Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan

Li Yan 1*, M.D., Hai-Tao Zhang 2*, Ph.D., Yang Xiao 2, Ph.D., Maolin Wang 2, Chuan Sun 2, Jing Liang 1, Shusheng Li 1, Mingyang Zhang 2, Yuqi Guo 2, Ying Xiao 2, Xiuchuan Tang 3, Haosen Cao 4, Xi Tan 4, Niannian Huang 4, Bo Jiao 4, Ailin Luo 4, M.D., Zhiguo Cao 2, Ph.D., Hui Xu 4, M.D., Ye Yuan 2, Ph.D.

1. Department of Emergency, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology

2. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

3. School of Mechanical Science and Engineering, Huazhong University of Science and Technology

4. Department of Anesthesiology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology

* Equal contribution.

Corresponding authors: Prof. Dr. Ye Yuan, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, P.R. China. Email: **yye@hust.edu.cn**. Prof. Dr. Hui Xu, Department of Anesthesiology, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology, Wuhan 430074, P.R. China. Email: sophia_wh@hotmail.com.

Abstract: The swift spread of COVID-19 epidemic has attracted worldwide attentions since Dec., 2019. Till date, 77,041 confirmed Chinese cases have been reported by National Health Commission of P.R. China with 9,126 critical cases whose survival rate is quite low. Meanwhile, COVID-19 epidemic emergence within the other countries (e.g., Korea, Italy, Japan and Iran) is also remarkable with the increasing spread speed. It plays a more and more important role to efficiently and precisely predict the survival rate for critically ill Covid-19 patients as more fatal cases can be targeted interfered in advanced. However, the survival rates of all the present critically ill COVID -19 patients are estimated manually from over 300 laboratory and clinical features, which inevitably leads to high misdiagnose and missed-diagnose rate due to lack of

experience and priori knowledge. As a remedy, we have developed a machine learning-based prognostic model that precisely predicts the survival for individual severe patients with more than 90% accuracy with clinical data collected from Tongji hospital, Wuhan. Significantly, such model only requires three key clinical features, i.e., lactic dehydrogenase (LDH), lymphocyte and High-sensitivity C-reactive protein (hsCRP) out of all 300+ features. The rationality of such mere three features may lie in that they are the representatives of tissue injury-, immunity- and inflammation-typed indices, respectively. From the COVID-19 patient diagnosis aspect, the work actualizes low-cost and prompt criticality classification and survival prediction before targeted intervening and diagnosis, especially for the triage of the large-scale explosive epidemic COVID -19 cases.

Keywords: COVID-19, Machine Learning, Clinical Biomarker

Funding: National Natural Science Foundation (NO. 61673189，91748112), National Key R&D Program of China (No. 2018YFB1004600).

Author's contribution: L. Y., H. Z., Y. X., H. X. and Y. Y. participated in study design; L. Y. collected data; M. W., C. S., Y. G., X. T., H. Z., Y. X., Z. C., Y. Y. performed data analysis, L. Y., H. Z., Y. X., H. X. and Y. Y. drafted the manuscript; and Y. Y., M. W. discovered the clinical features; all authors provided critical review of the manuscript and approved the final draft for publication.

Conflict of interest: None declared.

Introduction

December 2019 has witnessed the outbreak and swift spread of COVID-19 from Wuhan, China, to all over the world [1, 2]. By February 24th, 77,264 confirmed cases, 9,915 severe ill cases and 2,595 dead cases were confirmed and reported. The epidemic disease was caused by a coronavirus disease (COVID-19, previously termed as 2019-nCoV), given its severe infectivity and pathogenicity, which has been classified as a Public Health Emergency of

International Concern by the World Health Organization. The COVID-19 is an enveloped RNA virus which shares 88% identical genome sequence to that of bat-SARS-like coronavirus (bat-SL-CoVZC45) yet distinct from SARS-CoV and MERS-CoV[3-5].

The clinical features of patients infected with COVID-19 include fever, cough, shortness of breath, myalgia, fatigue, decrease of leukomonocyte and abnormal chest CT imaging [6-8]. With the improvement of diagnosis accuracy of COVID-19 patient, it becomes more urgent to identify and predict those seemly-mild cases promisingly progressing to critical cases due to their high death rate. Studies show 26.1-32% of patients exacerbate to critical illness such as acute respiratory disease syndrome (ARDS), shock, acute myopically injury (AMI) and acute renal failure (ARF) within a short time[2, 8]. It was reported that old patients are more prone to be infected COVID-19, especially for those with underlying diseases. According to the recent reports [7, 9, 10], Furthermore, Yang and colleagues induced the fatality rate of critically ill patients was 65.1%[11]. The severity of patients exerts great pressure on the shortage of intensive care resources. Therefore, we postulate reducing the transition from mild to severe and then to critically ill should be more sensible than trying to rescue endangered patients. The establishment of a novel model may accurately identify these patients, but no similar studies have yet been published.

Unfortunately, so far, the specific clinical features identifying different criticality stages of COVID-19 pneumonia still remain unclear, especially for those suffered from severe infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[11]. Due to the low survival-rate of critically ill cases, the current bottleneck of epidemic prevention has been shifted from cutting off the infection resource to prompt identification and prediction of critically ill cases prone to death by the clinical and laboratory features, and then intervene in advance before progressing to fetal cases. Traditionally, such critically ill and fatal cases are manually picked by doctors in accordance to the massive data from over 300 laboratory and clinical features, which inevitably leads to high misdiagnose and missed-diagnose rate dues to fatigue, personal

judgement inconsistency, and lack of experience and priori knowledge. Such a situation is severely intensified with sharp increase of patients crowded in ICU and emergent department.

As a remedy, nourished by the abundant inspection/clinical data, and the precious accumulated diagnosis experience, feature data-based machine learning becomes a promising method to help break through the bottleneck problem of survival-rate prediction. To this end, we have used an XGBoost machine learning method to establish a Prediction model according to the epidemiology and clinical feature data of 375 patients with confirmed COVID-19 infection admitted to Tongji Hospital, Wuhan. From technical point of view, the work helps to pave the way from machine learning method making full of the present available clinical and laboratory data to real applications in the triage of the large scale explosive epidemic COVID -19 cases.

Methods

Data resources

For this retrospective, single-center study, we collected 3,129 patients' electronic records confirmed or suspected COVID - 19 from January 10th to February 18th, 2020 at Tongji Hospital in Wuhan, China. We distilled epidemiological, demographic, clinical, laboratory, drugs, nursing record, outcome data from electronic medical record. The clinical outcomes were followed up to February 18th.

As show in Figure 1, of the 3,129 individuals retained in our hospital, 2,609 cases were excluded as they were still in treatment before February 19,2020. Per the other 520 cases, 375 ones with complete data material included 201 survivors. Pregnant or breast-feeding women, younger than 18 years older were excluded.

After February 19th, 2020, there were 26 new severe patients cleared, which were thus picked for test together with other 3 severe cleared patients from Ying Cheng People's Hospital for test. Note that all types of patients as samples for study, whereas just severe patients are selected for testing.

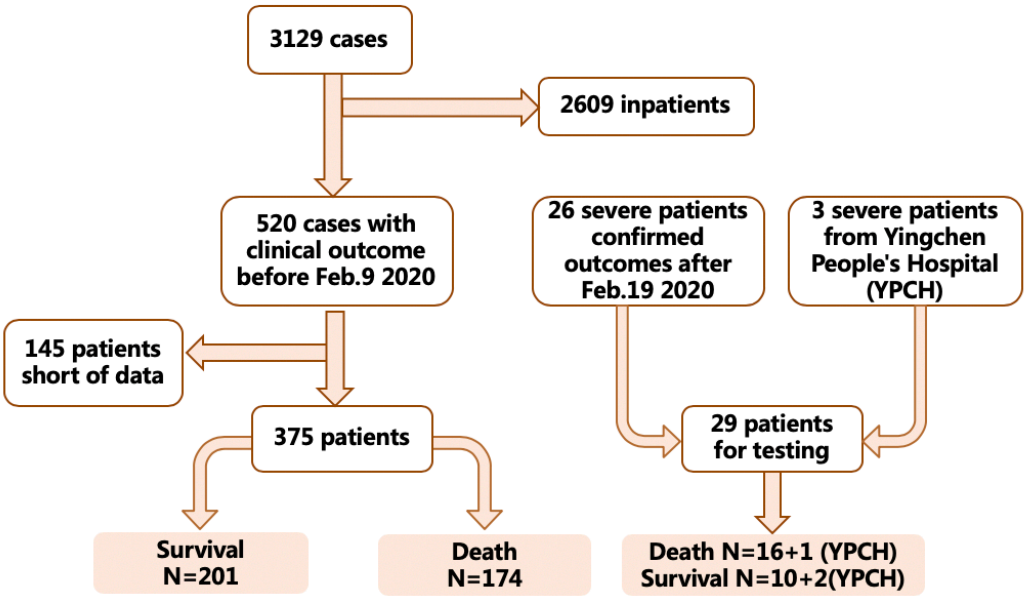

Figure 1. Patient enrollment flowchart.

The age distribution of the 375 patients was 43.59±18.59 years with 51.71% being males. Fever is the most common initial symptom (58.79%), followed by cough (20.79%), chest distress (6.66%), and fatigue (6.39%). The epidemiological history included Wuhan residents 38.39%, familial cluster 5.33%, only 0.8% were health worker. Others cannot get the information contact history. 375 patients were included in this study. Of the 375 patients, 20.27% were critical patients who identified by one of the three, (1) shock, (2) need mechanical ventilation and (3) admitted into ICU because of MODS. 33.86% patients were severe patients who identified by RR≥30bpm or SPO2 ≤93% on rest.

Table 1 Characteristics of the 375 patients

Characteristic		value
Age-yr		43.59±18.59
Male sex-no.(%)		51.71%
clinical symptoms		
	fever	58.79%
	cough	20.79%
	dyspnea	4.26%

	fatigue	6.39%
	chest distress	6.66%
	diarrhea	5.07%
Epidemiological history		
	Wuhan residents	38.39%
	familial cluster	5.33%
	Health worker	0.8%
severity of illness		
	critical	20.27%
	severe	33.86%
	normal	48.53%

Machine learning model

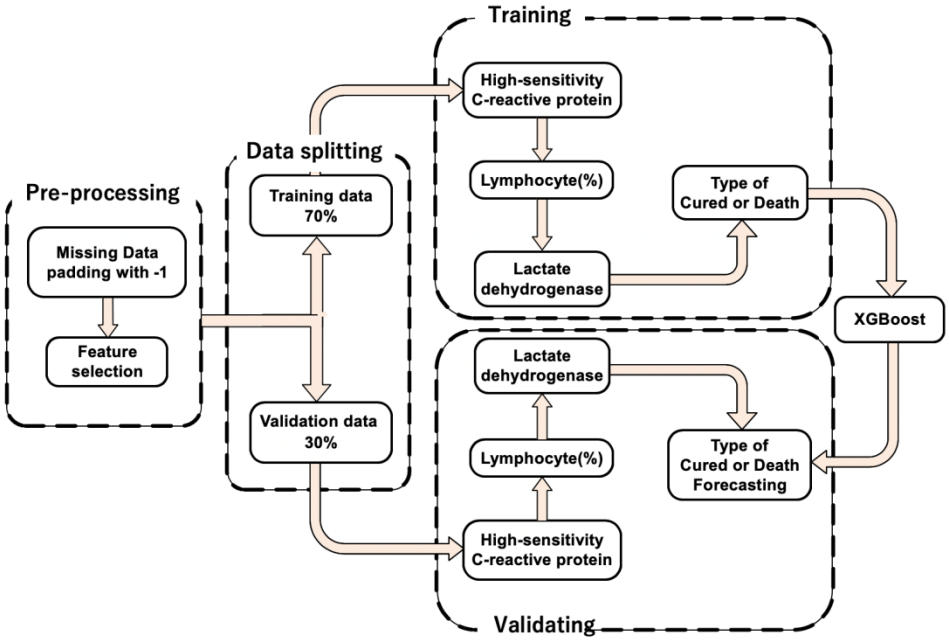

Figure 2. Flowchart of the XGBoost machine learning algorithm

In this study, a supervised XGBoost classifier [21] is used as the predictor, due to its superb pattern characterization and feature selection ability. As shown in Figure 2, the algorithm is detailed as below,

Data Pre-processing: Use “-1” padding method to complement the incomplete clinical measures in cases towards the patients with COVID-19.

Model Training: Randomly split the selected 2-category data is into a training set and a testing set, according to the ratio of 7:3. XGBoost is trained with the learning parameter setting as the max depth with 4, the tress number of estimators is set to 150, the values of the two regularization parameters α and β are set to 0.04 and 0.002 respectively, the subsample and max features both are set to 0.9.

Feature Selection: Select the top-3 key medical indicators by XGBoost, and afterwards retrain the XGBoost using such 3 key indicators for final prediction.

Model Prediction: Use the trained model to predict sample categories on the test set. Use the predicted and ground-truth label of test set, and then calculate the F1-socre for prediction performance evaluation.

Statistical Analysis

We first evaluate the performance of the model by assessing its predicted classification accuracy, equaling the ratio of the test samples predicted correctly. We calculated the area under the curve (AUC) of the receiver operating characteristic (ROC) curve of each class, as well as the sensitivity, specificity, and F1 score, with two sided 95% CIs. F1 score is the harmonic mean of the predictive positive value (PPV) and sensitivity, and was used to compare the relative performances of the XGBoost algorithm to the true labels. In terms of the prediction accuracy, the concept subset accuracy (i.e., exact match accuracy) was used to evaluate the ratio between the predicted label $\hat{y}_s$ and the ground-truth label y_s, according to

$$Accuracy = \frac{1}{S}\sum_{s=1}^{S}[\hat{y}_s = y_s] \quad (1)$$

where S denotes the sample number to be assessed. Besides, precision, sensitivity/recall, specificity, and F1 score of each class $n \in N$ represent the true positive (TP), true negative (TN), false positive (FP) and false negative rates (FN) calculated as below,

$$Precesion_n = \frac{TP}{TP+FP}, \quad (2)$$

$$Sensitivity_n/Recall_n = \frac{TP}{TP+FN}, \quad (3)$$

$$F1_n = \frac{2*Precision_n*Recall_n}{Precision_n+Recall_n}, \quad (4)$$

Results

Statistics of the Machine Learning model

We have carried out the prognostic AI model training with a randomly picked training dataset composed of 262 samples (70% of the whole dataset) with 126 death and 136 survival. It is observed that 97% (126/130) death prediction rate and 100 % (132/132) survival prediction rate are achieved, respectively. To show the AI model's performance on the training data more comprehensively, we demonstrate the precision, recall, F1-score and the corresponding support in Table 2, the score for survival and death prediction, accuracy, macro and weighted averages over all the samples keep larger than 0.98, which shows the prediction capability of the presently trained prognostic AI model.

Afterwards, we implement model validation with an independent validation dataset composed of 113 samples with 48 death. It is observed that still 97% death prediction and 97 % survival prediction accuracies are achieved, respectively, as well. Analogously to Table 2, to validate the AI model's performance on the testing data, we demonstrate the precision, recall, F1-score and the corresponding support in Table 3, the score for survival and death prediction, accuracy, macro and weighted averages over all the samples keep larger than 0.90 except the recall of survival (0.88) at the cost of the high recall of death (0.96). The effectiveness of the present prognostic AI model it thus verified.

Table 2. Performance of the proposed algorithm on training dataset.

	Precision	**Recall**	**F1-score**	**Support**
Survival	1.00	0.97	0.99	136
Death	0.97	1.00	0.98	126
accuracy			0.98	262
macro avg	0.98	0.99	0.98	262
weighted avg	0.99	0.98	0.98	262

Table 3: Performance of the proposed algorithm on validation dataset.

	Precision	**Recall**	**F1-score**	**Support**

Survival	0.97	0.88	0.92	65
Death	0.97	0.96	0.90	48
accuracy			0.91	113
macro avg	0.91	0.92	0.91	113
weighted avg	0.92	0.91	0.91	113

Outcomes: discovered three key clinical features

This method discovers that only three key features ('lactate dehydrogenase', 'lymphocyte (%)', 'High-sensitivity C-reactive protein'), are needed to distinct critical patients from two classes. From the feature importance sequence, one can figure out that: LDH > Lymphocyte (%)> Hs-CRP.

Prediction accuracy on the testing dataset

Finally, we test model with 29 patients specified in Figure 1 using only severe patients' three clinical features, whose outcome confirmed after February 19th. The confusion matrix of the testing data is shown in Figure 3, it is observed that still 100% death prediction accuracy and 90 % survival prediction accuracy are achieved, respectively. Analogously to Table 2, to validate the AI model's performance on the testing data, we demonstrate the precision, recall, F1-score and he corresponding support in Table 4, the score for survival and death prediction, accuracy, macro and weighted averages over all the samples keep larger than 0.90. We will discuss in the Discussion Section about these patients, whose outcome are predicted wrong.

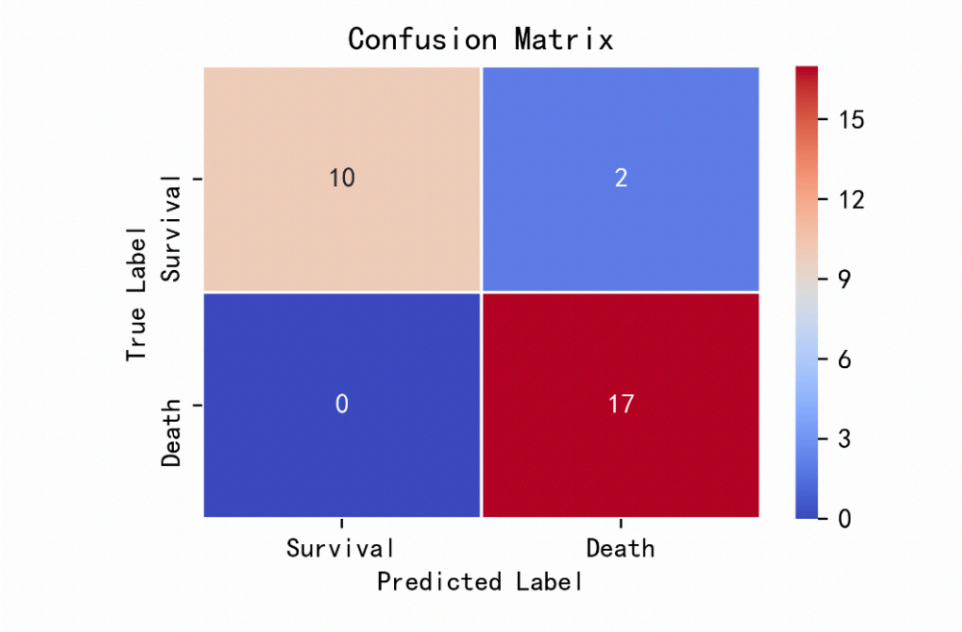

Figure 3. Confusion matrix on testing set.

Table 4: Performance of the proposed algorithm on testing dataset.

	Precision	Recall	F1-score	Support
Survival	1.00	0.83	0.91	12
Death	0.90	1.00	0.94	17
accuracy			0.93	29
macro avg	0.95	0.92	0.93	29
weighted avg	0.94	0.93	0.93	29

Discussion

Inspiringly, with the assistance of such a model, we have extracted merely three key clinical features, i.e., LDH, lymphocyte and hsCRP are extracted from all the 300+ features that precisely predict the survival rate with more than 90% accuracy. From the diagnosis aspect for COVID-19, these three discovered survival features realize low-cost and prompt criticality classification and survival rate prediction before targeted intervening and diagnosis, especially for the triage of the large scale explosive epidemic COVID -19 cases.

The increase of LDH reflects tissue destruction and is regarded as a common sign of tissue damage. LDH is expressed in a tissue-specific manner and

secreted during tissue injury. In patients with severe pulmonary interstitial disease, the increase of LDH is significant and is one of the most important prognostic factors of lung injury. Serum lactate dehydrogenase (LDH) has been identified as an important biomarker for the activity and severity of Idiopathic Pulmonary Fibrosis (IPF)[18]. The pathological features of COVID-19 are similar to those of SARS and MERS infection. Histological examination showed bilateral diffuse alveolar damage with cellular fibro myxoid Exudates, evident desquamation of pneumocytes and hyaline membrane formation, indicating acute respiratory distress syndrome(ARDS)[19] and then interstitial fibrosis. The increase of LDH level indicates that the activity and extent of lung injury in patients with COVID-19 are increased.

By investigating the clinical features of severely infected patients with COVID-19, our findings suggest that lymphocyte count play vital role in forecasting of progression from mild to critically ill and may serve as and may serve as a potential therapeutic target. The hypothesis can be validated with the results of clinical studies[7, 8]. ACE2 receptor binding is considered as the underlying affected pathogenic mechanism of COVID-19, owing to its large genetic diversity and frequent recombination, the clinical characteristics of COVID-19 remain largely unstable[12,13]. However, the injured alveolar epithelial cells would induce the infiltration of lymphocytes and lead to a persistent lymphopenia as SARS-CoV and MERS-CoV did, given they share the similar alveolar penetrating and antigen presenting cells (APC) impairing pathway[14,15]. Otherwise, the preferential accumulation of lymphocytes in the interstitial mononuclear inflammatory infiltrates is a prerequisite for ARDS, which could represent the severity and prognosis [16]. Jing and colleagues reported the lymphopenia is mainly related to the decrease of CD4+ and CD8+ T cells. Despite this, the biopsy study has provided strong evidence that the overactivation of CD4+ and CD8+ T cells is important for the accumulation of viral inclusion body[16]. Thus, selective activation of CD4+ and CD8+ T cells is presumably attributable to the profoundly cytokine storm, which deserve further investigation. In addition, the severity of cases between Hubei province

and other districts in the world seems diverse[17], which could be predicted by lymphopenia.

In the process of Acute lung injury(ALI) and ARDS, target cells are activated. The released inflammatory mediators activate with each other, forming SIRS and cascaded inflammatory storm. Age plays an important role in the role of hypersensitive CRP in ALI/ARDS. Most of the critically ill patients in COVID-19 are the elderly. Studies have shown that high hsCRP levels can predict higher mortality in elderly ALI patients[20], and plasma hsCRP levels in dead patients are significantly higher than those in surviving patients[20].

The predicted results for two patients from Yingcheng hospital were wrong. Yet, one of the patients had been admitted to the ICU because of an endangered condition and was recovered after emergency rescue. The other patient is in the cerebrovascular sequelae period with an extremely weak condition. Although currently alive, the prognosis is extremely poor.

This study has several limitations. First, this is a single-centered, retrospective study, which provides a preliminary assessment of the clinical course and outcome of critically ill patients. Second, although this database covers more than 3,000 patients, most clinical outcomes have not yet been presented due to its incompleteness. So the sample size is relatively limited, which could lead to bad performance of the proposed model. In this regard, we look forward to subsequent large sample and multicenter studies.

In summary, this study discovers three indicators (lymphocytes, lactate dehydrogenase, hypersensitive CRP), the accurate warning system makes it possible for the early detection, the early intervention and the reduction of mortality in high-risk patients with COVID-19. In the future exploration, we still need to consider more clinical confounding factors and Increase the sample size to further support these associations.

Reference

1. Chan, J.F.-W., et al., *A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster.* The Lancet, 2020. **395**(10223): p. 514-523.

2. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* The Lancet, 2020. **395**(10223): p. 497-506.
3. Corman, V.M., et al., *Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR.* Euro Surveill, 2020. **25**(3). doi: 10.2807/1560-7917.ES.2020.25.3.2000045.
4. Chen, Y., Q. Liu, and D. Guo, *Emerging coronaviruses: Genome structure, replication, and pathogenesis.* J Med Virol, 2020. **92**(4): p. 418-423.
5. Benvenuto, D., et al., *The 2019-new coronavirus epidemic: Evidence for virus evolution.* J Med Virol, 2020. **92**(4): p. 455-459.
6. Lei, J., et al., *CT Imaging of the 2019 Novel Coronavirus (2019-nCoV) Pneumonia.* Radiology, 2020: doi: 10.1148/radiol.2020200236. [Epub ahead of print]
7. Wang, D., et al., *Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China.* JAMA, 2020. doi: 10.1001/jama.2020.1585. [Epub ahead of print]
8. Chen, N., et al., *Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.* The Lancet, 2020. **395**(10223): p. 507-513.
9. Ronco, C., P. Navalesi, and J.L. Vincent, *Coronavirus epidemic: preparing for extracorporeal organ support in intensive care.* The Lancet Respiratory Medicine, 2020. doi: 10.1016/S2213-2600(20)30060-6. [Epub ahead of print]
10. Bassetti, M., A. Vena, and D.R. Giacobbe, *The novel Chinese coronavirus (2019-nCoV) infections: Challenges for fighting the storm.* Eur J Clin Invest, 2020: doi: 10.1111/eci.13209. [Epub ahead of print]
11. Lai, C.C., et al., *Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges.* Int J Antimicrob Agents, 2020: doi: 10.1016/j.ijantimicag.2020.105924. [Epub ahead of print]
12. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* The

Lancet, 2020. doi: 10.1016/S0140-6736(20)30251-8. Epub 2020 Jan 30.

13. Paraskevis, D., et al., *Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event.* Infect Genet Evol, 2020. **79**: doi: 10.1016/j.meegid.2020.104212. Epub 2020 Jan 29.
14. Li, F., et al., *Structure of SARS coronavirus spike receptor-binding domain complexed with receptor.* 2005. **309**(5742): p. 1864-1868.
15. Ge, X.-Y., et al., *Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor.* 2013. **503**(7477): p. 535-538.
16. Xu, Z., et al., *Pathological findings of COVID-19 associated with acute respiratory distress syndrome.* Lancet Respir Med, 2020. Feb. doi: 10.1016/S2213-2600(20)30076-X.
17. Xu, X.W., et al., *Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series.* BMJ, 2020. **368**: doi: 10.1136/bmj.m606.
18. Kishaba T , Tamaki H, Shimaoka Y , Fukuyama H, Yamashiro S. Staging of acute exacerbation in patients with idiopathic pulmonary fibrosis. Lung. 2014;192(1):141–9.
19. Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med. 2020 Feb 18. pii: S2213-2600(20)30076-X. doi: 10.1016/S2213-2600(20)30076-X. [Epub ahead of print]
20. Komiya K, Ishii H, Teramoto S, Takahashi O, Yamamoto H, Oka H, et al. Plasma C-Reactive protein levels are associated with mortality in elderly with acute lung injury. J Crit Care 2012;27(524):e521–6.
21. Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. (Submitted on 9 Mar 2016 (v1), last revised 10 Jun 2016 (this version, v3) arXiv:1603.02754 [cs.LG]