Healthcare Analytics 2 (2022) 100020

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An explanatory analytics framework for early detection of chronic risk factors in pandemics

Behrooz Davazdahemami [a,*], Hamed M. Zolbanin [b], Dursun Delen [c,d]

[a] Department of IT & Supply Chain Management, University of Wisconsin-Whitewater, United States
[b] Department of MIS, Operations & Supply Chain Management, Business Analytics, University of Dayton, United States
[c] Department of Management Science and Information Systems, Oklahoma State University, United States
[d] Center for Health Systems Innovation, Spears School of Business, Oklahoma State University, United States

ARTICLE INFO

Keywords:
Machine learning
Explanatory artificial intelligence
Pandemic risk analysis pandemic
Diagnostic analytics

ABSTRACT

Timely decision-making in national and global health emergencies such as pandemics is critically important from various aspects. Especially, early identification of risk factors of contagious viral diseases can lead to efficient management of limited healthcare resources and saving lives by prioritizing at-risk patients. In this study, we propose a hybrid artificial intelligence (AI) framework to identify major chronic risk factors of novel, contagious diseases as early as possible at the time of pandemics. The proposed framework combines evolutionary search algorithms with machine learning and the novel explanatory AI (XAI) methods to detect the most critical risk factors, use them to predict patients at high risk of mortality, and analyze the risk factors at the individual level for each high-risk patient. The proposed framework was validated using data from a repository of electronic health records of early COVID-19 patients in the US. A chronological analysis of the chronic risk factors identified using our proposed approach revealed that those factors could have been identified months before they were determined by clinical studies and/or announced by the United States health officials.

1. Introduction

1.1. Objective

Public health emergencies such as pandemics call for timely and efficient decision making and large-scale dissemination of resources [1]. Particularly, effective care for patients of a novel, contagious disease (especially if it has a potential for developing into an epidemic or a pandemic) depends highly on how much and when we know about the demographic and chronic risk factors of the disease. Such information not only can be used to identify the optimal treatment procedures for each individual regarding their characteristics and underlying medical conditions [2], but also informs the high-risk population to follow the recommended health protocols. For this reason, since the onset of the recent COVID-19 pandemic, many clinical studies have attempted to identify the risk factors for severe disease and mortality among the patients. While these studies eventually provided an adequate understanding of the major risk factors for COVID-19, they are collectively inadequate for future pandemics for several reasons. First, most of the studies that were conducted in the early stages of the pandemic were, unsurprisingly, based on small samples of patients and left the complete identification of risk factors to later meta-analytic studies. Second, several studies focused on certain groups of patients (e.g., [3–5]) rather than on diverse samples that represented the general population. Third, studies conducted to identify high-risk patients in the COVID-19 pandemic were mostly descriptive and did not provide a comparative assessment of the risk factors. Finally, these studies did not lead to the development of a comprehensive framework for future pandemics. This is crucial since it enables us to skip reinventing the wheel in a global health crisis and simply plug the data obtained from a consolidated sample of initial patients into a decision support system.

This paper seeks to fill these gaps by using data from the early stages of the COVID-19 pandemic in the United States. Since the specification of 'early stages' is open to questions and arbitrary, we shift our focus from the timeline to the number of infected patients to determine how high-risk comorbidities can be identified as early as possible in an ongoing pandemic. While we focus our analyses on the data from the COVID-19 pandemic, the framework we propose is independent of the focal disease and works for any novel virus.

1.2. Approach

In the following section, we review the existing literature on risk factors for severe illness or mortality among COVID-19 patients (later

* Correspondence to: Department of IT & Supply Chain Management, University of Wisconsin-Whitewater, 809 W Starin Rd. Hyland Hall 1222, Whitewater, WI, 53190, United States.
E-mail addresses: davazdab@uww.edu (B. Davazdahemami), hmzolbanin@udayton.edu (H.M. Zolbanin), dursun.delen@okstate.edu (D. Delen).

https://doi.org/10.1016/j.health.2022.100020
Received 13 November 2021; Received in revised form 29 December 2021; Accepted 3 January 2022
Available online xxxx

in the paper, we use this evaluation of prior work as a benchmark to validate our proposed framework). Next, regarding the lack of prior literature on a novel disease when a pandemic occurs, we employ descriptive analytics along with a prescriptive feature selection approach to determine the appropriate sample size and critical factors. Subsequently, we use the knowledge obtained from the descriptive analyses to build predictive models for high-risk patients and illustrate how each factor affects the odds of prognosis among the patients. We conclude the paper with a discussion of our contributions, limitations, and avenues for future research.

1.3. Background

Regarding the extent of the COVID-19 pandemic, a sizeable number of studies have evaluated the risk factors for severe illness and death using data from inpatient hospital admissions. According to these studies, it is more likely for male [6–8] or older patients [8,9] to become severely ill or die from COVID-19. Among the chronic health conditions, obesity [10–12], hypertension [8,13,14], diabetes [8, 14,15], cardiovascular disease [14,16], respiratory diseases such as COPD [8,14], cerebrovascular illnesses [14], cancer [8], immunosuppression [17], kidney diseases [18,19], liver disease [20], solid organ transplant [21], and malnutrition [22] have been found by meta-analytic studies to increase the risk of severe illness or death among the patients.

The Centers for Disease Control and Prevention (CDC) provides a more detailed list of COVID-19 risk factors; however, it does not identify the order of importance of these factors [5]. These conditions include cancer; chronic kidney disease; chronic lung diseases such as COPD, moderate-to-severe asthma, interstitial lung disease, cystic fibrosis, and pulmonary hypertension; neurological conditions such as dementia; type 1 or type 2 diabetes; down syndrome; heart conditions such as heart failure, coronary artery disease, or hypertension; weakened immune system; liver disease; obesity; hemoglobin blood disorders such as sickle cell disease or thalassemia; a history of smoking or substance abuse; solid organ or blood stem cell transplant; and cerebrovascular diseases. As expected, the CDC's list of risk factors conforms to the findings of the meta analytics studies.

2. Materials and methods

2.1. Data

Inpatient and emergency department (ED) visits recorded in the Cerner HealthFacts data warehouse between December 31, 2019 (around the date China announced initial cases of the novel Coronavirus in Wuhan), and June 9, 2020, were used. The data were made available to the authors after being de-identified by Cerner to avoid any violation of patients' privacy. We excluded visits by non-adult (<18 years of age) patients. All the 10,341 records (related to 10,189 unique patients) in the data set belonged to patients who were diagnosed with COVID-19 through at least one of the various diagnostic lab tests.

In the next step of data processing, we used patients' historical health records to perform one-hot encoding for their chronic conditions (based on the ICD-10 coding system). We excluded acute conditions from our analyses because they could be a consequence of COVID-19 rather than a cause for an elevated risk among the patients. This resulted in 1047 binary variables, where each variable denoted the existence of a chronic comorbidity in a patient. In addition to the diagnosis codes, patients' demographics and visit-specific information (i.e., age, gender, race, admission type, and payer type) were included for further analyses.

Finally, we derived a binary response variable to indicate whether a patient survived or died from COVID-19. Based on our data set, 982 out of 10,189 (9.6%) patients died due to COVID-19. This rate is much higher than the average mortality rate (2%–3%) reported for the disease [23]; whereas this difference could partly be due to the fact that our data do not contain asymptomatic patients or those with mild symptoms who did not go to any healthcare facilities, yet it could also be partly attributed to the novelty of COVID-19, increased vulnerability of high-risk patients, and lack of resources to manage the disease in the early stages of the global outbreak. Table 1 summarizes the demographics of the final data set. Proportions of comorbidities among the two groups of patients (i.e., survived vs. died) are depicted in Fig. 1.

Table 1
Summary statistics of the patients.

Variable	Average (Std Dev)	Proportion	
Age	54.16 (20.26)		
Race		White	41.0%
		Black	30.9%
		Other/Unknown	27.1%
Gender		Male	51.5%
		Female	48.4%
		Other/Unknown	0.1%

2.2. Methodology

2.2.1. Feature selection

Since our goal was to propose and evaluate a framework for the identification of risk factors at the beginning stages of pandemics, we assumed that the risk factors of COVID-19 are still unknown. As a result, we employed an exploratory approach to select the chronic comorbid conditions that affect the likelihood of an outcome in COVID-19 patients.

Due to the large number of features, selecting a reasonably small set of most relevant factors that optimally predict death of patients was too computationally complex a problem to be solvable using the exact optimization methods. Hence, we employed Genetic Algorithm (GA) [24] – an evolutionary heuristic search approach – to come up with a decent nearly optimal solution. GA formulates each potential solution (i.e., feature set in our problem) as a chromosome-like string; it then generates a large number of chromosomes randomly (i.e., a population). Through several iterations (i.e., generations), chromosomes in each population are evaluated using a fitness function (i.e., prediction performance in this problem); then the top ones are kept for the next generation and the others are replaced by new chromosomes, partly generated randomly and partly generated by mating the top existing chromosomes (i.e., using crossover and mutation processes). The process of creating new generations continues until the changes in the fitness value of the best solution in a few consecutive generations is negligible (i.e., the algorithm converges).

We used a forward feature selection GA, with a population size of 1000 solutions in each iteration (i.e., generation) to refine the feature set. Each solution in the initial population included a random set of 50 conditions from the 1407 features of the data set (i.e., around 3.5%). It should be noted that we ran the GA several times with different chromosome sizes (ranging from 30 to 200 features); however, for chromosome sizes above 50 features, the amount of improvement in the fitness function was minimal. Thus, we chose 50 as the final size of the feature set. In each iteration, using the selected set of features for that iteration, we trained a basic random forest (RF) model with multiple chronological splits of data for training and validation. The reason we chose RF for feature selection was a large number of binary features in the data, which makes the data an excellent candidate for tree-based algorithms as compared to other machine learning techniques. We also tried other tree-based models, such as gradient boosted trees and a simple decision tree; however, RF yielded the best prediction results with the whole set of features in the model validation process.

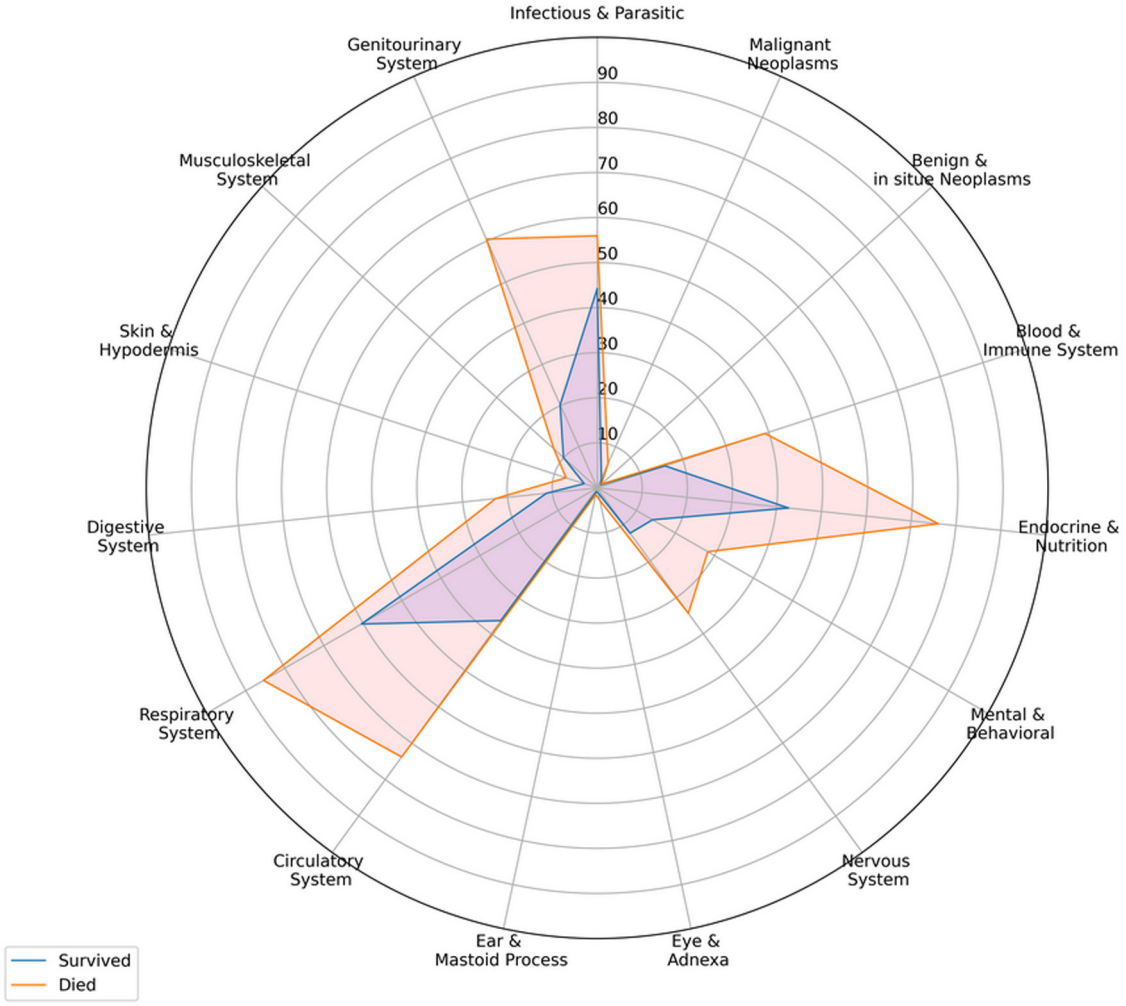

Fig. 1. Proportions of comorbidities by patient outcome.

Table 2
Chronological data splits.

Threshold date	Number (%) of training cases	Threshold date	Number (%) of training cases
01/31/2020	317 (3.12%)	03/22/2020	2481 (24.41%)
02/15/2020	500 (4.93%)	03/31/2020	3867 (38.8%)
02/29/2020	874 (8.63%)	04/07/2020	5241 (52.76%)
03/15/2020	1491 (14.77%)	04/15/2020	6751 (67.63%)
03/22/2020	2481 (24.41%)	04/22/2020	7958 (79.5%)

Table 2 shows the threshold dates used for splitting the data, as well as the proportion of training data associated with each threshold. Due to the slow rate of positive COVID-19 cases during January and February of 2020, we chose to go with jumps of two-week-long for the first few split thresholds but then decreased that to one week for the rest of the splits. A chronological split was critical since our goal was to identify the smallest number of patients (and therefore, the earliest possible time) for identifying a reliable set of risk factors.

In each run, we used the area under the receiver operating characteristic curve (AUC) of the trained model as the fitness function to identify the top feature sets of that generation. We chose AUC as the fitness function because it specifies the power of each feature set in distinguishing the deceased from the survived patients. We defined a tournament selection strategy for the GA with a 20% survival rate, 30% elitism rate, 30% crossover rate, and 10% mutation rate. A maximum of 100 generations was set for the algorithm, with the possibility of an early stop if no improvements were observed in AUC in 10 consecutive generations. For each split setting, the group of features associated with the highest AUC was retained to be used for fine-tuning the prediction models in the next step.

2.2.2. *Predictive modeling*

Since a majority of the predictors in our data were binary indicators of chronic comorbidities and given the high capability of tree-based models in handling categorical predictors, we selected RF to build the predictive models. Using the selected conditions for the split setting with the smallest training proportion (up to 01/31/2020), and by adding visit- and patient-related features, we fine-tuned the RF model hyperparameters to improve the predictive performance measures. To this end, we tried various split criteria (information gain, information gain ratio, and Gini index) for the decision trees and gradually increased the number of trees. An RF model with *information gain ratio* as the split criterion and 500 trees turned out as the winning configuration. We used this configuration to train the RF models for each of the data split settings (as pointed out in Table 2).

We compared the predictive powers of the RF models trained on each of the first nine data split settings (see Table 2) against the predictive power of the last setting (which included the highest training data proportion). The goal of this comparison was to determine the earliest split threshold that could yield a predictive model as good as the model that used around 80% of the data for training.

2.2.3. Model interpretation

Machine learning (ML) techniques are usually known as "black box" approaches with decent predictive power but little to no interpretability. In recent years, however, there has been a stream of research proposing various approaches, such as LIME [25], DeepLIFT [26], and Layer-Wise Relevance Propagation [27], to improve the interpretability aspect of ML (See [28,29] for a comprehensive review of such methods). Similarly, Lundberg and Lee [30] propose an intuitive approach called 'SHaply Additive exPlanations' (SHAP) to interpret a variety of ML models by weighing the marginal contribution of the feature values. SHAP combines the classic Shapley values approach [31] with a couple of other agnostic methods (including LIME and DeepLIFT) to assign each feature in the model an additive importance score for each data instance. The importance score for each feature represents the change in the expected model prediction when conditioning on that feature.

To calculate the score of any given instance i of feature X (i.e., $X(i)$), the SHAP approach considers all feature subsets (not including X itself) and computes the effect on predictions (i.e., deviation from the average of all predictions) of adding $X(i)$ to all those subsets. For a data set with N features, the method operates by training 2^N models, each with the whole data set and the same model hyperparameter settings, for all possible coalitions of features. Let us say M_1 and M_2 are two prediction models trained using the same subset of features, except one feature (f) which is present only in M_2. For each given instance in the data, the difference between its predicted value by M_1 and M_2 is considered as the marginal contribution of the feature f for that particular instance. Considering all marginal contributions of f across all possible coalitions of features, we may calculate a weighted average of those marginal contributions and attribute it to the feature f as its importance score (i.e., SHAP value) in predicting the outcome of that particular instance. In mathematical terms, the SHAP value of feature f for instance x is shown in Eq. (1) [30]:

$$SHAP_f(x) = \sum_{set\,:\,f \in set} \frac{P_{set}(x) - P_{set-f}(x)}{|set| \times \binom{F}{set}} \quad (1)$$

where set represents any subset of features containing f, $P_s(x)$ represents the predicted outcome for instance x using a model trained with the feature set s, $|set|$ indicates the size of a feature set, and F represents the total number of features. In fact, the importance score for each feature represents the change in the expected model prediction when conditioning on that feature.

Essentially, one major difference between SHAP and the classic Shapley values approach is the "local accuracy" property involved in SHAP, which enables it to explain every instance of a factor in the data by calculating a single marginal contribution for it; whereas Shapley values just assign an importance score to the whole factor (and not to each instance of data) [30]. As a result, using SHAP, we have an additive set of marginal contributions for each instance whose aggregation yields the prediction for that instance. Hence, in the specific context of this study, each patient's death probability (prediction) can be explained in terms of the additive marginal contributions of her corresponding health factors. Also, the overall importance score of each factor is simply the average of its marginal contributions across all instances (i.e., all patients).

In the last stage of our proposed framework, and after developing a predictive model (stage 2) using the optimally selected features (stage 1), we employ SHAP to interpret the selected RF models (i.e., the earliest decent model identified at stage 2 along with the model trained with around 80% of data) and to identify the major chronic conditions that contribute to one's death/survival after contracting COVID-19. Fig. 2 summarizes our proposed framework.

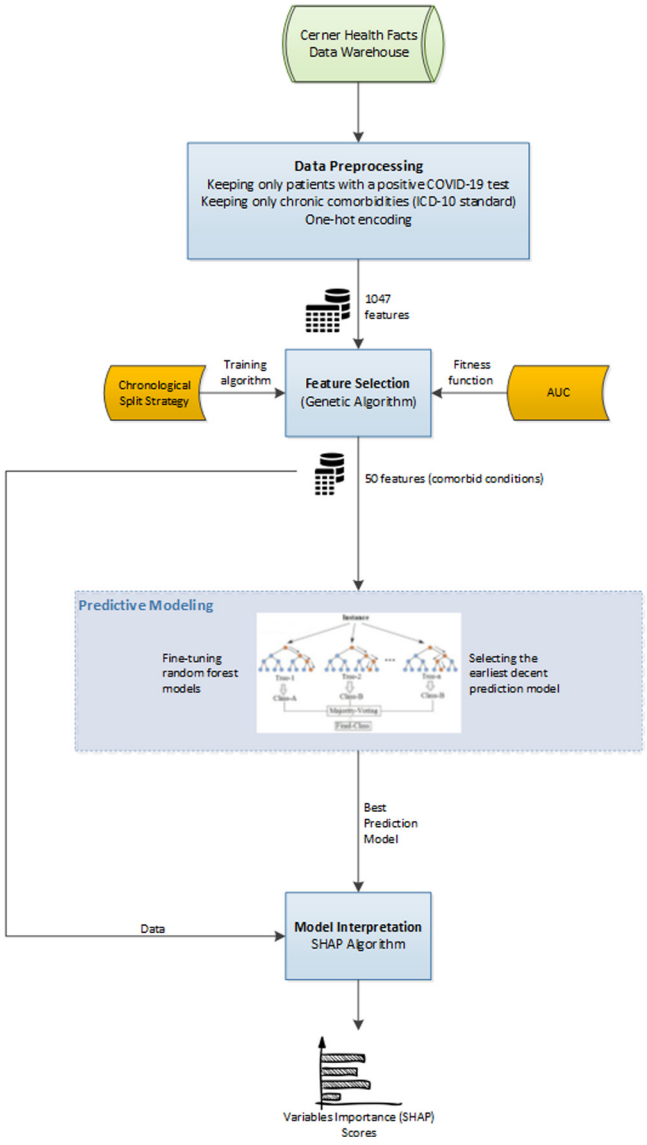

Fig. 2. The proposed framework.

3. Results and discussion

3.1. Feature selection

The GA models used for the different chronological splits were run on a computer with i9 2.90 GHz 8-Core processing power and 64 GB memory. The models converged after between 23 and 29 generations, which took around 3–4 h of processing in each case. The optimized feature set in each run involved 50 variables, each representing a chronic condition. We observed a 78% agreement rate across the different runs of the algorithm with different data splits. That is, 39 (out of 50) conditions were selected by the GA regardless of the data split settings. Also, the average best AUC across the different runs was 0.822 with a minimum of 0.814 and a maximum of 0.847. Table 3 contains a summary of the conditions commonly selected by all GA runs by their category (some conditions are common between multiple categories as they include complications).

In general, the categories of the risk factors identified in our study are the same as those found in meta-analytic studies; however, most of those studies were done several months after the beginning of the pandemic. Additionally, there are some conditions (categorized as "other")

Table 3
Selected chronic conditions.

Category	Number of conditions	ICD-10 codes
Heart	7	I13.0, I34.0, I42.4, I45.1, I48.9, Q23.4
Kidney	6	E11.2, I12.9, I13.0, N18.1, N31.9, Z94.0
Malignancy	2	C34.1, R18.0
Diabetes	3	E11.1, E11.2, E11.9
Hypertension	2	I12.9, I13.0
Blood and Circulatory system	8	D59.4, D69.6, D70.8, D72.8, E11.1, I66.0, I97.2, Q23.4, Z86.7
Nutritional and Metabolism	6	E03.9, E55.9, E78.1, E85.8, E89.3, M10.9
Respiratory	2	J30.9, J43.1
Other	11	F03.9, F40.0, G24.9, G47.2, G93.4, K26.4, L12.0, M13.8, Q85.0, R41.8, R53.8

that are most common among senior adults (e.g., dementia, duodenal ulcer, arthritis, and cognitive impairment), denoting that their selection could be related to patients' age rather than the condition itself. This is in line with the point noted by MacLeod and Hunter [32], who suggest that the age-dependent effects of COVID-19 must be taken into account in the analyses of the disease data.

3.2. Predictive model

We fine-tuned the RF model using the split configuration that included the smallest training portion (see Table 2) and used the same configuration to train the models with other split settings. Fig. 3 shows the changes in overall accuracy, recall, specificity, and AUC by increasing the training portion of the data chronologically.

For the first four split settings in Fig. 3, at least one of the four performance measures is relatively small. However, starting from the March 22nd split (associated with 24.4% training proportion), the models become more stable and the changes in performance measures are negligible. In other words, considering a chronological split strategy, a model trained as early as 80 days after the official announcement of the emergence of the novel coronavirus (but less than 10 days after the lockdown in the US) turned out good enough to correctly predict around 87% of deaths (i.e., recall) and 73% of survivals, mainly based on the patients' demographics and chronic conditions. With coordinated global efforts and established processes for data sharing and analysis [33], this period can potentially be shortened in future pandemics.

3.3. Models interpretation

While predicting survivability is a common application of data analytics in healthcare and involves various practical business implications [34,35], understanding the risk factors associated with the outcome is of even higher importance, especially for novel diseases. Therefore, in the last step, we employed the SHAP algorithm to look into the most relevant risk factors among the selected features. First, we ran the algorithm for the last prediction model (involving 79.5% training data). Fig. 4 (the blue bars) indicates the average SHAP scores for the top 20 features. We also ran the SHAP algorithm once more for the model trained with 24.4% of the data (i.e., March 22 split) to compare the score patterns. The patterned orange bars in Fig. 4 indicate the average SHAP scores obtained from that run.

A closer investigation of the two sets of bars shows several similarities: seven out of the top ten chronic risk factors are common between the two models. A relatively high rank-order correlation (i.e., 0.87) between the risk factor scores suggests that a majority of the insights obtainable from the April 2020 model could have been obtained a month earlier in March 2020. This is particularly important given that the US experienced its first major hike in COVID-19 cases in April with a 385% monthly increase [36] in the number of patients.

Expectedly, patients' *age* turned out as the top risk factor in both models. This risk factor is also identified by a large number of prior studies [37–40] as older people tend to have weaker immune systems. In addition, Race, Ethnicity, and Gender emerged as highly influential factors in both models. While there is little, if any, scientific evidence supporting the idea that some racial groups are more vulnerable to the COVID-19 infection, in line with Röösli et al. [41] we argue that the effect of race on the outcome in our models might be related to the socio-economic gaps in the US, as well as disparities in access to the required information and medical amenities in areas populated with minority racial groups. Unlike race and ethnicity, however, several prior studies on the risk factors of COVID-19 point to the greater vulnerability of men [6,40,42]. It should be noted that almost all the studies cited to validate our findings were published after March 2020. This means that the existence and application of the proposed framework in this study could have led to the identification of such risk factors weeks or months earlier than when they were noted in most scientific studies. Additionally, the number of historical diagnoses, as well as the number of diagnoses at the time of admission for each patient (both representing the general health of the patient), are among the top risk factors.

In terms of chronic risk factors, a history of the diseases of the circulatory system (ICD-10 code Z86.7) is identified as the top risk factor in both models. Also, blood disorders such as Thrombocytopenia and Lymphocytopenia have come out as the top risk factors among the COVID-19 patients in both models. A meta-analytic study performed by Lippi et al. [43] confirms that Thrombocytopenia is associated with more severe infections in COVID-19 patients. In addition, multiple studies mention Lymphocytopenia as a serious predictor of death in these patients [44–46]. Overall, in line with the findings of many studies published from mid-2020 to early 2021 (as shown in Table 1), our results suggest that chronic diseases of the circulatory systems, diabetes, kidney diseases, and chronic blood disorders are the major risk factors of COVID-19. A summary of the average SHAP scores for the major chronic conditions is shown in Table 4.

In addition to providing a high-level picture of the major risk factors and their relative importance, the SHAP algorithm offers individual-level analyses of the features and their contribution to the probability of an outcome for each patient. Such analyses could be particularly helpful for medical practitioners in making the best decision for each patient given their underlying conditions. Fig. 5, for instance, illustrates a waterfall chart for a given patient with separate bars for the risk factors. The first bar on the left indicates the average predicted probability (0.529) across the entire (balanced) training data set. Each of the other bars corresponds to one of the risk factors and specifies its contribution (positive = Blue, negative = Orange) to the predicted probability of death for a given patient.

As denoted on the horizontal axis, the patient shown in this chart is an 88 years old, white male with a large number of diagnoses (around 27) at the time of admission for COVID-19, suggesting poor general health. Among the chronic risk factors shown in Table 3, the patient has diabetes mellitus type II (ICD-10 = E11.9), Dementia (ICD-10 = F03.9), and age-related cognitive decline (ICD-10 = R41.8). The chart clearly shows that the main factors leading to the very high predicted probability of death for this patient (p = 0.904) are his age (+0.17), poor general health (+0.11), and diabetes (+0.14).

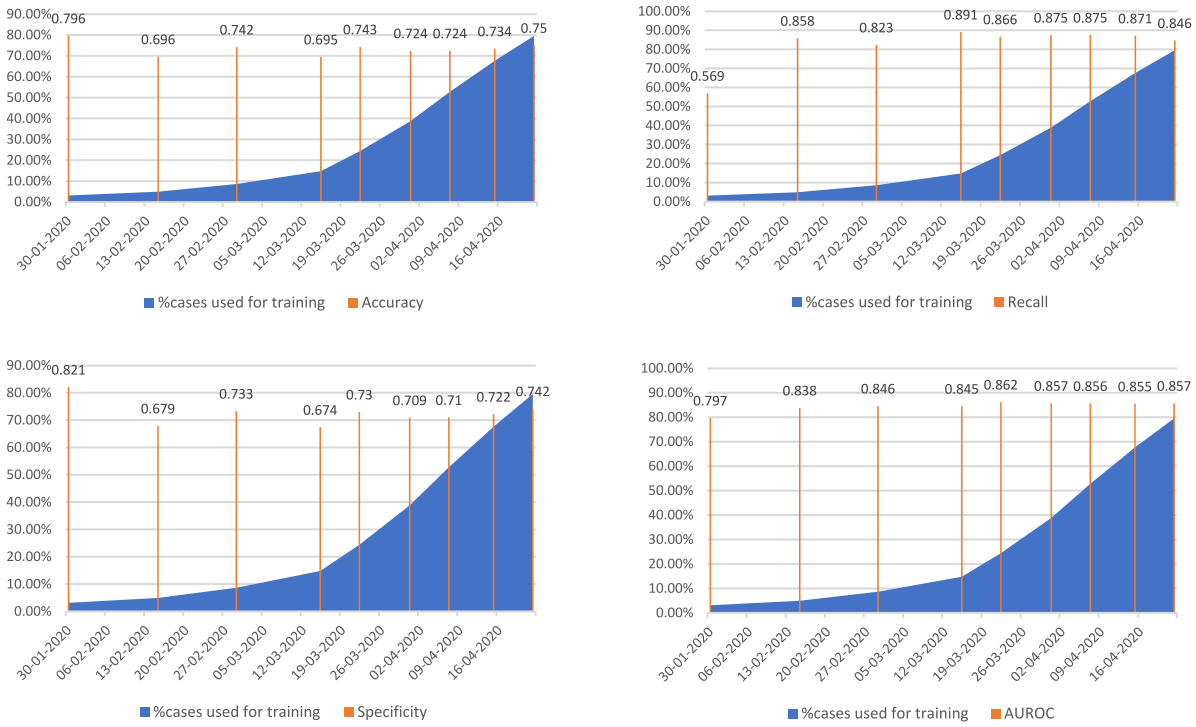

Fig. 3. Predictive performance of models with different data split settings.

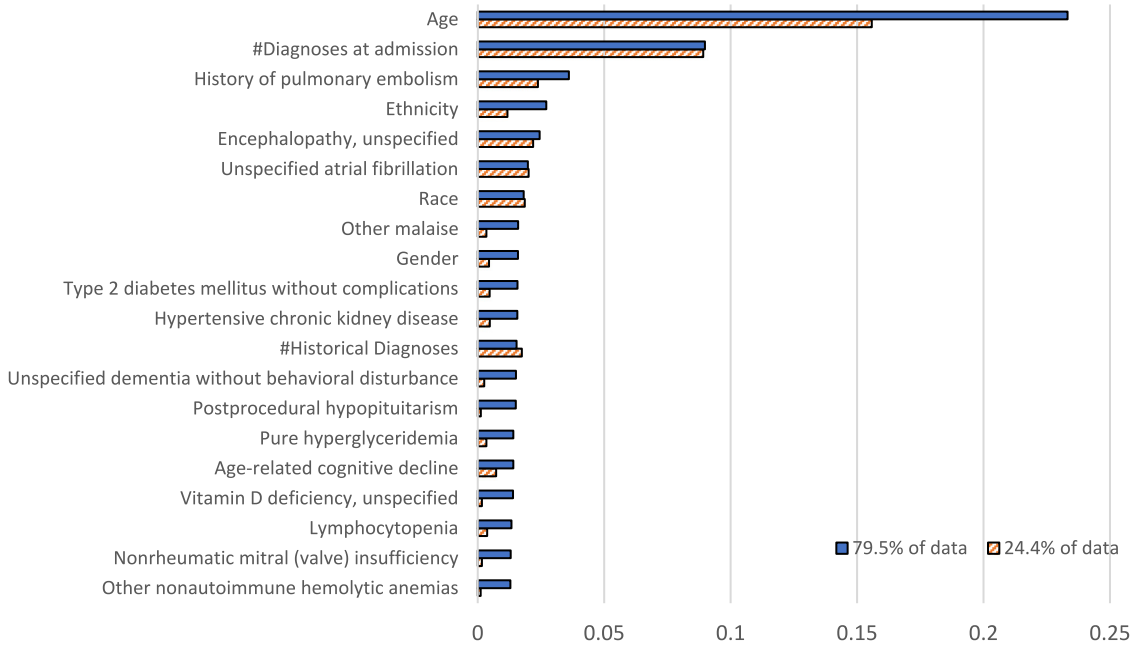

Fig. 4. Average SHAP scores with 79.5% (solid blue) vs. 24.4% (patterned orange) of data used for training.

4. Conclusion

One of the primary healthcare management concerns during a national or global outbreak of a novel infectious disease is to understand the main risk factors, which in turn, enables more effective use of limited resources and saves more lives. While controlled clinical trials are probably the most reliable approach in identifying such risk factors, the time required to complete them may lead to the loss of many lives. The present study leverages the power of predictive and prescriptive analytics along with the novel explanatory artificial intelligence algorithms to propose an applied framework for identifying major chronic risk factors at the time of epidemics or pandemics. We showcased the proposed framework by applying it to the data obtained from a US-based repository of EHR data that recorded COVID-19 incidents within the first 4 months of the pandemic.

Our results show that a majority of the chronic risk factors of COVID-19 could have been identified as early as mid-March 2020, around the time that WHO officially announced the disease as a global pandemic and much earlier than when those risk factors were recognized by peer-reviewed studies. Given the results of this study, as well as other studies that focus on the potential time savings obtainable by

Table 4
Average SHAP scores for each category of chronic conditions.

Category	Average SHAP Score (March 22)	Average SHAP Score (April 22)
Heart	0.0064	0.0065
Kidney	0.0029	0.0030
Malignancy	0.0001	0.0004
Diabetes	0.0044	0.0038
Hypertension	0.0067	0.0059
Blood and Circulatory system	0.0017	0.0013
Nutritional and Metabolism	0.0014	0.0016
Respiratory	0.0002	0.0003
Other	0.0027	0.0027

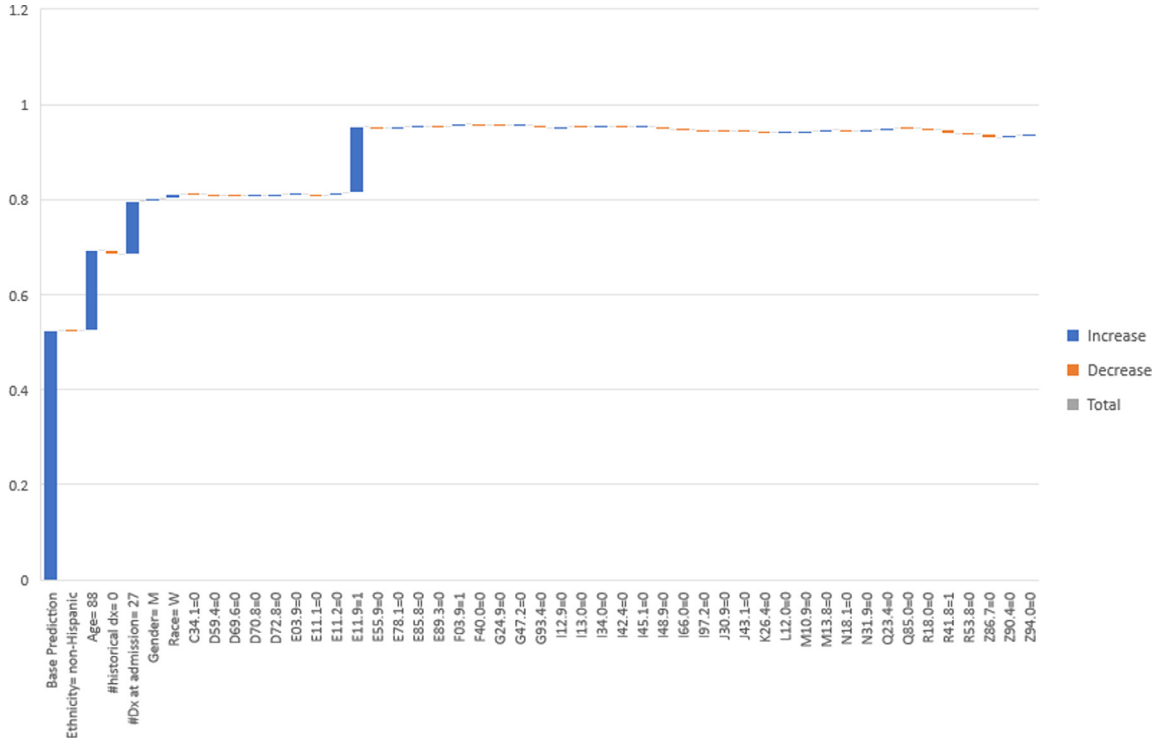

Fig. 5. SHAP scores for an individual patient (feature values are given on the horizontal axis). . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

employing data mining methodologies (e.g., [33] on early symptom detection), we believe that the national and international health agencies should coordinate their efforts and build the required infrastructure and capabilities to conduct data analytical investigations as early as possible in future large-scale health crises.

Although the risk factors identified by our proposed approach were previously recognized in the literature, what makes the present study stand out is the ability of the proposed framework to obtain the same results in a much shorter time. This period can be shortened even further if international organizations that are responsible for global health policies establish effective surveillance and data governance policies for future pandemics [33]. For instance, while we used data from the early US cases (until March 22, 2020) to reliably determine the risk factors of COVID-19, the same results could have been obtained earlier if there was an infrastructure that enabled data sharing between countries. We argue that the earlier identification of risk factors would have led to reduced overall mortality in two major ways. First, more accurate information could have been communicated to high-risk patients earlier in the pandemic. Second, limited care resources could have been prioritized and administered more efficiently for better outcomes.

As we pointed out earlier, a limitation of this study is the use of data from a single electronic health records (EHR) repository (i.e., not all the patients in the US). However, we believe that applying our proposed approach can save more time (and many lives) in pandemics, given the amount of data that are available to major national and international healthcare agencies. We encourage researchers who have access to multiple EHR sources to employ and validate our proposed framework, and to obtain a better estimate of the time or proportion of patients that are needed to identify risk factors of novel diseases reliably.

While our results suggest '*age*' as the top risk factor, looking into the chronic risk factors for each particular age group (e.g., 18–30, 30–45, etc.) and comparing the similarities and differences could be even more insightful. Unfortunately, the data set used in this study was too limited in terms of the number of positive instances (i.e., patients died from COVID-19) in the younger age groups to let us train separate predictive and explanatory models. Future research may adopt our proposed framework by employing a larger data set of COVID-19 patients to address that interesting research question.

Additionally, using COVID-19 data from other sources to replicate this study could help validating our findings and make them (fully or partially) more reliable for practical decision making. A comparison between the chronic risk factors of the initial variant of COVID-19 with its different variants, or with other similar viral diseases (e.g., Influenza) is another valuable interesting topic to be investigated by the future researchers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was conducted with data from the Cerner Corporation's HealthFacts datawarehouse of electronic medical records provided by the Oklahoma State University Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Cerner Corporation.

References

[1] E.K. Lee, C.-H. Chen, F. Pietz, B. Benecke, Modeling and optimizing the public-health infrastructure for emergency response, Interfaces (Providence) 39 (5) (2009) 476–490.
[2] X. Li, et al., Clinical characteristics of 25 death cases with COVID-19: a retrospective review of medical records in a single medical center, Wuhan, China, Int. J. Infect. Dis. 94 (2020) 128–132.
[3] N. Etienne, et al., HIV Infection and COVID-19: risk factors for severe disease, AIDS 34 (12) (2020) 1771.
[4] F. Peters, U. Marschall, C.-A. Behrendt, Prevalence of COVID-19 risk factors and risks of severe acute respiratory disease are markedly higher in patients with symptomatic peripheral arterial occlusive disease, Eur. J. Vasc. Endovasc. Surg. 61 (5) (2021) 859–860.
[5] B. Russell, et al., COVID-19 Risk factors for cancer patients: A first report with comparator data from COVID-19 negative cancer patients, Cancers (Basel). 13 (10) (2021) 2479.
[6] H. Peckham, et al., Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission, Nature Commun. 11 (1) (2020) 6317, http://dx.doi.org/10.1038/s41467-020-19741-6.
[7] T. Galbadage, et al., Systematic review and meta-analysis of sex-specific COVID-19 clinical outcomes, Front. Med. 7 (2020) 348.
[8] M. Parohan, S. Yaghoubi, A. Seraji, M.H. Javanbakht, P. Sarraf, M. Djalali, (COVID-19) Infection: a systematic review and meta-analysis of observational studies, Aging Male (2020) (2019) 1–9.
[9] K. Romero Starke, et al., The age-related risk of severe outcomes due to COVID-19 infection: a rapid review, meta-analysis, and meta-regression, Int. J. Environ. Res. Public Health 17 (16) (2020) 5974.
[10] M. Földi, et al., Obesity is a risk factor for developing critical condition in COVID-19 patients: a systematic review and meta-analysis, Obes. Rev. 21 (10) (2020) e13095.
[11] V.S. Malik, K. Ravindra, S.V. Attri, S.K. Bhadada, M. Singh, Higher body mass index is an important risk factor in COVID-19 patients: a systematic review and meta-analysis, Environ. Sci. Pollut. Res. 27 (33) (2020) 42115–42123.
[12] A. Hussain, K. Mahawar, Z. Xia, W. Yang, E.-H. Shamsi, Obesity and mortality of COVID-19. Meta-analysis, Obes. Res. Clin. Pract. (2020).
[13] Y. Du, N. Zhou, W. Zha, Y. Lv, Hypertension is a clinically important risk factor for critical illness and mortality in COVID-19: A meta-analysis, Nutr. Metab. Cardiovasc. Dis. 31 (3) (2021) 745–755.
[14] B. Wang, R. Li, Z. Lu, Y. Huang, Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis, Aging (Albany NY) 12 (7) (2020) 6049.
[15] A. Mantovani, C.D. Byrne, M.-H. Zheng, G. Targher, Diabetes as a risk factor for greater COVID-19 severity and in-hospital death: a meta-analysis of observational studies, Nutr. Metab. Cardiovasc. Dis. 30 (8) (2020) 1236–1248.
[16] F. Javanmardi, A. Keshavarzi, A. Akbari, A. Emami, N. Pirbonyeh, Prevalence of underlying diseases in died cases of COVID-19: A systematic review and meta-analysis, PLoS One 15 (10) (2020) e0241265.
[17] D. Tassone, et al., Immunosuppression as a risk factor for COVID-19: a meta-analysis, Intern. Med. J. 51 (2) (2021) 199–205.
[18] B.M. Henry, G. Lippi, Chronic kidney disease is associated with severe coronavirus disease 2019 (COVID-19) infection, Int. Urol. Nephrol. 52 (6) (2020) 1193–1194.
[19] R. Cai, J. Zhang, Y. Zhu, L. Liu, Y. Liu, Q. He, Mortality in chronic kidney disease patients with COVID-19: a systematic review and meta-analysis, Int. Urol. Nephrol. (2021) 1–7.
[20] F.M. Noor, M.M. Islam, Prevalence and associated risk factors of mortality among COVID-19 patients: A meta-analysis, J. Community Health 45 (6) (2020) 1270–1282.
[21] M.A. Raja, et al., COVID-19 In solid organ transplant recipients: A systematic review and meta-analysis of current literature, Transplant. Rev. (2020) 100588.
[22] S.M. Abate, Y.A. Chekole, M.B. Estifanos, K.H. Abate, R.H. Kabtyimer, Prevalence and outcomes of malnutrition among hospitalized COVID-19 patients: A systematic review and meta-analysis, Clin. Nutr. ESPEN (2021).
[23] Johns hopkins coronavirus resource center, 2021.
[24] J.H. Holland, Genetic algorithms, Sci. Am. 267 (1) (1992) 66–73.
[25] M.T. Ribeiro, S. Singh, C. Guestrin, 'Why should I trust you?' Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[26] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, 2017, arXiv Prepr. arXiv:1704.02685.
[27] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.
[28] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.
[29] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 80–89, http://dx.doi.org/10.1109/DSAA.2018.00018.
[30] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
[31] L.S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. 39 (10) (1953) 1095–1100.
[32] M.R. MacLeod, D.G. Hunter, The impact of age demographics on interpreting and applying population-wide infection fatality rates for COVID-19, INFORMS J. Appl. Anal. 51 (3) (2021) 167–178.
[33] H.M. Zolbanin, A. Hassan Zadeh, B. Davazdahemami, Miscommunication in the age of communication: A crowdsourcing framework for symptom surveillance at the time of pandemics, Int. J. Med. Inform. 151 (2021) 104486, http://dx.doi.org/10.1016/j.ijmedinf.2021.104486.
[34] M.F. Gensheimer, et al., Automated model versus treating physician for predicting survival time of patients with metastatic cancer, J. Am. Med. Inform. Assoc. (2020) http://dx.doi.org/10.1093/jamia/ocaa290.
[35] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst. 74 (2015) 150–161, http://dx.doi.org/10.1016/j.dss.2015.04.003.
[36] National geographic, 2021.
[37] R. Pastor-Barriuso, et al., Infection fatality risk for SARS-CoV-2: a nationwide seroepidemiological study in the non-institutionalized population of Spain, MedRxiv (2020) 2020.08.06.20169722, http://dx.doi.org/10.1101/2020.08.06.20169722.
[38] M. O'Driscoll, et al., Age-specific mortality and immunity patterns of SARS-CoV-2 infection in 45 countries, MedRxiv (2020) 2020.08.24.20180851, http://dx.doi.org/10.1101/2020.08.24.20180851.
[39] E. Eryarsoy, D. Delen, B. Davazdahemami, Adjusting COVID-19 reports for countries' age disparities: A comparative framework for reporting performances, 2020, http://dx.doi.org/10.1101/2020.08.31.20185223.
[40] N.D. Yanez, N.S. Weiss, J.-A. Romand, M.M. Treggiari, COVID-19 Mortality risk for older men and women, BMC Public Health 20 (1) (2020) 1742, http://dx.doi.org/10.1186/s12889-020-09826-8.
[41] E. Röösli, B. Rice, T. Hernandez-Boussard, Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19, J. Am. Med. Inform. Assoc. 28 (1) (2021) 190–192, http://dx.doi.org/10.1093/jamia/ocaa210.
[42] S.S. Bhopal, R. Bhopal, Sex differential in COVID-19 mortality varies markedly by age, Lancet 396 (10250) (2020) 532–533, http://dx.doi.org/10.1016/S0140-6736(20)31748-7.
[43] G. Lippi, M. Plebani, B.M. Henry, Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis, Clin. Chim. Acta 506 (2020) (2019) 145–148, http://dx.doi.org/10.1016/j.cca.2020.03.022.
[44] Q. Zhao others, Lymphopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A systemic review and meta-analysis, Int. J. Infect. Dis. 96 (2020) (2019) 131–135, http://dx.doi.org/10.1016/j.ijid.2020.04.086.
[45] A. Ziadi, et al., Lymphopenia in critically ill COVID-19 patients: A predictor factor of severity and mortality, Int. J. Lab. Hematol. 43 (1) (2021) e38–e40, http://dx.doi.org/10.1111/ijlh.13351.
[46] J. Liu, et al., Lymphopenia predicted illness severity and recovery in patients with COVID-19: A single-center, retrospective study, PLoS One 15 (11) (2020) e0241659.