Survival analysis & machine learning to investigate prognostic factors of overall survival, progression-free survival and perediction of severity of COVID-19

Análise de sobrevivência & inteligência artificial para investigação dos fatores prognósticos de sobrevida global, sobrevida livre de progressão e previsão de severidade de COVID-19

Alexandre de Fátima Cobre[1], Beatriz Böger[1], Mariana M. Fachi[1], Raquel O. Vilhena[2], Fernanda S. Tonin[1], Roberto Pontarolo[2]

[1]Pharmaceutical Sciences Postgraduate Program, Federal University of Paraná, Curitiba, Brasil; [2]Department of Pharmacy, Federal University of Paraná, Curitiba, Brazil.

Abstract

Objective – COVID-19 remains an emergency public health problem. We aimed to investigate the prognostic factors of Overall survival (OS), Progression-free survival (PFS) in patients with COVID-19 and develop two machine learning (ML) models to predict disease severity. **Methods** – A retrospective cohort study was performed including patients with a positive diagnostic for COVID-19 in Mexico (January-November 2020). Cox regression analysis was used to investigate the prognostic factors of OS and PFS. Artificial neural network (ANN) and decision tree (TD) models were also developed to predict disease severity. **Results** – We included 95,839 patients. The medians OS and PFS were, 9.087 days (8.887-9,287) and 8.233 (8.036-8.430), respectively. Women aged >40 years, intubated patients and previous comorbidities were significantly associated with shorter OS and PFS times. The ANN and DT models were able to predict the severity of COVID-19, with an accuracy (area under the ROC curve) of 99.5% and 99.8%, respectively. The DT model showed better performance, because it showed less training time and less prediction error. **Conclusion** – In Mexico, elderly women that needed intubation and had comorbidities were associated with a worse prognosis of OS and PFS for COVID-19. The DT model could be used as a tool to predict the hospitalization of COVID-19 outpatients.

Descriptors: Coronavirus, comorbidities, Cox regression, Artificial neural network, Decision tree, severity

Resumo

Objetivo – COVID-19 continua sendo um problema emergencial de saúde pública. Nosso objetivo foi investigar os fatores prognósticos de sobrevida global (OS), sobrevida livre de progressão (PFS) em pacientes com COVID-19 e desenvolver dois modelos de aprendizado de máquina (ML) para prever a gravidade da doença. **Métodos** – Um estudo de coorte retrospectivo foi realizado incluindo pacientes com diagnóstico positivo para COVID-19 no México (janeiro a novembro de 2020). A análise de regressão de Cox foi usada para investigar os fatores prognósticos de OS e PFS. Rede neural artificial (RNA) e modelos de árvore de decisão (DT) também foram desenvolvidos para prever a gravidade da doença. **Resultados** – Incluímos 95.839 pacientes. As medianas de OS e PFS foram, 9,087 dias (8,887-9,287) e 8,233 (8,036-8,430), respectivamente. Mulheres com idade> 40 anos, pacientes intubados e comorbidades anteriores foram significativamente associados a tempos de OS e PFS mais curtos. Os modelos ANN e DT foram capazes de predizer a severidade da COVID-19, com acurácia (área sob a curva ROC) de 99,5% e 99,8%, respectivamente. O modelo DT apresentou melhor desempenho, pois apresentou menor tempo de treinamento e menor erro de predição. **Conclusão** – No México, mulheres idosas que precisaram de intubação e tinham comorbidades foram associadas a um pior prognóstico de OS e PFS para COVID-19. O modelo DT pode ser usado como uma ferramenta para prever a hospitalização de pacientes ambulatoriais com COVID-19.

Descritores: Coronavírus, Comorbidades, Regressão de Cox, Rede neural artificial, Árvore de decisão, Gravidade

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), caused by coronavirus 2019 (COVID-19), was originally reported in China (December 2019), and now is considered a global threat. It has already caused 2.1 million deaths worldwide, with more than 97.3 million infected people[1–3].

The real-time PCR exam is the gold standard for the diagnosis of COVID-19[4]. Patients with COVID-19 present with fever and cough, and a minority have chest discomfort or difficulty breathing, diagnosed as pneumonia by chest X-rays or computed tomography (CT)[5,6]. Data suggest that advanced age and comorbidities play a vital role in the severity of the disease and on negative clinical outcomes[3,5]. Patients with previous cardiovascular metabolic comorbidities may have a worse pneumonia prognosis, as already demonstrated in previous studies assessing Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome coronavirus (MERS) viruses[5,7–9]. In a trial evaluating 637 MERS-CoV patients, diabetes and hypertension were prevalent in about 50% of patients, and heart diseases were reported in 30% of cases[9]. Diabetes is already considered a risk factor for both mortality and morbidity in patients with SARS and COVID-19[8,10], and obesity has been associated as a significant risk for mortality in H1N1 influenza cases. These data suggest the need for further monitoring and more tailored treatment for these patients[11,12].

The delay in the diagnosis of the disease, the lack of hospital beds caused by overcrowding of the units,

are also one of the main factors that contribute to the increase in the number of deaths from the disease[13,14].

Mexico currently has 694,121 confirmed cases of the disease and more than 73,258 deaths[1]. In 2019, the country, that accounts for more than 60 cities, received more than 15 million visitors by air transport. This condition may have aggravated the spread of the virus, especially considering the high connectivity between regions and due to the high population density of most of the territories[15].

With the increase in the number of cases of COVID-19, especially in regions where healthcare resources are insufficient, individuals with comorbidities are susceptible to a worse prognosis. To date, little is known of the main risk factors for intensive care unit (UCI) admission and fatal outcomes associated with the infection in patients with further comorbidities living in underdeveloped countries[3,11,12]. Based on this, the present study has two objectives: the first objective was to investigate the factors that regulate (increase or decrease) the time from diagnosis to death from COVID-19 or other causes (overall survival) and the time from the beginning from treatment to disease progression (progression-free survival), using methods of survival analysis (Kaplan-Meier method, Cox Regression, Log rank test, Breslow test, Tarone-Ware test and Hazard ratio); The second objective of the study was to develop a machine learning model of artificial neural network (ANN) and decision tree (DT) to predict the severity of COVID-19.

Methods

Study design and data collection

A retrospective study was carried out involving patients with a positive diagnosis for COVID-19 in Mexico (Latin America). Patient data were collected in the open database of the Ministry of Health of the Government of Mexico that included patients registered from January 8 to November 4, 2020[16]. Patients with a negative diagnosis for the disease were excluded from the study. The study did not require an ethics committee to be carried out.

Statistical analysis

Categorical variables as presented as frequency. Continuous variables that did not have a normal distribution by the Kolmogorov-Smirnov test are presented as medians and interquartile ranges (IQR), and those that showed normal distributions are presented as means and standard deviations (SD).

Survival analysis

Overall survival (OS) was considered to be the primary outcome. The OS time (registered in days) was estimated from the date of diagnosis of COVID-19 until the patient's death by the disease or by other causes and was classified as the dependent variable. Progression-free survival was considered to be the secondary outcome. The PFS time (also registered in days) was estimated from the date of the beginning of the treatment until the progression of the disease (hospitalization in an intensive or semi-intensive care unit) or the death of the patient and was classified as the dependent variable.

The independent variables used for the analysis of OS and PFS were: patient sociodemographic characteristics (age, sex and location/state), type of admission (outpatient or inpatient), inpatient in ICU, need for intubation, presence of pneumonia, presence of comorbidities (i.e. diabetes chronic obstructive pulmonary disease (COPD), asthma, immunodepression, hypertension, cardiovascular disease and obesity chronic kidney disease (CKD)).

The OS and PFS times were estimated using the Kaplan-Meier method. The statistical comparison of the Kaplan-Meier survival curves was performed using the Log Rank test (beginning of follow-up) , Breslow test (middle of follow-up) and Tarone-Ware (end of follow-up)[17–19]. The investigation of prognostic factors associated with the OS and PFS of COVID was performed using univariate and multivariate analysis of time-dependent covariate Cox regression. In order to minimize possible risks of bias in OS and PFS prognostic factors, a sensitivity analysis of the time-dependent covariate Cox regression model controlling for age was additionally performed, in which a high risk group model (patients with advanced age) and a low-risk group model (young patients) were constructed and prognostic factors were compared. Hazard ratios (HR) with 95% confidence intervals (CI) were used to quantify the size of the effect of the time-dependent covariate Cox regression analyses. Statistical analyses were performed using SPSS® Software version 20, and $p<0.05$ was considered statistically significant.

Development of machine learning models: artificial neural network and decision tree

The independent variables used to build the artificial neural network and decision tree models were: patient sociodemographic characteristics (age, sex and location/state), type of admission (outpatient or inpatient), inpatient in ICU, need for intubation, presence of pneumonia, presence of comorbidities (i.e. diabetes chronic obstructive pulmonary disease (COPD), asthma, immunodepression, hypertension, cardiovascular disease and obesity chronic kidney disease (CKD)).

The dependent variable was the severity of the disease (binary variable) in which patients were classified into: severe disease and non-severe disease. Patients with non-severe disease were considered, those patients who were treated on an outpatient basis, while patients with severe disease were considered, those hospitalized in an intensive or semi-intensive care unit. For the construction of the models, 70% of the samples for the training set (calibration) and 30% of the samples for

the test set were used and these samples were selected using the Kenran-Stone algorithm[20].

The ANN and DT machine learning models were validated using the following parameters: specificity, sensitivity and accuracy, using ROC curves (receiver operator characteristic curve). In machine learning, a sample is considered to be true negative (TN), when it is correctly classified by the ML model as being negative, given that the person is free of disease; whereas, a sample is considered false negative (FN), when it is wrongly classified by the ML model as negative, given that the patient has the disease. True positive sample (TP) is that correctly classified by the model as positive, given that the person has the disease; whereas a false positive sample (PF) is one that is wrongly classified as positive, since the person is free from the disease[21–23].

In machine learning models, sensitivity is the model's ability to correctly classify true positive samples and is calculated using the formula in equation 1:

$$\text{Sensitivity} = TP/(TP+FN) \quad (1)$$

Specificity is the ability of an ML model to correctly classify true negative samples, and is calculated by the following equation 2:

$$\text{Specificity} = TN/(TN+FP) \quad (2)$$

Accuracy is the ML model's ability to correctly classify both samples (negative and positive), and express and model efficiency. Accuracy is obtained by integrating the area under the ROC curve or applying the formula in equation 3:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FN+FP) \quad (3)$$

Results

Patient characteristics

A total of 95,839 patients with a positive diagnosis of COVID-19 were included in the study during the evaluated period. The median age was 41 years (IQR, 30-53 years) and more than half of the patients were men (50.8%, n = 48720). The data on the clinical characteristics of the patients are shown in table 1.

Survival analysis

In the present study, 27% (n = 25867) of the patients had disease progression, of which 3% (n= 3435) died. The median OS and PFS were, 9,087 days (8,887-9,287) and 8,233 (8,036-8,430), respectively. The Kaplan-Meier curves of OS and PFS are shown in figure S1 (supplementary material).

Young patients aged <40 years had significantly the highest OS (9.172 vs 8.412 days) and PFS (8.525 vs 6.929) times when compared to other age group (log rank test, p = 0.001). Male patients had greater OS (9.302 vs 8.691 days) than females (Log Rank, p<0.001; Breslow, p<0.001 and Tarone-Ware; p<0.001) (Table S1).

Patients admitted to the hospital had shorter OS (10.247 vs 8.980 days) and PFS (10.247 vs 7.891 days) times when compared to outpatients (Log Rank, p<0.001; Breslow, p<0.001 and Tarone-Ware; p<0.001). Moreover, intubated patients (OS = 8.7 days; PFS=7.964 days) and patients that did not attend the ICU (OS = 8.798 days) also survived less (p<0.0001). Patients presenting comorbidities had significantly (Log rank, p<0.005, Breslow, p<0.005 and Tarone-Ware, p<0.005) shorter OS and PFS times compared to healthy individuals. OS and PFS rates among groups were not different for other clinical variables asthma and pneumonia (Table S1).

Prognostic factors of OS and PFS: multivariate and univariate analysis of Cox regression time-dependent covariate

The multivariate analysis of Cox regression covariable time-dependent demonstrated that the factors that were associated with the worst OS prognosis were: patients aged > 40 years (HR = 1.189 ; CI95% [1.066-1.325]; p = 0.002), Hospitalized (HR = 1.252; CI95% [1.107-1.415]; p = 0.000), diabetis (HR = 1.125; CI95% [1.045-1.211]; p = 0.002), COPD (HR = 1.192; CI95% [1.057-1.344]; p = 0.004), cardiovascular disease (HR = 1.131; CI95% [1.002-1.278]; p = 0.047). However, univariate analysis showed that female (HR = 1.093; CI95% [1.019-1.173]; p = 0.012), intubated patient (HR = 1.107; CI95% [1.021-1.201]; p = 0.014), hypertension (HR = 1,116; CI95% [1,042-1,195]; p = 0.002) and patient with CKD (HR = 1,481; CI95% [1,204-1,823]; p = 0.000) were also associated with the worst OS prognosis (Table 2).

The multivariate analysis showed that the factors that were associated with the worst prognosis for progression-free survival were: female (HR = 1.110; CI95% [1.038-1.188]; p = 0.002), >40 years (HR = 1.385; CI95% [1.261-1.521]; p = 0.000), Intubated patient (HR = 1.193; CI95% [1.098-1.296]; p = 0.000), Diabetis (HR = 1.081; CI95% [1.008-1.160]; p = 0.003), COPD (HR = 1.171; CI95% [1.044-1.313]; p = 0.007) and cardiovascular disease (HR = 1.139; CI95% [1.015-1.278]; p = 0.027). However, univariate analysis showed that patient with CKD (HR = 1.093; CI95% [1.019-1.173]; p = 0.012) were also associated with the worst PFS prognosis (table 2).

Sensitivity analysis

The sensitivity analysis of Cox's model was performed by controlling for age, in which high-risk patients (> 40 years) and low-risk patients (<40 years) were grouped. The results of the sensitivity analysis for overall survival showed that no factor was associated with a worse OS prognosis in patients under the age of 40 years. For patients over 40, the results were similar to the global

Tabela 1. Sociodemographic and clinical characteristics of patients with COVID-19

Characteristics		Frequency (n)	Percentage (%)
Age (years)	<40 years	45419	47,4
	>40 years	50420	52,6
Gender	Female	47119	49,2
	Male	48720	50,8
Type of patient	Ambulatory	70268	73,3
	Hospitalized	25571	26,7
	Intubated patient	1934	2,0
	Patients with pneumonia	17628	18,4
	Diabetic patients	12878	13,4
	Patient with COPD	2462	2,6
	Patient with asthma	4328	4,5
	Immunodepressed patients	2314	2,4
	Hypertensive patients	16716	17,4
	Cardiovascular disease	2986	3,1
	Obese patient	15597	16,3
	Patient with CKD	2287	2,4
	Smoking patient	9311	9,7
	Other complications	4642	4,8
	ICU patient	2180	2,3
	Deaths	3435	3,6

Note: COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; CKD, chronic kidney disease

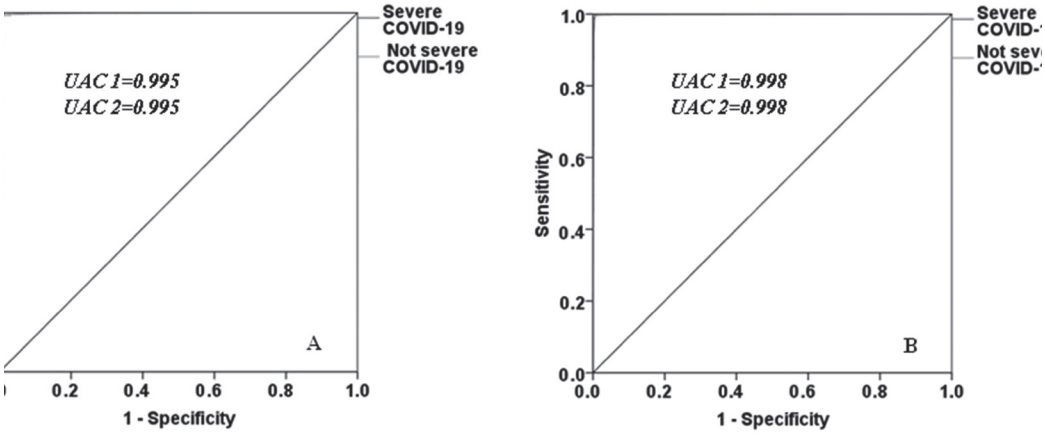

Figura 1. ROC curves of the accuracy of machine learning models. Artificial neural network model (A) and decision tree model (B). *UAC1*- Area under the curve of patients with severe disease. *UAC2*- Area under the curve of patients with non-severe disease

Cox regression results shown in Table 2. This shows that advanced age over 40 years is the most important factor in the prognosis of OS (Table S2 in supplementary material)

In the global Cox regression model, asthma was not associated with the PFS prognosis, however the sensitivity analysis for PFS showed that asthma is a factor associated with the worst PFS prognosis in the group of patients younger than 40 years (HR = 1.568; 95% CI [1.073-2,291]; p = 0.002) (Table S2 in supplementary material).

Machine learning models: general considerations

For the calibration (training) of the ML models (ANN and DT), 12933 samples from patients with non-severe disease and 12933 samples from patients with severe disease were used, totaling 25866 samples belonging to the training set. For model the prediction (test), 12933 samples from patients with non-severe disease and 12933 from patients with severe disease were also used, totaling 25866 samples from the test set.

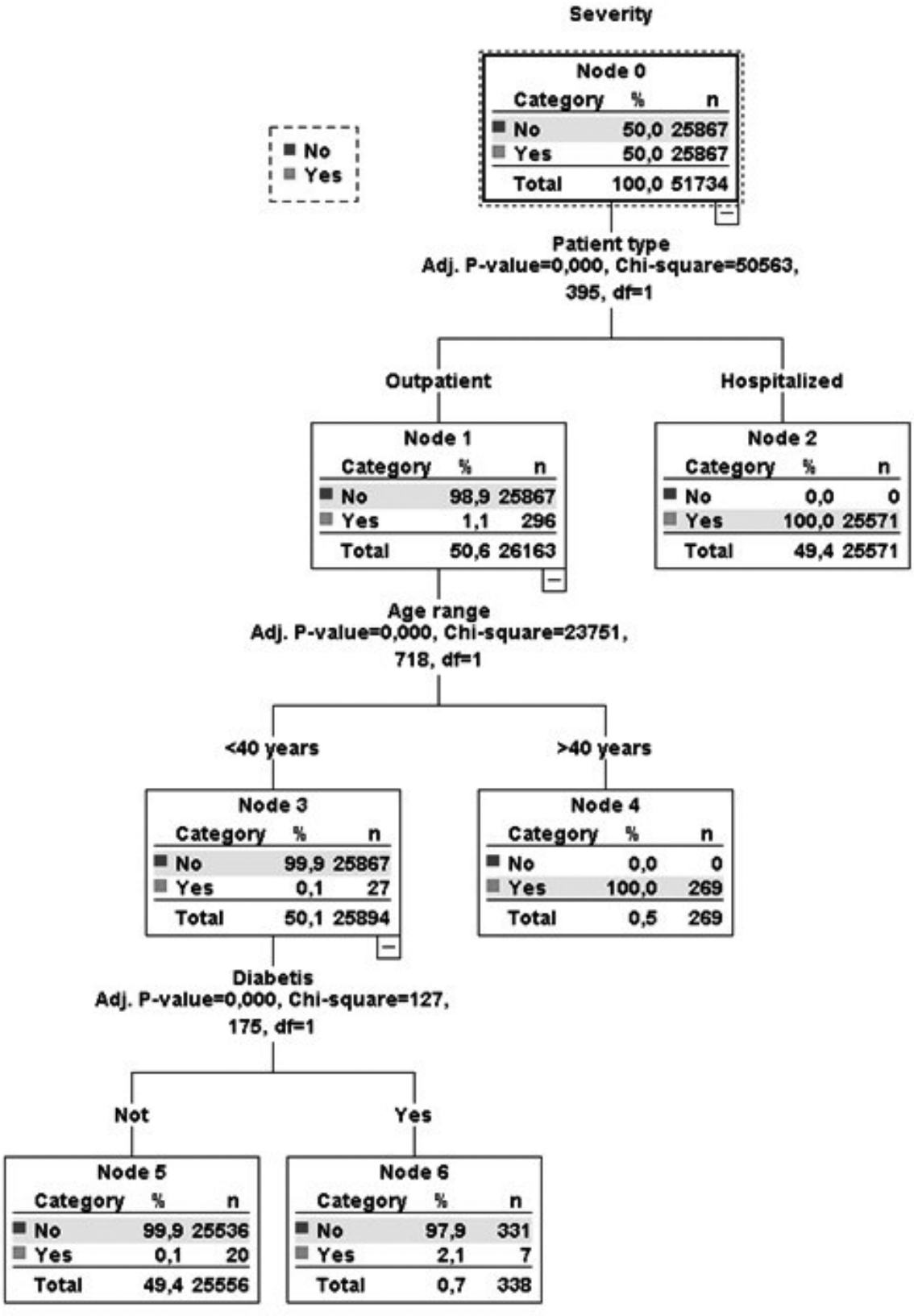

Figura 2. Decision tree model for predicting COVID-19 severity

Tabela 2. Results of the univariate and multivariate analysis of Cox regression covariable time dependent on the prognostic factors associated with overall survival and progression free-survivall of COVID-19 between January to November 2020 in Mexico

Covariates	Overall survival (OS)						Progression free survival (PFS)					
	Univariate			Multivariate			Univariate			Multivariate		
	HR	95CI	p	HR	95CI	p	HR	95CI	p	HR	95CI	p
Female	1.1	1.0-1.2	0.012	1.1	0.9-1.1	0.105	1.1	1.1-1.2	0.000	1.1	1.0-1.2	0.002
>40 years	1.1	1.0-1.2	0.042	1.2	1.1-1.3	0.002	1.3	1.2-1.4	0.000	1.4	1.2-1.5	0.000
Hospitalized	1.2	1.1-1.4	0.002	1.2	1.1-1.4	0.000						
Intubated patient	1.1	1.0-1.2	0.014	0.9	0.8-1.1	0.495	1.2	1.1-1.3	0.000	1.2	1.0-1.3	0.000
Pneumonia patient	0.9	0.8-1.0	0.310	1.0	0.9-1.1	0.586	1.0	0.9-1.1	0.323	1.0	0.9-1.1	0.586
Diabetes patients	1.2	1.1-1.2	0.000	1.1	1.0-1.2	0.002	1.1	1.0-1.2	0.023	1.1	1.0-1.2	0.030
COPD	1.2	1.1-1.4	0.001	1.2	1.1-1.3	0.004	1.2	1.1-1.3	0.002	1.1	1.0-1.3	0.007
Asthma	0.9	0.8-1.2	0.697	1.1	0.9-1.4	0.281	1.1	0.9-1.3	0.269	1.0	0.9-1.2	0.824
Patients with ID	1.1	0.9-1.3	0.276	1.0	0.9-1.1	0.975	1.0	0.9-1.2	0.543			
Hipertension	1.1	1.0-1.2	0.002	0.9	0.8-1.0	0.100	1.1	0.9-1.1	0.096	1.0	0.9-1.1	0.277
Patient with CD	1.2	1.1-1.4	0.001	1.1	1.0-1.3	0.047	1.2	1.1-1.3	0.002	1.1	1.0-1.3	0.027
Obese patient	1.0	0.9-1.1	0.847	1.0	0.9-1.3	0.926	0.9	0.9-1.0	0.242	0.9	0.8-1.0	0.154
Patient with CKD	1.5	1.2-1.8	0.000	0.7	0.3-1.9	0.512	1.3	1.1-1.5	0.005	1.2	1.0-1.5	0.118
Smoking patient	1.1	0.9-1.2	0.062	1.1	0.9-1.2	0.154	1.0	0.9-1.2	0.145	1.1	0.9-1.2	0.158
ICU patient	1.3	1.2-1.4	0.000	1.3	1.2-1.5	0.000	.....	.....	.....	.....	.....	.....

Note: HR, hazard ratio; CI, confidence interval; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit; CKD = chronic kidney disease

Artificial neural network model

The ANN model developed was *multilayer percetron*. The input layer of the neural network was formed by 27 neurons (categories of predictor variables), the hidden layer was formed by 8 neurons, and the output layer was formed by two neurons (outcome variables: severe or non-severe disease). The neurons activation functions of the hidden layer and the output layer were *hyperbolic tamgent* and *softamx*, respectively. The model's error function was entropy-error.

For training the ANN model, 2 minutes and 13 seconds were used. The calibration error was 4.1%. Of the 25,866 samples from patients with non-severe disease, 25,836 were correctly classified, and 30 samples were wrongly classified (false positive samples) representing a specificity (percentage of correctness) of 99.8%. Of the 25,866 samples from patients with severe disease, 25676 samples were correctly classified by the ANN model as being samples from patients with severe disease, and 191 samples were wrongly classified (false negative samples), representing a sensitivity (percentage of correctness) of 99.2%. The accuracy of the model was calculated using the ROC curve, where the area under the curve (UAC) calculated for the prediction of negative samples and positive samples were both 99.5%, both for the severe disease group and for the group of non-severe disease, which proves the good efficiency of the model in predicting severity (figure 1A).

The type of patient (inpatient or outpatient), the patient's age, having pneumonia, were the three main most impacting variables in predicting severity by the ANN model, in which they had a normalized importance of 100%, 59.9% and 18.7%, respectively.

Decision tree model

For training the DT model, 1 minutes and 49 seconds were used. The calibration error was 2.3%. The DT model was built using seven (7) nodes, four (4) terminal nodes and three (3) depths. The DT model showed that the severity of COVID-19 was associated with: (i) outpatients ($p < 0.001$, chi-squared = 50563.395; df=1) older than 40 years ($p<0.001$; chi-squared= 23751.718, df=1); (ii) or outpatients under 40 years of age suffering from diabetes ($p<0.001$, chi-squared=127.175, df=1), as can be seen in figure 2.

Among 25,866 patients with severe disease, 25,840 patients were correctly classified, representing a sensitivity (percentage of correctness) of the model of 99.8%. All 25867 samples of patients with non-severe disease were correctly classified representing a 100% correctness percentage (specificity). The model's accuracy (calculated using equation 3) was 99%. The area under the ROC curve was 99.8% for both classes (severe disease and non-severe disease), which proves the good performance of the model (figure 1b).

Discussion

In this study we applied independent models of Cox regression and machine learning (artificial neural networks and decision tree) to predict the prognosis of patients with COVID-19.

There is an increasing tendency to apply machine learning models to predict Covid-19 prognosis, and several studies are available in the literature[24–26]. Although the existence of several machine learning models (for example, principal component analysis, artificial neural networks, decision tree, random forest),

the performance of these models depends on the available sample size. The larger the sample size, the more robust the machine learning model[27–30]. As an example, it is artificial neural networks multilayer or artificial neural networks deep learning that require a very large sample size for your training[29–31].

In the case of COVID-19, several studies have been conducted using a small sample size, which compromises the performance of the models[32,33].

For example, Hu (2020) used machine learning to predict the severity of COVID-19 using only 183 patients (115 survivors and 68 deaths), whose sensitivity and specificity were 83.9% and 79.4%, respectively[34]. The values of sensitivity and specificity of study Hu (2020), are lower than those found in our study, which were both 99%; the reason for this difference can be explained by the fact that our study used a large sample size (n = 25,867) than Hu's study (n = 183), which makes our models more robust.

A study conducted by Xu (2021), who also developed several machine learning models to predict the severity of COVID-19, the best model (decision tree) showed an accuracy similar to that found in our study, which was 99%[35]. Despite this similarity, there is a difference in the sample size used for the development of the models (n = 659 patients in the study by Xu and 25,867 patients in our study), which obviously also compromises the performance of the models.

In this study, the majority of patients infected with COVID-19 in Mexico had a mean age of 41 years and were men. We confirmed the association of increased age with higher rates of death by COVID-19. Previous studies evaluating monkeys inoculated with SARS-CoV demonstrated that older animals had a stronger response to virus infection than younger adults, with an increase in the expression of genes associated with inflammation, while the expression of type I interferon-β was reduced[36,37]. Moreover, COVID-19 is associated with exacerbated release of cytokines by the overactivated immune system, as in SARS and MERS. Aberrant and excessive immune responses lead to long-term lung function and structure damage in patients released from the ICU[38,39]. Recent prospective and retrospective human cohort studies conducted in China (Wuhan) and the United States (New York and Michigan) have also found an association between advanced age and increased mortality and a worse prognosis of the disease[5,36,40–42].

We also found that some pre-existing clinical conditions including pneumonia, diabetes, hypertension, cardiovascular disease and CKD were associated with shorter DSS in Mexico. In other studies, these comorbidities were also prevalent in patients admitted to the ICU, with rates varying from 17.1% to 18.6% for hypertension, 1.9% for chronic obstructive pulmonary disease, 3% for chronic kidney disease, 9.7-11.9% for diabetes and 14.4-16.4% for cardiovascular disease[3,5,12,43]. These data suggest that these specific clinical conditions are important risk factors for poor prognosis in COVID-19. In a recent systematic review of epidemiological data, the authors found a 36.8% rate of patients with comorbidities (mean age of 51.97 years; majority males) with the most significant being hypertension, cardiovascular disease and diabetes[3]. In a study by Guo *et al*. (2020)., 80% of patients hospitalized in the United States for COVID-19 were 65 years of age or older; around one third (28.0%) had some coronary event, and were almost twice as likely to die compared to those patients without a cardiovascular events[7]. A study conducted in China found that diabetes was associated with a three-fold increase in the risk of death in patients with COVID-19[10].

Other studies also stress that obesity may be a reason for the increase in mortality among infected patients, given its association with higher hospitalization rates and need for mechanical ventilation[11,44]. Obesity is related to a decrease in the volume of expiratory reserve, functional capacity and compliance of the respiratory system. In addition, in patients with increased abdominal obesity, lung function is further compromised. There is also an increase in inflammatory cytokines, which may contribute to the increase in obesity-associated morbidity in COVID-19[11,45]. Although the effects of COVID-19 in these patients are still not well-established, caution in their care is needed. Previous experience with H1N1 influenza in California in 2009 showed that 58% of patients who were hospitalized or died due to the infection were obese (BMI ≥ 30), of which 67% were severely obese (BMI ≥ 40). Most people (66%) also had underlying diseases, such as chronic lung disease, including asthma, heart problems or diabetes[11,44]. In that same year, of all patients hospitalized in New Mexico for H1N1 that needed mechanical ventilation, around 56% were obese[11,45].

Concomitant respiratory diseases play an important role in patients diagnosed with COVID-19, because these conditions contribute to enhance the severity of the respiratory symptoms. Death rates in these patients that may reach 6.3%, which is three times higher compared to healthy patients (rates around 2.3%)[46]. A study demonstrated that patients previously diagnosed with pneumonia are more likely to die due COVID-19[47]. In our analyses, we did not find a clear association between COPD or asthma and patient survival. However, in Mexico, the rate of individuals with these comorbidities was low (around 3%), which may explain this situation. Studies have shown that asthmatic patients have a type 2immune response, with the accumulation of eosinophils and involvement of type 2 cytokines (e.g. IL-4, IL-13), which may have protective effects against COVID-19 infection. Asthma treatment also involves the use of corticosteroids, monoclonal anti-IgE antibody and allergic immunotherapy, which are drugs that significantly contribute to the prevention of SARS-VOC-2 infection, improve immune defense or reduce inflammation[48].

The great limitation of our study is the fact that we have given retrospectives, which may not reflect the real scenario. Prospective studies are needed.

In Mexico, elderly women patients that were admitted to a hospital, needed intubation and had comorbidities (e.g. diabetes, pneumonia or hypertension) were associated with a worst prognosis for OS and PFS of COVID-19. The ANN and DT models were able to predict the severity of COVID-19, both models with an accuracy greater than 80%, which is the minimum recommended in most machine learning models. The results of survival analysis highlight the need for more tailored treatments towards other diseases besides the viral infection. Patients should be monitored according to their healthcare limitations. The Machine Learning models developed could be used as a tool to support the clinical team in making decisions regarding the need for hospitalization of outpatients with COVID-19.

References

1. World Health Organization. Donate WHO Coronavirus Disease (COVID-19) Dashboard. 2021.

2. Ahn DG, Shin HJ, Kim MH, Lee S, Kim HS, Myoung J, *et al*. Current status of epidemiology, diagnosis, therapeutics, and vaccines for novel coronavirus disease 2019 (COVID-19). J Microbiol Biotechnol 2020;30:313–24. https://doi.org/10.4014/jmb.2003.03011.

3. Rodriguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holguin-Rivera Y, Escalera-Antezana JP, *et al*. Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. Travel Med Infect Dis 2020;34:101623. https://doi.org/10.1016/j.tmaid.2020.101623.

4. Böger B, Fachi MM, Vilhena RO, Cobre AF, Tonin FS, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. Am J Infect Control 2021;49:21–9. https://doi.org/10.1016/j.ajic.2020.07.011.

5. Li B, Yang J, Zhao F, Zhi L, Wang X, Liu L, *et al*. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clin Res Cardiol 2020;109:531–8. https://doi.org/10.1007/s00392-020-01626-9.

6. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al*. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 2020;382:727–33. https://doi.org/10.1056/nejmoa2001017.

7. Guo T, Fan Y, Chen M, Wu X, Zhang L, He T, *et al*. Cardiovascular Implications of Fatal Outcomes of Patients with Coronavirus Disease 2019 (COVID-19). JAMA Cardiol 2020;5:811–8. https://doi.org/10.1001/jamacardio.2020.1017.

8. Badawi A, Ryoo SG. Prevalence of comorbidities in the Middle East respiratory syndrome coronavirus (MERS-CoV): a systematic review and meta-analysis. Int J Infect Dis 2016;49:129–33. https://doi.org/10.1016/j.ijid.2016.06.015.

9. Yang JK, Feng Y, Yuan MY, Yuan SY, Fu HJ, Wu BY, *et al*. Plasma glucose levels and diabetes are independent predictors for mortality and morbidity in patients with SARS. Diabet Med 2006;23:623–8. https://doi.org/10.1111/j.1464-5491.2006.01861.x.

10. Wu Z, McGoogan JM. Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention. JAMA - J Am Med Assoc 2020;323:1239–42. https://doi.org/10.1001/jama.2020.2648.

11. Dietz W, Santos-Burgoa C. Obesity and its Implications for COVID-19 Mortality. Obesity 2020;28:1005. https://doi.org/10.1002/oby.22818.

12. Muhammad F. Association Between Age, Sex, and Pre-Existing Health Conditions and Death of COVID-19 Patients. Ann Mil Healh Sci Res 2020;18. https://doi.org/10.5812/amh.104270.

13. Cobre AF, Böger B, Vilhena RO, Fachi MM, Santos JMMF, Tonin FS. Uma análise multivariada dos fatores de risco associados à morte por Covid-19 em EUA, Itália, Espanha e Alemanha. Z Gesundh. 2022;30(5): 1189-95. doi: 10.1007/s10389-020-01397-7.

14. Cobre AF, Böger B, Fachi MM, Vilhena RO, Domingos EL, Tonin FS, *et al*. Risk factors associated with delay in diagnosis and mortality in patients with Covid-19 in the city of Rio de Janeiro, Brazil. Cienc e Saude Coletiva 2020;25:(suppl 2). https://doi.org/10.1590/1413-812320202510.2.26882020.

15. Díaz-Quiñonez JA. Emergence of novel coronavirus SARS-CoV2 in China and the response in Mexico. Gaceta Med México. 2020; 156(2). doi: 10.24875/GMM.M20000345.

16. Mexico. Database of the Ministry of Health of the Government of Mexico 2020 2020.

17. Wakounig S, Heinze G, Schemper M. Non-parametric estimation of relative risk in survival and associated tests. Stat Methods Med Res 2015;24:856–70. https://doi.org/10.1177/0962280211431022.

18. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ. 1998;5,317(7172):1572. doi: 10.1136/bmj.317.7172.1572.

19. Cobre AF, Pedro CAA, Fachi MM, Vilhena RO, Marson BM, Nicobue V, *et al*. Five-year survival analysis and predictors of death in HIV-positive serology patients attending the Military Hospital of Nampula, Mozambique. AIDS Care - Psychol Socio-Medical Asp AIDS/HIV 2020;32:1379–87. https://doi.org/10.1080/09540121.2020.1761938.

20. Kennard RW, Stone LA. Technometrics computer aided design of experiments. Technometric 1969;11:137–48. doi: 10.2307/1266770.

21. Ruiz-Perez D, Guan H, Madhivanan P, Mathee K, Narasimhan G. So you think you can PLS-DA? BMC Bioinformatics 2020;21: (Suppl 1): 2. doi:10.1186/s12859-019-3310-7.

22. WISE BM. PLS Toolbox 4.0 for use with MATLAB. Eigenvector Research Inc., Manson, WA, 2015.

23. Ballabio D; Consonni V. Ferramentas de classificação em química. Parte 1: Métodos Anal. 2013;5(16):3790-8. doi: 10.1039/C3AY40582F.

24. Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, *et al*. Role of machine learning techniques to tackle the covid-19 crisis: Systematic review. JMIR Med Inform 2021;9:1–19. https://doi.org/10.2196/23811.

25. Albahri AS, Hamid RA, Alwan J k., Al-qays ZT, Zaidan AA, Zaidan BB, *et al*. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. J Med Syst 2020;44. https://doi.org/10.1007/s10916-020-01582-x.

26. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons and Fractals 2020;139:110059. https://doi.org/10.1016/j.chaos.2020.110059.

27. SUTHAHARAN S. Machine learning models and algorithms for big data: Thinking with examples for effective learning. New York: Springer; 2016. https://doi.org/10.1007/978-1-4899-7641-3.

28. Réda C, Kaufmann E, Delahaye-Duriez A. Machine learning applications in drug development. Comput Struct Biotechnol J 2019;18:241–52. https://doi.org/10.1016/j.csbj.2019.12.006.

29. Liu B, Wei Y, Zhang Y, Yang Q. Deep neural networks for high dimension, low sample size data. IJCAI Int Jt Conf Artif Intell 2017;0:2287–93. https://doi.org/10.24963/ijcai.2017/318.

30. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. J Choice Model 2018;28:167–82. https://doi.org/10.1016/j.jocm.2018.07.002.

31. Shallue CJ, Lee J, Antogniniy J, Sohl-Dickstein J, Frostig R, Dahl GE. Measuring the effects of data parallelism on neural network training. ArXiv 2018;20:1–49.

32. Wu P, Ye H, Cai X, Li C, Li S, Chen M, *et al*. An effective machine learning approach for identifying non-severe and severe coronavirus disease 2019 patients in a rural Chinese population: The wenzhou retrospective study. IEEE Access 2021;9:45486–503. https://doi.org/10.1109/ACCESS.2021.3067311.

33. Dai W, Ke PF, Li ZZ, Zhuang QZ, Huang W, Wang Y, *et al*. Establishing classifiers with clinical laboratory indicators to distinguish COVID-19 from community-acquired pneumonia: Retrospective cohort study. J Med Internet Res 2021;23. https://doi.org/10.2196/23390.

34. Hu C, Liu Z, Jiang Y, Shi O, Zhang X, Xu K, *et al*. Early prediction of mortality risk among patients with severe COVID-19, using machine learning. Int J Epidemiol 2021;49:1918–29. https://doi.org/10.1093/ije/dyaa171.

35. Xu W, Sun NN, Gao HN, Chen ZY, Yang Y, Ju B, *et al*. Risk factors analysis of COVID-19 patients with ARDS and prediction based on machine learning. Sci Rep 2021;11:2933. https://doi.org/10.1038/s41598-021-82492-x.

36. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020;395:1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3.

37. Smits SL, De Lang A, Van Den Brand JMA, Leijten LM, Van Ijcken WF, Eijkemans MJC, *et al*. Exacerbated innate host response to SARS-CoV in aged non-human primates. PLoS Pathog 2010;6. https://doi.org/10.1371/journal.ppat.1000756.

38. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

39. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, *et al*. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med 2020;8:420–2. https://doi.org/10.1016/S2213-2600(20)30076-X.

40. Palaiodimos L, Kokkinidis DG, Li W, Karamanis D, Ognibene J, Arora S, *et al*. Severe obesity is associated with higher in-hospital mortality in a cohort of patients with COVID-19 in the Bronx, New York. Metabolism 2020;108:154262. https://doi.org/10.1016/j.metabol.2020.154262.

41. Imam Z, Odish F, Gill I, O'Connor D, Armstrong J, Vanood A, *et al*. Older age and comorbidity are independent mortality predictors in a large cohort of 1305 COVID-19 patients in Michigan, United States. J Intern Med 2020;288:469–76. https://doi.org/10.1111/joim.13119.

42. Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, *et al*. Predictors of mortality for patients with COVID-19 pneumonia caused by SARSCoV- 2: A prospective cohort study. Eur Respir J 2020;55. https://doi.org/10.1183/13993003.00524-2020.

43. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, *et al*. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N Engl J Med 2020;382:1199–207. https://doi.org/10.1056/nejmoa2001316.

44. Louie JK, Acosta M, Winter K, Jean C, Gavali S, Schechter R, *et al*. Factors associated with death or hospitalization due to pandemic 2009 influenza A(H1N1) infection in California. JAMA - J Am Med Assoc 2009;302:1896–902. https://doi.org/10.1001/jama.2009.1583.

45. Dannefer R, Williams DA, Baronberg S, Silver L. Healthy bodegas: Increasing and promoting healthy foods at corner stores in New York City. Am J Public Health 2012;102(10):e27-e31. doi: 10.2105/AJPH.2011.300615.

46. Halpin DMG, Singh D, Hadfield RM. Inhaled corticosteroids and COVID-19: A systematic review and clinical perspective. Eur Respir J 2020;55. https://doi.org/10.1183/13993003.01009-2020.

47. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, *et al*. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020;395:507–13. https://doi.org/10.1016/S0140-6736(20)30211-7.

48. Liu S, Zhi Y, Ying S. COVID-19 and Asthma: Reflection During the Pandemic. Clin Rev Allergy Immunol 2020;59:78–88. https://doi.org/10.1007/s12016-020-08797-3.

Endereço para correspondência:

Beatriz Böger
Federal University of Paraná, *Campus* III
Pharmaceutical Sciences Postgraduate Program
Av. Pref. Lothário Meissner, 632, Jardim Botânico
Curitiba – PR, Zipe code 80210-170
Brazil

E-mail: beatrizboger@ufpr.br

Received Jun 16, 2021
Accepted August 14, 2025