Letter to the Editor

COVID-19 epidemic in a Respiratory Diseases Unit: predictor ranking and mining

Giulia M. Stella[1]^, Davide Piloni[1], Giulia Accordino[1], Amelia Grosso[1], Federica Albicini[1], Erica Gini[1], Silvia Mancinelli[1], Matteo Della Zoppa[1], Andrea P. G. Marchelli[2], Chandra Bortolotto[3], Angelo G. Corsico[1]

[1]Department of Medical Sciences and Infective Diseases, Unit of Respiratory Diseases, IRCCS Policlinico San Matteo Foundation and University of Pavia Medical School, Pavia, Italy; [2]Analog and Power FMT and Smart Power Technology, ST Microelectronics, Agrate Brianza, Italy; [3]Department of Intensive Medicine, Unit of Radiology, IRCCS Policlinico San Matteo Foundation and University of Pavia Medical School, Pavia, Italy

Correspondence to: Giulia M. Stella, MD, PhD. Department of Medical Sciences and Infective Diseases, Unit of Respiratory Diseases, IRCCS Policlinico San Matteo Foundation and University of Pavia Medical School, Piazzale Golgi 19, 27100 Pavia, Italy. Email: g.stella@smatteo.pv.it.

Submitted Sep 21, 2020. Accepted for publication Oct 28, 2020.
doi: 10.21037/jtd-20-2934
View this article at: http://dx.doi.org/10.21037/jtd-20-2934

Novel coronavirus related disease (COVID-19) has profoundly influenced hospital organization and structures worldwide. In Italy, the Lombardy Region, with almost 17% of the Italian population (2019 data) (1), rapidly became the most severely affected area, and till now the region comprises the 48% of all COVID-19 related deaths in Italy and 40% of the confirmed cases (data from Protezione Civile at 1st, June 2020, http://www.protezionecivile.gov.it/). In the province of Pavia 5293 COVID-19 diagnosis have been reported, with more than 1000 deaths. The manifestations of *SARS-CoV-2* infection in humans range from mild respiratory symptoms to severe acute respiratory syndrome (2,3). Moreover, a variety of general symptoms can occur concomitantly. Among them, cardio-vascular disease, mainly represented by pulmonary thromboembolism is a major cause of death in infected patients (4-6). Starting from the end of February 2020, the Pneumology Unit at San Matteo Hospital Foundation received more 150 COVID-19 positive patients. The present study evaluates the demographic and clinical features of the patient population that came to our observation with the aim of ranking the many risks factors and identifying outcome predictive markers. From the 1st of March until the 1st of May 2020, 133 patients were hospitalized, 85 men and 48 women. The main age at diagnosis of COVID-19 was 74 years; most patients [98] lived in the province of Pavia, 24 come from the red zone in province of Lodi, the remaining from other Lombardy provinces. In 35 out of the 133 cases analysed, the nasal swab was negative for real time (RT) PCR detection of *SARS-CoV-2* RNA but clinical, laboratory and imaging features were highly coherent to the diagnostic suspect of viral infection. In only three cases, patient respiratory conditions and performance status allowed the execution of bronchoscopy and RT-PCR test performed on the obtained broncho-alveolar fluid identified novel coronavirus.

It should be noted that the diagnostic imaging approach to COVID patients is very complex. Some guidance is provided by guidelines and statement. Taking into account Fleishner society (7) and similar recommendation CT can be considered a powerful aid in diagnosis of COVID-19 but it is not necessary nor the solely imaging modality adequate to this task.

A database with exhaustive data of the patient population in study has been constructed. The statistical analysis of all the data has been conducted by using the JMP partition algorithm (JMP-Statistical Discoveries. From SAS, website at www.jmp.com) which is able to search all possible splits of best response predictors. These splits (or partitions) of the data are done recursively to form a tree of decision rules. The partition algorithm chooses optimum splits from many possible trees, making it a powerful modelling, and data discovery tool. The technique is often considered as a data mining approach since it can explore relationships

^ ORCID: 0000-0003-0929-4394.

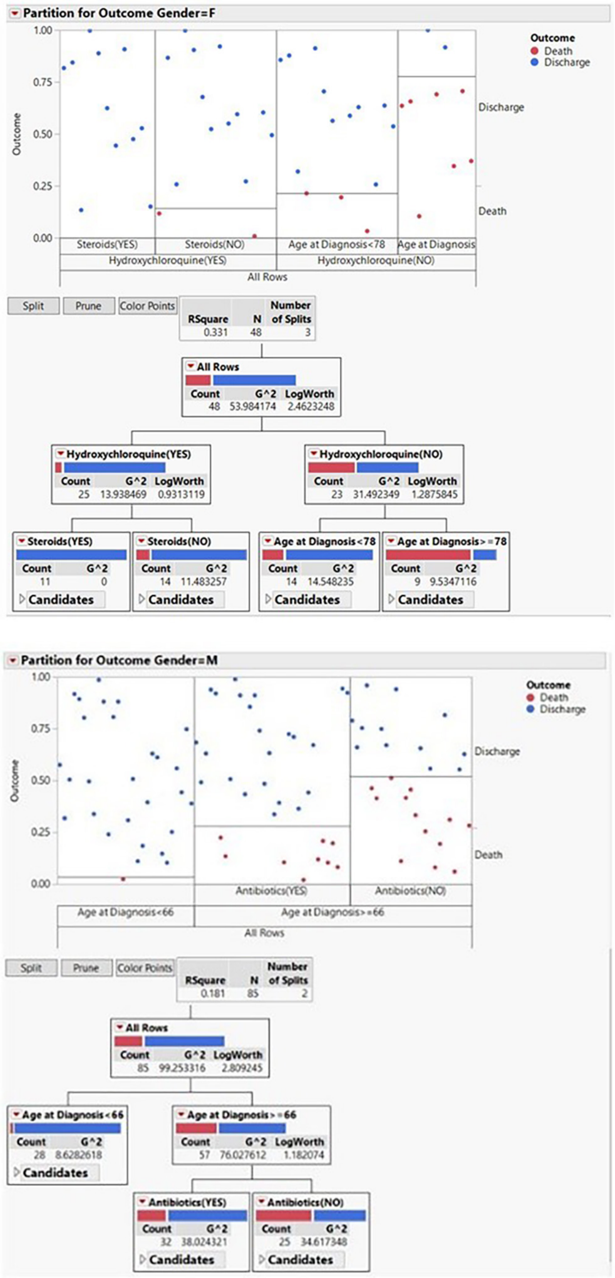
in absence of a good prior model. Moreover, it is able to reduce big problems to easier interpretable results. A useful application of partitioning is to create a diagnostic heuristic for a disease. Given symptoms and outcomes for a population, partitioning can be used to generate a hierarchy of questions to help diagnose new patients. Predictors can be either continuous or categorical (nominal or ordinal). If a predictor is continuous, then the splits are created by a cutting value. The sample is divided into values below and above the cutting one. If a predictor is categorical, then the sample is divided into two groups of levels. The response can also be either continuous or categorical (nominal or ordinal). If the response is continuous, then the platform fits the means of the response values. If the response is categorical, then the fitted value is a probability for the levels of the response. To properly estimate the residual uncertainty of classification, several numerical indexes are used. Among them the Gini heterogeneity index (I_G) which is calculable from relative frequencies (p_i) of each of the M total classes (8). Another one is entropy (H) which is associated to the concept of quantity of uncertainty (9). By using one of the above-mentioned indexes, the algorithm is able to take a decision on which partition make at each step of tree construction. Data regarding the analysed cohort were processed according to the above-described approach. Respiratory failure affected most of the analysed patients and more than 66% of them required ventilatory support with either high-flow nasal cannula (HFNC) or non-invasive ventilation with continuous positive airway pressure (C-PAP) with a fraction of inspired oxygen (FiO_2) included between 60% and 100%. Therapeutic schedules included the antiviral combination lopinavir/ritonavir 42.10% patients (10), hydroxychloroquine (45.11% of cases) (11); steroids were added in about 20% of patients. In 81 (60.9%) out of the 133 patients, antibiotic treatment was introduced, mainly in case of lung consolidation patterns, fever recurrence, and increased levels of reactive C-protein and procalcitonin. Overall 16 (12.1%) out of 133 patients met the criteria to be enrolled in clinical trials with experimental therapies: 7 patients were treated with the antiviral remdesivir, which acts by blocking the coronavirus's RNA polymerase (12), 7 with the anti-IL-6 tocilizumab (13), and the remaining 3 with plasma obtained by convalescent donors (14). Within respect to the thromboembolic risk management (median D dimer value 4,573.6 ng/L at hospital admission *vs*. normal range <500 ng/L, and higher median value at 4,864.8 ng/L), treatment with low molecular weight heparin was started in 55 patients (41.3%); 37 of them received prophylactic doses, whereas in the remaining cases a full dosage was started after diagnosis of venous thrombosis and/or pulmonary embolism (18 cases). The mean duration of hospitalization was 16.3 days. Overall mortality rate reached 21.8%, significantly higher in males (17.2%) *vs*. females (9.02%). The most relevant predictor of survival in men was the age, being only one event (death) occurred in a male younger than 66 years (39 subjects) *vs*. 22 events in older patients. Among male patients older than 66 years, antibiotic treatment seemed to be associated to positive outcome. Interestingly among females, survival was significantly associated to treatment with hydroxychloroquine and steroids and all treated patients are alive. In the subgroup of patients not treated with hydroxychloroquine, the age higher than 78 years significantly affects mortality irrespective of smoking habit. Results are shown in *Figure 1*. To deeper query the database and detect strongest predictors, we moved to apply partition analysis approach to the ranking of main predictive parameters: WBC, ANC, ALC, LDH, RCP and a panel of cytokines (IL-6, IL-8, IL-10), which can be used to predict progression of COVID-19 (15). Overall, none of the reported parameters emerged *per se* as relevant predictors in the partition analysis. Interestingly, some differences emerged based on patient gender (*Figure 2*). In females, better survival was associated to lower LDH levels, and when LDH was higher than 486 mU/mL, leucocytosis was associated to worst prognosis and patient death. In males, the level of leukocytes was highly predictive of outcome, and in case of moderate leucocytosis ($>9.2\times10^3$/uL), patient age is confirmed to significantly affect outcome. These quite unexpected results confirmed the peculiarity of COVID-19 infection and pointed out that immune/inflammatory markers might play a role in predict outcome only in context-specific setting and not as single variable.

Taken together these findings, although limited to the population studied, allow some considerations. Firstly, it should be underlined that treatment guidelines evolved during the study interval and this fact is reflected in some temporal heterogeneity in therapeutic approach to COVID-19. Therefore, only 26 out of the 133 patients studied underwent steroid treatment and corresponded to those hospitalized later, whereas 60, mainly among the earlier hospitalized patients, assumed hydroxychloroquine. A second observation regards the role played by comorbidities, since the partition algorithm found that none of them (neoplastic, diabetes, cardio-pulmonary, rheumatic) impact on outcome. However, the concomitant increase of

A

Patients (total number)	132
Gender	
Male	86
Female	46
Median age at diagnosis	
Smoking history	
Current	35
Never	55
Past	42
Nasal swab (RT-PCR)	
Positive	97
Negative	35
Days between symptom start and swab	11.3
Comorbidities	
Cardiovascular	67
Pulmonary	28
Endorcrinologic/Rheumatologic	24
Neoplastic	21
Others	75
Complications	
TEP	15
Venous thrombosis	7
infections/sepsis	3
Cardiac failure, acute oedema	9
Lab tests	
CRP medium	11.27
CRP max	17.49
Leukocytes	8.46
leukocytes max	14.6
lynphocytes	1.38
lynphocytes min	0.93
LDH medium	383.93
LDH max	438.55
D dimer	1599.27
D dimer max	1852.19
CD4	207-79
Outcome	
Death	29
Intensive care	19
Hospital discharge	84
Duration hospitalization	16.3

B



RSquare	N	Number of Splits
0.331	48	3

All Rows

Count	G^2	LogWorth
48	53.984174	2.4623248

Hydroxychloroquine(YES)

Count	G^2	LogWorth
25	13.938469	0.9313119

Hydroxychloroquine(NO)

Count	G^2	LogWorth
23	31.492349	1.2875845

Steroids(YES)

Count	G^2
11	0

Steroids(NO)

Count	G^2
14	11.483257

Age at Diagnosis<78

Count	G^2
14	14.548235

Age at Diagnosis>=78

Count	G^2
9	9.5347116

RSquare	N	Number of Splits
0.181	85	2

All Rows

Count	G^2	LogWorth
85	99.253316	2.809245

Age at Diagnosis<66

Count	G^2
28	8.6282618

Age at Diagnosis>=66

Count	G^2	LogWorth
57	76.027612	1.182074

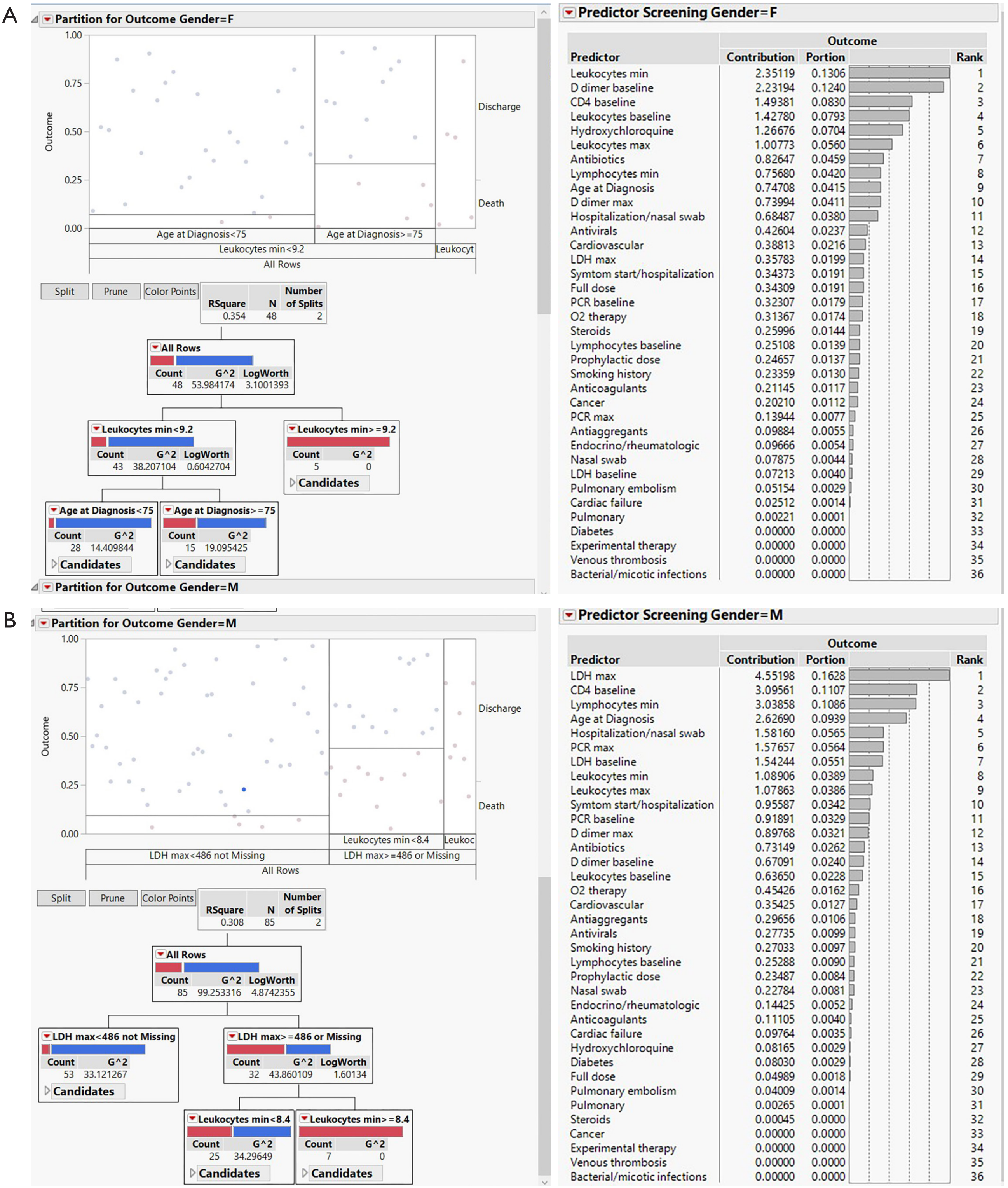
Antibiotics(YES)

Count	G^2
32	38.024321

Antibiotics(NO)

Count	G^2
25	34.617348

Figure 1 Patient characteristic and data mining analysis. (A) Demographic and clinical features of the cohort evaluated. (B) Partition analysis in male and female patients. Lower values indicate better fit. O_2, oxygen; TEP, pulmonary thrombo-embolism; CPE, cardio-pulmonary oedema; CRP, C reactive protein; Count, number of training observation; G2, Gini index.



Predictor Screening Gender=F

Predictor	Outcome		Rank
	Contribution	Portion	
Leukocytes min	2.35119	0.1306	1
D dimer baseline	2.23194	0.1240	2
CD4 baseline	1.49381	0.0830	3
Leukocytes baseline	1.42780	0.0793	4
Hydroxychloroquine	1.26676	0.0704	5
Leukocytes max	1.00773	0.0560	6
Antibiotics	0.82647	0.0459	7
Lymphocytes min	0.75680	0.0420	8
Age at Diagnosis	0.74708	0.0415	9
D dimer max	0.73994	0.0411	10
Hospitalization/nasal swab	0.68487	0.0380	11
Antivirals	0.42604	0.0237	12
Cardiovascular	0.38813	0.0216	13
LDH max	0.35783	0.0199	14
Symtom start/hospitalization	0.34373	0.0191	15
Full dose	0.34309	0.0191	16
PCR baseline	0.32307	0.0179	17
O2 therapy	0.31367	0.0174	18
Steroids	0.25996	0.0144	19
Lymphocytes baseline	0.25108	0.0139	20
Prophylactic dose	0.24657	0.0137	21
Smoking history	0.23359	0.0130	22
Anticoagulants	0.21145	0.0117	23
Cancer	0.20210	0.0112	24
PCR max	0.13944	0.0077	25
Antiaggregants	0.09884	0.0055	26
Endocrino/rheumatologic	0.09666	0.0054	27
Nasal swab	0.07875	0.0044	28
LDH baseline	0.07213	0.0040	29
Pulmonary embolism	0.05154	0.0029	30
Cardiac failure	0.02512	0.0014	31
Pulmonary	0.00221	0.0001	32
Diabetes	0.00000	0.0000	33
Experimental therapy	0.00000	0.0000	34
Venous thrombosis	0.00000	0.0000	35
Bacterial/micotic infections	0.00000	0.0000	36

Predictor Screening Gender=M

Predictor	Outcome		Rank
	Contribution	Portion	
LDH max	4.55198	0.1628	1
CD4 baseline	3.09561	0.1107	2
Lymphocytes min	3.03858	0.1086	3
Age at Diagnosis	2.62690	0.0939	4
Hospitalization/nasal swab	1.58160	0.0565	5
PCR max	1.57657	0.0564	6
LDH baseline	1.54244	0.0551	7
Leukocytes min	1.08906	0.0389	8
Leukocytes max	1.07863	0.0386	9
Symtom start/hospitalization	0.95587	0.0342	10
PCR baseline	0.91891	0.0329	11
D dimer max	0.89768	0.0321	12
Antibiotics	0.73149	0.0262	13
D dimer baseline	0.67091	0.0240	14
Leukocytes baseline	0.63650	0.0228	15
O2 therapy	0.45426	0.0162	16
Cardiovascular	0.35425	0.0127	17
Antiaggregants	0.29656	0.0106	18
Antivirals	0.27735	0.0099	19
Smoking history	0.27033	0.0097	20
Lymphocytes baseline	0.25288	0.0090	21
Prophylactic dose	0.23487	0.0084	22
Nasal swab	0.22784	0.0081	23
Endocrino/rheumatologic	0.14425	0.0052	24
Anticoagulants	0.11105	0.0040	25
Cardiac failure	0.09764	0.0035	26
Hydroxychloroquine	0.08165	0.0029	27
Diabetes	0.08030	0.0029	28
Full dose	0.04989	0.0018	29
Pulmonary embolism	0.04009	0.0014	30
Pulmonary	0.00265	0.0001	31
Steroids	0.00045	0.0000	32
Cancer	0.00000	0.0000	33
Experimental therapy	0.00000	0.0000	34
Venous thrombosis	0.00000	0.0000	35
Bacterial/micotic infections	0.00000	0.0000	36

Figure 2 Main predictors with their value and rank. Predictors with higher contribution are likely to impact on patient outcome. (A) Partition analysis applied to predictor ranking in females; (B) Partition analysis applied to predictor ranking in males.

leukocytes, as in bacterial infection/inflammatory reaction, significantly impacts on patient outcome. Interestingly, the smoking history did not correlate with prognosis. Although the role of smoking in onset and progression of COVID-19 is not fully clarified (16,17), in the analysed population, the occurrence of the viral infection in current smokers was considerably lower (12.7%) than that among previous and never smokers subjects (87%).

It should be noted that this is a preliminary report and that the limited population included could impact on significant differences. Overall, based on predictor ranking analysis, differences emerged based on patient gender. Combination of age and bacterial co-infection and the consequent antibiotic treatment could define the most powerful prognostic model in men; whereas treatment whit hydroxychloroquine should be preferred in those younger women carrying lower LDH, namely those who would present better outcomes.

Acknowledgments

All the authors would like to thank Luigi Redaelli for fruitful discussion and statistical analysis and data revision. GMS would like to thank Elena Morganti and Benedetta Marchelli for invaluable support during COVID-19 pandemic.

Funding: None.

Footnote

Provenance and Peer Review: This article was a standard submission to the journal. The article was sent for external peer review.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jtd-20-2934). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. ISTAT. Demo-Geodemo. - Maps, Population, Demography of ISTAT - Italian Institute of Statistics. Accessed 29 May 2020. Available online: http://demo.istat.it/index_e.html
2. Vancheri SG, Savietto G, Ballati F, et al. Radiographic findings in 240 patients with COVID-19 pneumonia: time-dependence after the onset of symptoms. Eur Radiol 2020;30:6161-9.
3. Goh KJ, Choong MC, Cheong EH, et al. Rapid Progression to Acute Respiratory Distress Syndrome: Review of Current Understanding of Critical Illness from COVID-19 Infection. Ann Acad Med Singap 2020;49:108-18.
4. Bompard F, Monnier H, Saab I, et al. Pulmonary embolism in patients with COVID-19 pneumonia. Eur Respir J 2020;56:2001365.
5. Connors JM, Levy JH. COVID-19 and its implications for thrombosis and anticoagulation. Blood 2020;135:2033-40.
6. Poissy J, Goutay J, Caplan M, et al. Pulmonary Embolism in Patients With COVID-19: Awareness of an Increased Prevalence. Circulation 2020;142:184-6.
7. Rubin GD, Ryerson CJ, Haramati LB, et al. The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement From the Fleischner Society. Chest 2020;158:106-16.
8. Capecchi S, Iannario M. Gini heterogeneity index for detecting uncertainty in ordinal data surveys. Metron 2016;74:223-32.
9. Eliazar I, Sokolov I M. Maximization of statistical heterogeneity: From Shannon's entropy to Gini's index. Physica A 2010;389:3023-38.
10. Cao B, Wang Y, Wen D, et al. A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19. N Engl J Med 2020;382:1787-99.
11. Patil VM, Singhal S, Masand N. A systematic review on use of aminoquinolines for the therapeutic management of COVID-19: Efficacy, safety and clinical trials. Life Sci 2020;254:117775.
12. Grein J, Ohmagari N, Shin D, et al. Compassionate Use

of Remdesivir for Patients with Severe Covid-19. N Engl J Med 2020;382:2327-36.

13. Colaneri M, Bogliolo L, Valsecchi P, et al. Tocilizumab for Treatment of Severe COVID-19 Patients: Preliminary Results from SMAtteo COvid19 REgistry (SMACORE). Microorganisms 2020;8:695.
14. Perotti C, Del Fante C, Baldanti F, et al. Plasma from donors recovered from the new Coronavirus 2019 as therapy for critical patients with COVID-19 (COVID-19 plasma study): a multicentre study protocol. Intern Emerg Med 2020;15:819-24.
15. Tang Y, Liu J, Zhang D, et al. Cytokine Storm in COVID-19: The Current Evidence and Treatment Strategies. Front Immunol 2020;11:1708.
16. Vardavas CI, Nikitara K. COVID-19 and smoking: A systematic review of the evidence. Tob Induc Dis 2020;18:20.
17. Cattaruzza MS, Zagà V, Gallus S, et al. Tobacco smoking and COVID-19 pandemic: old and new issues. A summary of the evidence from the scientific literature. Acta Biomed 2020;91:106-12.

Cite this article as: Stella GM, Piloni D, Accordino G, Grosso A, Albicini F, Gini E, Mancinelli S, Della Zoppa M, Marchelli APG, Bortolotto C, Corsico AG. COVID-19 epidemic in a Respiratory Diseases Unit: predictor ranking and mining. J Thorac Dis 2021;13(4):2569-2574. doi: 10.21037/jtd-20-2934