Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 136 (2021) 104753

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Risk profiles for negative and positive COVID-19 hospitalized patients

Fahimeh Nezhadmoghadam *, Jose Tamez-Peña

School of Medicine, Tecnologico de Monterrey, Monterrey, Nuevo Leon, Mexico

A R T I C L E I N F O

Keywords:
Consensus clustering
COVID-19
Cluster analysis
Decision trees
Reproducibility of results
Risk factors
Unsupervised machine learning

A B S T R A C T

COVID-19 is a viral infection that affects people differently, where the majority of cases develop mild symptoms, some people require hospitalization, and unfortunately, a small number of patients perish. Hence, identifying risk factors is critical for physicians to make treatment decisions. The purpose of this article is to determine whether unsupervised analysis of risk factors in positive and negative COVID-19 subjects can aid in the identification of a set of reliable and clinically relevant risk profiles. Positive and negative patients hospitalized were randomly selected from the Mexican Open Registry between March and May 2020. Thirteen risk factors, three distinct outcomes, and COVID-19 test results were used to categorize registry patients. As a result, the dataset was reported via 6144 different risk profiles for each age group. The unsupervised learning method is proposed in this study to discover the most prevalent risk profiles. The data was partitioned into discovery (70%) and validation (30%) sets. The discovery set was analyzed using the partition around medoids (PAM) method, and the stable set of risk profiles was estimated using robust consensus clustering. The PAM models' reliability was validated by predicting the risk profile of subjects from the validation set and patients admitted in November 2020. In the validation set, the clinical relevance of the risk profiles was evaluated by determining the prevalence of three patient outcomes: pneumonia diagnosis, ICU admission, or death. Six positive and five negative COVID-19 risk profiles were identified, with significant statistical differences between them. As a result, PAM clustering with consensus mapping is a viable method for discovering unsupervised risk profiles in subjects with severe respiratory health problems.

1. Introduction

Due to the rapid spread of the SARS-CoV-2 virus worldwide, the Coronavirus Disease 2019 (COVID-19) pandemic outbreak has become a public health emergency of international concern. The high mortality risk associated with COVID-19, which ranges between 2% and 20% depending on the availability and quality of medical resources and economic conditions [1,2], is one of the pandemic's primary concerns. Another issue is that many recovered patients experience long-term sequelae that impact their lives and may have economic consequences [3,4]. As a result, effective treatments are needed to improve or cure COVID-19 cases and control the disease's effects.

Identifying and characterizing the various risk profiles of infected subjects is a critical task in managing COVID-19. The accurate characterization of a subject's risk profile is critical for the prompt selection of effective treatment for that particular patient. Additionally, it may facilitate effective medical resource allocation and provide critical information to identify and protect the most vulnerable populations [5]. Numerous studies have been conducted in this risk profiling field. COVID-19 studies identified the most critical disease severity risk factors, including advanced age, male gender, obesity, and smoking, as well as comorbidities such as hypertension, diabetes, hematologic, renal, cardiovascular, and respiratory diseases, all of which may have a significant impact on the prognosis of COVID-19 infected subjects [5–12]. Additionally, Gansevoort et al. discovered that subjects with chronic kidney disease have an extremely high risk of mortality from COVID-19 [13].

As previously stated, it is critical to identify risk factors, and more importantly, have tools that can predict disease severity in at-risk populations; thus, various supervised approaches have been proposed to identify risk factors for COVID-19 progression. The most frequently used methods for modeling risk factors for disease severity prediction in patients with COVID-19 have been univariate and multivariate ordinal logistic regression models [14]. In comparison, Ji et al. used multivariate Cox regression to investigate the risk factors for COVID-19 progressing to a critical or fatal state [15]. These efforts have been

* Corresponding author.
E-mail address: f.nejad.moghadam@gmail.com (F. Nezhadmoghadam).

https://doi.org/10.1016/j.compbiomed.2021.104753
Received 16 April 2021; Received in revised form 5 August 2021; Accepted 5 August 2021
Available online 9 August 2021

conducted in various settings or with limited clinical data [16–18].

Additionally, supervised approaches are limited because many possible risk factors can be associated with the severity of the outcome. Each risk factor and its combination generate various possible COVID-19 risk profiles; thus, an extensive data set is required to train complex statistical models accurately. A novel approach is proposed to address the risk factor combination issue, based on unsupervised data clustering for identifying robust patterns in subjects' risk presentations that are easily associated with disease severity and outcome [19]. This study aims that by using clustering to identify patients' risk profiles, data analysis for treatment decisions can be streamlined.

There are numerous algorithms for data clustering [20–23]. Certain algorithms can be thought of as statistical clustering strategies [24–26]. They are robust approaches that result in models that adequately describe data, with each model containing explicit factors that aid in data comprehension [27,28]. Additionally, novel algorithmic advances facilitate the discovery of robust data clusters in multidimensional data sets. Consensus clustering is one such technique [29]. Consensus clustering utilizes multiple iterations of the clustering method of choice to discover the most reliable partitions from multidimensional data sets. Additionally, Partitioning Around Medoids (PAM) is a robust statistical clustering algorithm that aims to find K-medoids that minimize the sum of the observations' dissimilarities to their nearest medoid [30]. The proposed method utilizes consensus clustering and the PAM clustering algorithm to determine the risk profiles of patients.

The purpose of this study is to determine whether the unsupervised discovery of risk profiles for COVID-19 and non-COVID-19 patients seeking medical attention can aid in identifying a subset of hospitalized patients at increased risk of either: 1) developing pneumonia; 2) requiring admission to an intensive care unit (ICU), or 3) perishing as a result of the infection. To accomplish this, we used the Open Mexican Repository, which collects COVID-19 test results, outcomes (pneumonia diagnosis, hospitalization, and death), and known risk factors such as age, gender, pregnancy, smoking, obesity, and common comorbidities such as hypertension and diabetes.

2. Material and methods

2.1. Data preparation

The preliminary data for this study were obtained on May 9, 2020, from the COVID-19 Mexican Open Repository, maintained by the Mexican government's General Directorate of Epidemiology [31]. On June 8, the dataset was updated to ensure the best possible patient outcome. The dataset contained 128,148 subjects and included the following variables: patient ID, age, sex, exposure history, obesity, smoking, pregnancy, patient type (ambulatory/hospitalized), and other underlying comorbidities (diabetes, hypertension, cardiovascular disease, chronic obstructive pulmonary disease (COPD), asthma, immunosuppression, chronic kidney failure, and other diseases) (pneumonia, ICU, intubation, and date of death).

The study was limited to only hospitalized subjects due to the Mexican COVID-19 sentinel testing strategy [32]. Among hospitalized patients, 13,367 subjects tested positive for COVID-19, while 19,958 subjects tested negative. Each patient was described using 35 characteristics, but for this study, the focus was on the set of 13 risk factors associated with illness severity, such as age, sex, obesity, smoking, and underlying comorbidities. Other reported data pertain to patients' personal information and exposure history, such as the date on which the patient's symptoms began, the date on which the patient was admitted to the care unit, the patient's nationality, and whether the patient speaks an indigenous language. Since these characteristics had no discernible effect on the severity of illness, they were excluded from this study. As a result, the selected data set generated 6144 different risk profiles for each age group.

Moreover, hospitalized subjects in November 2020 were used as a new cohort of patients to validate the discovered risk profiles' accuracy. This test set included 31,987 patients with positive COVID-19 test results, while 18,170 had negative test results. This study initially took place in June 2020. The trained model was then validated using the January 2021 testing set. Table 1 contains descriptive statistics on selected characteristics and outcomes for hospitalized patients with positive and negative COVID-19 test results. Additionally, Table 2 illustrates the characterization of selected test set features and outcomes.

2.2. Initial statistical analysis

The selected characteristics of the positive and negative groups were compared to determine if there were any differences between the positive and negative COVID-19 subjects (Tables 1 and 2). Additionally, each group was defined by the number of recovered and deceased patients. Cohen's d (Z) and odds ratio (OR) was used to determine the effect size of all features for continuous and discrete variables, respectively [33]. Tables S1–S4 summarizes the characteristics of positive and negative COVID-19 subjects from the discovery set (March to May hospitalized patients) and the validation set (November hospitalized patients). Finally, the prevalence of the top ten major risk factors identified in the discovery set was calculated in males and females with positive/negative test results stratified by age group: young (20–40), middle (40–60), and elderly (>60). In other words, this report details the frequency with which the top 120 risk profiles occur. Supplementary material and Fig. S1 are available.

2.3. Consensus clustering and the PAM clustering model

Fig. 1 summarizes the overall methodology used to discover clusters and model risk profiles. Initially, hospitalized patients were considered from the beginning of the COVID-19 pandemic through May, and the data set was divided randomly into discovery/training and validation

Table 1
The characteristics of subjects from the COVID-19 Mexico hospitalization data set (from March to May, at the start of the COVID-19 pandemic). The values indicate the number of subjects (percentage) and the mean (SE) for Age. The OR was calculated with a 95% confidence interval for positive vs negative COVID-19. *, **, and *** denote a small effect size (between 0.2 and 0.5 for Z and 1.5 to 2 for OR), a medium effect size (between 0.5 and 0.8 for Z and 2 to 3 for OR), and a large effect size (greater than 0.8 for Z and more than 3 for OR), respectively.

Feature	Positive COVID	Negative COVID	Effect Size
Subjects (male ratio)	13367 (65.75%)	19958 (55.92%)	OR = 1.18 (1.13–1.22)
Age	53.75 (0.13)	44.43 (0.16)	Z = 0.3*
Pregnancy	58 (0.43%)	287 (1.44%)	OR = 0.3 (0.23–0.4)
Diabetes	4099 (30.66%)	5288 (26.50%)	OR = 1.16 (1.11–1.21)
COPD	549 (4.11%)	1387 (6.95%)	OR = 0.59 (0.53–0.65)
Asthma	335 (2.51%)	769 (3.85%)	OR = 0.65 (0.57–0.74)
Immunosuppression	370 (2.77%)	1273 (6.38%)	OR = 0.43 (0.39–0.49)
Hypertension	4313 (32.27%)	6126 (30.69%)	OR = 1.05 (1.01–1.1)
Cardiovascular	588 (4.40%)	1433 (7.18%)	OR = 0.61 (0.56–0.68)
Obesity	3307 (24.74%)	3497 (17.52%)	OR = 1.41 (1.34–1.49)
Chronic kidney	607 (4.54%)	1488 (7.45%)	OR = 0.61 (0.55–0.67)
Smoking	1251 (9.36%)	2151 (10.78%)	OR = 0.87 (0.81–0.93)
Other diseases	584 (4.37%)	1482 (8.63%)	OR = 0.59 (0.53–0.65)
Outcome			
ICU	1596 (11.94%)	1457 (7.30%)	OR = 1.64 (1.52–1.76) *
Deaths	5610 (41.97%)	2510 (12.58%)	**OR = 3.34 (3.17–3.51)****
Pneumonia	9490 (71.00%)	11342 (56.83%)	OR = 1.25 (1.21–1.29)

Table 2
The characteristics of subjects from the COVID-19 Mexico November hospitalization test data set. The values indicate the number of subjects (percentage) and the mean (SE) for Age. The OR was calculated with a 95% confidence interval for positive vs negative COVID-19. *, **, and *** denote a small effect size (between 0.2 and 0.5 for Z and 1.5 to 2 for OR), a medium effect size (between 0.5 and 0.8 for Z and 2 to 3 for OR), and a large effect size (greater than 0.8 for Z and greater than 3 for OR), respectively.

Feature	Positive COVID	Negative COVID	Effect Size
Subjects (male ratio)	31987 (58.64%)	18170 (52.23%)	OR = 1.3 (1.25–1.35)
Age	58.93 (0.09)	48.92 (0.18)	Z = 0.48*
Pregnancy	163 (0.51%)	487 (2.68%)	OR = 0.19 (0.16–0.22)
Diabetes	10919 (34.13%)	5017 (27.61%)	OR =1.36 (1.31–1.41)
COPD	1159 (3.62%)	770 (4.24%)	OR = 0.85 (0.77–0.93)
Asthma	639 (2.01%)	388 (2.13%)	OR = 0.93 (0.82–1.06)
Immunosuppression	685 (2.14%)	695 (3.82%)	OR = 0.55 (0.49–0.61)
Hypertension	13207 (41.29%)	6039 (33.24%)	OR = 1.41 (1.36–1.47)
Cardiovascular	1450 (4.53%)	1204 (6.63%)	OR = 0.67 (0.62–0.72)
Obesity	6740 (21.07%)	2560 (14.09%)	OR = 1.63 (1.55–1.71) *
Chronic kidney	1605 (5.96%)	1907 (8.83%)	OR = 0.45 (0.42–0.48)
Smoking	2517 (7.87%)	1460 (8.03%)	OR = 0.98 (0.91–1.05)
Other diseases	1631 (5.10%)	1301 (7.16%)	OR = 0.7 (0.65–0.75)
Outcome			
ICU	2390 (7.47%)	1077 (5.93%)	OR = 1.28 (1.19–1.30)
Deaths	13653 (42.68%)	3502 (19.27%)	**OR = 3.12 (2.99–3.26)****
Pneumonia	20299 (63.46%)	8231 (45.30%)	**OR = 2.1 (2.02–2.18)** **

sets. This strategy eliminated biases associated with discovery/training in the risk assessment of each patient's risk profile. 70% of subjects were randomly selected to participate in the cluster discovery and training of the final risk profile prediction model. The corresponding risk profiles were predicted for the remaining 30% of patients after estimating all the data transformation parameters, the optimal number of clusters, and the final cluster parameters via the training data set. Additionally, the corresponding risk profiles of the November hospitalized patients were predicted to evaluate the discovered risk profiles and the training models' accuracy. Finally, using Classification and Regression Tree (CART) analysis, the role of each risk characteristic in each of the risk profiles was described [34].

Each feature was assigned a value between 0 and 1. Between the minimum and maximum ages, the age factor was normalized [35]. Males were assigned a code of one, while females were assigned zero. The remaining risk categorical features were set to 1 for risk factor presence and 0 for risk factor absence. Dimensionality was reduced using the principal components analysis (PCA) transform by selecting PCA feature vectors that captured more than 80% of total variance [36,37]. The risk profile was discovered in the following manner. First, the Partitioning Around Medoids (PAM) algorithm was selected as a method for clustering [30]. PAM is insensitive to data distribution differences, and the user provided the initial K-medoids. Consensus clustering was used to determine the optimal number of *K*-medoids.

Consensus clustering is based on multiple random replications of the determined clustering method, enabling a robust assessment of the clustering approach's sensitivity to input variation [38–40]. Additionally, the repeated random repetition makes the K-medoids selection more robust to random changes in the input parameters. The randomness of the approach was increased by randomly selecting 70% of subjects for medoid discovery, and the holdout discovery samples were used to assess the predictability of clustering labels predicted on the holdout set. The procedure was repeated 100 times to obtain a reliable evaluation of training-holdout-sample clustering with different clustering numbers ($K = 2,3,4,5,6,7$). The computation of the cluster co-association matrix (CCAM) is used to assess the reliability/stability of consensus clustering [41]. The CCAM is a matrix in which each column and row represents a subject from the discovery set, and it stores the counts of the number of times two holdout subjects shared a cluster label. As a result, stable data partitions produce sharp checkerboard patterns, whereas unstable data partitions produce fuzzy patterns. The

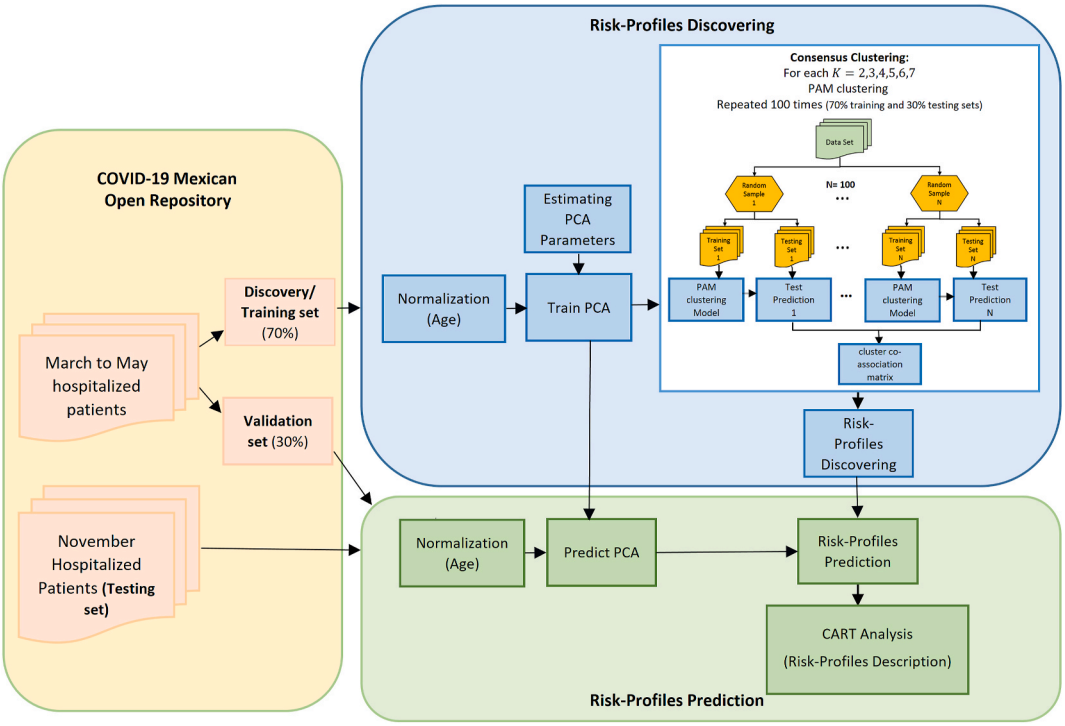

Fig. 1. The overall methodology of risk-profile's classification of Mexico COVID-19 data set. The multimodal data is split into training and testing sets and the results of the testing set are used to describe the association of disease risk-profiles to clinically relevant outcomes.

CCAM's clarity is determined by the proportion of ambiguous clustering (PAC). As a result, low PAC values indicate a very robust clarification scheme that is insensitive to changes in the discovery set. As a result of repeating the consensus clustering for various K values, the optimal data partition has the lowest PAC number and thus represents the most robust and reliable data clustering.

2.4. Statistical and CART analysis of the discovered risk-profiles

After computing the PCA transform and determining the optimal number of clusters and their associated medoids for each discovered risk profile, the risk profiles for each validation and test set samples were predicted. The validation set consisted of 30% of March to May hospitalized patients, while the test set consisted of 30% of November hospitalized patients. Three steps are involved in predicting risk profiles: Initially, normalizing the patients' age; second, forecasting the magnitude of each subject's principal component and third, labeling the validation and test samples' risk profiles. This risk profile prediction algorithm generates a unique class label for each subject in the validation and test sets.

After predicting risk profiles, the prevalence of adverse outcomes associated with each discovered risk profile was analyzed. Three adverse events were examined: pneumonia diagnosis, intensive care unit (ICU) admission, and patient death, and the data set only contained these negative outcomes. As a result, the risk profile associated with the highest prevalence of adverse outcomes is most critical. Moreover, the discovered risk profiles of the validation set and test set were compared and then analyzed for the difference of the selected features and outcomes together in both testing sets. Finally, the objective was to develop simple decision rules for categorizing each new patient according to the discovered risk profiles. The classification and regression trees (CART) analysis were selected for this purpose. CART generates decision tree algorithms automatically for problems involving classification or predictive regression modeling [42]. For continuous and discrete values, either the ANOVA or chi-square test was used to infer the statistical significance of each discovered risk profile. Additionally, the statistical difference between the validation and test sets for selected features of each predicted risk profile was computed using the proportions test and the *t*-test test for discrete and continuous values, respectively. Significant values were defined as those less than 0.05, and no attempt was made to correct for false discovery.

Implementation and data used are available on GitHub (https://github.com/FahimehN/COVID-19-Risk-Profiles-Discovering).

3. Results

As the discovery set, a cohort of 33,325 patients with positive and negative COVID-19 tests were analyzed who were hospitalized from March to May. Additionally, hospitalized subjects with positive and negative COVID-19 tests in November (N = 50157) were investigated to add a new patient group to the test set. The characteristics of positive and negative COVID-19 hospitalized patients in the discovery and test sets are summarized in Tables 1 and 2 Differences in their statistical significance were expressed as effect sizes with 95% confidence intervals (95% CI). The mortality rate was significantly different between positive and negative COVID-19 in both sets: OR = 3.34 (95% CI = 3.17 to 3.51) and OR = 3.12 (95% CI = 2.99 to 3.26) respectively. In other words, subjects infected with COVID-19 had a higher mortality rate than other patients with respiratory problems.

Tables S1–S4 detail the characteristics of infected and non-infected subjects based on deaths and recoveries from the discovery and test sets, respectively. The findings indicated a moderate difference in age, COPD, chronic kidney disease, and ICU hospitalization between those who perished and those who recovered with positive COVID-19 test results in the discovery set. In comparison, age has a negligible effect size (Z = 0.64) on the difference between the deaths and the recovered groups in the test set (Table S3). Between March and May, deceased patients were 2.25 and 2.35 times (95% confidence intervals, 1.89 to 2.68 and 1.98 to 2.78) more likely to have COPD or chronic kidney disease, respectively, than recovered patients (Table S1). Similarly, moderate effect sizes were observed for age (Z = 0.68) and ICU admission during the same period (OR of 2.41, 95% CI, 2.16 to 2.68). Table S2 and Table S4 show the recovered-death analysis of negative COVID-19 patients from March to May and November. In the discovery set, chronic kidney disease and advanced age were the most significant risk factors for death (Table S2). There were marginal differences in diabetes, COPD, immunosuppression, hypertension, and cardiovascular disease between the two groups, with ORs ranging between 1.5 and 2. However, subjects over the age of 65, those with diabetes, and those with hypertension had the highest risk of death among non-infected hospitalized patients in November (Table S4).

Additionally, the prevalence of the top 120 risk profiles was determined. The frequency distributions of the top 10 risk profiles by age/gender and the COVID-19 test results are shown in Fig. S1. The combined analysis results revealed that most men and women aged 60 and over suffer from hypertension, diabetes, or both, whereas obesity is highly prevalent in the younger age group of 20–40. As a result, the majority of hospitalized patients (both males and females with positive and negative COVID) in the middle age range of 40–60 years have hypertension, diabetes, or obesity, indicating that these three comorbidities are significant risk factors for seeking medical care following a respiratory illness.

Afterward, consensus clustering and the PAM clustering model were used to determine the risk profiles of positive and negative hospitalized COVID-19 subjects in the validation set (30% of March to May hospitalized patients) and test set (November hospitalized patients). Figs. 2 and 3 depict the optimal CCAM partitioning and PAC analysis for the hypothesis of 2–7 different risk profiles for positive and negative COVID-19 subjects in the discovery set, respectively. The optimal partition for positive COVID-19 patients consisted of 6 clusters, whereas the optimal partition for negative COVID-19 patients consisted of 5 clusters. As a result, the PAM clustering model was trained using the discovery set's optimal number of clusters for both positive (K = 6) and negative (K = 5) groups. Then, using the trained models, the risk profiles of the validation and test samples were determined. Tables S5–S8 present descriptive statistics about the investigated features stratified by risk profiles for subjects with positive and negative COVID-19 test results in the validation and test sets. In Tables S5 and S7, three risk profiles with a high death risk were labeled accordingly (risk profiles 4, 5, and 6). The analysis of risk profiles revealed that the distribution of features was significantly different for each risk profile. Risk profile #6 posed the most significant risk of death. It was primarily composed of elderly males who were hypertensive or diabetic. Risk profile #4 included subjects who were hypertensive but did not have diabetes. While risk profile #5 included individuals with diabetes. The analysis of the 5 risk profiles for the negative COVID-19 group in the validation and test sets is shown in Table S6 and Table S8. Subjects who tested negative for COVID-19 had a better chance of surviving their respiratory condition. The risk group with the highest mortality rate was risk profile #5, with 23.73% and 31.14% of patients perishing in the March to May and November sets, respectively. It was composed entirely of men (100%) with diabetes (100%), and most of them suffered from hypertension.

Moreover, Figs. 4 and 5 illustrate the percentage of selected features in each risk profile for subjects with positive and negative COVID-19 test results in the validation set (March to May) and the test set (November), respectively. Hypertension and diabetes are more prevalent in high-risk groups, as illustrated in Fig. 4. Additionally, the rate of COVID infection among women increased in November compared to the early months of the pandemic.

Between March and May, the percentage of non-infected patients with hypertension and diabetes increased significantly compared to hospitalized patients (Fig. 5).

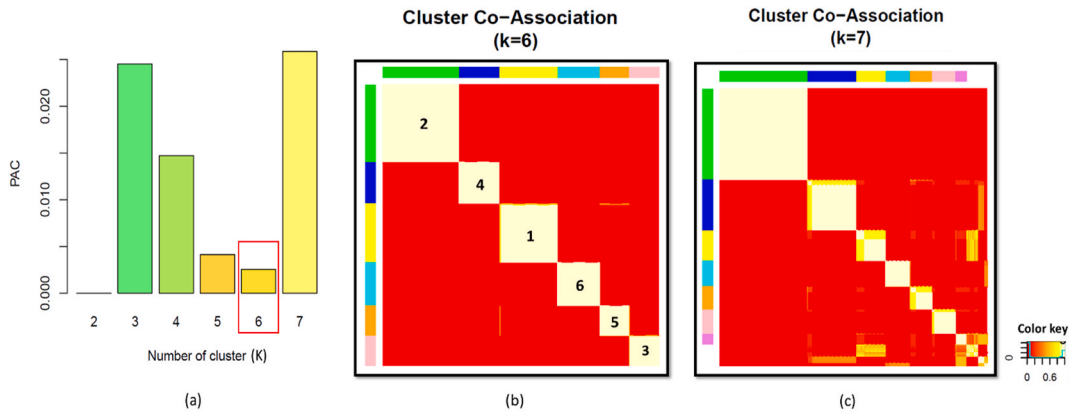

Fig. 2. Result of consensus clustering applied to the discovery set of subjects with positive COVID-19 test results. (a) The comparison of PAC (the lower the number, the better) between the cluster numbers from 2 to 7, (b) the best result of Consensus mapping for K = 6, and (c) the worst result of Consensus mapping for K = 7 with the highest PAC value.

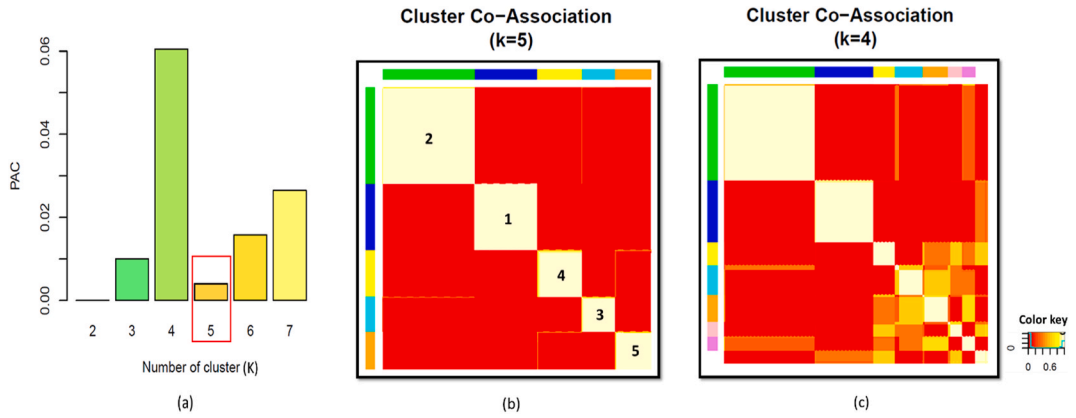

Fig. 3. Results of consensus clustering applied to the discovery set of subjects with negative COVID-19 test results. (a) The comparison of PAC (the lower the number, the better) between the cluster numbers from 2 to 7, (b) the best result of Consensus mapping for K = 5, and (c) the worst result of Consensus mapping for K = 4 with the highest PAC value.

Fig. 6 shows the violin plots of age in the validation and test sets for each risk profile. Fig. 6(a) and (b) illustrate the age distributions of positive and negative subjects, respectively. Age analysis of patients with positive test results in November revealed that they are older than patients with negative test results between March and May, whereas subjects with negative test results are younger in a higher number of risk profiles in the November test set.

In both sets, the three adverse outcomes associated with identified risk profiles were analyzed. The validation set (30% of March to May) and the test set (November) outcomes are compared in each risk profile of positive and negative hospitalized patients and shown in Fig. 7. The percentage of positive subjects who perish or are hospitalized in the intensive care unit was higher in the high-risk groups (Risk profiles #4, #5, and #6) than in the low-risk groups (Fig. 7(a)). However, patients who tested positive for COVID-19 in November had a lower death rate than those who tested positive from March to May. Additionally, they were less likely to develop pneumonia and require ICU admission.

Additionally, while negative COVID-19 subjects in November had a lower ICU and pneumonia hospitalization rate than subjects from March to May, the November patients' mortality rate increased significantly (Fig. 7(b)).

The results of the CART analysis on the validation set and the test set are depicted in Fig. S2 and Fig. S3, respectively. The figures illustrate the relationship between risk factors and newly discovered risk profiles using decision trees constructed from the validation and test sets for positive and negative COVID-19. According to Figs. S2(a) and S3(a), 40% of positive subjects in November and 48% of positive subjects between March and May were in high-risk groups (the total of the percentage of observation of risk profiles 4, 5, and 6 with a higher probability of mortality). Patients with hypertension and diabetes were classified as high-risk (Risk profile 6), whereas women without hypertension were classified as low-risk (Risk profile 1). The validation set's negative risk profile decision trees analysis revealed that risk profiles 4 and 5 have distinct decision rules, and the predicted probability of risk profiles included mixture values. Surprisingly, CART analysis eliminates age as a significant factor for COVID-19 positive patients (Fig. S2 (b)). However, between March and May, there were differences in the decision rules for subjects with different risk profiles (Fig. S3 (b)). Finally, the primary risk factors associated with positive COVID-19 patients by region were identified. Fig. S4 and Tables S9 and S10 illustrate the geographical distributions of risk factors and how they changed from March to May to November.

4. Discussion

This study discovered, described, and classified the risk profiles of hospitalized COVID-19 positive and negative subjects. Initially, a detailed combinatory analysis of 6144 different risk profiles of hospitalized Mexican patients stratified by age was conducted. The detailed analysis identified the risk factors associated with the top ten profiles by gender, COVID-19 test result, and age. According to the analysis of positive patients, hypertension, diabetes, and obesity were prevalent

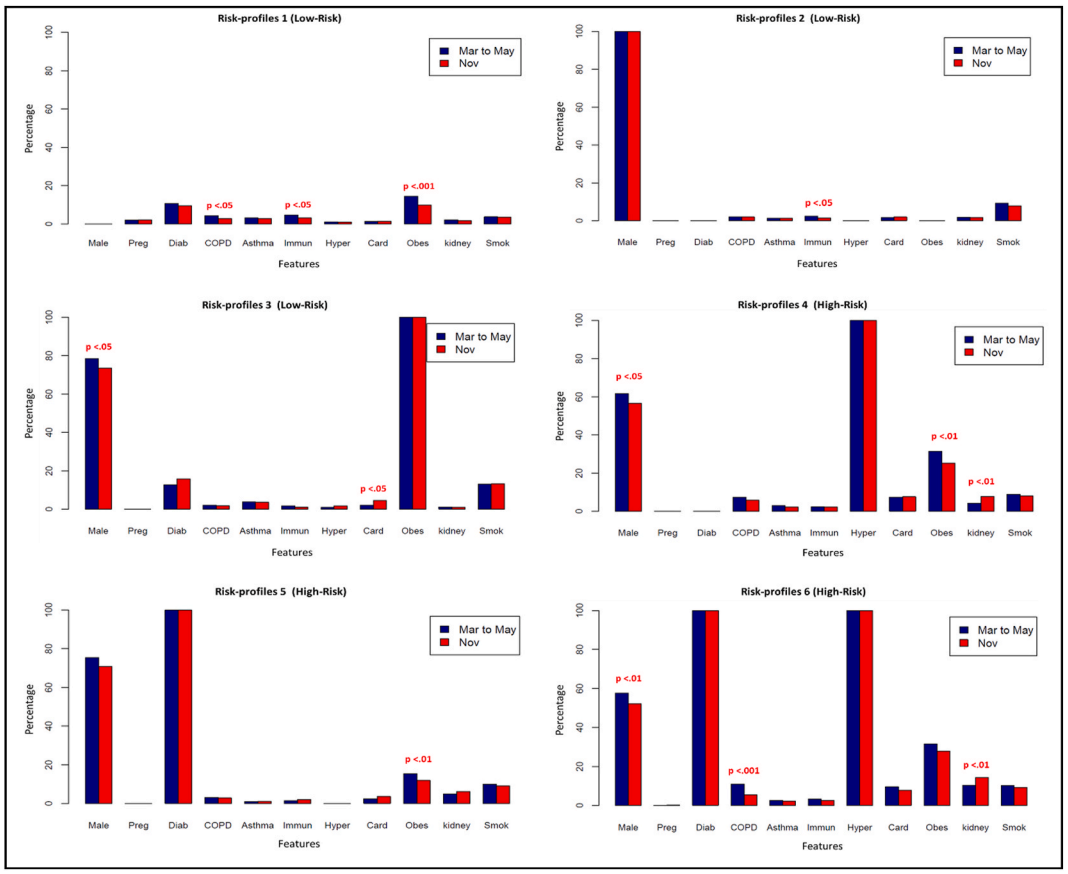

Fig. 4. The comparison between the percentage of the selected features in each risk-profile for hospitalized patients with positive COVID-19 test results in the validation set (30% of March to May) and the test set (November). The p-value was measured by the proportion test. The values that were less than 0.05 were mentioned at the top of the features. The abbreviations include; Preg: Pregnancy, Diab: Diabetes, COPD: Chronic obstructive pulmonary disease, Immun: Immunosuppression, Hyper: Hypertension, Card: Cardiovascular, Obes: Obesity, Kidney: Chronic kidney, Smok: Smoking.

among 40-year-old women, while smoking was also more prevalent among men in the same age group. Smoking, obesity, diabetes, and hypertension in younger men were prevalent in the 20–40 age group, whereas smoking was less prevalent in women in this age group. These latter findings confirm previous findings that seeking medical care is strongly associated with health comorbidities and smoking [43,44].

Unsupervised learning via consensus clustering was used to discover the major risk profiles, and six and five risk profiles for infected and non-infected COVID-19 patients, respectively, were discovered. These profiles were discovered and reproduced consistently using a small training set and were validated using holdout cross-validation and an independent set. The risk profiles were identified using a representative training set of positive and negative COVID-19 individuals. The classes were consistently predicted on an independent validation set by modeling these risk profiles with PAM clustering. Following discovery, supervised decision trees were used to rank each risk profile's discovered risks.

The association of the discovered positive COVID-19 groups to a severe-outcome analysis identified three high-risk profiles. The majority of vulnerable subjects were 60 years of age or older and had pre-existing medical conditions such as hypertension, diabetes, or obesity. Additionally, men were predisposed to severe conditions. The decision rule analysis revealed that the most significant risk factor is hypertension combined with diabetes (risk profile #6). Other risk groups included men predominantly with hypertension or diabetes (risk profiles #4 and #5). However, women without hypertension who were infected with COVID-19 were in the lowest risk group (risk profile #1). It is essential to highlight that CART analysis revealed that age did not significantly impact stratifying COVID-19 patients into the six risk profile groups. A significant implication was that hypertension, obesity, diabetes, and gender were the primary factors that characterized the top six risk profiles regardless of age. The findings corroborate previous reports that patients with hypertension and diabetes suffer from more severe illnesses and have a higher mortality rate than those without hypertension or diabetes [45–47]. Overall, the identified risk factors for people in high-risk groups corroborated previous research, demonstrating that age, obesity, diabetes, and hypertension are all significantly associated with severe COVID-19 [48–50]. On the other hand, unsupervised clustering models can be used to classify newly diagnosed patients associated with COVID-19 risk factors into known subgroups to aid in the treatment process.

The analysis of negative COVID-19 subjects' decision rules for both the validation and test sets revealed that certain nodes contain a mixture of risk profiles with no significantly predicted probabilities. By contrast, there were significant differences in the characteristics of individuals with various negative risk profiles. The results indicated that individuals with negative non-confirmed COVID-19 subjects were slightly different between November and validation. The most stringent validation set included women over 63 who had diabetes or were hypertensive without diabetes. According to the November set, diabetic men faced the worst outcomes. Although the geographical analysis indicated a shift in risk factors from March–May to November, the prevalence of adverse outcomes in COVID-19 positive patients did not change significantly during the first six months.

In contrast, negative subjects with high-risk profiles had a higher prevalence of adverse outcomes. Additionally, it was observed that hospitalized patients who did not contract COVID-19 and had negative test results were more susceptible to other conditions. The disease presenting in these patients with respiratory symptoms could be a bacterial

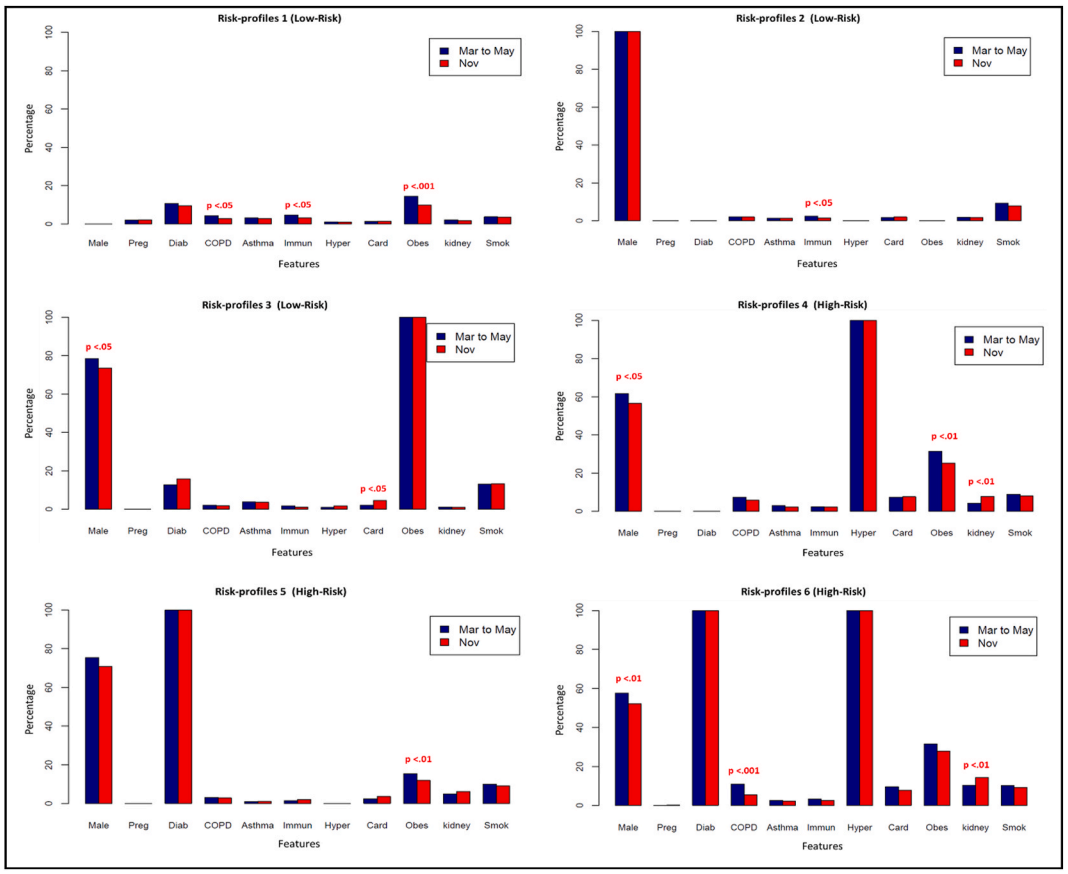

Fig. 5. The comparison between the percentage of the selected features in each risk-profile for hospitalized patients with negative COVID-19 test results in the validation set (30% of March to May) and the test set (November). The p-value was measured via the proportion test. The values that were less than 0.05 were mentioned at the top of the features. The abbreviations include; Preg: Pregnancy, Diab: Diabetes, COPD: Chronic obstructive pulmonary disease, Immun: Immunosuppression, Hyper: Hypertension, Card: Cardiovascular, Obes: Obesity, Kidney: Chronic kidney, Smok: Smoking.

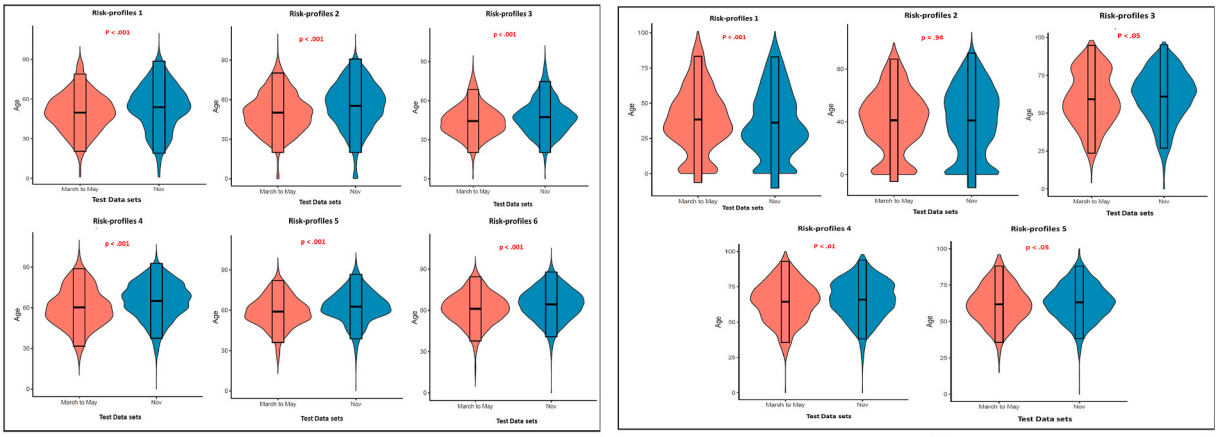

Fig. 6. The comparison of age of the validation set (30% of March to May) and the test set (November) in each risk-profile for hospitalized patients with a) Positive COVID-19 test results. b) Negative COVID-19 test results. The p-value was measured by the *t*-test test.

infection, influenza, or another respiratory infection that presents similarly to COVID-19 [51]. Additionally, while several respiratory symptoms may be associated with smoking, the percentage of smokers in each negative risk profile was insignificant.

On the other hand, the most frequently encountered complication in hospitalized COVID-19 patients was severe pneumonia caused by viral infections, bacterial infections, or other conditions [52]. However, in some individuals, coronavirus infection can progress to pneumonia. Additionally, pneumonia and respiratory disorders are caused by a variety of diverse sources. Thus, the increased risk of pneumonia in COVID-19-negative individuals may be related to other health problems. Additionally, comparing the outcomes of positive and negative COVID-19 hospitalized cases revealed significant differences in mortality and ICU admission rates between the two data sets, indicating that infected COVID-19 patients are more likely to become critically ill, and some will perish.

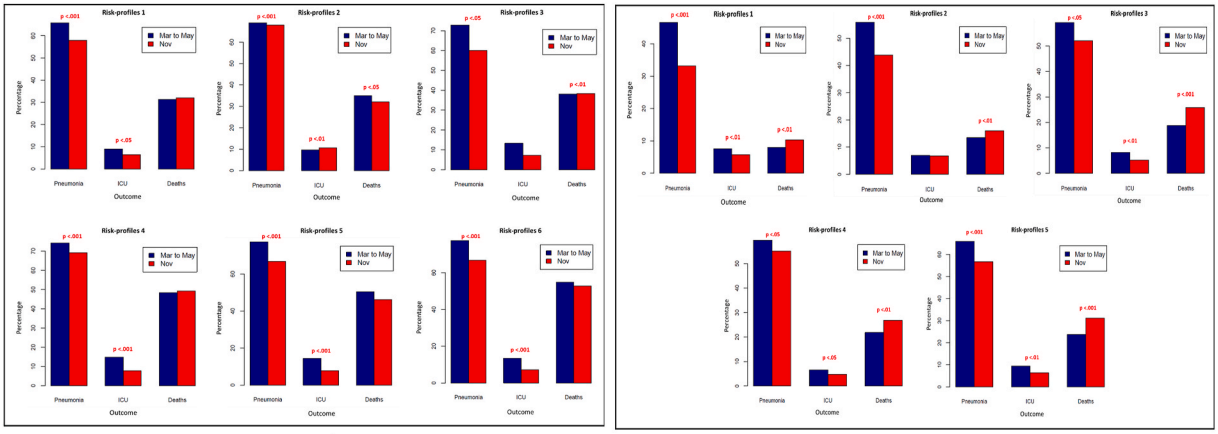

Fig. 7. The comparison of the outcomes of the validation set (30% of March to May) and the test set (November) in each risk-profile for hospitalized patients with a) Positive COVID-19 test results. b) Negative COVID-19 test results. The p-value was measured by the proportion test. The values that were less than 0.05 were mentioned at the top of the outcomes.

Additionally, when the validation set and the test set were compared in each risk profile, it was found that the positive COVID-19 subjects in November were less likely to perish, require ICU care, or even progress to pneumonia than the March–May subjects. It could be related to the development of effective coronavirus treatments. However, the mortality rate for negative subjects increased significantly in November compared to the validation set's non-infected patients. Fear of Covid-19 may be causing other patients to forego necessary treatment. In other words, patients who require additional medical care urgently delay or forego critical procedures that could save their lives. Additionally, hospitals may ask patients with underlying diseases to discontinue treatment in order to minimize risk. This could increase the mortality rate of patients who were not infected with Covid-19.

The clustering method had the advantage of predicting clusters of patients associated with various combinations of risk factors for both positive and negative COVID data sets. Simultaneously, supervised decision trees failed to discover consistent decision rules from the discovered risk profiles.

Multiple limitations applied to this study. The findings of this study were based on a Mexican cohort that is skewed toward those seeking medical care and those who have been hospitalized. As a result, they cannot be applied to the entire population. As a result, the findings must be validated in a separate cohort. A second limitation was that because the cluster-based analysis was used to identify the significant risk profiles, many different conditions were missed, and thus the simple decision-making rules presented in this study cannot be used to make clear treatment decisions. A third limitation was that outcomes changed throughout the pandemic. Numerous treatments were evaluated, and hospital saturation varied significantly between patients. As a result, the risk association findings presented in this paper are most likely to be valid only for the population studied.

5. Conclusion

This study demonstrated the use of consensus clustering in conjunction with PAM models to identify the most consistent risk profiles among COVID infected and non-infected patients. Additionally, CART analysis was used to describe the relationship between newly discovered risk factors and each risk profile. The findings demonstrated that the proposed method could identify a small set of the most prevalent risk profiles for both data sets, and it may be a valuable tool for filtering out the most prevalent risk profiles in larger multidimensional datasets. The findings indicated that regardless of age group, gender, hypertension, diabetes, and obesity may be the primary high-risk factors for COVID-19 mortality.

Acknowledgments

This research was supported with funding from the Mexican National Council for Science and Technology (CONACYT). The authors are thankful to Dr. Víctor Treviño, Dr. Emmanuel Martínez and Dr. Santiago Conant-Pablos for all valuable comments and suggestions, which helped us to improve the quality of the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104753.

References

[1] Kaiyuan Sun, Jenny Chen, Cecile Viboud, Early epidemiological analysis of the 2019-nCoV outbreak based on a crowdsourced data, medRxiv, 2020.
[2] Yang Yang, et al., Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China, MedRxiv, 2020.
[3] Raul D. Mitrani, Nitika Dabas, Jeffrey J. Goldberger, COVID-19 cardiac injury: implications for long-term surveillance and outcomes in survivors, Heart Rhythm 17 (11) (2020) 1984–1990.
[4] Sana Salehi, Sravanthi Reddy, Ali Gholamrezanezhad, Long-term pulmonary consequences of coronavirus disease 2019 (COVID-19): what we know and what to expect, J. Thorac. Imag. 35 (4) (2020) W87–W89.
[5] Yuetian Yu, et al., Identification of risk factors for mortality associated with COVID-19, PeerJ 8 (2020) e9885.
[6] Wei-jie Guan, et al., Comorbidity and its impact on 1590 patients with covid-19 in China: a nationwide analysis, Eur. Respir. J. 55 (2020) 5.
[7] Giacomo Grasselli, et al., Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy, JAMA Inter. Med. (2020).
[8] Annemarie B. Docherty, et al., Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study, bmj (2020) 369.
[9] Lindsay Kim, et al., Risk factors for intensive care unit admission and in-hospital mortality among hospitalized adults identified through the US coronavirus disease 2019 (COVID-19)-associated hospitalization surveillance network (COVID-NET), Clin. Infect. Dis. (2020).
[10] Tao Liu, et al., Risk factors associated with COVID-19 infection: a retrospective cohort study based on contacts tracing, Emerg. Microb. Infect. 9 (1) (2020) 1546–1553.
[11] Zhaohai Zheng, et al., Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis, J. Infect. (2020).
[12] Elizabeth R. Lusczek, et al., Characterizing COVID-19 clinical phenotypes and associated comorbidities and complication profiles, PloS One 16 (3) (2021), e0248956.
[13] Ron T. Gansevoort, Luuk B. Hilbrands, CKD is a key risk factor for COVID-19 mortality, Nat. Rev. Nephrol. (2020) 1–2.

[14] Shan-Yan Zhang, et al., Clinical characteristics of different risk-profiles and risk factors for the severity of illness in patients with COVID-19 in Zhejiang, China, Infect. Dis. Poverty 9 (1) (2020) 1–10.
[15] Dong Ji, et al., Prediction for progression risk in patients with COVID-19 pneumonia: the CALL Score, Clin. Infect. Dis. (2020).
[16] Char Leung, Risk factors for predicting mortality in elderly patients with COVID-19: a review of clinical data in China, Mech. Ageing Dev. (2020) 111255.
[17] Qiao Shi, et al., Clinical characteristics and risk factors for mortality of COVID-19 patients with diabetes in Wuhan, China: a two-center, retrospective study, Diabetes Care (2020).
[18] Ling Hu, et al., Risk factors associated with clinical outcomes in 323 coronavirus disease 2019 (COVID-19) hospitalized patients in Wuhan, China, Clin. Infect. Dis. 71 (16) (2020) 2089–2098.
[19] Fahimeh Nezhadmoghadam, et al., Robust Discovery of Mild Cognitive impairment subtypes and their Risk of Alzheimer's Disease conversion using unsupervised machine learning and Gaussian Mixture Modeling, medRxiv, 2020.
[20] Mayra Z. Rodriguez, et al., Clustering algorithms: a comparative approach, PloS One 14 (1) (2019), e0210236.
[21] Divya Pandove, Shivan Goel, Rinkl Rani, Systematic review of clustering high-dimensional and large datasets, ACM Trans. Knowl. Discov. Data 12 (2) (2018) 1–68.
[22] M. Emre Celebi, Hassan A. Kingravi, Linear, Deterministic, and Order-Invariant Initialization Methods for the K-Means Clustering algorithm." Partitional Clustering Algorithms, Springer, Cham, 2015, pp. 79–98.
[23] Julien Jacques, Cristian Preda, Functional data clustering: a survey, Adv. Data Anal. Classif. 8 (3) (2014) 231–255.
[24] Toshiro Tango, "Disease Mapping: Visualization of Spatial Clustering." Statistical Methods for Disease Clustering, Springer, New York, NY, 2010, pp. 33–47.
[25] Rainer Dangl, Friedrich Leisch, Effects of resampling in determining the number of clusters in a data set, J. Classif. (2019) 1–26.
[26] Soham Sarkar, Anil K. Ghosh, On perfect clustering of high dimension, low sample size data, IEEE Trans. Pattern Anal. Mach. Intell. 42 (9) (2019) 2257–2272.
[27] A. Colin Cameron, Douglas L. Miller, A practitioner's guide to cluster-robust inference, J. Hum. Resour. 50 (2) (2015) 317–372.
[28] García-Escudero, Luis Angel, et al., A review of robust clustering methods, Adv. Data Anal. Classif. 4 (2–3) (2010) 89–109.
[29] Matthew D. Wilkerson, D. Neil Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (12) (2010) 1572–1573.
[30] Aruna Bhat, K-medoids clustering using partitioning around medoids for performing face recognition, Int. J. Soft Comput. Math. Contr. 3 (3) (2014) 1–12.
[31] The general directorate of Epidemiology of the Mexico government, Retrieved from, https://www.gob.mx/salud/documentos/datos-abiertos-152127.
[32] Joseph Friedman, et al., Excess out-of-hospital mortality and declining oxygen saturation: the sentinel role of emergency medical services data in the COVID-19 crisis in Tijuana, Mexico, Ann. Emerg. Med. 76 (4) (2020) 413–426.
[33] Gail M. Sullivan, Richard Feinn, Using effect size—or why the P value is not enough, J. Grad. Med. Educ. 4 (3) (2012) 279–282. .
[34] Joyce N. Barlin, et al., Classification and regression tree (CART) analysis of endometrial carcinoma: seeing the forest for the trees, Gynecol. Oncol. 130 (3) (2013) 452–456.
[35] S. Patro, Kishore Kumar Sahu, Normalization: a preprocessing stage, 2015 arXiv preprint arXiv:1503.06462.
[36] Neil R. Clark, Avi Ma'ayan, Introduction to statistical methods to analyze large data sets: principal components analysis, Sci. Signal. 4 (2011) 190, tr3-tr3.
[37] Sasan Karamizadeh, et al., An overview of principal component analysis, J. Signal Inf. Process. 4 (3B) (2013) 173.
[38] Ramazan Ünlü, Petros Xanthopoulos, Estimating the number of clusters in a dataset via consensus clustering, Expert Syst. Appl. 125 (2019) 33–39.
[39] F. Li, et al., Clustering ensemble based on sample's stability, Artif. Intell. 273 (2019) 37–55.
[40] Sandro Vega-Pons, José Ruiz-Shulcloper, A survey of clustering ensemble algorithms, Int. J. Pattern Recogn. Artif. Intell. 25 (3) (2011) 337–372.
[41] Y. Şenbabaoğlu, G. Michailidis, J.Z. Li, Critical limitations of consensus clustering in class discovery, Sci. Rep. 4 (1) (2014) 1–13.
[42] Niko Speybroeck, Classification and regression trees, Int. J. Publ. Health 57 (1) (2012) 243–246.
[43] Yue Zhou, et al., Comorbidities and the risk of severe or fatal outcomes associated with coronavirus disease 2019: a systematic review and meta-analysis, Int. J. Infect. Dis. (2020).
[44] Adekunle Sanyaolu, et al., "Comorbidity and its Impact on Patients with COVID-19." SN Comprehensive Clinical Medicine, 2020, pp. 1–8.
[45] Weina Guo, et al., Diabetes is a risk factor for the progression and prognosis of COVID-19, Diabetes (2020), e3319.
[46] Matteo Apicella, et al., COVID-19 in people with diabetes: understanding the reasons for worse outcomes, Lancet Diabetes Endocrinol. (2020).
[47] Giuseppe Lippi, Johnny Wong, Brandon Michael Henry, Hypertension and its severity or mortality in Coronavirus Disease 2019 (COVID-19): a pooled analysis, Pol. Arch. Intern. Med. 130 (4) (2020) 304–309.
[48] Cyrielle Caussy, et al., Prevalence of obesity among adult inpatients with COVID-19 in France, Lancet Diabetes Endocrinol. 8 (7) (2020) 562–564.
[49] Daisuke Miyazawa, Why obesity, hypertension, diabetes, and ethnicities are common risk factors for COVID-19 and H1N1 influenza infections, J. Med. Virol. (2020).
[50] Edgar Denova-Gutiérrez, et al., The association of obesity, type 2 Diabetes, and hypertension with severe coronavirus disease 2019 on admission among Mexican patients, Obesity 28 (10) (2020) 1826–1832.
[51] Marianna Sockrider, et al., COVID-19 infection versus influenza (Flu) and other respiratory illnesses, Am. J. Respir. Crit. Care Med. (2020).
[52] Jordan Cates, et al., Risk for in-hospital complications associated with COVID-19 and influenza—veterans health administration, United States, october 1, 2018–may 31, 2020, MMWR (Morb. Mortal. Wkly. Rep.) 69 (42) (2020) 1528.