Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Demystifying COVID-19 Mortality Causes with Interpretable Data Mining

Xinyu Qian
Central South University
Zhihong Zuo
Xiangya Hospital Central South University
Danni Xu
Central South University
Shanyun He
Central South University
Conghao Zhou
University of Waterloo
Zhanwen Wang
Xiangya Hospital Central South University
Shucai Xie
Xiangya Hospital Central South University
Yongmin Zhang
Central South University
Fan Wu
Central South University
Feng Lyu
fenglyu@csu.edu.cn
Central South University
Lina Zhang
Xiangya Hospital Central South University
Zhaoxin Qian
Xiangya Hospital Central South University

Article

Keywords:

Posted Date: February 19th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3912968/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on May 2nd, 2024. See the published version at https://doi.org/10.1038/s41598-024-60841-w.

Demystifying COVID-19 Mortality Causes with Interpretable Data Mining

Xinyu Qian[1,+], Zhihong Zuo[2,+], Danni Xu[1], Shanyun He[1], Conghao Zhou[3], Zhanwen Wang[2], Shucai Xie[2], Yongmin Zhang[1], Fan Wu[1,*], Feng Lyu[1,*], Lina Zhang[2,*] and Zhaoxin Qian[2]

[1]School of Computer Science and Engineering, Central South University, Changsha, Hunan, China
[2]Department of Critical Care Medicine, Xiangya Hospital, Central South University, Changsha, Hunan, China
[3]Department of Electrical and Computer Engineering, University of Waterloo, Canada
[*]Corresponding author email: fenglyu@csu.edu.cn, zln7095@csu.edu.cn, wfwufan@csu.edu.cn
[+]These authors contributed equally to this work

ABSTRACT

While SARS-CoV-2 infection rates are declining, older adults remain vulnerable to severe disease with high mortality. Although there have been some studies on revealing different risk factors affecting the death of COVID-19 patients, such as bilirubin, organ failure, patient age, and underlying disease, they fail to provide a comprehensive analysis to reveal their relationships and interactive effects on the risk of death. Based on the demographic information, inspection indicators, and underlying diseases of 1917 patients (102 were dead) admitted to Xiangya Hospital over a 4-month period, we used the association rule mining method to identify the risk factors leading causes of death among elderly Omicron patients. Firstly, we used the Affinity Propagation clustering to extract key features such as blood parameters, liver function indicators, renal function indicators, coagulation function indicators, and underlying diseases affecting death from the dataset. Then, we applied the Apriori to obtain 7 groups of abnormal feature combinations with significant increments in mortality rate. The results showed a relationship between the number of abnormal feature combinations and mortality rates within different groups. For instance, patients with "C-reactive protein > 8 mg/L", "neutrophils percentage > 75.0 %", "lymphocytes percentage < 20 %", and "albumin < 40 g/L" have a 2x mortality rate than the basic one. If the characteristics of "D-dimer > 0.5 mg/L" and "WBC > $9.5 * 10^9$ /L" are continuously included in this foundation, the mortality rate can be increased to 3x or 4x. In addition, we also found that liver and kidney diseases significantly affect patient mortality. Given patients with liver and renal diseases associated with other abnormal features, their mortality rate can be as high as 100 %. These findings can support auxiliary diagnosis and treatment to, facilitate early intervention in patients, thereby reducing patient mortality.

Introduction

According to the recent World Health Organization (WHO) report on SARS-CoV-2 infections, the probability of death remains high in vulnerable populations with reinfection, especially the elderly[1,2]. Existing studies suggest that deceased patients were generally older than survivors, and the proportion was higher over 60 years of age, with a 58% increase in risk associated with every 10-year increase in age[3]. As individuals aged over 60 respond to Omicron infection, a decline in immune capacity, attributable to their age, may increase the likelihood of experiencing severe inflammation[4]. In addition, the death of elderly patients may arise from the deterioration of their basic diseases, chronic diseases, and comorbidities that they had already suffered due to COVID-19 infection[3,5,6]. Moreover, many studies have highlighted the relationship between COVID-19 and multi-organ damage, including effects on the heart, lung, liver, kidney, pancreas, spleen, and the progression of underlying disease[7].

At present, there are some noteworthy challenges in the diagnosis and treatment of COVID-19. First, the condition of individual patients significantly differs, and many risk factors may cause death, each of which may be related to others. Therefore, it is difficult to identify the exact risk factors causing the death of patients. Second, due to a large number of possible risk factors, existing studies narrow their focus to specific risk factors, such as physical patient indicators or lung infections. As a result, the existing studies may not provide a comprehensive understanding for each individual patient and cannot quickly reveal the relationship among different risk factors. Third, the advancement of machine learning techniques facilitates the identification of relationships among these risk factors, the obtained black-box results do not always offer clear explanations.

In this study, we gathered data of 1917 COVID-19 patients from Xiangya Hospital, Central South University. The data contain these patients' demographic information, inspection indicators, and underlying diseases. To investigate the risk factors influencing mortality, we initially employed the Affinity Propagation clustering method, a machine learning technique, to identify potentially significant indicators, including immune markers, liver and kidney function parameters, coagulation indices,

hematological parameters, etc. Subsequently, we employed the Apriori algorithm to mine data association rules. The classic example of beer and diapers can illustrate data association rule mining, where individuals who purchased beer are also likely to purchase diapers. Similarly, in our context, we considered patients' characteristics as "items like beer and diapers" and their examination conditions, such as "WBC > $9.5 * 10^9$ /L" and "Cancer," as potential "items" for their acquisition. Each patient was then associated with either the item of "deceased" or "survivor" based on their individual outcomes. By applying the mined association rule, we aim to identify patients who were associated with an increased occurrence of the "death" item, for identifying the factors affecting the fatality rate across COVID-19 patients. For instance, our analysis revealed that patients with characteristics such as "C-reactive protein > 8 mg/L", "neutrophils percentage > 75.0 %", "lymphocytes percentage < 20 %" and "albumin <40 g/L" exhibited a mortality rate was twice the baseline. Furthermore, the presence of "D-dimer > 0.5 mg/L" in conjunction with the aforementioned characteristics was associated with a mortality rate three times higher than the baseline. Through these feature combinations, we categorized the patients into seven groups, each showing an escalating trend in mortality as the number of relevant features increased. Notably, abnormal markers including age, CRP, lymphocyte and neutrophil levels, albumin, D-dimer, hemoglobin, and comorbidities, particularly liver and renal diseases, played a significant role in the escalation of mortality. These findings can aid clinicians in promptly identifying and intervening with high-risk COVID-19 patients, thereby improving patient care and resource allocation.

Results

Dataset Introduction: Our dataset is from COVID-19 patients admitted to Xiangya Hospital of Central South University between December 2022 and March 2023（Fig.1a). This dataset includes the records of 1917 patients and involves 46 laboratory indicators, basic disease information, and demographic details. Following meticulous data preprocessing (Fig.1b), we retained 1181 patients with complete information, among which, 102 patients were deceased and the overall mortality rate was 8.6%. Our dataset comprises comprehensive records spanning from admission to discharge (or demise), as detailed in Appendix Table 1. Since the patients are primarily from Hunan province, a non-response bias cannot be entirely eliminated.

Key Relevant Features Selection and Analysis: Initially, we conducted an analysis to compare the various indicators of the deceased with those of surviving patients. Significant disparities were observed in all fundamental indicators (refer to Appendix Table 1). Nevertheless, we cannot pinpoint the specific key indicators influencing mortality through this approach. Therefore, we employed the AP clustering algorithm to achieve a more refined patient classification, which can identify four patient groups with different mortality rates (Fig.1c).

Table 1. Differences in four groups of indicators

Indicator	Class1	Class2	Class3	Class4	P-value
C-reactive protein (mg/L)	20.21±29	57.14±58	48.54±51	116.02±82	<0.05
Hematocrit (%)	37.22±5	37.59±5	30.73±6	30.2±7	<0.05
WBC (10^9/L)	5.82±3	9.14±4	5.09±2	11.15±8	<0.05
Hemoglobin (g/L)	122.17±18	124.13±17	102.2±20	98.24±24	<0.05
Lymphocyte count (10^9/L)	1.62±2	0.89±1	0.75±0	0.7±1	<0.05
Percentage of lymphocytes (%)	27.33±12	10.81±6	17.13±11	7.06±6	<0.05
Uric Acid (umol/L)	285.79±103	275.17±116	269.52±116	460.78±194	<0.05
Creatinine (umol/L)	75.45±61	85.24±53	99.88±122	427.52±381	<0.05
D-dimer (mg/L)	0.38±1	0.63±1	0.75±1	4.05±5	<0.05
Age	59.82±22	68.04±18	72.26±16	69.05±16	<0.05
Death rate (%)	0.01±0	0.09±0	0.05±0	0.31±0	<0.05
Data are presented as mean ± SD, and P values were calculated with the use of the Kruskal-Wallis test.					

A subsequent statistical analysis for these four groups identified some indicators with significant differences. Based on the evaluation of these indicators, an experienced physician further evaluated these indicators and confirmed that patients in each of the four groups exhibited abnormalities in a total of 10 indicators: C-reactive protein, hematocrit, white blood cell count, hemoglobin, lymphocyte differential count, lymphocyte percentage, uric acid, creatinine, and D-dimer (refer to Table 1). Building upon previous research, we incorporated additional features such as albumin, total bilirubin, neutrophils, activated

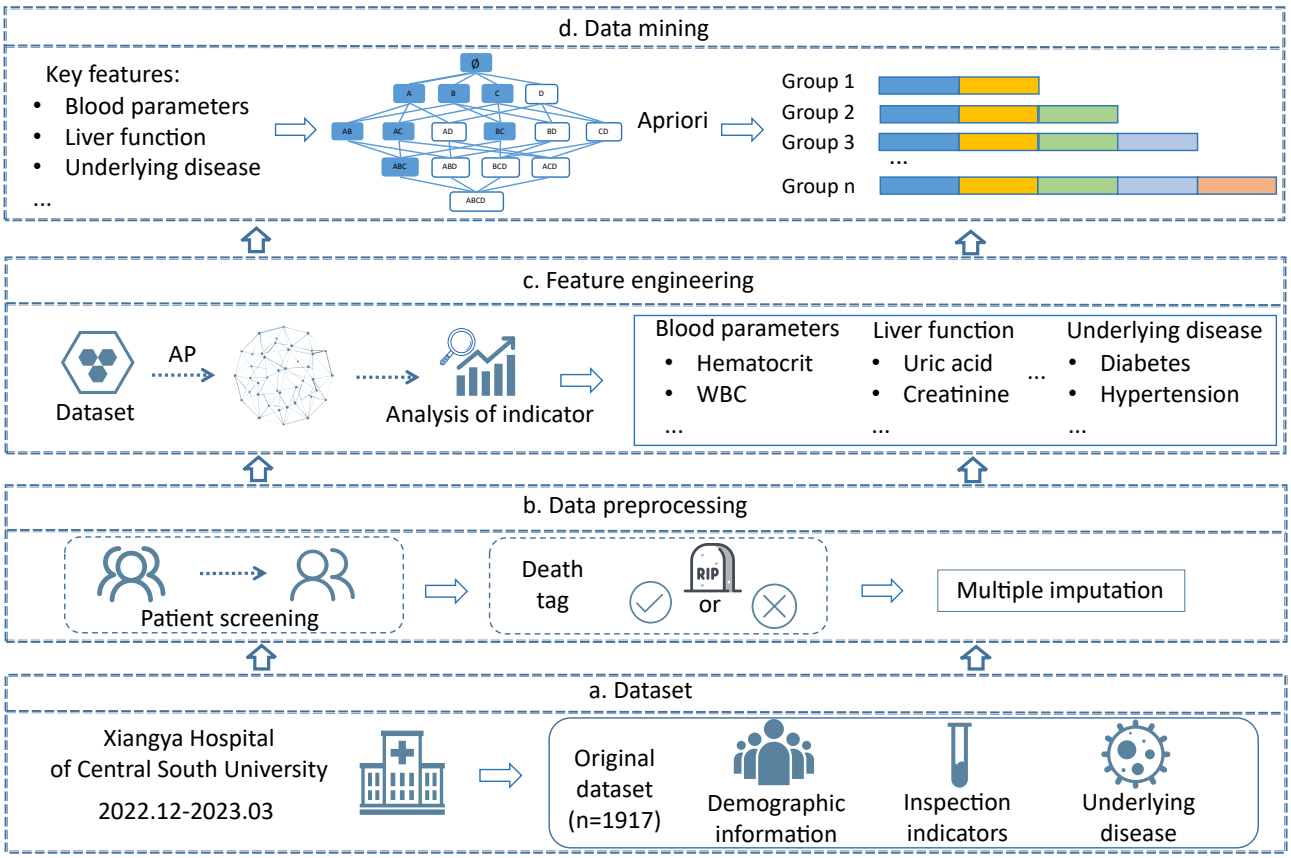


Figure 1. Frame Diagram. The diagram illustrates a comprehensive architecture designed to analyze the combined mortality indicators of COVID-19 patients. This overall mining process can be broken down into four primary stages: data collection, data preprocessing, feature selection, and association rule mining. In the initial stage, a detailed review of the examination data of COVID-19 patients is conducted, including screening of basic disease information and demographic details, and data integrity checks. Following data preprocessing, the processed dataset was utilized to identify key features affecting COVID-19 patient mortality. This selection process leveraged both machine learning clustering techniques and established medical knowledge. Lastly, employing the association rule mining method, we derived specific mortality patterns for COVID-19 patients. These patterns not only align with existing research but also reveal fresh insights, thereby supporting clinical in diagnosis and treatment protocols.

partial thromboplastin time, and underlying diseases. We established a set of 25 features potentially influencing the probability of death in COVID-19 patients.

Data Mining Results: We initiated an association rule data mining analysis on the feature set that may impact the mortality probability of COVID-19 patients. The primary objective of this study is to identify which combinations of abnormal features present at the time of admission exerted a significant influence on patient mortality (Fig.1d).

The mining results are detailed in Table 2. According to the results of association rule mining, we identified six distinct groups characterized by abnormal feature combinations associated with a substantial increase in mortality rate. To clarify this point, we clustered patients in group 1, who had the following abnormal features, including elevated CRP levels, heightened neutrophil percentages, reduced lymphocyte percentages, and lower albumin levels. Group 1 accounted for 84% of total deaths, with a mortality rate within this group of 17.3%. Notably, upon analysis and observation, we observed that patients in Groups 2-6 of abnormal patient features emerged by introducing additional abnormal features based on the establishment of the first group. As these features added, the mortality rate of patients within these groups demonstrated a noteworthy increase.

While the results for abnormal feature combinations can indicate mortality rates, they did not explicitly include underlying diseases for several reasons. Firstly, it is possible that the abnormal indicators, to some extent, represented the underlying diseases of the patients, leading to a degree of overlap, which obscured the distinct impact of underlying diseases. Secondly, the impact of underlying diseases on patient mortality may be comparatively weaker when contrasted with the influence of abnormal examination indicators. Lastly, the limited number of patients within this subset may result in the lack of underlying

Table 2. Six groups categorized by indicators

Indicator	Group1	Group2	Group3	Group4	Group5	Group6
C-reactive protein> 8 mg/L	✓	✓	✓	✓	✓	✓
neutrophils percentage> 75.0 %	✓	✓	✓	✓	✓	✓
lymphocytes percentage< 20 %	✓	✓	✓	✓	✓	✓
albumin< 40 g/L	✓	✓	✓	✓	✓	✓
D-dimer> 0.5 mg/L		✓	✓	✓	✓	✓
WBC> $9.5 * 10^9$ /L			✓	✓	✓	✓
Age> 70				✓	✓	✓
Age> 80						✓
lymphocyte count< $1.1 * 10^9$ /L				✓	✓	✓
hCT< 40 %				✓	✓	✓
creatinine> 111 umol/L					✓	✓
Uric Acid> 428.0 umol/L						✓
hemoglobin< 130 g/L						✓
Mortality within the group	17.30 %	27.10 %	35.40 %	48.20 %	66.70 %	71.40 %
Lift	2.1	3.2	4.2	5.7	7.9	8.5
Proportion of deaths	84 %	65 %	44 %	26 %	13 %	5 %

diseases, thereby causing these disease combinations to be overlooked by the association rule mining algorithm. To investigate the impact of underlying diseases on mortality, we analyzed the underlying diseases across different age groups.

As depicted in the table Table 3, (additional details available in Appendix Table 2), we observed that different combinations of underlying diseases yielded different effects on mortality. For the same set of underlying diseases, a gradual increase in mortality is correlated with the advancing age of patients. Meanwhile, most combinations of primary disease characteristics, such as cancer and lung diseases, exhibited a lower impact on mortality compared with the combinations of abnormal indicator characteristics. Notably, liver and kidney diseases had a more significant impact on mortality than that of other underlying diseases, with the mortality rate escalated further when patients had both liver and kidney diseases. However, the proportion of these patients is relatively small, which had previously been excluded from the association rule mining due to lower patient proportion thresholds. Consequently, we adjusted the patient proportion threshold to find a combination of abnormal patient characteristics involving liver and kidney diseases alongside other indicators. This finding, identified as combination 7, encompasses C-reactive protein> 8 mg/L, neutrophils percentage> 75.0 %, lymphocytes percentage< 20 %, albumin< 40 g/L, D-dimer> 0.5 mg/L, WBC> $9.5 * 10^9$ /L, Age> 80, lymphocyte count< $1.1 * 10^9$ /L, hCT< 40 %, creatinine> 111 umol/L, Uric Acid> 428.0 umol/L, hemoglobin< 13.g /L, totaling 12 indicators.

Table 3. Disease analysis results

Feature	Age>60	Age>70	Age>80	Cancer	PD	RFD	LFD	CI	HTN	Proportion of deaths	Death rate	Lift
Category6	✓	✓	✓	✓						8 %	27.60 %	3.3
Category10	✓	✓	✓		✓					10 %	11.80 %	1.4
Category23						✓				25 %	16.60 %	2
Category27							✓			10 %	14.50 %	1.7
Category36	✓					✓	✓			6.00 %	54.50 %	6.5
Category38	✓	✓	✓			✓	✓			3.50 %	57.10 %	6.8

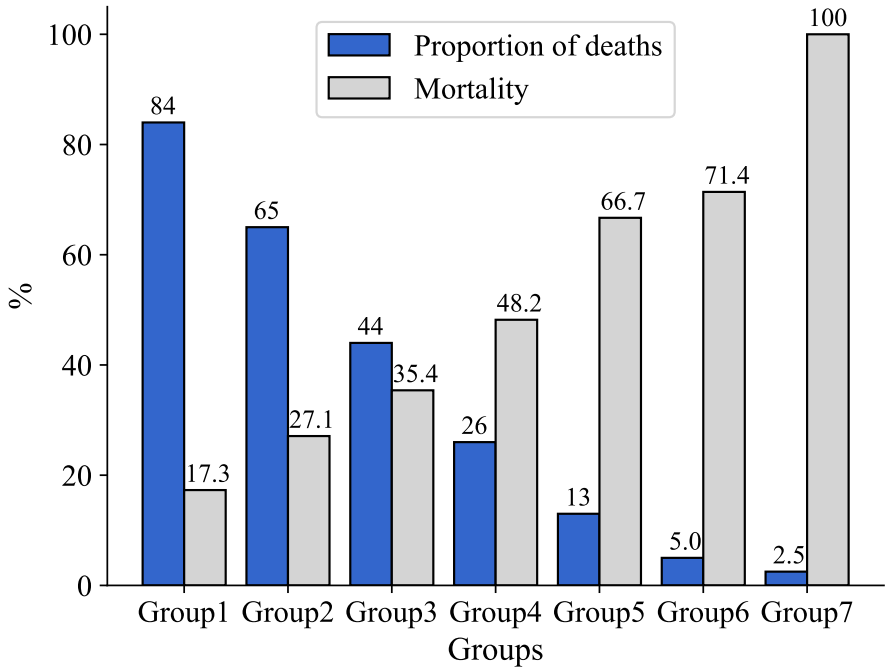


Figure 2. The mortality rate and the proportion of deaths in the seven groups.

We show bar plots to illustrate the proportion of deaths among patients and the mortality rate within each of the seven combinations of abnormal characteristics (Fig 2). The figure demonstrates that, with the increase of relevant features, the number of screened patients within each group exhibited a decreasing trend, while the mortality rate within the group continued to rise. This underscores the substantial impact of these selected indicators on patient mortality. Group 7, in particular, showed a higher incidence of liver and kidney diseases. Although the proportion of deaths in this group was relatively low, the internal mortality rate reached 100%, providing further evidence of the heightened impact of liver and kidney diseases on mortality.

We also plotted survival curves for each group (Fig 3). Notably, the rate of decline in the probability of survival over time varied considerably among the groups. As we progressed from group 1 to group 6, the risk of death and the rate of deterioration increased gradually with the accumulation of more abnormal indicators. Conversely, Group 7 demonstrated a sharp decline in the probability of survival within a short period, particularly when patients with liver and kidney diseases exhibited abnormalities in other characteristics.

Discussion

We used a combination of machine-learning methods and prior medical knowledge to screen for possible features on mortality. In the process of screening features using machine learning methods, we clustered 46 lab indicators of patients to obtain several groups of patients with different mortality rates. After evaluation by experienced physicians, which included the features impacting mortality, a subsequent data mining study was conducted. At the same time, combined with previous studies, we included albumin, total bilirubin, neutrophils, activated partial thromboplastin, and underlying diseases which include diabetes mellitus, hypertension, coronary heart disease, pulmonary disease, heart failure, myocardial infarction, cerebral infarction, arrhythmia, cancer, renal function disease and liver function disease. We used the selected features to mine the association rules of patient death, and finally obtained 7 groups of abnormal feature combinations that had obvious hints for mortality. In addition, we found that except for liver and kidney function diseases, other diseases had a smaller effect on mortality than examination features.

Association rule mining generated a basic group characterized by high-level CRP, high neutrophils, low lymphocytes, and low albumin. Basically, patients in this group were in the hyperinflammation and malnutrition status. The elevated level of CRP, a plasma protein released by the liver, indicates the presence of infection, especially bacterial pathogen or inflammatory diseases, particularly cardiovascular diseases[8]. The increase of CRP has been reported in several viral infections, such as H1N1[9] and recent SARS-CoV2[10–12]. Patients with elevated CRP level concentrations at the time of initial presentation were more likely to be associated with VTE, AKI, critical illness, and in-hospital mortality during subsequent hospital stays compared with those with lower initial measurements. Patients with the highest CRP level are the worst in this group[13]. This study clearly

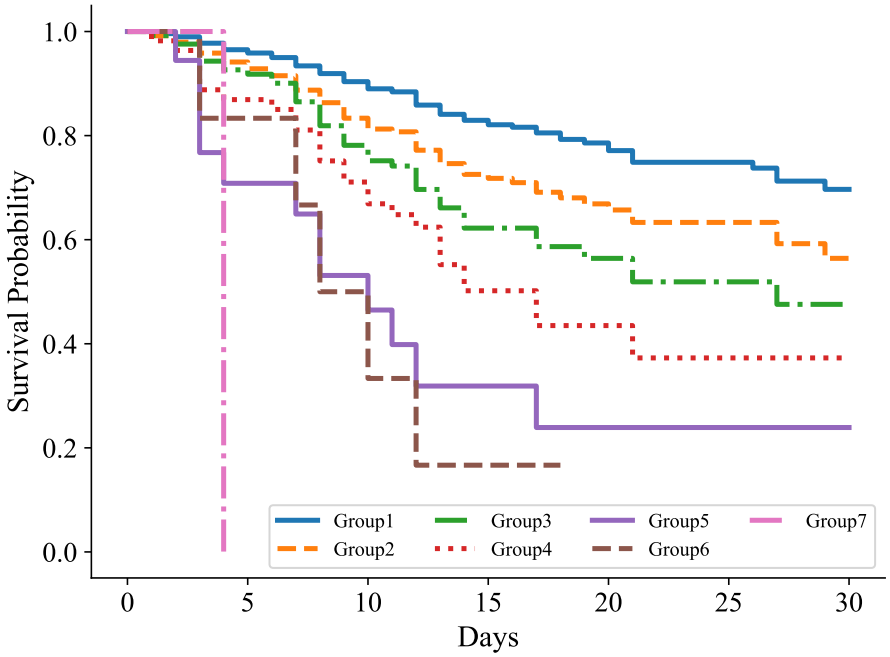


Figure 3. Seven groups of survival curves.

demonstrated the relationship between CRP level and COVID-19 severity. A retrospective study showed CRP measured both on admission and during the course of the disease in patients with COVID-19 was helpful for guiding therapy[14], which was further validated by a meta-analysis[15]. Additionally, both observational study and meta-analysis supported CRP as a prognostic factor in assessing disease lethality for COVID-19 patients[16–18]. Compared with patients without disease progression, those with disease progression presented persistently low lymphocyte counts and elevated CRP levels[19]. When CRP was combined with lymphocytes, it turned out to be a significant predictive factor in poor short-term clinical outcomes of SARS-CoV-2 BA.2.2 patients[20]. However, a multicenter study did not support the lymphocyte-to-CRP ratio as an accurate factor to predict disease severity or mortality[21], which indicated other factors should be coupled to achieve a more reliable prediction model. Moreover, the negative correlation between lymphocyte subset and CRP could be a useful tool to predict patients' responses to therapy, particularly for patients with relatively lower WBC[22]. Our study combined high CRP, high neutrophils, low lymphocytes, and low albumin as a group and defined a population at high risk of death, providing a new predictive strategy for COVID-19.

When further adding the level of D-dimer into evaluation, we identified a subset of the basic group with a higher mortality rate. D-dimer is a blood protein, released upon the breakdown of clots. The elevated level of D-dimer is believed to be associated with a high incidence of VTE formation, including PE and DVT, which might be useful to evaluate cardiovascular and hematologic diseases[23,24]. It was shown that serum D-dimer concentrations in severe COVID-19 patients were significantly higher than those presented with non-severe forms[25]. A systematic review and meta-analysis presented that the elevated level of D-dimer was positively related to the prevalence of thromboembolic events occurring in COVID-19 patients[26]. Moreover, the variation in the level of D-dimer was positively correlated with the prognosis of COVID-19 regardless of whether patients had cardiovascular diseases[27]. Irrespective of venous thromboembolism, the level of D-dimer at hospital admission for COVID-19 was associated with in-hospital mortality[28]. Hence, the level D-dimer is indicated as a significant factor for predicting the severity and the mortality of COVID-19[29,30]. The D-dimer level not only was closely related to COVID-19 severity but also might influence the likelihood of rapid onset of organ injury after admission[31]. However, a cohort study in Italy concluded that the predictive value of D-dimer alone was moderate, and adding D-dimer to a model with the consideration of age, sex, and co-morbidities as predictors did not lead to any meaningful improvement[32]. International guidelines suggest an age-adjusted D-dimer test threshold, rather than a simple D-dimer, must be considered for people older than 50 years[33]. Furthermore, D-dimer can guide anti-coagulation therapy for vein thrombosis and arterial embolism in COVID-19[23,34,35]. In our model, based on a high level of CRP, increased neutrophil, low lymphocytes, and low albumin, patients with a high level of D-dimer were with a higher mortality, indicating that the measurement of D-dimer is helpful for COVID-19 prediction, which is consistent with previous findings.

Individuals of any age are susceptible to SARS-CoV2, but older individuals are more vulnerable and suffering from an aggressive form of COVID-19. Studies have shown that longer viral shedding was tightly associated with advanced age[36].

Additionally, the age distribution of deaths in younger age groups (less than 65 years of age) was shown very consistent across different settings, whereas the mortality in individuals of age over 65 is significantly heterogeneous[37]. In the analysis of 72,134 cases, an overall case fatality rate (CFR) was 2.3%, while CFR was 8% in patients ageing from 70 to 79 years and 14.5% in patients aged 80 or/and over[38]. The similar patterns were suggested in studies conducted in Asia, Europe, and North America[39–41]. Essential explanations for this scenario would be the dynamic remodeling of immune response. As age advances, the responses of T cells and B cells decrease, and the dysfunction of innate immunity is observed as well, which is called immunosenescence[42]. Innate immunity, as the first defense, protects bodies against pathogens by secreting interferon to achieve an antimicrobial state to limit viral replication[43]. However, due to immunosenescence characterized by delayed IFN reaction, pathogens could escape from innate immunity at the early stage of infection, and older individuals would be at high risk of severe infection[44]. Delayed and insufficient activation of innate immunity contributed to weakened adaptive immunity, which is also a part of immunosenescence. All mentioned above postponed the clearance of virus and lead to dysregulated immune response, thereby causing the cytokine storm and severe conditions in older patients[45]. Most older individuals develop inflammaging, a condition characterized by elevated levels of blood inflammatory markers that carries a high susceptibility to chronic morbidity, disability, frailty, and premature death[46]. Under general circumstances, inflammaging functions as a protective mechanism in the older population. When balance is disturbed, the dysregulation of anti-inflammatory elements would bring about an excessive inflammatory response[45,47]. This high inflammatory status is consistent with high level CRP, as shown above, and the high production of IL-6 and tumor necrosis factor-a[48,49].

Based on previous studies of COVID-19, the presence of comorbidities usually indicates poor outcomes. As a meta-analysis shows, patients with a history of organ dysfunction, such as cerebrovascular disease, hypertension, liver dysfunction, and renal dysfunction, have a higher risk of developing severe COVID-19, which in turn can promote acute multiorgan injury[50,51]. In this model that includes common comorbidities, only liver and renal dysfunction are identified as significant factors for predicting high mortality. Liver function tests were found to be associated with systemic inflammation and organ dysfunction, which was an independent predictor of death or transfer to ICU in COVID-19[52,53]. The incidence of COVID-19 in people with chronic kidney disease (CKD) might be higher than in those without CKD[54], and COVID-19 related mortality is about ten times higher than that of CKD patients without COVID-19[55]. It is speculated that COVID-19 is influenced by two mechanisms. First, compared to other organ injuries, liver and renal dysfunction were more prevalent among the population, especially individuals of advanced age, who were at a relatively malnutrition and suppressed immune status[56]. Second, the overproduction of hyperinflammatory cytokines and chemokines is more likely to cause a lethal outcome in SARS and MERS[57]. It has also been found that patients with severe COVID-19 commonly had higher levels of inflammatory factors than those with non-severe forms, indicating that a cytokine storm was associated with disease severity[58]. Individuals with chronic diseases have been in a proinflammatory state with impaired immunity thus, are vulnerable to severe COVID-19.

Lung is a major target organ infected with SARS-CoV. Lung injury causes inflammation and infiltration, thereby affecting body oxygenation. Decreased hemoglobin levels can aggravate cell hypoxia and damage to tissue and organs. Hemoglobin levels were found to be significantly lower in severe COVID-19 for opposed to moderate case[59]. A retrospective cohort study implied that anemia was an independent risk factor for COVID-19[60], which was further verified by a prospective study in Iran[61]. Furthermore, hemoglobin of <12g/dL for females and < 13g/dL for males and were significantly associated with the risk of mortality[62]. All these results are consistent with what we found in the seven groups of patients with different mortality rates obtained by clustering factors associated with death. Therefore, healthcare providers should closely monitor the hemoglobin level of patients with COVID-19 on admission.

Conclusions The current study has effectively summarized the medical characteristics of the patients and has identified seven distinct patient groups at different levels of risk of mortality. Healthcare providers should heighten their awareness of features such as age, CRP, lymphocyte and neutrophil levels, albumin, D-dimer, hemoglobin, and comorbidities, particularly liver and renal diseases. Clinicians can identify individuals and implement tailored treatment strategies, ultimately facilitating reducing the incidence of severe COVID-19 cases and overall mortality.

Advantages and limitations Due to the limited patient sample size and the presence of a long-tail distribution in much of the medical research data, relying solely on techniques like LASSO and CPH to identify critical features for mortality research is inefficient. Specifically, it is challenging to identify whether the selected features truly bear significance for understanding mortality patterns or if they merely enhance the performance of mortality prediction models. Consequently, we adopted a multi-faceted approach, combining machine learning, expert physician evaluation, and existing research findings in our feature selection process. Instead of accepting the machine learning clustering results as the definitive patient classification; instead, we engaged clinicians to scrutinize and extract meaningful features. Subsequently, we incorporated these indicators in conjunction with prior research insights to facilitate interpretable data mining. In contrast to models like neural networks that directly yield patient mortality probabilities, these conclusions offer enhanced interpretability. They are more adept at guiding clinicians in formulating tailored treatment approaches. It is worth noting that the majority of our patients originated from Hunan Province, China. This may introduce certain demographic disparities and statistical biases into our analysis.

Methods

Ethical Approval: This study received approval from the Ethical Review Board of Xiangya Hospital, Central South University. All methods were performed in accordance with relevant guidelines and regulations. Informed consent from patients was obtained in this research context.

Study Design: We employed multiple imputations using chained equations to handle missing data below a specified threshold. The machine learning technique, Affinity Propagation (AP), was utilized to categorize patients into groups with significant differences in mortality risk. Subsequently, we selected statistically significant and clinically relevant features in each group, incorporating insights from relevant research papers to finalize our data features. The Apriori algorithm was then employed to unearth association rules related to mortality, identifying feature combinations that exhibited markedly higher mortality rates compared to the baseline and accounted for a substantial portion of patient fatalities.

Data Preprocessing: To ensure data integrity, we excluded patients who contracted the virus after admission, including patients only with COVID-19 at admission. Furthermore, we categorized deceased patients into two groups: patients passed away during hospitalization and patients succumbed after treatment withdrawal and discharge. The latter group underwent further scrutiny, with patients discharged within 1 day following ICU discharge or still under invasive intubation evaluated for potential treatment abandonment leading to post-discharge mortality. The final patient outcomes were cross-verified with medical records to ensure accurate death labels.

Data Filling: The dataset encompassed both numerical and binary features. Since there were no missing values in binary features, our focus was on completing the numerical features. We applied multiple imputations by chained equations to fill in missing numerical data points in cases where less than 25% of a patient's numerical features were missing.

AP Clustering: We employed the AP clustering algorithm for patient classification, utilizing all completed features as input and applying normalization. We set the preference parameter of the AP algorithm to -1300, which yielded four distinct patient groups with significant differences in mortality rates.

Apriori Algorithm: To identify abnormal feature combinations associated with substantial increases in mortality, we used the Apriori algorithm, a renowned association rule mining technique relying on key metrics such as support, confidence, and lift.

In this study, these three metrics correspond to the proportion of deaths, the mortality rate within each group, and the rate of mortality improvement, respectively. Notably, if an abnormal feature combination lacks relevance to mortality, the value of lift remains at 1. Otherwise, the value of lift exceeds 1. By defining the minimum support and the minimum confidence thresholds, we identified specific abnormal feature combinations. This indicates that, when patients were stratified based on these combinations, both the number of deaths and the group mortality rate within each group exceeded the defined threshold values. The calculation formulas are as follows: assuming X is a collection of various abnormal features, $X = \{feature_{i_1}, feature_{i_2}, feature_{i_3} \ldots\}$, X^+ is the collection of deaths who can fit a certain set of characteristics, $X^+ = \{feature_{i_1}, feature_{i_2}, feature_{i_3} \ldots, die\}$, and X^- denotes the collection of surviving patients who can fit a certain set of characteristics, $X^- = \{feature_{i_1}, feature_{i_2}, feature_{i_3} \ldots, survive\}$.

$$\text{Support(the proportion of deaths)} = \frac{X^+}{\text{all deaths}} \geq \text{threshold} \tag{1}$$

$$\text{Confidence(the mortality rate within the group)} = \frac{X^+}{X^+ + X^-} = \frac{X^+}{X} \geq \text{threshold} \tag{2}$$

$$\text{Lift} = \frac{(P(X^+))}{(P(X) * P(Death))} > 1 \tag{3}$$

Initially, during the exploration of the correlation between abnormal feature combinations and mortality, we identified six sets of feature combinations, each with a substantial increase in the mortality rate. Recognizing the importance of the connection between underlying diseases and mortality, we conducted a comprehensive analysis involving age and basic diseases. Building on the insights from mining basic diseases and their relationship with mortality, we adjusted the support and confidence thresholds to lower values. This allowed us to revisit the association between abnormal feature combinations and mortality, yielding a seventh set of results.

Data availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the confidential policy of National Health Commission of China, but are available from the corresponding author Lina Zhang upon reasonable request.

Author contributions statement

X.Q., Z.Z., F.W., F.L. and L.Z. contributed to the study design. X.Q., Z.Z., D.X., S.H., Z.W. and S.X. contributed to the literature search. Z.Z., Z.W. and S.X. contributed to collect data. X.Q. and Z.Z. contributed to data preprocess and data cleaning. X.Q., Z.Z., Z.W., L.Z., D.X., S.H. and F.L. contributed to data analysis, model implementation, and the creation of tables and figures. L.Z., F.W., F.L., Z.Q. and Y.Z. verified data and contributed to data interpretation. X.Q., Z.Z, D.X. and S.H. drated the initial manuscript, which was reviewed by C.Z., F.L., L.Z., Z.Q. and Y.Z., and all other co-authors contributed scientific inputs equally toward the interpretation of the finds and the final draft of the manuscript. All authors reviewed and edited the final draft. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Funding

This work was supported by the National Key Research and Development Program of China (2022YFC2009800 to ZQ) and Foundation of Shanghai Municipal Health Commission (202240204 to YL). The funders of the study had no role in the study design, data collection, analysis and interpretation, the writing of the manuscript, or the decision to submit the manuscript for publication.

Competing interests

The authors declare no competing interests.

References

1. Organization, W. H. *et al.* Who policy brief: Covid-19 surveillance, 11 april 2023. In *WHO policy brief: COVID-19 surveillance, 11 April 2023* (2023).
2. Abdullah, F. *et al.* Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, south africa. *Int. J. Infect. Dis.* **116**, 38–42 (2022).
3. Wang, Y. *et al.* Clinical course and outcomes of 344 intensive care patients with covid-19. *Am. journal respiratory critical care medicine* **201**, 1430–1434 (2020).
4. Zhang, L. *et al.* Characterizing distinct profiles of immune and inflammatory response with age to omicron infection. *Front. Immunol.* **14** (2023).
5. Du, Y. *et al.* Clinical features of 85 fatal cases of covid-19 from wuhan. a retrospective observational study. *Am. journal respiratory critical care medicine* **201**, 1372–1379 (2020).
6. Hawerkamp, H. C. *et al.* Characterisation of the pro-inflammatory cytokine signature in severe covid-19. *Front. Immunol.* **14**, 1170012 (2023).
7. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long covid: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
8. Sproston, N. R. & Ashworth, J. J. Role of c-reactive protein at sites of inflammation and infection. *Front. immunology* **9**, 754 (2018).
9. Vasileva, D. & Badawi, A. C-reactive protein as a biomarker of severe h1n1 influenza. *Inflamm. Res.* **68**, 39–46, DOI: 10.1007/s00011-018-1188-x (2019).
10. Luo, X. *et al.* Prognostic value of c-reactive protein in patients with coronavirus 2019. *Clin. Infect. Dis.* **71**, 2174–2179 (2020).
11. Gao, Y. *et al.* Diagnostic utility of clinical laboratory data determinations for patients with the severe covid-19. *J. medical virology* **92**, 791–796 (2020).
12. Liang, W. *et al.* Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19. *JAMA internal medicine* **180**, 1081–1089 (2020).
13. Smilowitz, N. R. *et al.* C-reactive protein and clinical outcomes in patients with covid-19. *Eur. Hear. J.* **42**, 2270–2279, DOI: 10.1093/eurheartj/ehaa1103 (2021).
14. Pink, I. *et al.* C-reactive protein and procalcitonin for antimicrobial stewardship in covid-19. *Infection* **49**, 935–943, DOI: 10.1007/s15010-021-01615-8 (2021).

15. Yong, S. J. *et al.* Inflammatory and vascular biomarkers in post-covid-19 syndrome: A systematic review and meta-analysis of over 20 biomarkers. *Rev. medical virology* **33**, e2424 (2023).

16. Sahu, B. R., Kampa, R. K., Padhi, A. & Panda, A. K. C-reactive protein: a promising biomarker for poor prognosis in covid-19 infection. *Clin. chimica acta* **509**, 91–94 (2020).

17. Stringer, D. *et al.* The role of c-reactive protein as a prognostic marker in covid-19. *Int. journal epidemiology* **50**, 420–429 (2021).

18. Paranga, T. G. *et al.* Comparison of c-reactive protein with distinct hyperinflammatory biomarkers in association with covid-19 severity, mortality and sars-cov-2 variants. *Front. Immunol.* **14**, 1213246 (2023).

19. Lin, S.-M. *et al.* Clinical and laboratory predictors for disease progression in patients with covid-19: A multi-center cohort study. *Biomed. J.* **46**, 100–109 (2023).

20. Xiao, B. *et al.* C-reactive protein to lymphocyte ratio is a significant predictive factor for poor short-term clinical outcomes of sars-cov-2 ba. 2.2 patients. *Front. Public Heal.* **11**, 1168375 (2023).

21. Abensur Vuillaume, L. *et al.* Lymphocyte-to-c-reactive protein (lcr) ratio is not accurate to predict severity and mortality in patients with covid-19 admitted to the ed. *Int. J. Mol. Sci.* **24**, 5996 (2023).

22. Mangano, C., Oliva, B. M. & Group, G.-C.-. W. Relationship between lymphocyte subsets values and c-reactive protein in covid-19 patients. *Cytom. Part A* **99**, 462–465 (2021).

23. Kleinegris, M.-C. F., ten Cate, H. & ten Cate-Hoek, A. J. D-dimer as a marker for cardiovascular and arterial thrombotic events in patients with peripheral arterial disease. *Thromb. haemostasis* **110**, 233–243 (2013).

24. van Hylckama Vlieg, A. *et al.* The risk of a first and a recurrent venous thrombosis associated with an elevated d-dimer level and an elevated thrombin potential: results of the the-vte study. *J. Thromb. Haemostasis* **13**, 1642–1652 (2015).

25. Paliogiannis, P. *et al.* D-dimer concentrations and covid-19 severity: a systematic review and meta-analysis. *Front. public health* **8**, 432 (2020).

26. Wu, T. *et al.* Venous thromboembolic events in patients with covid-19: A systematic review and meta-analysis. *Age Ageing* **50**, 284–293 (2021).

27. Li, Y. *et al.* Dynamic relationship between d-dimer and covid-19 severity. *Br. journal haematology* **190**, e24 (2020).

28. Chocron, R. *et al.* D-dimer at hospital admission for covid-19 are associated with in-hospital mortality, independent of venous thromboembolism: Insights from a french multicenter cohort study. *Arch. cardiovascular diseases* **114**, 381–393 (2021).

29. Ruetzler, K. *et al.* D-dimer levels predict covid-19 severity and mortality. *Kardiologia Polska (Polish Hear. Journal)* **79**, 217–218 (2021).

30. Popovska Jovičić, B. *et al.* Vitamin d, albumin, and d-dimer as significant prognostic markers in early hospitalization in patients with covid-19. *J. Clin. Medicine* **12**, 2825 (2023).

31. Zhang, W. *et al.* Association of d-dimer elevation with inflammation and organ dysfunction in icu patients with covid-19 in wuhan, china: a retrospective observational study. *Aging (Albany NY)* **13**, 4794 (2021).

32. Hassan, S. *et al.* The usefulness of d-dimer as a predictive marker for mortality in patients with covid-19 hospitalized during the first wave in italy. *PLoS One* **17**, e0264106 (2022).

33. Zuin, M., Rigatelli, G., Zuliani, G. & Roncon, L. Age-adjusted d-dimer cutoffs to guide anticoagulation in covid-19. *The Lancet* **398**, 1303–1304 (2021).

34. Stawiarski, K., Loutoo, A., Velardi, L. & Zarich, S. D-dimer driven deep vein thrombosis prophylaxis strategy for hospitalized patients with covid-19. *Thromb. Res.* **201**, 151–153 (2021).

35. Gatto, M. C. *et al.* Efficacy and safety of anticoagulant therapy in covid-19-related pulmonary embolism with different extension. *Biomedicines* **11**, 1282 (2023).

36. Aldhaeefi, M., Tahir, Z., Cote, D. J., Izzy, S. & El Khoury, J. Comorbidities and age are associated with persistent covid-19 pcr positivity. *Front. cellular infection microbiology* **11**, 650753 (2021).

37. O'Driscoll, M. *et al.* Age-specific mortality and immunity patterns of sars-cov-2. *Nature* **590**, 140–145 (2021).

38. Wu, Z. & McGoogan, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *jama* **323**, 1239–1242 (2020).

39. Kang, S.-J. & Jung, S. I. Age-related morbidity and mortality among patients with covid-19. *Infect. & chemotherapy* **52**, 154 (2020).

40. Mehra, M. R., Desai, S. S., Kuy, S., Henry, T. D. & Patel, A. N. Cardiovascular disease, drug therapy, and mortality in covid-19. *New Engl. J. Medicine* **382**, e102 (2020).

41. Salje, H. *et al.* Estimating the burden of sars-cov-2 in france. *Science* **369**, 208–211 (2020).

42. Castle, S. C. Clinical relevance of age-related immune dysfunction. *Clin. Infect. Dis.* **31**, 578–585 (2000).

43. De Wit, E., Van Doremalen, N., Falzarano, D. & Munster, V. J. Sars and mers: recent insights into emerging coronaviruses. *Nat. reviews microbiology* **14**, 523–534 (2016).

44. Agrawal, A. Mechanisms and implications of age-associated impaired innate interferon secretion by dendritic cells: a mini-review. *Gerontology* **59**, 421–426 (2013).

45. Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A. & Ng, L. F. The trinity of covid-19: immunity, inflammation and intervention. *Nat. Rev. Immunol.* **20**, 363–374 (2020).

46. Ferrucci, L. & Fabbri, E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat. Rev. Cardiol.* **15**, 505–522 (2018).

47. Fulop, T. *et al.* Immunosenescence and inflamm-aging as two sides of the same coin: friends or foes? *Front. immunology* **8**, 1960 (2018).

48. Tegethoff, S. A. *et al.* Tnf-related apoptosis-inducing ligand, interferon gamma-induced protein 10, and c-reactive protein in predicting the progression of sars-cov-2 infection: a prospective cohort study. *Int. J. Infect. Dis.* **122**, 178–187 (2022).

49. Wu, C. *et al.* Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, china. *JAMA internal medicine* **180**, 934–943 (2020).

50. Wu, T. *et al.* Multi-organ dysfunction in patients with covid-19: a systematic review and meta-analysis. *Aging disease* **11**, 874 (2020).

51. Gracia-Ramos, A. E., Jaquez-Quintana, J. O., Contreras-Omana, R. & Auron, M. Liver dysfunction and sars-cov-2 infection. *World J. Gastroenterol.* **27**, 3951 (2021).

52. Piano, S. *et al.* Abnormal liver function tests predict transfer to intensive care unit and death in covid-19. *Liver Int.* **40**, 2394–2406 (2020).

53. Li, X. Association of liver steatosis and fibrosis with clinical outcomes in patients with sars-cov-2 infection (covid-19). *Annals Hepatol.* **22**, 100295 (2021).

54. Chung, E. Y. *et al.* Incidence and outcomes of covid-19 in people with ckd: a systematic review and meta-analysis. *Am. J. Kidney Dis.* **78**, 804–815 (2021).

55. Gibertoni, D. *et al.* Covid-19 incidence and mortality in non-dialysis chronic kidney disease patients. *PloS one* **16**, e0254525 (2021).

56. Wang, M.-K. *et al.* Covid-19 and liver dysfunction: What nutritionists need to know. *World J. Gastroenterol.* **28**, 1526 (2022).

57. Channappanavar, R. & Perlman, S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. In *Seminars in immunopathology*, vol. 39, 529–539 (Springer, 2017).

58. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet* **395**, 497–506 (2020).

59. Taneri, P. E. *et al.* Anemia and iron metabolism in covid-19: a systematic review and meta-analysis. *Eur. journal epidemiology* **35**, 763–773 (2020).

60. Tao, Z. *et al.* Anemia is associated with severe illness in covid-19: a retrospective cohort study. *J. medical virology* **93**, 1478–1488 (2021).

61. Faghih Dinevari, M., Somi, M. H., Sadeghi Majd, E., Abbasalizad Farhangi, M. & Nikniaz, Z. Anemia predicts poor outcomes of covid-19 in hospitalized patients: a prospective study in iran. *BMC infectious diseases* **21**, 1–7 (2021).

62. Raman, N. *et al.* Ferritin and hemoglobin as predictors of fatal outcome in covid-19: two sides of the same coin. *J Assoc Physicians India* **69**, 11–12 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryinformationV2.pdf