VIA MEDICA

ORIGINAL ARTICLE

Cardiology Journal
2024, Vol. 31, No. 4, 512–521
DOI: 10.5603/cj.98489
Copyright © 2024 Via Medica
ISSN 1897–5593
eISSN 1898–018X

Phenotype clustering of hospitalized high-risk patients with COVID-19 — a machine learning approach within the multicentre, multinational PCHF-COVICAV registry

Mateusz Sokolski[1], Sander Trenson[2], Konrad Reszka[1], Szymon Urban[1], Justyna M. Sokolska[1], Tor Biering-Sørensen[3], Mats C. Højbjerg Lassen[3], Kristoffer Grundtvig Skaarup[3], Carmen Basic[4], Zacharias Mandalenakis[4], Klemens Ablasser[5], Peter P. Rainer[5], Markus Wallner[5–7], Valentina A. Rossi[8], Marzia Lilliu[9], Goran Loncar[10], Huseyin A. Cakmak[11], Frank Ruschitzka[8], Andreas J. Flammer[8]

[1]Wroclaw Medical University, Faculty of Medicine, Institute of Heart Diseases, Wroclaw, Poland and Institute of Heart Diseases, University Hospital, Wroclaw, Poland
[2]Department of Cardiology, Sint-Jan Hospital Bruges, Bruges, Belgium
[3]Department of Cardiology, Copenhagen University Hospital — Herlev & Gentofte, Copenhagen, Denmark
[4]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[5]Division of Cardiology, Medical University of Graz, Austria
[6]Cardiovascular Research Center, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, United States
[7]Center for Biomarker Research in Medicine, CBmed GmbH, Graz, Austria
[8]Department of Cardiology, University Heart Center, University Hospital Zurich, Switzerland
[9]Division of Infectious Diseases, Azienda ULSS 9, M. Magalini Hospital, Villafranca di Verona, Verona, Italy
[10]Institute for Cardiovascular Diseases Dedinje, Faculty of Medicine, University of Belgrade, Belgrade, Serbia
[11]Department of Cardiology, Mustafakemalpasa State Hospital, Bursa, Turkey

Abstract

Introduction: *The high-risk population of patients with cardiovascular (CV) disease or risk factors (RF) suffering from COVID-19 is heterogeneous. Several predictors for impaired prognosis have been identified. However, with machine learning (ML) approaches, certain phenotypes may be confined to classify the affected population and to predict outcome. This study aimed to phenotype patients using unsupervised ML technique within the International Postgraduate Course Heart Failure Registry for patients hospitalized with COVID-19 and Cardiovascular disease and/or RF (PCHF-COVICAV).*

Methods: *Patients from the eight centres with follow-up data available from the PCHF-COVICAV registry were included in this ML analysis (K-medoids algorithm).*

Address for correspondence: Mateusz Sokolski, MD, PhD, Institute of Heart Diseases, Wroclaw Medical University, ul. Borowska 213, 50–556 Wrocław, Poland, tel: + 48 717 331 112, e-mail: mateusz.sokolski@umw.edu.pl

Received: 10.12.2023 Accepted: 10.04.2024 Early publication date: 04.06.2024

Results: *Out of 617 patients included into the prospective part of the registry, 458 [median age: 76 (IQR: 65–84) years, 55% male] were analyzed and 46 baseline variables, including demographics, clinical status, comorbidities and biochemical characteristics were incorporated into the ML. Three clusters were extracted by this ML method. Cluster 1 (n = 181) represents mainly women with the least number of overall comorbidities and cardiovascular RF. Cluster 2 (n = 227) is characterized mainly by men with non-CV conditions and less severe symptoms of infection. Cluster 3 (n = 50) mainly represents men with the highest prevalence of cardiac comorbidities and RF, more extensive inflammation and organ dysfunction with the highest 6-month all-cause mortality risk.*

Conclusions: *The ML process has identified three important clinical clusters from hospitalized COVID-19 CV and/or RF patients. The cluster of males with severe CV disease, particularly HF, and multiple RF presenting with increased inflammation had a particularly poor outcome.* (Cardiol J 2024; 31, 4: 512–521)

Keywords: clustering, machine learning, artificial intelligence, COVID-19, SARS-CoV-2, cardiovascular disease

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become one of the biggest health-care crises globally [1].

While the majority of SARS-CoV-2 infections have a mild or even asymptomatic course, approximately 14% of COVID-19 cases were classified as severe, and 5% were critical [2, 3]. Critically ill patients often develop severe pneumonia, respiratory failure that requires mechanical ventilation, and other organ failures, and thus they require admission to intensive-care units (ICU) [2, 4]. The mortality rate is estimated at 0.09–2.54% [5].

Besides classical cardiovascular (CV) risk factors (RF), particularly pre-existing CV diseases, notably heart failure (HF), hypertension and coronary artery disease portends a high risk for adverse outcome in COVID-19. Moreover, conditions such as diabetes mellitus (DM), chronic obstructive pulmonary disease (COPD), and chronic kidney disease (CKD), which often coexist with CV diseases, are also linked with a poor prognosis [4, 6–8].

COVID-19 can exacerbate pre-existing HF or even cause HF de novo [9, 10]. Patients with HF often have longer, and more complicated hospital stays, are more susceptible for developing acute HF during hospitalization, and face a significantly higher mortality rate both during their hospital stay and after discharge [10].

The pandemic was overcome through the widespread availability of vaccinations and on 5th May 2023 World Health Organisation declared the end to COVID-19 as a global health emergency [11, 12]. Yet, due to the high mutation rate of SARS-CoV-2, there is still a risk of new variants emerging [13]. Hence, meticulous assessment and analysis of course, outcomes, and acknowledgment of RF may be vital to enhance treatment in the event of future pandemic.

In recent years, the rapid growth of Artificial Intelligence (AI) and Machine Learning (ML) has been observed, not only in our daily routines, but also in clinical practice [14]. AI and ML in healthcare are valuable tools in the decision-making process across various conditions and disease stages. They can assist in suggesting appropriate screening methods, facilitating specific diagnoses, and recommending suitable treatment options [14].

ML, particularly statistical clustering, is a technique designed to learn the inherent structure within a dataset [15]. Clustering is the unsupervised ML technique that segments the population into smaller subgroups, which are internally similar and distinct from the other ones. The aim of such analyses is to gain more detailed insight into the heterogeneity of the studied population [16].

The current study has implemented ML algorithms for CV patients hospitalized due to COVID-19 and their clinical variables obtained during hospitalization. The aim was to identify a subgroup of patients with confirmed SARS-CoV-2 infection and CV comorbidities who are at higher risk of an unfavourable in-hospital and mid-term prognosis.

Methods

The study characteristic is included in the previous publication of the registry [10]. The study was registered in the ClinicalTrials.gov database as The Global PCHF-COVICAV Registry (PCHF-COVICAV), Identifier NCT04390555. In the above registry, 28 centers participated, while prospective data was available only from 8 centers: Austria (Graz), Denmark (Copenhagen), Italy (Verona), Poland (Wroclaw), Sweden (Gothenburg), Switzerland (Zurich), Serbia (Belgrade), Turkey (Bursa). Only data from these centers were used in the current study.

Study population

This multicentre, international cohort study included 617 hospitalized adult patients (≥ 18 years old) with laboratory confirmed COVID-19 defined as positive result by polymerase chain reaction testing of a nasopharyngeal sample or a positive blood antigen. After the non-CV and non-RF patients were excluded (159 patients), 458 patients were incorporated in the clustering analysis.

Clustering and data analysis

The initial dataset consisted of 355 variables, which underwent a comprehensive screening process. Variables with missing values exceeding the threshold of 50% were excluded from subsequent analysis to ensure data integrity. This step resulted in a refined dataset comprising 181 variables for further investigation.

From the remaining variables, a rigorous selection process was conducted to identify 46 parameters that encompass key aspects of demographics, clinical status, comorbidities, and biochemical characteristics. All the variables were manually screened to identify and remove the outliers, then the spreadsheet was implemented into the RapidMiner software (RapidMiner Studio 9.1).

Automated pre-processing was performed to eliminate variables which were correlated with $r > 0.6$, but none of the variables fulfilled the criteria. Missing values were imputed using the mean values, as clustering algorithms are unable to process data with missing values. Additionally, nominal values were converted into numerical representations, and all numerical parameters were normalized to a range of 0 to 1. This normalization ensured that each variable had equal influence on the calculated distance, facilitating unbiased clustering analysis.

The segmentation process can be approached using various algorithms, each employing distinct grouping strategies. In the analysis, the k-medoids algorithm were utilized, which relies on centroids representing existing data points for segmentation [17]. This algorithm performs clustering by iteratively assigning examples to clusters based on minimizing the distance to the centroid and subsequently recomputing the centroid.

To optimize the clustering process and achieve the highest quality clusters, the automated optimization functionality of RapidMiner was leveraged. This involved adjusting selected parameters to maximize cluster quality as measured by the Davies-Bouldin index [18].

Within this optimization framework, the system was allowed to determine the optimal number of clusters, considering a range from 2 to 5, and to select the most appropriate numeric distance/similarity measure the EuclideanDistance, CamberraDistance, ChebychevDistance, CorrelationSimilarity, CosineSimilarity, DiceSimilarity, DynamicTimeWarpingDistance, InnerProductSimilarity, JaccardSimilarity, KernelEuclideanDistance, ManhattanDistance, MaxProductSimilarity, and OverlapSimilarity from the provided list of variables. To prevent excessive fragmentation, a maximum of 5 clusters was considered, as exceeding this threshold may lead to overly granular segmentation results.

Statistical analysis

The disparities in clinical parameters across different clusters were examined. Prior to analysis, the normality of the parameter distribution was assessed using the Shapiro-Wilk test. Variables conforming to a normal distribution were reported as mean ± standard deviation, while parameters with skewed distribution were presented as median [interquartile range].

Statistical significance was determined by applying appropriate tests such as variance analysis, ANOVA Kruskal-Wallis test, and Chi-square Pearson test. These tests were chosen based on the nature and characteristics of the variables being investigated. Furthermore, the impact of the identified clusters on all-cause mortality until 6 months was assessed using Kaplan-Meier curves and Cox proportional-hazards regressions. Moreover, in-hospital all-cause death, intensive care hospitalization, the duration of hospitalization, acute HF events during hospitalization were also used to describe the prognosis. All statistical analyses were conducted using STATISTICA software (TIBCO Statistica, v. 13.3, TIBCO Software Inc.).

Results

Data from 458 patients hospitalized with COVID-19 (median age: 76 [IQR: 65–84] years, 55% male) were included in the present clustering analysis. Figure 1 presents the flowchart of the variables and patients included in the analysis. Baseline 46 parameters were incorporated into the model after pre-processing (Table 1).

Clustering

The algorithm created three clusters, enumerated from 1 to 3, which differ in demographics, comorbidities, signs and symptoms, laboratory and lifestyle features (see Table 2).

Cluster 1 (n = 181)

Cluster 1 represents mainly women with the least number of comorbidities and cvRF. This cluster had the lowest prevalence of HF, chronic kidney disease, history of smoking, and lowest body mass index (BMI). On admission, they presented with the highest baseline systolic blood pressure and less often had dyspnoea. The lactates, procalcitonin, NT-proBNP, creatinine, potassium and INR levels were the lowest in this cluster. Moreover, haemoglobin levels were the highest.

Cluster 2 (n = 227)

This Cluster is characterized mainly by men with non-CV conditions and less severe symptoms of infection. This was the largest cluster with the lowest prevalence of the following comorbidities: peripheral artery disease, stroke or transient ischemic attack, dyslipidemia, arterial hypertension. On admission, they presented least frequently with cough and had the lowest body temperature. This cluster had the highest alanine aminotransferase (ALT) and lowest hemoglobin and albumins levels.

Cluster 3 (n = 50)

Cluster 3 mainly represents men with the highest prevalence of cardiac comorbidities and RF, more extensive inflammation and organ dysfunction with the highest 6-month all-cause mortality risk. This cluster had the highest prevalence of HF, myocardial infarction, atrial fibrillation, peripheral artery disease, stroke or transient ischemic attack, history of smoking, dyslipidemia, DM, arterial hypertension, CKD. On admission, this cluster had the highest body temperature, lowest systolic blood pressure and oxygen saturation. Among laboratory parameters lactates, procalcitonin, NT-proBNP, creatinine, potassium, INR levels were the highest in this group.

Prognostic significance of clusters

The overall in-hospital and 6-month mortality in the entire study sample were 152 (33%) and 180 (39%) respectively. The median of hospital stay was 11 [6–22] days. The in-hospital mortality from cluster 1 to cluster 3 was: 28% vs. 34% vs. 48%,

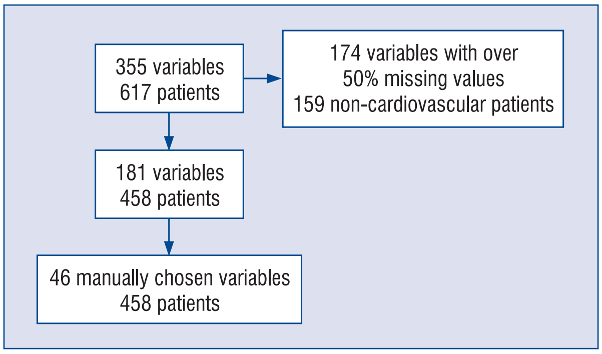

Figure 1. Flowchart of the variables and patients included in the analysis

Table 1. Variables included in the analysis

Category	Variables
Demographics	Age, sex
Comorbidities	Heart failure, myocardial infarction, significant valvular heart disease, atrial fibrillation, peripheral artery disease, stroke or transient ischemic attack, smoking, dyslipidemia, type 2 diabetes, arterial hypertension, COPD or asthma, malignant neoplasms
Clinical status	BMI, body temperature, respiratory rate, heart rate, systolic blood pressure, cough, dyspnea, chest pain, gastrointestinal symptoms, altered smell or taste sensation, central congestion, edema
Lifestyle factors	History of smoking
Laboratory parameters	Oxygen saturation, pH, arterial pCO_2, lactate, CRP, procalcitonin, hemoglobin, white blood cell count, blood platelet count, creatine kinase, NT-proBNP, creatinine, ALT, albumin, D-dimers, INR, ferritin, sodium, potassium

ALT — alanine aminotransferase; AST — aspartate aminotransferase; BMI — body mass index; CRP — C-reactive protein; eGFR — estimated glomerular filtration rate; INR — international normalized ratio; NT-proBNP — n-terminal pro-b-type natriuretic peptide; pCO_2 — partial pressure of carbon dioxide

Table 2. Baseline characteristics of the included population and created clusters

Parameter	All cohort	Cluster 1	Cluster 2	Cluster 3	P-value
N [%]	458	181 (39%)	227 (50%)	50 (11%)	
Age (years)	75.7 [65–83.5]	75.5 [64–84.4]	75.7 [64–83.1]	76.5 [70–83]	0.779
BMI kg/m^2	26.3 [23.9–30.5]	25.8 [23.3–28.8]	26.3 [23.9–31]	28.4 [25.1–30.7]	0.064
Body temperature (degrees celsius)	37.4 [36.8–38.2]	37.5 [36.6–38]	37.3 [36.8–38.1]	37.9 [37–38.6]	0.037
Respiratory rate (per minute)	20 [18–25]	20 [18–25]	20 [18–24]	22 [18–28]	0.235
Heart rate (beats per minute)	85 [75–100]	85 [75–100]	85 [75–99]	84 [76–100]	0.852
Systolic blood pressure (mmHg)	128 [115–145]	130 [120–150]	127 [114–142]	124 [110–137]	0.012
Oxygen saturation [%]	94 [9 –96]	94 [91–96]	94 [91–96]	92 [87–95]	0.011
PH	7.45 [7.41–7.49]	7.45 [7.42–7.48]	7.45 [7.41–7.49]	7.42 [7.38–7.48]	0.123
Arterial pCO$_2$ (kPa)	45 [39.9–52]	44.1 [39–50.3]	45.51 [40.53–53]	43.26 [38–52]	0.214
Lactate (mmol per liter)	1.3 [1–1.9]	1.2 [1–1.8]	1.4 [1–1.9]	1.8 [1–1.8]	0.037
CRP (mg per liter)	67 [29–134]	64 [26–122]	64 [26–122]	64 [35–130]	0.321
Procalcitonin (ng per milliliter)	0.19 [0.09–0.49]	0.16 [0.07–0.31]	0.23 [0.1– 0.72]	0.25 [0.11–0.72]	0.017
Hemoglobin (g per liter)	11.6 [8.1–13.4]	12.6 [10.5–13.7]	9.8 [7.5–13.2]	11.2 [9–12.9]	< 0.001
White blood cell count (× 10^9 per liter)	7 [5.1–9.9]	6.9 [5– 9.9]	7 [5.1–9.9]	7.1 [4.5– 9.5]	0.645
Blood platelet count (× 10^9 per liter)	196 [157.5–273]	199 [164– 291]	204.5 [155– 269]	185 [144– 211]	0.078
Creatine kinase (U per liter)	106 [69–98]	104 [62–172]	112 [69–263]	105 [79–170]	0.397
NT-proBNP (n-terminal pro-b-type natriuretic peptide)	546.3 [153–683]	264 [102–925]	555 [152–2760]	928 [619–1932]	0.006
Creatinine (mg per deciliter)	1.02 [0.75–1.5]	0.93 [0.7–1.2]	1.05 [0.78–1.82]	1.31 [1.0–1.9]	< 0.001
ALT (U per liter)	27 [17–45]	26 [17–45]	30 [18–47]	18 [1 –35]	0.045
Albumin (mg per deciliter)	3.33 ± 0.64	3.48 ± 0.62	3.19 ± 0.65	3.19 ± 0.6	0.007
D-dimers (microgram per milliliter)	1.21 [0.75–2.2]	1 [0.7–1.74]	1.48 [0.79–2.4]	1.26 [0.52–2.13]	0.188
INR	1.09 [1–1.23]	1.03 [0.99–1.13]	1.1 [–1.3]	1.12 [1–1.46]	0.015
Ferritin (g per liter)	570 [203–1118]	542 [210–1280]	591 [294–1060]	127 [66–629]	0.052

cont. →

Table 2. (cont.) Baseline characteristics of the included population and created clusters

Parameter	All cohort	Cluster 1	Cluster 2	Cluster 3	P–value
Sodium (mmol per liter)	138 [135–141]	137 [133–141]	139 [136–141]	138 [135–141]	0.054
Potassium (mmol per liter)	4 [3.6–4.3]	3.9 [3.5–4.2]	4 [3.6–4.4]	4.1 [3.8–4.5]	0.013
Sex (female), n [%]	208 (45%)	131 (72%)	64 (28%)	13 (26%)	< 0.001
Heart failure (yes), n [%]	109 (24%)	29 (16%)	40 (18%)	40 (80%)	< 0.001
Myocardial infarction (yes), n [%]	57 (18%)	21 (14%)	18 (14%)	22 (55%)	< 0.001
Significant valvular heart disease (yes), n [%]	43 (9%)	13 (7%)	22 (10%)	8 (16%)	0.168
Atrial fibrillation (yes), n [%]	131 (29%)	43 (24%)	55 (24%)	33 (66%)	< 0.001
Peripheral artery disease (yes) n ,%	47 (10%)	15 (8%)	13 (6%)	19 (38%)	< 0.001
Stroke or transient ischemic attack (yes), n [%]	62 (19%)	25 (16%)	16 (12%)	21 (51%)	< 0.001
History of smoking (yes), n [%]	146 (43%)	42 (33%)	73 (42%)	31 (82%)	< 0.001
Dyslipidemia (yes), n [%]	212 (47%)	118 (66%)	54 (25%)	40 (80%)	<0.001
Type 2 diabetes (yes), n [%]	154 (34%)	50 (28%)	63 (28%)	41 (82%)	<0.001
Arterial hypertension (yes), n [%]	351 (77%)	145 (81%)	161 (72%)	45 (90%)	0.011
Copd or asthma (yes), n [%]	68 (21%)	25 (16%)	32 (24%)	11 (27%)	0.160
Malignant neoplasms (yes), n [%]	53 (16%)	26 (17%)	22 (17%)	5 (12%)	0.752
Chronic Kidney disease (eGFR < 60) (yes), n [%]	76 (17%)	21 (12%)	43 (19%)	12 (24%)	0.046
Cough (yes), n [%]	148 (47%)	91 (59%)	31 (25%)	26 (65%)	< 0.001
Dyspnea (yes), n [%]	184 (58%)	57 (38%)	94 (75%)	33 (80%)	< 0.001
Chest pain (yes), n [%]	14 (4%)	7 (5%)	6 (5%)	1 (2%)	0.805
Gastrointestinal symptoms (yes), n [%]	23 (7%)	13 (9%)	8 (6%)	2 (5%)	0.634
Altered smell or taste sensation (yes), n [%]	6 (2%)	4 (3%)	2 (2%)	0,00	0.543
Central congestion (yes), n [%]	37 (12%)	15 (11%)	16 (13%)	6 (16%)	0.704
Edema, n [%]	35 (12%)	14 (10%)	14 (11%)	7 (18%)	0.432

ALT — alanine aminotransferase; AST — aspartate aminotransferase; BMI — body mass index; CRP — C-reactive protein; eGFR — estimated glomerular filtration rate; INR — international normalized ratio; NT-proBNP — n-terminal pro-b-type natriuretic peptide; pCO_2 — partial pressure of carbon dioxide p = 0.028, respectively, while 6-month mortality was: 34% vs. 41% vs. 52%, p = 0.056. The groups also differed in terms of acute HF during hospitalization from cluster 1 to cluster 3: 5% vs. 7% vs. 14%, p = 0.032 (Table 3). The risks for 6-month mortality compared with the rest of the population were calculated for each cluster.

Cluster 3 had the highest 6-month all-cause mortality with hazard ratio (95% confidence interval): 1.53 [1.01–2.32], p = 0.045. There were no significant differences compared to the rest of the population for clusters 1 and 3 (Table 4). Figure 2 shows the Kaplan-Meier curves for the 6-month all-cause mortality risks by clusters.

Table 3. Outcome across the clusters

Parameter	All cohort	Cluster 1	Cluster 2	Cluster 3	P-value
Death during hospitalization, n [%]	152 (33%)	50 (28%)	78 (34%)	24 (48%)	0.028
Mortality until 6 months, n [%]	180 (39%)	61 (34%)	93 (41%)	26 (52%)	0.056
ICU stay during hospitalization, n [%]	85 (19%)	32 (18%)	41 (18%)	12 (24%)	0.611
Days in hospital	11 [6–22]	11 [6–20]	11 [5–23]	10 [8–26]	0.511
Acute hf during hospitalization, n [%]	31 (7%)	9 (5%)	15 (7%)	7 (14%)	0.032

HF — heart failure; ICU — intensive care unit

Table 4. Hazard ratios for 6-month all-cause mortality. Each cluster was compared with the rest of the population

Cluster	HR, 95% CI	P-value
Cluster 1	0.74 [0.54–1.01]	0.053
Cluster 2	1.11 [0.83–1.49]	0.481
Cluster 3	1.53 [1.01–2.32]	0.045

CI — confidence interval; HR — hazard ratio

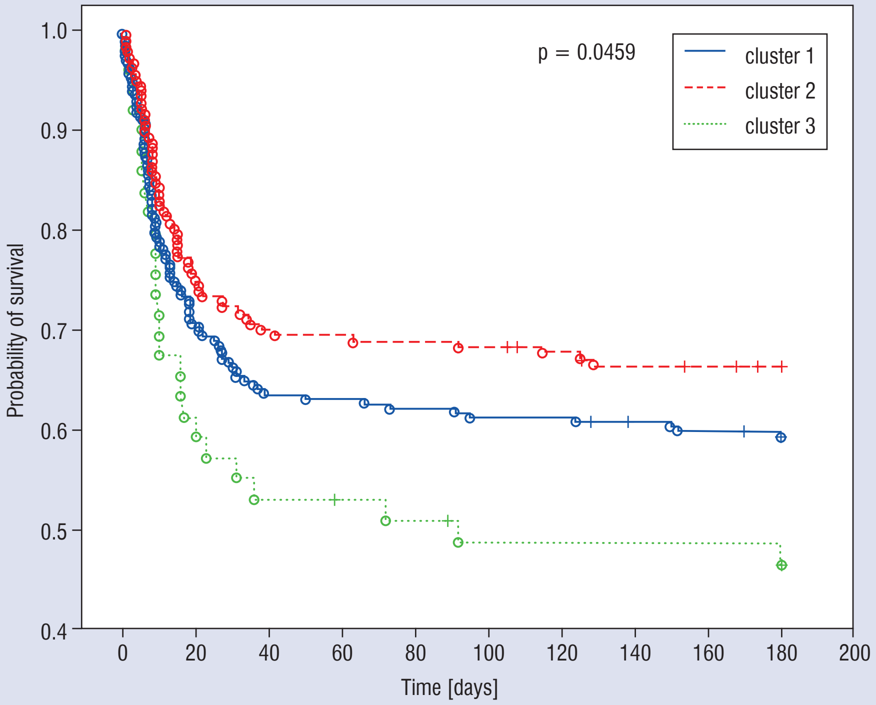

Figure 2. Kaplan-Meier curves for 6-month all-cause mortality across the clusters

Discussion

The key finding of the present study is that ML can be incorporated into COVID-19 high risk population identifying specific subgroups with varying outcomes. With the help of ML, a particular high-risk cluster was identified (cluster 3). Although all patients in this registry have very high mortality risk, this cluster of patients was extremely endangered. These are particularly patients with severe CV disease and heart failure.

This is the typical population characterized by impaired vascular function. The vascular endothelium is extremely important for the regulation of vascular tone and the maintenance of vascular homeostasis. With vascular dysfunction, a shift towards vasoconstriction with organ ischemia, inflammation with tissue edema and a pro-coagulant state is induced. As COVID-19 affects the vasculature (endotheliitis) patients with pre-existing endothelial dysfunction are particularly vulnerable to adverse outcomes [19].

ML has been reported as an effective and innovative tool in clinical practice, demonstrated favourable performance in various conditions such as CV diseases, cancer, sepsis, and depression [20–25]. During COVID-19 pandemic AI and ML applications have been developed for both clinical and non-clinical purposes. These technologies were designed to assist healthcare providers and aid public health officials in controlling the pandemic outside the hospitals [15].

In a clinical setting, machine learning (ML) has been utilized to identify patients with potential undesired outcomes or predict course of the disease, enhance clinical diagnosis, and assist with image recognition [15].

In the presented study, the ML model identified a subgroup of CV COVID-19 patients with a high risk of in-hospital complications and the poorest prognosis (cluster 3). Many of the features observed in this subgroup align with previous studies, as these patients were more likely to be male, obese and smokers [2–4, 6–7,10]. Furthermore, they had a higher prevalence of HF, which was accompanied by comorbidities. Among the laboratory findings associated with worse outcomes were: higher concentrations of lactates, NT-proBNP, creatinine, potassium, international normalized ratio (INR) and procalcitonin, as well as lower haemoglobin and albumins levels. The third cluster integrated HF comorbidities and laboratory abnormalities which are components of the pathophysiological mechanisms linking COVID-19 and HF. COVID-19 through enhanced inflammatory response and endothelial damage was causing organ dysfunction including cardio-renal-hepatic axis, especially in conditions with primarily affected endothelium, like is reported in HF [6–7, 26–30]. It should be emphasized the high mortality rate at 6-month follow-up in the whole study population, which exceeded 50% in cluster 3. This is probably due to the selection of the more elderly population with CV disease.

Advanced age is associated with a higher prevalence of comorbidities, weaker immune system, and elevated levels of proinflammatory cytokines [4, 10]. Many studies reported that advanced age is linked with worse outcomes of SARS-CoV-2 infection [2, 4, 10]. However, in the present study, the mean age was comparable in all clusters and did not have a significant impact on the prognosis.

The previous clustering analyses showed similar results compared with the current study [26–29]. Usually three or four clusters are identified and have found correlations for a poor prognosis, like high comorbidity scores [28], being male, high lymphocytes high neutrophil count [26], and albumin level [27].

The study was conducted on a group of 7,606 COVID-19-positive patients hospitalized in Hong Kong is particularly noteworthy for its significant statistical power and generalizability. This study demonstrated a high ability to differentiate a cluster by capturing 86.6% of fatal cases and encompassing their clinical characteristics and correlations. The authors reported that old age, male gender, lower levels of hemoglobin, hematocrit, lymphocytes, albumin, elevated lactate dehydrogenase, higher levels of neutrophils, urea, and C-reactive protein (CRP), as well as higher comorbidity scores, were correlated with a higher mortality rate. However, it should be noted that this study analysed the general population of Chinese ethnicity and did not specifically provide an analysis of the subgroup of CV patients [28].

While the present study was meant to identify a subgroup of CV patients with a high risk of in-hospital complications, it should be noted that there are also ML models dedicated to predicting patient outcomes and mortality [32, 33]. The model incorporating three biomarkers (i.e. lactic dehydrogenase (LDH), lymphocytes, and high-sensitivity CRP), demonstrated an accuracy of over 90% and an area under the curve (AUC) of 95.06% in predicting the outcome within a 10-day period [32].

Similarly, a model trained on a dataset of patients retrospectively recruited from 30 clinical centers across Italy was able to predict in-hospital mortality within a median follow-up of 13 days with a sensitivity of 95.2%, specificity of 30.8%, and a classification accuracy of 83.4% [33]. Analysis was based on C-reactive protein concentration, renal function, and age [33]. The main advantage of these models is that they rely on biomarkers that are typically collected from patients with COVID-19, making them feasible and applicable in healthcare settings.

Conclusions

Using the ML three phenotypes from COVID-19 CV and/or RF patients with distinct clinical characteristics and outcomes were extracted. High-risk phenotype male patients with severe CV disease were identified, particularly HF, and multiple RF presenting with increased inflammatory and organ dysfunction parameters on admission. The results show that ML techniques, incorporating classical clinical parameters can be a useful tool in distinguishing risk groups among COVID-19 patients with CV disease and/or RF presented at emergency departments. The results may be used in clinical practice for early detection of patients at high risk for in-hospital mortality or within 6 months so that in the next step physicians may adjust appropriate management strategies for these patients.

Acknowledgements: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contributions: Concept and design: Sokolski, Trenson, Sokolska, Urban; acquisition, analysis, interpretation of data, administrative, technical, or material support; revision of the manuscript: all authors; drafting of the manuscript: Sokolski, Reszka; critical revision of the manuscript for important intellectual content: all authors; statistical analysis: Sokolski, Urban; supervision: Sokolski, Trenson, Ruschitzka, Flammer; final approval of the version to be published: All authors.

Funding: University of Zurich.

Conflict of interest: Authors report no competing interests.

References

1. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int (09.07.2023).
2. Pepera G, Tribali MS, Batalik L, et al. Epidemiology, risk factors and prognosis of cardiovascular disease in the Coronavirus Disease 2019 (COVID-19) pandemic era: a systematic review. Rev Cardiovasc Med. 2022; 23(1): 28, doi: 10.31083/j.rcm2301028, indexed in Pubmed: 35092220.
3. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020; 323(13): 1239–1242, doi: 10.1001/jama.2020.2648, indexed in Pubmed: 32091533.
4. Gao YD, Ding M, Dong X, et al. Risk factors for severe and critically ill COVID-19 patients: A review. Allergy. 2021; 76(2): 428–455, doi: 10.1111/all.14657, indexed in Pubmed: 33185910.
5. Ritchie H, Mathieu E, Rodés-Guirao L, et al. Coronavirus Pandemic (COVID-19). Our World Data. 2020. Available at: https://ourworldindata.org/covid-vaccinations (09.07.2023).
6. AlShahrani I, Hosmani J, Shankar VG, et al. COVID-19 and cardiovascular system-a comprehensive review. Rev Cardiovasc Med. 2021; 22(2): 343–351, doi: 10.31083/j.rcm2202041, indexed in Pubmed: 34258902.
7. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol. 2020; 146(1): 110–118, doi: 10.1016/j.jaci.2020.04.006, indexed in Pubmed: 32294485.
8. Gao C, Cai Y, Zhang K, et al. Association of hypertension and antihypertensive treatment with COVID-19 mortality: a retrospective observational study. Eur Heart J. 2020; 41(22): 2058–2066, doi: 10.1093/eurheartj/ehaa433, indexed in Pubmed: 32498076.
9. Sokolski M, Trenson S, Sokolska JM, et al. Heart failure in COVID-19: the multicentre, multinational PCHF-COVICAV registry. ESC Heart Fail. 2021; 8(6): 4955–4967, doi: 10.1002/ehf2.13549, indexed in Pubmed: 34533287.
10. Sokolski M, Reszka K, Suchocki T, et al. History of Heart Failure in Patients Hospitalized Due to COVID-19: Relevant Factor of In-Hospital Complications and All-Cause Mortality up to Six Months. J Clin Med. 2022; 11(1), doi: 10.3390/jcm11010241, indexed in Pubmed: 35011982.
11. Magri MMC, Uip DE, Rodrigues FK, et al. Impact of the Vaccination Against COVID-19 on Frontline Health Workers. Curr Drug Saf. 2023; 18(4): 511–515, doi: 10.2174/157488631766622062012 0444, indexed in Pubmed: 35726430.
12. United Nations. https://news.un.org/en/story/2023/05/1136367 (09.07.2023).
13. Malik JA, Ahmed S, Mir A, et al. The SARS-CoV-2 mutations versus vaccine effectiveness: New opportunities to new challenges. J Infect Public Health. 2022; 15(2): 228–240, doi: 10.1016/j.jiph.2021.12.014, indexed in Pubmed: 35042059.
14. Aldhoayan MD. The Role of Artificial Intelligence and Machine Learning During the Covid-19 Pandemic: A Review. Stud Health Technol Inform. 2022; 295: 28–32, doi: 10.3233/SHTI220651, indexed in Pubmed: 35773797.
15. Urban S, Błaziak M, Jura M, et al. Novel Phenotyping for Acute Heart Failure-Unsupervised Machine Learning-Based Approach. Biomedicines. 2022; 10(7), doi: 10.3390/biomedicines10071514, indexed in Pubmed: 35884819.

16. Tan, Pang-Ning, Michael Steinbach, Vipin Kumar. Introduction to data mining. Pearson Education India. 2016.
17. Kaufman L, Rousseeuw PJ. Partitioning Around Medoids (Program PAM). Finding Groups in Data. 2008: 68–125, doi: 10.1002/9780470316801.ch2.
18. Davies D, Bouldin D. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979; PAMI-1(2): 224–227, doi: 10.1109/tpami.1979.4766909.
19. Varga Z, Flammer AJ, Steiger P, et al. Endothelial cell infection and endotheliitis in COVID-19. Lancet. 2020; 395(10234): 1417–1418, doi: 10.1016/S0140-6736(20)30937-5, indexed in Pubmed: 32325026.
20. Krittanawong C, Virk HU, Bangalore S, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020; 10(1): 16057, doi: 10.1038/s41598-020-72685-1, indexed in Pubmed: 32994452.
21. Lu W, Fu D, Kong X, et al. FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. Cancer Med. 2020; 9(4): 1419–1429, doi: 10.1002/cam4.2786, indexed in Pubmed: 31893575.
22. Castaldo R, Cavaliere C, Soricelli A, et al. Radiomic and Genomic Machine Learning Method Performance for Prostate Cancer Diagnosis: Systematic Literature Review. J Med Internet Res. 2021; 23(4): e22394, doi: 10.2196/22394, indexed in Pubmed: 33792552.
23. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020; 46(3): 383–400, doi: 10.1007/s00134-019-05872-y, indexed in Pubmed: 31965266.
24. Lee Y, Ragguett RM, Mansur RB, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. J Affect Disord. 2018; 241: 519–532, doi: 10.1016/j.jad.2018.08.073, indexed in Pubmed: 30153635.
25. Xiong Y, Ma Y, Ruan L, et al. National Traditional Chinese Medicine Medical Team. Comparing different machine learning techniques for predicting COVID-19 severity. Infect Dis Poverty. 2022; 11(1): 19, doi: 10.1186/s40249-022-00946-4, indexed in Pubmed: 35177120.
26. Li WT, Ma J, Shende N, et al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inform Decis Mak. 2020; 20(1): 247, doi: 10.1186/s12911-020-01266-z, indexed in Pubmed: 32993652.
27. Ye W, Lu W, Tang Y, et al. Identification of COVID-19 Clinical Phenotypes by Principal Component Analysis-Based Cluster Analysis. Front Med (Lausanne). 2020; 7: 570614, doi: 10.3389/fmed.2020.570614, indexed in Pubmed: 33282887.
28. Lau KYY, Ng KS, Kwok KW, et al. An Unsupervised Machine Learning Clustering and Prediction of Differential Clinical Phenotypes of COVID-19 Patients Based on Blood Tests-A Hong Kong Population Study. Front Med (Lausanne). 2021; 8: 764934, doi: 10.3389/fmed.2021.764934, indexed in Pubmed: 35284429.
29. Ilbeigipour S, Albadvi A, Akhondzadeh Noughabi E. Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making. Inform Med Unlocked. 2022; 32: 101005, doi: 10.1016/j.imu.2022.101005, indexed in Pubmed: 35813016.
30. Nägele MP, Haubner B, Tanner FC, et al. Endothelial dysfunction in COVID-19: Current findings and therapeutic implications. Atherosclerosis. 2020; 314: 58–62, doi: 10.1016/j.atherosclerosis.2020.10.014, indexed in Pubmed: 33161318.
31. Cui W, Robins D, Finkelstein J. Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records. Stud Health Technol Inform. 2020; 272: 1–4, doi: 10.3233/SHTI200478, indexed in Pubmed: 32604585.
32. Yan Li, Zhang HT, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence. 2020; 2(5): 283–288, doi: 10.1038/s42256-020-0180-7.
33. Di Castelnuovo A, Bonaccio M, Costanzo S, et al. COvid-19 RISk and Treatments (CORIST) collaboration. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutr Metab Cardiovasc Dis. 2020; 30(11): 1899–1913, doi: 10.1016/j.numecd.2020.07.031, indexed in Pubmed: 32912793.