4]	21	4.30E-06
Neutrophils	Log_e	0.82	2.3	[1.1 - 4.6]	5.2	0.022
RDW-SD	Log_e	9.3	11000	[250 - 510000]	23	1.70E-06
RDW-CV	Log_e	6.4	600	[13 - 28000]	11	0.0012
MCV	Log_e	10	36000	[41 - 3.2e+07]	9.2	0.0024
Glicemia	Log_e	1.7	5.2	[2.3 - 12]	16	5.50E-05
RDW-CV	Box-Cox	0.21	1.2	[1.1 - 1.4]	9.3	0.0023

Source: authors' own elaboration.

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

Further, multivariate models were created with assistance of the forward-backward step-wise regression. This method selects outperforming models with a lower Akaike Information Criterion (AIC). When the step-wise method reached an optimum model (with minimum AIC), a variable with infinite estimator and/or p-value > 0.05 was removed from the set of variables, and a new model was refitted.

Table 3. Summary of univariate Cox Proportional-Hazards regression for each variable transform in late-stage subset with β coefficients, Hazard Ratios (HR), HR 95%CI, Wald tests scores and Wald p-value scores. Results displayed were filtered by p-value < 0.05 and HR > 1.1.

Variable	Data transform	β	HR	95%CI for HR	Wald test	p-value
RDW-CV	-	0.24	1.3	[1.2 - 1.4]	34	6.40E-09
MPV	-	0.79	2.2	[1.7 - 2.8]	42	7.50E-11
Total bilirubin	-	0.23	1.3	[1.2 - 1.4]	26	4.40E-07
Direct bilirubin	-	0.31	1.4	[1.2 - 1.5]	30	4.00E-08
Cancer	-	1	2.7	[1.2 - 6.4]	5.3	0.021
Cardiovascular disease	-	1	2.8	[1.6 - 5]	13	0.00033
Respiratory disease	-	0.7	2	[1.1 - 3.8]	4.7	0.031
Cronic Renal Failure	-	1.2	3.4	[1.6 - 7.1]	11	0.00096
Obesity	-	0.91	2.5	[1.2 - 5.1]	6.1	0.014
Anemia	-	0.64	1.9	[1.1 - 3.4]	4.5	0.033
Neutrophilia	-	1.1	2.9	[1.6 - 5.1]	13	0.00029
Thrombocytopenia	-	1.1	3.1	[1.4 - 6.9]	7.6	0.006
Lymphocytopenia	-	1.9	6.5	[3.3 - 13]	28	1.30E-07
WBC	Square-root	0.7	2	[1.5 - 2.8]	18	2.90E-05
Neutrophils	Square-root	0.82	2.3	[1.7 - 3]	31	2.70E-08
RDW-SD	Square-root	1.7	5.6	[3.6 - 8.8]	59	1.70E-14
RDW-CV	Square-root	2.6	14	[5.9 - 33]	35	2.60E-09
MCV	Square-root	1.4	4.1	[1.5 - 11]	8	0.0046
MPV	Square-root	7.3	1500	[160 - 14000]	42	1.10E-10
Total bilirubin	Square-root	1.3	3.6	[2.3 - 5.8]	30	4.50E-08
Direct bilirubin	Square-root	1.5	4.5	[2.8 - 7.3]	38	7.20E-10
LDH	Square-root	0.23	1.3	[1.2 - 1.4]	40	2.40E-10
AST	Square-root	0.17	1.2	[1.1 - 1.3]	19	1.40E-05
Glicemia	Square-root	0.17	1.2	[1.1 - 1.3]	8.4	0.0037
WBC	Loge	1.3	3.8	[1.9 - 7.5]	15	0.00011
Neutrophils	Loge	1.4	4.1	[2.4 - 7.1]	26	3.30E-07
RDW-SD	Loge	8.2	3600	[450 - 29000]	60	1.20E-14
RDW-CV	Loge	7.3	1500	[140 - 15000]	37	1.20E-09
MCV	Loge	8.6	5500	[12 - 2600000]	7.5	0.006
MPV	Loge	17	2.20E+07	[120000 - 3.9e+09]	41	1.70E-10
Total bilirubin	Loge	1.6	4.8	[2.7 - 8.5]	30	4.70E-08
Direct bilirubin	Loge	1.6	4.8	[3 - 7.7]	43	5.60E-11
LDH	Loge	2.6	13	[5.7 - 30]	37	1.40E-09
AST	Loge	0.68	2	[1.4 - 2.8]	14	0.00017
Glicemia	Loge	1.4	4.1	[1.6 - 10]	9.3	0.0023
RDW-CV	Box-Cox	0.24	1.3	[1.2 - 1.4]	34	6.40E-09

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

MPV	Box-Cox	0.79	2.2	[1.7 - 2.8]	42	7.50E-11
Total bilirubin	Box-Cox	0.23	1.3	[1.2 - 1.4]	26	4.40E-07
Direct bilirubin	Box-Cox	0.31	1.4	[1.2 - 1.5]	30	4.00E-08

Source: authors' own elaboration.

Two models for assessing relative risk at the early stage of the disease showed promising performance, with high Concordance, Wald, and Likelihood test results (Table 4).

Table 4. Summary results of Cox-Proportional Hazards models #73 and #75, respectively.

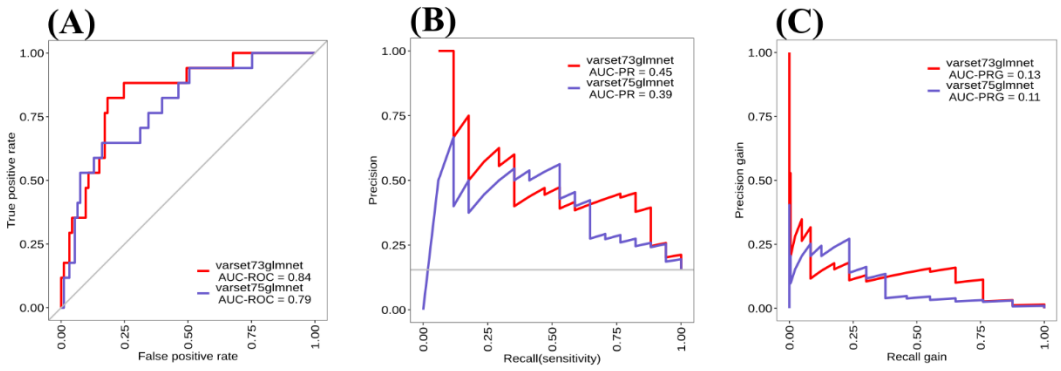
Variables	Concordance	Likelihood test	Wald test	Score test
RDW-SD, Lymphocytopenia, Age, Obesity	0.819 (se = 0.035)	41.83 (p ≤ 2E-8)	36.41 (p ≤ 2E-7)	43.15 (p ≤ 1E-8)
Cardiovascular disease, Lymphocytopenia, RDW-SD (Square-root)	0.808 (se = 0.033)	37.53 (p ≤ 4e-08)	35.00 (p ≤1e-07)	41.82 (p ≤4e-09)

Source: authors' own elaboration.

3.1 VALIDATION OF VARIABLE SELECTION

To determine the 'Early-stage' variable selection, machine learning models were created using the *caret* framework. The combined metrics results of Cox Proportional Hazards and the Machine Learning models (Figure 1), Schoenfeld's and Power effect tests were paramount in concluding that the variables that are more likely to predict severity at symptom onset are age, lymphocytopenia, obesity, and RDW-SD.

Figure 1. Receiver Operating Characteristic (ROC) curves (AUC-ROC), Precision-Recall (PR) curves (AUC-PR), and PR gain curves (AUC-PRG) outputs from logistic regression models using variable sets #73 and #75 are presented. The GLMnet models outperformed therefore others were removed from this graph.

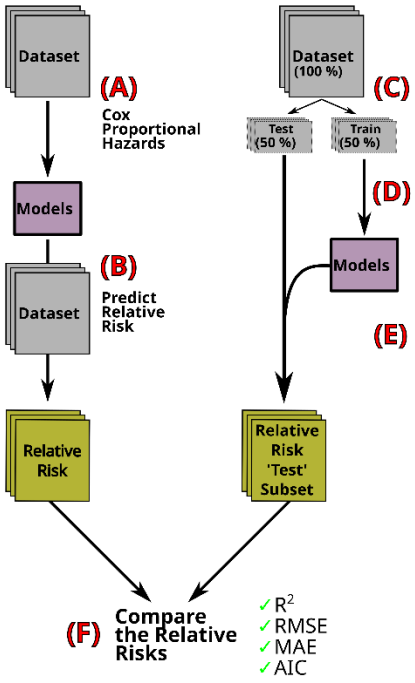


Source: authors' own elaboration.

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

The modelling effort for the late stage of the disease resulted in up to 15 models. For the Late-stage variable selection, relative risk was calculated for each patient using 15 models and used as the ground truth. The data were then split into training (n = 193) and testing (n = 192) subsets. New models were generated using the train set, and relative risks were calculated for patients in the test set. Model metrics such as RMSE, MAE, R^2, and AICs were calculated by comparing the ground truth and predictions from the testing set Table 8. A workflow is depicted in Figure 2. Results are shown at Table 5.

Figure 2. Workflow for assessing Late-stage models. Models were created using a set of variables as listed in Table 7 (A) using all patient's data (n = 385). Risk scores were calculated for all patients (B). The dataset was split into two subsets named 'Train' and 'Test' (C). Both sets were split into half (~50%/50%, n = 193/192 respectively) stratified by the factor 'Death'. New models were created using data from the 'Train' subset (D). Risk scores were calculated for each patient in the 'Test' subset (E). Finally, risk scores calculated at step B were compared for the same patients at step E (Table 5).



Source: authors' own elaboration.

Later, variable selection was tested with a screening using machine learning models using caret framework. Late-stage models' options were narrowed for two candidates (Model #5 and Model #13) (Figure 3).

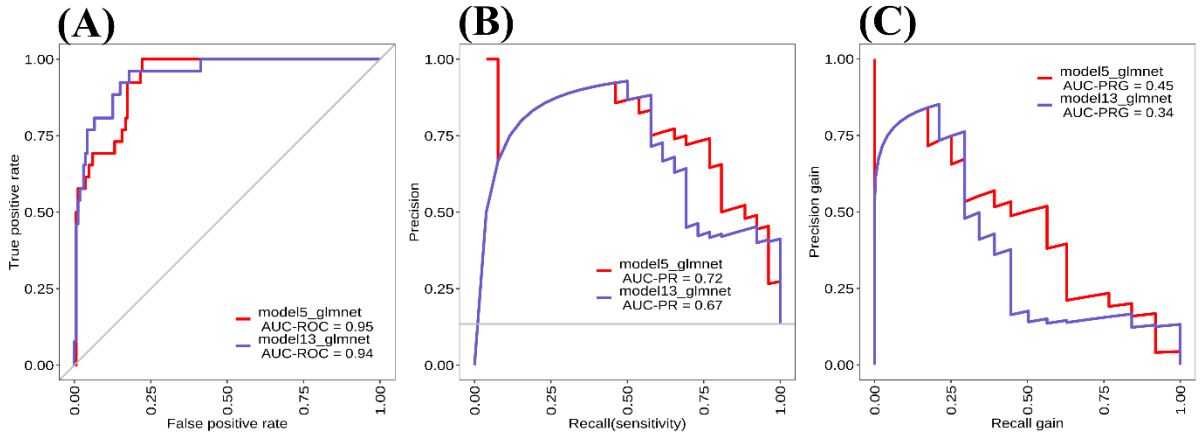
Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

Table 5. Metrics comparing relative risk calculated by train models with ones calculated with models built with the complete dataset.

Model #	Variables	R^2	RMSE	MAE	AIC
10	WC (Loge), Glicemia (Loge), AST (Loge)	0.951324	0.647082	0.253971	202.0144
9	Chronic kidney disease, WBC (Loge), Glicemia (Loge), AST (Loge)	0.923687	0.795856	0.30737	200.8062
8	WBC (Loge), Glicemia (Loge), Cardiovascular disease, AST (Loge)	0.947096	1.186644	0.417198	197.5699
13	LDH (Loge), RDW-CV (Square-root)	0.715595	1.680189	0.460104	190.4549
11	MPV (Loge), Lymphocytes (Box-Cox)	0.921783	3.425232	0.967792	183.5541
14	Direct bilirubin (Loge), RDW-SD (Square-root), Neutrophils (Loge)	0.883494	3.526217	0.499747	190.5306
12	Monocytes (Box-Cox), MPV (Square-root), RDW-SD (Loge)	0.974116	3.689849	0.671614	176.2663
5	Neutrophils (Square-root), RDW-SD (Square-root), LDH (Square-root), Obesity, Age	0.788557	4.628773	1.070254	173.0173
1	MPV, Total Bilirubin, Cardiovascular disease, Obesity	0.619826	4.730657	1.173511	177.3043
4	Neutrophils (Loge), LDH (Loge), Cardiovascular disease, Cronic kidney disease, Obesity	0.847539	5.092064	1.226934	180.9646
7	MCV (Square-root), Cardiovascular disease, Obesity, AST (Loge)	0.686565	5.267525	1.487993	186.2132
6	LDH (Box-Cox), Cardiovascular disease, Obesity, Glicemia (Loge), AST (Loge)	0.957716	5.703737	1.28384	183.7481
15	RDW-CV (Loge), MCV (Loge), MPV (Loge)	0.91672	7.984673	1.135523	181.8466
3	Lymphocytes (Box-Cox), Monocytes (Box-Cox), Cardiovascular disease, Obesity	0.742512	16.24481	3.608015	178.1428
2	Neutrophils, RDW-SD, MPV, Obesity, Age	0.921336	24.12355	4.500868	160.7398

Source: authors' own elaboration.

Figure 3. Receiver Operating Characteristic (ROC) curves (AUC-ROC), Precision-Recall (PR) curves (AUC-PR), and PR gain curves (AUC-PRG) outputs from logistic regression models using variable sets #5 and #13 are presented. The GLMnet models outperformed therefore others were removed from this graph.



Source: authors' own elaboration.

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

3.2 FINAL COX PROPORTIONAL HAZARDS MODELS

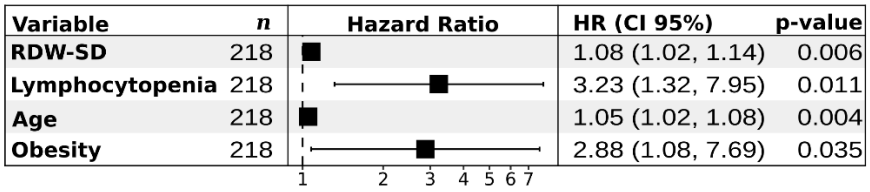
Both models were checked for Cox Proportional Hazards assumption violation with Schoenfeld's test and Power effect analysis. None of the models violates the assumptions and were considered for final decision. The coefficients were extracted and equations were deducted. For more details on Hazard Ratios and 95% confidence interval, see Forest Plots (Figures 4 & 5).

Early-stage *RR* (ERR) may be calculated using Equation 1 as follows:

$$ERR = \frac{e^{([0.0749 \times RDWSD] + [1.1735 \times Lymphocytopenia] + [0.048 \times Age] + [1.06 \times Obesity])}}{487}$$

(1), where RDW-SD (fL), Lymphocytopenia (1 if Lymphocytes' count lower than 1), Age (in years) and Obesity (Yes = 1).

Figure 4. Forest plots for the Early-stage model are displayed below. The first column contains variables, the second shows Hazard Ratio (HR) for each variable in each multivariate model, and the third column shows HR with confidence intervals and p-values.



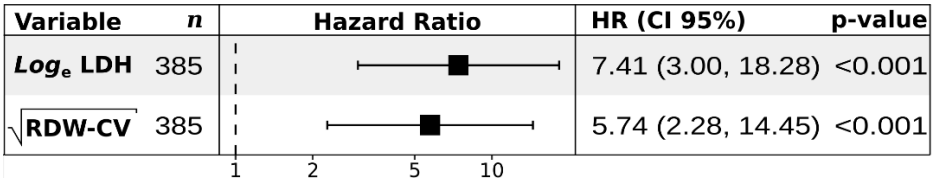
Source: authors' own elaboration.

The Late-stage *RR* (*LRR*) may be calculated using Equation 2 below:

$$LRR = \frac{e^{([2.002302 \times ln(LDH+1)] + [1.747124 \times \sqrt{RDWCV+1}])}}{676840037}$$

(2), where LDH (U/L) and RDW-CV (%). A higher relative risk (RR) implies a higher chance of the event (death) occurring in a shorter period.

Figure 5. Forest plots for the Late-stage model are displayed below. The first column contains variables, the second shows Hazard Ratio (HR) for each variable in each multivariate model, and the third column shows HR with confidence intervals and p-values.



Source: authors' own elaboration.

Additionally, power analysis was calculated for both models using a specific method for Cox Proportional Hazards. The Early-stage and Late-stage models present a Power effect of 91.28% and 97.8%, respectively. The correlation between covariates was low, implying controlled collinearity, if any and low variances (Table 6).

Revista Thema
V.24 | n.2 | 2025
p.1-17
DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

Table 6. Results from the power effect calculation for Cox Proportional Hazards models considering the present data.

	Early-stage model	Late-stage model
n	218	385
θ	$e^{1.199269}$	$e^{2.002302}$
α	0.05	0.05
Power	0.913	0.990
ρ^2	0.034	0.068
σ^2	0.250	0.146
Ψ	0.151	0.132

Source: authors' own elaboration.

4 DISCUSSION

To date, COVID-19 remains a significant global concern, requiring innovative responses to the escalating worldwide scenario caused by the novel coronavirus. The initial strategy proposed was social and physical distancing. Early on, Response Teams developed reports forecasting the rate of spread in several countries, highlighting the importance of high-quality public health data for data-driven policy and decision-making (Bertozzi et al., 2020; Ros et al., 2021; Subbian et al., 2021).

However, due to the lack of raw data readily available in public repositories, we only considered the data mentioned in the Methods section. Despite the good quality data available, this study has limitations. There are many missing biomarkers of higher relevance constantly referred to in studies, such as serum ferritin, CRP, and D-Dimer, as well as other demographic characteristics and hospitalization/ICU data (SpO2, mechanical ventilation etc.). Nonetheless, studies between 2020 and 2022 conducted in different hospitals worldwide have identified recurring variables similar to those adjusted in the models proposed in this study.

By fitting Cox Proportional Hazards models using different time ranges (5-8 days for Early-stage and the last follow-up day for Late-stage), we obtained different variables/models, which are in agreement with guidelines for patient admission/discharge (Stasi et al., 2020) and stages of the disease (Griffin et al., 2021). We developed two models to assist in the Early-stage and Late-stage of the disease. Both models were created independently using computer-aided variable removal, resulting in a different set of variables, which align with clinical characteristics previously reported for each scenario.

The Early-stage model was built using a time-frame of 5-8 days along with variables such as RDW, Lymphocytopenia, Age, and Obesity. These variables have been previously identified as important for risk assessment in COVID-19 patients by several studies (Terpos et al., 2020; Yuan et al., 2020; Zeng et al., 2020) and also Machine Learning approaches (Djakpo et al., 2020; Ma et al., 2020; Pan et al., 2020; Zheng et al., 2020).

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

The Late-stage model used the last lab data reported, which best fitted with RDW and LDH variables. Most studies that use haematological data highlight these two variables as important for patient evaluation (Ferrari et al., 2020; Guan et al., 2021; Vaid et al., 2020).

The Red Blood Cell distribution width (RDW), a variable commonly reported in complete blood count (CBC) exams, has been adjusted alongside other covariates in both models. RDW is useful in determining the aetiology of different types of anaemia is found to be higher in COVID-19 patients, especially in more severe cases. The anisocytosis may be due to the hyperinflammatory state caused by diverse cytokines, mainly IL-6, which can alter haematopoiesis and increase oxidative stress (Guaní-Guerra et al., 2022; Moreno-Torres et al., 2021).

Lymphocytopenia was evaluated for each patient and also fitted the model. It has been described as a marker for disease progression and an increase in severity (Stasi et al., 2020; Terpos et al., 2020; Yuan et al., 2020). It is unclear why COVID-19 patients develop immunosuppression related to cytokine storm but similar studies point that low level of lymphocytes on the first week of symptom's onset predicts SARS, therefore, treatments with immunomodulators might trigger lymphocytes multiplication (Zou et al., 2021).

Obesity has been found to increase the risk of both hospitalization and death, showing a linear relationship between BMI indexes and outcomes, especially in the early stage of the disease (Albashir, 2020; Sawadogo et al., 2022).

Lactate dehydrogenase (LDH) is a ubiquitous enzyme present in cells that catalyses the interconversion of pyruvate and lactate, playing a crucial role in cell metabolism. Its high presence in the serum is related to cell and tissue damage (Ali et al., 2022; Henry et al., 2020; Jurisc; Radenkovic; Konjevic, 2015). There are five different LDH isozymes expressed in various cell and organ types, with LDH-3 being prevalent in lung tissue. However, levels of LDH-3 did not increase significantly compared to other isoforms (Serrano-Lorenzo et al., 2021). This suggests that the serum increase of LDH is related to systemic cell damage rather than specific tissue or organ damage, given an increase in LDH-3 would mean pulmonary tissue damage. Nonetheless, LDH is found to be a strong predictor of mortality in the Late-stage model, which aligns with the Multisystem Inflammatory Phase, indicating the possibility of multiple organ failure (Griffin et al., 2021).

5. CONCLUSION

The COVID-19 Ranking Criteria (CRC-19) comprises two models to be used according to the first day of symptom onset. For up to 10 days after symptom onset, the Early-stage model is used to calculate relative risk. After 10 days or when presenting a critical condition, the patient may be in the Primary/Secondary Infection and Multisystem Inflammatory Phases, and the Late-stage model is recommended. Relative risks can be calculated daily to monitor disease progression. A higher relative risk implies an increased probability of an event occurring in a shorter time.

In conclusion, our study introduces a workflow for assessing the evolution of COVID-19. This workflow can be easily implemented using an electronic calculator, a spreadsheet processor, or a patient monitoring system in healthcare facilities. By incorporating variables such as age, lymphocytopenia, obesity, RDW-SD, and LDH, our models provide a method for calculating relative

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

risk at different stages of the disease. These models can assist healthcare professionals in patient follow-up and decision-making, ultimately improving patient care and outcomes.

6. DATA AVAILABILITY

Software Workspace backup containing Cox models, Machine Learning models and results in their original nature are available at the Author's GitHub <https://github.com/candidosobrinhosa/CRC-19> allowing audit, reproducibility, further improvement and use for educational and scientific purposes.

7. REFERÊNCIAS

ALI, Noman *et al.* Utility of biomarkers in predicting complications and in-hospital mortality in patients with COVID-19. **Pakistan Journal of Medical Sciences**, v. 38, n. 5, p. 1–6, 2022.

ALVES, Lise. Brazilian ICUs short of drugs and beds amid COVID-19 surge. **Lancet (London, England)**, v. 397, n. 10283, p. 1431–1432, 2021.

BERTOZZI, Andrea L. *et al.* The challenges of modeling and forecasting the spread of COVID-19. **Proceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 29, p. 16732–16738, 2020.

BOTTON, Jérémie *et al.* Reduced risk of severe COVID-19 in more than 1.4 million elderly people aged 75 years and older vaccinated with mRNA-based vaccines. **Vaccine**, v. 40, n. December 2020, p. 414–417, 2021.

BRASIL. **RESOLUÇÃO Nº 510, DE 07 DE ABRIL DE 2016**. Disponível em: <http://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf>. Acesso em: 17 fev. 2021.

COLLINS, Gary S. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. **BMC Medicine**, v. 13, n. 1, p. 1–10, 2015.

DAFALLAH ALBASHIR, Ahmed Abdalazim. The potential impacts of obesity on COVID-19. **Clinical Medicine, Journal of the Royal College of Physicians of London**, v. 20, n. 4, p. E109–E113, 2020.

DJAKPO, Dodji Kossi *et al.* Blood routine test in mild and common 2019 coronavirus (COVID-19) patients. **Bioscience Reports**, v. 40, n. 8, p. 1–5, 2020.

FERRARI, Davide *et al.* Routine blood tests as an active surveillance to monitor COVID-19 prevalence. A retrospective study. **Acta Biomedica**, v. 91, n. 3, p. 1–9, 2020.

FRENCH, Geoffrey *et al.* Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021. **MMWR. Morbidity and Mortality Weekly Report**, v. 70, n. 46, p. 1613–1616, 2021.

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

GARIBALDI, Brian T. *et al.* Patient trajectories among persons hospitalized for covid-19. **Annals of Internal Medicine**, v. 174, n. 1, p. 33–41, 2021.

GRIFFIN, Daniel O. *et al.* The importance of understanding the stages of covid-19 in treatment and trials. **AIDS Reviews**, v. 23, n. 1, p. 40–47, 2021.

GUAN, Xin *et al.* Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. **Annals of Medicine**, v. 53, n. 1, p. 257–266, 2021.

GUANÍ-GUERRA, Eduardo *et al.* Diagnostic Accuracy of the RDW for Predicting Death in COVID-19. **Medicina (Lithuania)**, v. 58, n. 5, p. 1–9, 2022.

HENRY, Brandon Michael *et al.* Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis. **American Journal of Emergency Medicine**, v. 38, n. 9, p. 1722–1726, 2020.

JOHN, Christopher R. **MLeval: Machine Learning Model Evaluation**. CRAN, 2020.

JONES, Aaron *et al.* External validation of the 4C mortality score among COVID-19 patients admitted to hospital in Ontario, Canada: a retrospective study. **Scientific Reports**, v. 11, n. 1, p. 4–10, 2021.

JURISIC, Vladimir; RADENKOVIC, Sandra; KONJEVIC, Gordana. The actual role of LDH as tumor marker, biochemical and clinical aspects. **Advances in Experimental Medicine and Biology**, v. 867, p. 115–224, 2015.

KASSAMBARA, Alboukadel; KOSINSKI, Marcin; BIECEK, Przemyslaw. **survminer: Drawing Survival Curves using 'ggplot2**. CRAN, 2021. Disponível em: <https://cran.r-project.org/package=survminer>. Acesso em: 14 mar. 2021

KENNEDY, Nick. **forestmodel: Forest Plots from Regression Models**. CRAN, 2020. Disponível em: <https://cran.r-project.org/package=forestmodel>. Acesso em: 20 out. 2023

KNIGHT, Stephen R. *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. **The BMJ**, v. 370, n. September, p. 1–13, 2020.

KUHN, Max. **caret: Classification and Regression Training**. CRAN, 2021.

LANINI, Simone *et al.* COVID-19 disease - Temporal analyses of complete blood count parameters over course of illness, and relationship to patient demographics and management outcomes in survivors and non-survivors: A longitudinal descriptive cohort study. **PLoS ONE**, v. 15, n. 12 December, p. 1–17, 2020.

LEVER, Jake; KRZYWINSKI, Martin; ALTMAN, Naomi. Points of Significance: Model selection and overfitting. **Nature Methods**, v. 13, n. 9, p. 703–704, 2016.

MA, Xuedi *et al.* Development and Validation of Prognosis Model of Mortality Risk in Patients with COVID-19. **Epidemiology and Infection**, 2020.

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

MIDDLEMAN, Amy B.; KLEIN, Judy; QUINN, Jane. Vaccine Hesitancy in the Time of COVID-19 : Attitudes and Intentions of Teens and Parents Regarding the COVID-19 Vaccine. p. 1–14, 2022.

MORENO-TORRES, Víctor *et al.* Red blood cell distribution width as a marker of hyperinflammation and mortality in COVID-19. **Annals of Palliative Medicine**, v. 0, n. 0, p. 0–0, 2021.

PAN, Pan *et al.* Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: Model development and validation. **Journal of Medical Internet Research**, v. 22, n. 11, 2020.

QIU, Weiliang *et al.* **powerSurvEpi: Power and Sample Size Calculation for Survival Analysis of Epidemiological Studies**. CRAN, 2021.

R CORE TEAM. **R: A language and environment for statistical computing**. Disponível em: <https://www.r-project.org/>. Acesso em: 19 fev. 2020.

ROB J. HYNDMAN; YEASMIN KHANDAKAR. Automatic Time Series Forecasting: The forecast Package for R. **Journal of Statistical Software**, v. 27, n. 3, p. 22, 2008.

ROS, Francisco *et al.* Addressing the Covid-19 pandemic and future public health challenges through global collaboration and a data-driven systems approach. **Learning Health Systems**, v. 5, n. 1, p. 1–12, 2021.

SAWADOGO, Wendemi *et al.* Overweight and obesity as risk factors for COVID-19-associated hospitalisations and death: systematic review and meta-analysis. **BMJ Nutrition, Prevention & Health**, v. 5, n. 1, p. 10–18, 2022.

SCHMITT, Peter; MANDEL, Jonas; GUEDJ, Mickael. A Comparison of Six Methods for Missing Data Imputation. v. 6, n. 1, p. 1–6, 2015.

SERRANO-LORENZO, Pablo *et al.* Plasma LDH: A specific biomarker for lung affectation in COVID-19? **Practical Laboratory Medicine**, v. 25, 1 maio 2021.

SHAKEEL, Choudhary Sobhan *et al.* Global COVID-19 Vaccine Acceptance: A Systematic Review of Associated Social and Behavioral Factors. **Vaccines**, v. 10, n. 1, p. 110, 12 jan. 2022.

STACKLIES, Wolfram *et al.* pcaMethods - A bioconductor package providing PCA methods for incomplete data. **Bioinformatics**, v. 23, n. 9, p. 1164–1167, 2007.

STASI, Cristina *et al.* Treatment for COVID-19: An overview. **European Journal of Pharmacology**, v. 889, n. July, p. 173644, 2020.

SUBBIAN, Vignesh *et al.* Ethics and informatics in the age of COVID-19: challenges and recommendations for public health organization and public policy. **Journal of the American Medical Informatics Association: JAMIA**, v. 28, n. 1, p. 184–189, 2021.

TERPOS, Evangelos *et al.* Hematological findings and complications of COVID-19. **American Journal of Hematology**, v. 95, n. 7, p. 834–847, 2020.

THERNEAU, Terry. **survival**. Rochester, MinnesotaCRAN, , 2021. Disponível em: <https://github.com/therneau/survival>. Acesso em: 24 mar. 2021

DOI: http://doi.org/10.15536/revistathema.24.2025.4002
ISSN: 2177-2894 (online)

VAID, Akhil *et al.* Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: Model development and validation. **Journal of Medical Internet Research**, v. 22, n. 11, p. 1–19, 2020.

VELAVAN, Thirumalaisamy P.; MEYER, Christian G. Mild versus severe COVID-19: Laboratory markers. **International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases**, v. 95, n. January, p. 304–307, jun. 2020.

VENABLES, WN; RIPLEY, BD. **Modern Applied Statistics with S**. Fourth edi ed. *[S.l.]*: Springer, 2002.

WHO. **Tracking SARS-CoV-2 variants**. Disponível em: <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>.

WICKHAM, Hadley *et al.* **dplyr: A Grammar of Data Manipulation**, 2021.

WOLKEWITZ, Martin; PULJAK, Livia. Methodological challenges of analysing COVID-19 data during the pandemic. **BMC Medical Research Methodology**, v. 20, n. 1, p. 4–7, 2020.

YEK, Christina *et al.* Risk Factors for Severe COVID-19 Outcomes Among Persons Aged ≥ 18 Years Who Completed a Primary COVID-19 Vaccination Series — 465 Health Care Facilities, United States, December 2020 – October 2021. v. 71, n. 1, p. 19–25, 2022.

YOSHIDA, Kazuki; BARTEL, Alexander. **tableone: Create "Table 1" to Describe Baseline Characteristics with or without Propensity Score Weights**. CRAN, , 2022. Disponível em: <https://cran.r-project.org/package=tableone>. Acesso em: 15 fev. 2021

YUAN, Xiaohong *et al.* Changes of hematological and immunological parameters in COVID-19 patients. **International Journal of Hematology**, v. 112, n. 4, p. 553–559, 2020.

ZENG, Furong *et al.* Can we predict the severity of coronavirus disease 2019 with a routine blood test? **Polish Archives of Internal Medicine**, v. 130, n. 5, p. 400–406, 2020.

ZHANG, Bicheng *et al.* Immune Phenotyping Based on the Neutrophil-to-Lymphocyte Ratio and IgG Level Predicts Disease Severity and Outcome for Patients With COVID-19. **Frontiers in Molecular Biosciences**, v. 7, n. July, p. 1–7, 2020.

ZHENG, Yichao *et al.* A Learning-Based Model to Evaluate Hospitalization Priority in COVID-19 Pandemics. **Patterns**, v. 1, n. 6, p. 100092, 2020.

ZOU, Zhi Ye *et al.* Persistent lymphopenia after diagnosis of COVID-19 predicts acute respiratory distress syndrome: A retrospective cohort study. **European Journal of Inflammation**, v. 19, 2021.

Submission: May 6, 2024

Accepted: October 23, 2025