Prediction of COVID-19 severity from clinical and biochemical markers: a single-center study from Saudi Arabia

H.M. ALSHANBARI[1], W. SAMI[2,3], T. MEHMOOD[4], M. ABOUD[5], T. ALANAZI[6], M.A HAMZA[7,9], I. BREMA[8], B. ALOSAIMI[9]

[1]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh 11671, Saudi Arabia
[2]Department of Community Medicine and Public Health, College of Medicine, Majmaah University, Almajmaah, Saudi Arabia
[3]Azra Naheed Medical College, Superior University, Lahore, Pakistan
[4]School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[5]King Abdullah bin Abdulaziz University Hospital, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[6]Department of Pathology and Laboratory Medicine, King Abdullah bin Abdulaziz University Hospital, Princess Nourah bint University, Riyadh, Saudi Arabia
[7]Faculty of Medicine, King Fahad Medical City, Ministry of Health, Riyadh, Saudi Arabia
[8]Obesity, Endocrine & Metabolism Centre, King Fahad Medical City, Riyadh, Saudi Arabia
[9]Research Center, King Fahad Medical City, Ministry of Health, Riyadh, Saudi Arabia

Abstract. – OBJECTIVE: It is known that the severity of COVID-19 is linked to the prognosis of patients; therefore, an early identification is required for patients who are likely to develop severe or critical COVID-19 disease. The purpose of this study is to propose a statistical method for identifying the severity of COVID-19 disease by using clinical and biochemical laboratory markers.

PATIENTS AND METHODS: A total of 48 clinically and laboratory-confirmed cases of COVID-19 were obtained from King Fahad Hospital, Medina (KFHM) between 27th April 2020 to 25th May 2020. The patients' demographics and severity of COVID-19 disease were assessed using 39 clinical and biochemical features. After excluding the demographics, 35 predicting features were included in the analysis (diabetes, chronic disease, viral and bacterial co-infections, PCR cycle number, ICU admission, clot formation, cardiac enzymes elevation, hematology profile, sugar levels in the blood, as well as liver and kidney tests, etc.). Logistic regression, stepwise logistic regression, L-2 logistic regression, L-2 stepwise logistic regression, and L-2 best subset logistic regression were applied to model the features. The consistency index was used with kernel Support-Vector Machines (SVM) for the identification of associated markers.

RESULTS: L-2 best subset logistic regression technique outperformed all other fitted models for modeling COVID-19 disease severity by achieving an accuracy of 88% over the test data. Consistency index over L-2 best subset logistic regression identified 14 associated markers that can best predict the COVID-19 severity among COVID-19 patients.

CONCLUSIONS: By combining a variety of laboratory markers with L-2 best subset logistic regression, the current study has proposed a highly accurate and clinically interpretable model of predicting COVID-19 severity.

Key Words:
Severity, Logistic regression, L2 Norm, COVID-19, Supervised learning, Saudi Arabia.

Introduction

Globally, by June 2021, there were approximately 176 million confirmed COVID-19 infections with over 3.8 million reported deaths. In Saudi Arabia, more than 464 thousand confirmed cases were reported along with approximately 7.6 thousand deaths, https://www.worldometers.info/coronavirus.

The missed COVID-19 diagnosis may cause a delay in providing the necessary therapeutic treatment, thus increasing the likelihood of a bad

Corresponding Author: Waqas Sami, Ph.D; e-mail: w.mahmood@mu.edu.sa; waqas_sami@hotmail.com

outcome. On the other hand, the treatment of a severe or critically ill COVID-19 patient necessitates the use of extensive medical resources; in this case, multiple misdiagnoses will only exhaust those resources and exacerbate the medical burden. As a result, early identification of individuals at risk of developing severe COVID-19 infection is crucial for the clinical management and epidemic control.

People over the age of 65 years, as well as those with pre-existing medical issues such as diabetes, heart disease, and asthma, tend to be more sensitive to falling extremely unwell with the COVID-19 virus. When individuals with diabetes gets a viral infection, it might be more difficult to treat the infection because of fluctuations in blood glucose and, perhaps, due to the existence of diabetes associated complications[1]. This explains that complications have been observed in COVID-19 patients with diabetes. Similarly, in Denmark, psychosocial consequences of the COVID-19 pandemic with diabetes were also observed[2,3].

One of the key goals of this study is to identify the clinical and biochemical markers that can early predict the patients who are likely to develop severe/critical COVID-19. The effects of different demographics and related factors on COVID-19 severity have been previously investigated. In particular, one study[4] explored the prevalence, pathophysiological causes, and effects of COVID-19 infection on type 2 diabetic patients. The link between COVID-19 severity and diabetes pathology, implying that underlying difficulties or pathologies in individuals may exacerbate infection progression, are reported also in other studies[5].

For identifying the associated markers, reference to COVID-19 patients through the patient's history and clinical information logistic regression can be used[6,7]. To deal with the clinical correlated predictors, the regularized logistic regression provides the potential solution[8]. For regularization in logistic regression, L-2 norm is used, which penalizes the regression coefficients of non-important markers. Again, for identifying the associated markers, stepwise[9] and best subset[10], logistic regression with L-2 norm is suggested. The use of the L-2 norm with stepwise and best subset logistic regression in medical science is an emerging frontier.

The goal of this study is to the investigate the COVID-19 severity-related indicators and develop an evaluation model for predicting it among COVID-19 patients by using the stepwise and best subset logistic regression with L-2 norm. With the use of these methods, the tasks can be accomplished with minimal involvement of humans and somehow with more accuracy. This leads us to develop a strong interest in proposing an L-2 norm logistic model that can help in modeling severity among COVID-19 patients in Saudi hospitals. The accuracy of each of the methods was evaluated in this study and is presented in the subsequent sections.

Patients and Methods

The specimens for the study were obtained retrospectively from a population of 48 people hospitalized at King Fahad Hospital (KFHM), Medina, Saudi Arabia between 27th April 2020 to 25th May 2020. There were 14 critical cases that required ICU hospitalization and 34 were mild cases. Nine patients died (all of them were admitted to the ICU), while the remaining patients survived. Thirteen patients were Saudi citizens, whereas the rest were non-Saudis. The target population was both male and female COVID-19 patients of all ages. Level of precision formula was used to calculate the sample size based on GPower software (p=0.45, 1- p=0.55, d=0.05). The minimum required sample size as calculated was 40. Demographic and clinical data (Table I) were obtained without any personally identifiable information, including the following clinical laboratory results: gender, age, nationality, and risk factors including chronic disease, viral and bacterial co-infections, PCR cycle number (measured by CT value), ICU admission (severe or mild), clot formation measured by D-dimer, cardiac enzymes for the diagnosis of myocardial infarction measured by CK, CK-MB, and troponin test, hematology profile, including RBCs, WBCs, platelets, HB, Neutrophile and Lymphocytes counts, sugar levels in blood measured by glucose, degree of inflammation measured by CRP and ESR, liver enzymes measured by AST, ALT, Albumin, Urea, and total protein, blood biochemistry tests for kidney functions, including creatinine.

The research was approved by the Institutional Review Board (IRB), Princess Nourah bin Abdulrahman University, Riyadh, KSA via registration Number with KACST, KSA: HAP-01-R-059.

Modeling COVID-19 Severity

For modelling the severity of COVID-19, several demographic and risk factors comprise the data matrix X. Different variants of logistic regression were used in supervised statistical learning. This

Table I. The demographic and clinical characteristics of patients in relation to COVID-19 severity (critical and mild) are summarized. Moreover, the chi-square-based *p*-values indicates the significance of respective factors.

Baseline variable	All patients N = 48	Critical N = 14 (29%)	Mild N = 34 (71%)	*p*-value
Demographic				
Age				
Range	1-92	25-74	1-92	
Gender				
Male	37 (77%)	11 (30%)	26 (70%)	
Female	11 (23%)	3 (27%)	8 (73%)	1.000
Saudi				
Yes	13 (27%)	6 (46%)	7 (54%)	
No	35 (73%)	8 (23%)	27 (77%)	0.222
Co-infection				
Yes	34 (71%)	9 (26%)	25 (74%)	
No	14 (29%)	5 (36%)	9 (64%)	0.771
Chronic disease				
Yes	29 (60%)	4 (14%)	25(86%)	
No	19 (40%)	10 (53%)	9 (47%)	0.010*
Viral				
Yes	27 (56%)	9 (34%)	18 (66%)	
No	21 (44%)	5 (24%)	16 (76%)	0.689
Bacterial				
Yes	17 (35%)	2 (12%)	15 (88%)	
No	31 (65%)	12 (39%)	19 (61%)	0.103
Diabetes				
Yes	26 (54%)	4 (15%)	22 (85%)	
No	22 (45%)	10 (45%)	12 (54%)	0.049*

*Statistically significant at 5% level of significance.

includes the standard logistic regression, stepwise logistic regression, L-2 logistic regression, L-2 stepwise logistic regression, and L-2 best subset logistic regression.

Logistic Regression

Logistic regression is a standard statistical discrimination method that assumes that the demographic and risk factors which form the data matrix X_{nxq} with $x = x_1, . . . , x_q$ can discriminate the response severity, i.e., critical *y,* by estimating the probability *p* (*y = critical*) in the log-odds form:

$$logit(p) = log\frac{p(y = Critical)}{1 - p(y = Critical)} = \alpha + \sum_{j=1} \beta_j x_j$$

Here α is the intercept, is the change in severity level *y,* i.e., critical and mild with respect to per unit change in explanatory marker for j = 1, 2, *q* where q=39, i.e., number of explanatory markers *X* The above logit function can be written as:

$$p(y = Critical) = \frac{1}{1 + \exp\,(\alpha + \sum_{j=1} \beta_i x_j)}$$

The cost function used in logistic regression is defined as:

$$j(\beta) = \frac{1}{q}[\sum_{j=1} y\log(p) + (1 - y)\log\,(1 - p)]$$

L-2 Logistic Regression

The demographic and clinical markers which construct the explanatory matrix ***X*** are expected to be colinear. In the presence of multicollinearity, the standard logistic regression estimate becomes overfitted. The problem gets worse, especially when the sample size is smaller, which is the case for the current study. To avoid the overfitting, penalized regression is the way forward, where *L*2 norm, i.e., regularization, is implemented as:

$$j(\beta) = \frac{1}{q}[\sum_{j=1} y\log(p) + (1 - y)\log\,(1 - p)] + \frac{\lambda}{q}\sum_{j=1} w_j^2$$

Where is a shrinkage parameter, the regularization term heavily penalizes large . This effect is generally less on smaller , which can be tuned through cross-validation. Larger value of λ will shrink closer to 0, which might lead to underfitting, and λ =0 will have no regularization effect.

Stepwise L-2 Logistic Regression

L-2 The overfitting effect induced by multicollinearity is resolved *via* regularized logistic regression, although none of the regression coefficients are set to zero. As a result, it does not offer the associated marker selection. For interpretation, we seek the most relevant demographic and clinically associated markers that are best model for COVID-19 severity. The stepwise strategy is utilized here for associated marker selection, which is based on forward selection followed by back-ward elimination.

Backward deletion is repeated until only one element remains in the model. The cost-complexity statistic *C = deviance + cpxdf*, where *cp* stands for "complexity parameter" is used to determine which markers to add or remove in each phase. For the Akaike information criterion (AIC) and Bayesian information criterion (BIC), popular values are *cp* = 2 and *cp = log* (*sample size*), respectively.

Best subset L-2 Logistic Regression

To tackle the best subset selection problem, the primal-dual active set (PDAS) methodology is applied. It employs an active set update technique to fit the sub-models using complimentary primal and dual markers. The generalized PDAS approach for logistic regression loss functions with the best subset constraint is employed in this case. The algorithm performs the stages listed below:

1. Set the active set's cardinality k and the maximum number of iterations *mc*. Set A to be a random k-subset of 1, . . . , p and I = Ac.

2. For m = 1, 2, . . . , mmax, do.
- Estimate L-2 logistic regression coefficients β.
- Compute gradient gr = $\Delta\beta$ and Hessian *Hs* = diag ($\Delta^2\beta$).
- Update $A = j: , \Delta j > \Delta k$.
- (2.d) Stop if is *A* invariant.

3. Output A, β, Δ.

In reality, the subset size (*k*) is frequently unknown; hence, it must be determined using data-driven methods such as cross-validation.

Monte Carlo Validation

The accuracy of COVID-19 severity models must be validated, which implies how well the COVID-19 severity models will place a new patient into severity level as critical and mild and is defined as:

$$\text{Accuracy} = \frac{\text{(Number of correct COVID-19 severity)}}{\text{(Total number of predictions)}}$$

The severity model runs from 0 to 100 percent, with a high percent indicating that the COVID-19 severity models are desirable. We utilized Monte Carlo validation method to validate COVID-19 severity models, with data separated into training (70 %) and test (30 %). Training data was used to fit the logistic, stepwise logistic, L-2 logistic, L-2 stepwise logistic, and L-2 best subset logistic regression. The tuning parameters were tuned over test data, and model accuracy was computed over both test and training data. Each Monte Carlo run was repeated 100 times.

Results

For modelling the COVID-19 severity, 35 predicting features were extracted from 48 patients in this dataset, the demographic and clinical characteristics of patients in relation to COVID-19 severity are summarized in Table I. Notably, two blood profile tests were conducted for each patient. The first blood profile that was conducted at admission is named postfix with '1', while the second one that was conducted after 48 hours is named postfix with '2'.

The dataset was then screened for acquisition errors which are common to occur. The errors, traditionally known as outliers, may occur because of data entry typos or related to the sampling procedure. Before applying the proposed models, it is recommended that the outliers are removed from the dataset. In this study, the logistic regression deletion diagnostics[11] was used, where Cooks distance d_c on logistic regression residuals was applied. The samples having Cooks distance $d_c > 4\bar{d}c$ were considered as outliers. The Cooks distance computed from logistic regression residuals is presented in Figure 1. Case No. 38 was identified as outlier, it was deleted from the dataset, resulting in a total of 47 patients.

By using the information from 47 patients and applying Monte Carlo validation, the sample of 30 was used to train the COVID-19 severity model, including logistic, stepwise logistic, L-2 logistic, L-2 stepwise logistic, and L-2 best subset logistic regression. The comparison of the test

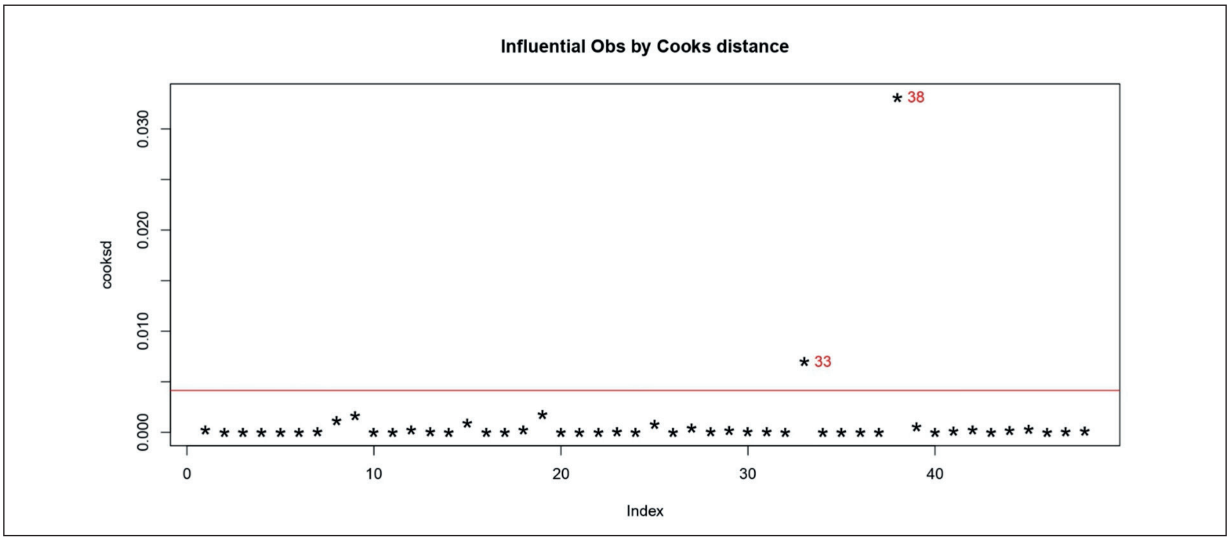

Figure 1. The Cook's Distance computed from logistic regression residuals is plotted against each sample. Outliers having $\mathrm{d}_{\mathrm{c}} > 4d\bar{c}$ are highlighted in red.

and training accuracy of these classifiers is presented in Figure 2. It seemed that all COVID-19 severity models on training data revealed around 100% accuracy. The average test accuracy of logistic regression was about 61% which was the lowest, followed by L-2 logistic with the average test accuracy of about 63%. The stepwise logistic regression average accuracy on test data was around 72%, the L-2 stepwise logistic average accuracy on test data was 80%, and the L-2 best subset logistic regression's average accuracy on test data was about 88%. Hence, the L-2 best subset logistic regression best models the COVID-19 severity. The L-2 best subset logistic regression defines the COVID-19 severity model by the threshold *k*, which defines the associated markers being selected in the model. The threshold reflects which best COVID-19 severity models on test data is marked as optimal. Optimal threshold was extracted in each Monte Carlo run. The optimal threshold and the respective number of selected associated markers against accuracy on test data are presented in Figure 3. The second order response surface is fitted to indicate the overall

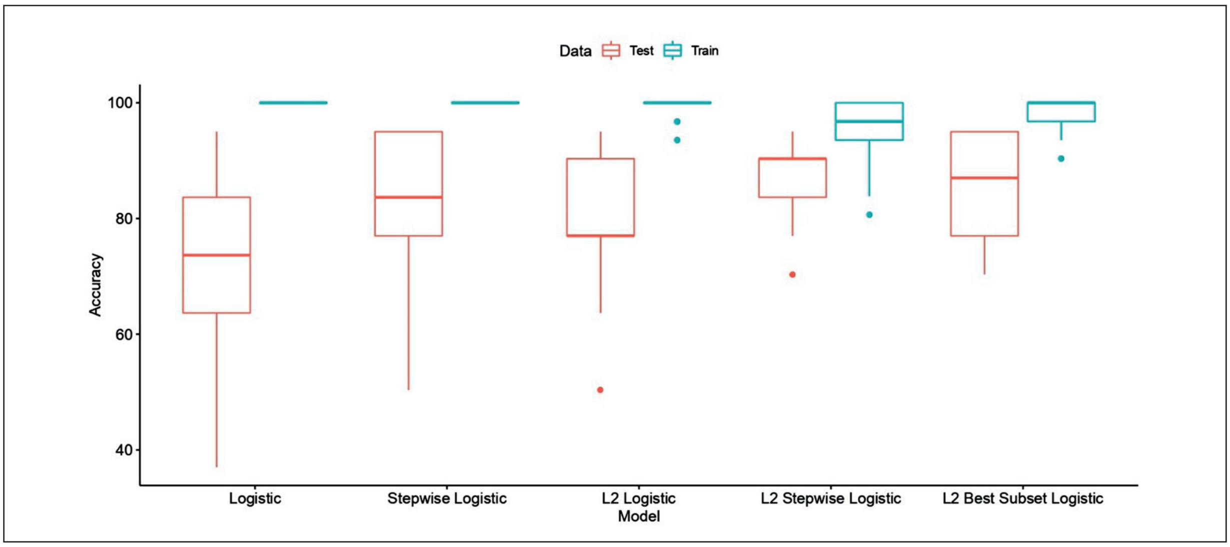

Figure 2. The comparison of training and test accuracy of severity including logistic, stepwise logistic, L-2 logistic, L-2 stepwise logistic, and L-2 best subset logistic regression is presented.

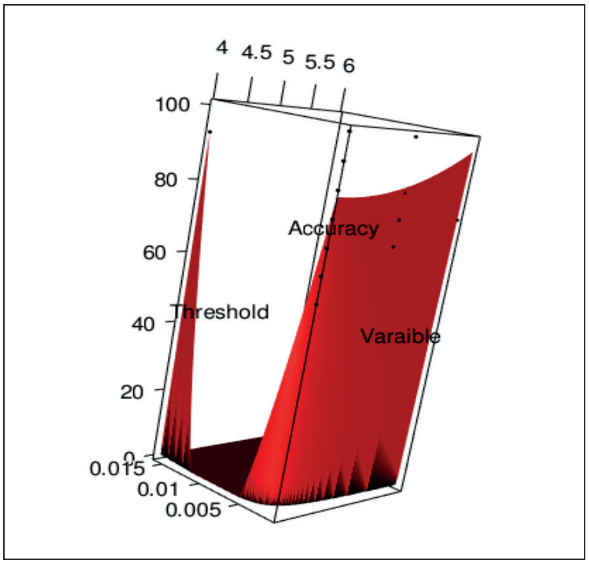

Figure 3. The optimal threshold and a respective number of selected associated markers against accuracy on test data are presented. Second order response surface in red color is fitted to indicate the overall behavior of the L-2 best subset logistic regression model.

behavior of the L-2 best subset logistic regression model. This indicates that the best COVID-19 severity models can be achieved with a lower level of threshold, which is closer to 0.001. Moreover, the optimal threshold resulted in 6 variables.

To identify the associated markers, the Monte Carlo simulation was applied. Each of the Monte Carlo runs the list of associated markers, which may vary from iteration to iteration. Hence, we have computed the consistency index, which is simply the count of each variable selected as important. A marker having a consistency index greater than 10% is considered associated[12]. The consistency index of the markers is presented in Figure 4.

The associated selected markers at 10% consistency index with COVID-19 severity model includes chronic, glucose, AST, COVID-19 disease severity, age, Saudi nationality, bacterial, RBC1, creatinine, total protein, lymph2, platalet1, CK, and WBC1. However, there are 3 variables that reach 20% consistency index, which are marked as the most important variables that highly affected the COVID-19 patients' severity index. The distribution of associated markers against COVID-19 severity, i.e., critical and mild, is presented in Figure 5.

Figure 5 indicates, as glucose increases, the chances of being a critically ill COVID-19 patients also increase (odds ratio= 10.4). The average glucose in patients with mild COVID-19 disease is 7.3 mmol/L (SD = 4.5), while the average glucose in critical patients is 13.9 mmol/L (SD= 8.6). In comparison, as AST increases, the chances of being a critically ill COVID-19 patient decrease (odds ratio=0.442). The average AST in patients with COVID-19 disease is 93.7 (SD=160.9), and the average AST in critical patients is 38.8 (SD=17.1). As age increases, the chances of being a critically ill COVID-19 patient also increases (odds ratio=1.21). The average age in mild COVID-19 patients is 47.0 (SD= 20.2), while the average age in critical COVID-19 patients is 53.8 (SD=15.9). As RBC1 increases, the chances of being a critically ill COVID-19 patients also gets higher (odds ratio=10.37). The av-

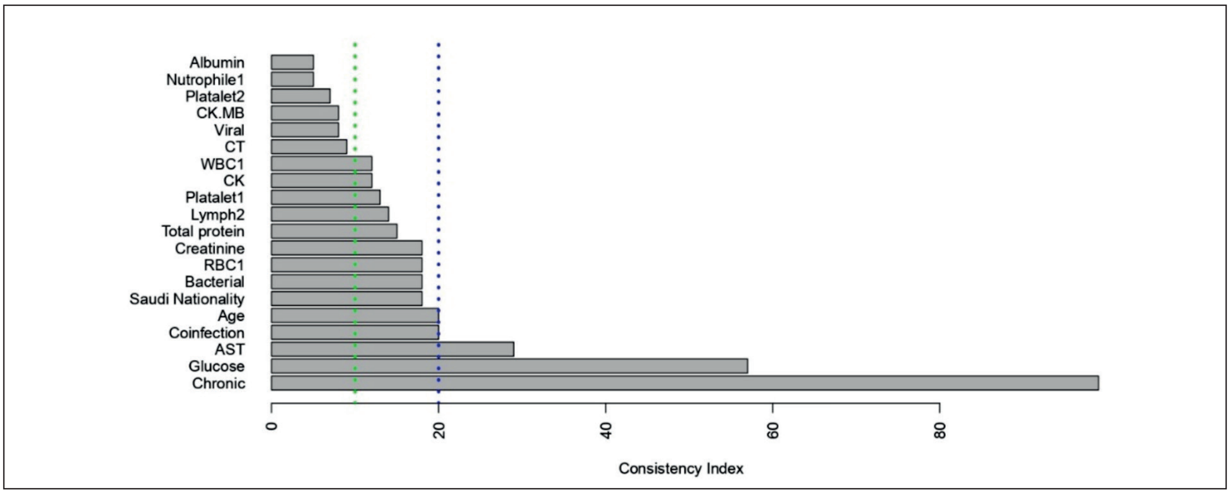

Figure 4. The consistency index of the markers is presented. The markers having index values greater than 10 are considered as associated and are indicated by a green dotted line.

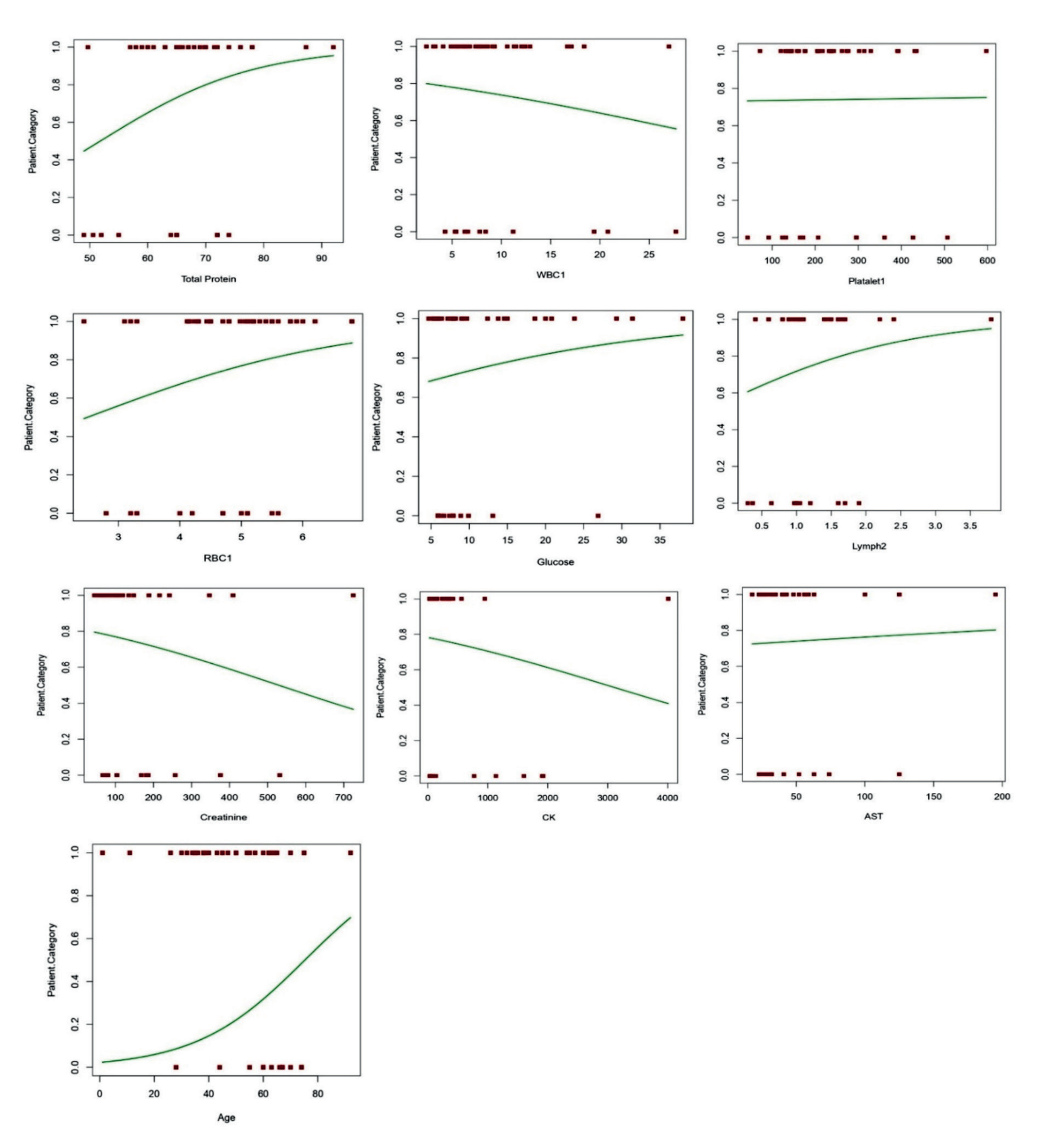

Figure 5. The distribution of associated continuous markers against COVID-19 severity, i.e., critical and mild, is presented

erage RBC1 in mild COVID-19 patient is 4.5 (SD = 0.8) while average RBC1 in critical COVID-19 patient is 4.9 (SD = 1.1). As creatinine increases, the chances of being a critically ill COVID-19 patient decrease (odds ratio=0.987). The average creatinine in mild COVID-19 patients is 139.0 (SD =140.3), while the average creatinine in critical COVID-19 patients is 193.5 (SD= 206.5). As total protein increases, the chances of being a critically ill COVID-19 patients increases as well (odds ratio=3.562). The average total protein in mild COVID-19 patients is 64.3 (SD=8.3), while the average total protein in critical COVID-19 patients is 68.1 (SD=9.3). As lymph2 (lymphocytes in the blood) increases, the chances of being a critically ill COVID-19 patients increase as well (odds ratio=1.781). The average lymph2 in mild COVID-19 patients is 1.3 (SD = 0.5) while the average lymph2 in critical COVID-19 patient is 1.1 (SD = 0.7). As platelet 1 increases, the chances of being a critically ill COVID-19 patients also decrease (odds ratio=0.852). The average Platelet1 in mild COVID-19 patients is 241.6 (SD= 101.8), while the average Platelet 1 in critical COVID-19 patients is 221.4 (SD= 137.4). As K increases, the chances of being a critically ill COVID-19 patients decrease (odds ratio=0.271). The average CK in mild COVID-19 patients is 800.8 (SD= 1221.0), while the average CK in critical COVID-19 patients is 247.8 (SD= 346.6). As BC1 increases, the chances of being a critically COVID-19 patients also increases (odds ratio=3.641). The average WBC1 in mild COVID-19 patient patients is 10.4 (SD= 5.5), while the average WBC1 in critical COVID-19 patients is 8.5 (SD= 5.9).

Figure 5 also indicates the connection between patients with chronic disease who are infected with COVID-19, it depicts that such patients are more likely to develop severe COVID-19 disease (odds ratio=9.45). Among 14 critical COVID-19 patients, there were 86% who do not suffer from chronic diseases, while 14% suffer from the chronic disease. Results show that Saudi patients are less likely to fall in the category of severe COVID-19 infection. (odds ratio=0.241). The patients who had bacterial infections are less likely to fall in the critical COVID-19 infection category (odds ratio=0.315). Among the critical COVID-19 patients, only 12% of them were infected by bacterial infections.

Discussion

Coronavirus disease 2019 (COVID-19) is an acute respiratory illness that is caused by infection with Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Unfortunately, COVID-19 infection causes a high rate of hospitalizations, intensive care unit admissions, and deaths. Therefore, identifying patients at the highest risk of developing a severe infection is of utmost importance to start strict treatments and procedures at an early stage and hence reduce the severity of the disease by avoiding devastating consequences. It is well known that COVID-19 infection is associated with lymphopenia, thrombocytopenia, and leukopenia in the case of hospitalized patients[13,14]. Nonetheless, it has been shown that the clinical course of hospitalized patients may significantly vary from one patient to another, with the most widely reported evidence showing that ICU admission and mortality risk are associated with elevated D-dimer level and a reduced lymphocyte count[15]. However, despite having approved vaccinations that have dramatically changed the course of the disease, there is still a desperate need for additional routine biomarkers for patient risk stratification in order to tailor interventions and use hospital resources efficiently.

In this connection, the current study has proposed a highly accurate and clinically interpretable model for predicting the COVID-19 severity. We applied logistic regression, stepwise logistic, L-2 logistic, L-2 stepwise logistic, and L-2 best subset logistic with 100 Monte Carlo validation runs. L-2 best subset logistic outperformed other methods by achieving an accuracy of 88%. The baseline characteristics listed in Table I showed that a total number of subjects constituted 48 patients, of whom (29%) had critical COVID-19 infection that required ICU admission while 71% had a mild COVID-19 infection that was treated in a general ward. Males were more predominant in this cohort, being about 77.1%. On the other hand, patients with diabetes constituted 54.2% of this patient population. Patients of Saudi ethnic group constituted 27.1% only, while most of the patients were non-Saudi (72.9%).

As shown in the model from Figure 5, the increase in plasma glucose increased the odds of being in the critical patient category by more than 10-fold (odds ratio= 10.4) and patients in this category had mean plasma glucose of 13.9 mmol/l in the critical category and 7.3 mmol/l in the mild disease category, respectively. Our findings are consistent with many published studies in the literature inside Saudi Arabia[16,17] and in other parts of the world, all of which showed that both diabetes and high serum glucose levels increased the severity of COVID-19 infections. In our study, the mean glucose value appeared to be positively associated with COVID-19 severity, as previously reported[18]. COVID-19 severity-related glucose metabolism disorders profoundly affect the morphological structure and physiological functions of erythrocytes, resulting in insufficient microcirculation perfusion, hypoxia, and oxidative stress, promoting the occurrence of critical COVID-19 patient complications and lowering patients' quality of life[19]. Given the importance of erythrocytes in the pathological development of complications, erythrocyte count correlated with the occurrence and progression of these complications.

Unexpectedly, the model indicated that lower rather than higher plasma levels of AST were associated with critical COVID-19 infection, with average serum AST 38.8 in the critical COVID-19 infection cases *vs.* 93.7 in the mild infection group with an odds ratio of 0.442. In contrast, most studies[20,21] have shown that liver damage and raised liver enzymes, both AST and ALT, were associated with moderate to severe COVID-19 cases admitted to the hospital. Another important factor that appeared to increase the severity of COVID-19 and ICU admission was age, and we replicate a linear relationship that has been consistently reported in most of the published studies. In the current study cohort, the model showed that the mean age of critically ill patients with COVID-19 was 53.8 years *vs.* 47 years, respectively, with an odds ratio of 1.21. A recent meta-analysis from several European countries showed COVID- 19 related deaths and ICU admissions in Europe across different ages. Patients aged less than 40 years old represented about 0.1 and 5% of COVID-19 related deaths and ICU admissions, respectively, whereas those more than 70 years old represented about 85 and 40%, respectively[22]. Moreover, the model identified a high RBC count proportionally associated with COVID-19 critical infection, with the odds ratio increased by 10.37-fold. RBC is positively associated with the chances of critical COVID-19 patients and is also observed in a study conducted in China[23]. RBC Adhesion of red blood cells in COVID-19 severity is mediated by the advanced glycation end product receptor[24]. Few studies have investigated the changes in hematological parameters with COVID- 19 severity and reached similar findings. The recent study by Zhu et al[25], which systematically investigated the effect of WBC count on mortality, showed that the death risk was associated with the WBC count at admission, although the index was at the normal range, those with higher WBC count patients had a much higher probability of death.

Limitations

We acknowledge the small sample size and the lack of detailed data regarding morality, which are known limitations of this study. Findings may vary with variation in the demographic characteristics, location, culture, and other variables.

Conclusions

By combining a variety of laboratory markers with L-2 best subset logistic regression, the current study has proposed a highly accurate and clinically interpretable model of predicting COVID-19 severity. L-2 Best subset logistic regression has outperformed the other classifiers by achieving an accuracy of 88%. The algorithm also identified 14 significant clinical and biochemical markers that can predict the potential COVID-19 patient to be mild or critically ill. Further larger studies using similar methodology are needed to replicate our findings, which may help physicians dealing with COVID-19 patients in clinical decision making and risk stratification for earlier interventions.

Conflict of Interest

The Authors declare that they have no conflict of interests.

Acknowledgements

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R299), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would like to thank the Research Center at King Fahd Medical City, Riyadh, for their valuable help and support.

References

1) Moore JT, Pilkington W, Kumar D. Diseases with health disparities as drivers of COVID-19 outcome. J Cell Mol Med 2020; 24: 11038-11045
2) Cuschieri S, Grech S. COVID-19 and diabetes: The why, the what and the how. J Diabetes Complications 2020;34: 107637.
3) Joensen LE, Madsen KP, Holm L, Nielsen KA, Rod MH, Petersen AA, Rod NH, Willing I. Diabetes and COVID-19: psychosocial consequences of the COVID-19 pandemic in people with diabetes in Denmark-what characterizes people with high levels of COVID-19-related worries? Diab Med 2020; 37: 1146-1154.
4) Corrao S, Pinelli K, Vacca M, Raspanti M, Argano C. Type 2 Diabetes Mellitus and COVID-19: A Narrative Review. Front Endocrinol 2021; 12: 609470.
5) Feldman EL, Savelieff MG, Hayek SS, Pennathur S, Kretzler M, Pop-Busui R. COVID-19, and Diabetes: A Collision and Collusion of Two Diseases. Diabetes 2020; 69: 2549-2565.
6) Bhandari S, Tak A, Singhal S, Shukla J, Shaktawat AS, Gupta J, Patel B, Kakkar S, Dube A,

Dia S, Dia M, Wehner TC. Patient Flow Dynamics in Hospital Systems During Times of COVID-19: Cox Proportional Hazard Regression Analysis. Front Public Health 2020; 8: 585850.

8) Table BP, Herman WH. A m multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care 2002; 25: 1999-2003.

9) G. Leontidis, B. Al-Diri, A. Hunter, Exploiting the retinal vascular geometry in identifying the progression to diabetic retinopathy using penalized logistic regression and random forests, in: Emerging trends and advanced technologies for computational intelligence. Springer 2016; 381-400.

10) Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics 2008; 9: 30-50.

11) Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med 2016; 4: 136.

12) J. Fox, S. Weisberg, An r companion to applied regression. sag, Thousand Oaks.

13) Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. A partial least squares-based algorithm for parsimonious variable selection. Algorithms Mol Biol 2011; 6: 1-12.

14) Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, Zhao Y, Li Y, Wang X, Peng Z. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA 2020; 323: 1061-1069.

15) Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC, Du B, Li LJ, Zeng G, Yuen KY, Chen RC, Tang CL, Wang T, Chen PY, Xiang J, Li SY, Wang JL, Liang ZJ, Peng YX, Wei L, Liu Y, Hu YH, Peng P, Wang JM, Liu JY, Chen Z, Li G, Zheng ZJ, Qiu SQ, Luo J, Ye CJ, Zhu SY, Zhong NS; China Medical Treatment Expert Group for Covid-19. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med 2020; 382: 1708-1720.

16) Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020; 395: 1054-1062.

17) Varikasuvu SR, Dutt N, Thangappazham B, Varshney S. Diabetes and COVID-19: A pooled analysis related to disease severity and mortality. Prim Care Diabetes 2021; 15: 24-27.

18) Varikasuvu SR, Varshney S, Dutt N. Markers of coagulation dysfunction and inflammation in diabetic and non-diabetic COVID-19. J Thromb Thrombolysis 2021; 51: 941-946.

19) Moyer J, Wilson D, Finkelshtein I, Wong B, Potts R. Correlation between sweat glucose and blood glucose in subjects with diabetes. Diabetes Technol Ther 2012; 14: 398-402.

20) Wang Y, Yang P, Yan Z, Liu Z, Ma Q, Zhang Z, Wang Y, Su Y. The Relationship between Erythrocytes and Diabetes Mellitus. J Diabetes Res 2021; 2021: 6656062.

21) Youssef M, H Hussein M, Attia AS, M Elshazli R, Omar M, Zora G, S Farhoud A, Elnahla A, Shihabi A, Toraih EA, S Fawzy M, Kandil E. COVID-19 and liver dysfunction: A systematic review and meta-analysis of retrospective studies. J Med Virol 2020; 92: 1825-1833.

22) Zhang C, Shi L, Wang FS. Liv r injury in COVID-19: management and challenges. Lancet Gastroenterol Hepatol 2020; 5: 428-430.

23) Cohen JF, Korevaar DA, Matczak S, Chalumeau M, Allali S, Toubiana J. COVID-19-Related Fatalities and Intensive-Care-Unit Admissions by Age Groups in Europe: A Meta-Analysis. Front Med (Lausanne) 2021; 7: 560685.

24) Li Q, Li L, Li Y. Enhanced RBC Aggregation in Type 2 Diabetes Patients. J Clin Lab Anal 2015; 29: 387-389

25) Zhu B, Feng X, Jiang C, Mi S, Yang L, Zhao Z, Zhang Y, Zhang L. Correlation between white blood cell count at admission and mortality in COVID-19 patients: a retrospective study. BMC Infect Dis 2021; 21: 574.