

1 **Machine Learning for COVID-19 Patient Management:**

2 **Predictive Analytics and Decision Support**

3 Christopher El Hadi^{1†}, Rindala Saliba^{1,2}, Georges Maalouly^{1,3}, Moussa Riachy^{1,4}, Ghassan
4 Sleilaty^{1,5}

5 ¹ Faculty of Medicine, Saint Joseph University of Beirut, Beirut, Lebanon.

6 ² Clinical Microbiology Department, University Medical Center Hôtel-Dieu de France
7 Hospital, Faculty of Medicine, Saint Joseph University of Beirut, Beirut, Lebanon.

8 ³ Internal Medicine Department, University Medical Center Hôtel-Dieu de France Hospital,
9 Faculty of Medicine, Saint Joseph University of Beirut, Beirut, Lebanon.

10 ⁴ Pulmonary and Critical Care Department, University Medical Center Hôtel-Dieu de France
11 Hospital, Faculty of Medicine, Saint Joseph University of Beirut, Beirut, Lebanon.

12 ⁵ Cardiovascular Department, University Medical Center Hôtel-Dieu de France Hospital,
13 Faculty of Medicine, Saint Joseph University of Beirut, Beirut, Lebanon.

14

15 **Word count:** Abstract = 250 words, Text = 3111 words

16

17 **Author contribution**

18 C.H. drafted the manuscript and analyzed the data collected and supplied by M.R. and G.S.

19 G.M. supervised the work and analyzed the outputs. R.S. extensively revised all versions of
20 the article and contributed to the final manuscript.

1 **Abstract**

2 **Background.** The global impact of severe acute respiratory syndrome coronavirus 2 (SARS-
3 CoV-2) has profoundly affected economies and healthcare systems around the world,
4 including Lebanon. While numerous meta-analyses have explored the systemic
5 manifestations of COVID-19, few have linked them to patient history. Our study aims to fill
6 this gap by using cluster analysis to identify distinct clinical patterns among patients, which
7 could aid prognosis and guide tailored treatments.

8 **Methods.** We conducted a retrospective cohort study at Beirut's largest teaching hospital on
9 556 patients with SARS-CoV-2. We performed cluster analyses using K-prototypes,
10 KAMILA and LCM algorithms based on 26 variables, including laboratory results,
11 demographics and imaging findings. Silhouette scores, concordance index and signature
12 variables helped determine the optimal number of clusters. Subsequent comparisons and
13 regression analyses assessed survival rates and treatment efficacy according to clusters.

14 **Results.** Our analysis revealed three distinct clusters: "resilient recoverees" with varying
15 disease severity and low mortality rates, "vulnerable veterans" with severe to critical disease
16 and high mortality rates, and "paradoxical patients" with a late presentation but eventual
17 recovery.

18 **Conclusions.** These clusters offer insights for prognosis and treatment selection. Future
19 studies should include vaccination data and various COVID-19 strains for a comprehensive
20 understanding of the disease's dynamics.

21

22 **Keywords:** COVID-19, Clustering, Machine learning, K-prototypes, KAMILA, LCM

1 **1. Background**

2 Over the past three years, the epidemic of severe acute respiratory syndrome caused by the
3 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread around the
4 world, affecting multiple economies and healthcare systems, and rubbing salt into the wound
5 of Lebanon's economy and status quo [1]. As an infectious disease of the respiratory tract,
6 SARS-CoV-2, which causes coronavirus disease (COVID-19), generally manifests itself with
7 common symptoms such as fever, fatigue, headache, cough and sore throat [2]. However,
8 clinical presentations and disease severity in patients with COVID-19 can vary considerably,
9 depending on the circulating viral strain, comorbidities and the patient's immune constitution.
10 Symptoms can range from none, as in over a third of infected individuals, to the life-
11 threatening, as in those with underlying acute respiratory failure [3]. Predicting a patient's
12 reaction to the virus and administering the appropriate treatments to avoid an unfavorable and
13 potentially fatal outcome may seem an avant-garde approach, but it's possible thanks to
14 modern statistical algorithms for clustering patients.

15 Cluster analysis is a fundamental technique in data mining, designed to reveal patterns that
16 may be hidden by the complexity of the data, and extract knowledge from them. It has many
17 applications in a variety of fields, e.g. socio-economic and medical, and has proved
18 particularly useful for uncovering patterns in clinical data that might not be easily discerned
19 by human analysis alone [4,5]. It involves the use of algorithms to divide data into groups of
20 observations or "clusters", based on increasing the similarity between the components of a
21 cluster, while reinforcing the dissimilarity between clusters [6]. Such advance have led to a
22 paradigm shift in medicine, where precision medicine is becoming tantamount to evidence-
23 based medicine, particularly in areas such as cancer and metabolic diseases [7]. For
24 challenging diseases such as COVID-19, it may be of interest to identify distinct patient
25 categories to enable a more personalized and rigorous approach to patient care.

1 We thus sought to classify patients with COVID-19 treated at the Hôtel Dieu de France
2 Hospital in Beirut, on the basis of their medical history and the biochemical and radiological
3 results obtained on hospital presentation. The representative criteria of each class obtained
4 were then compared, enabling the clusters to be calibrated in order to adopt proactive
5 approaches for each Lebanese admitted patient newly diagnosed with COVID-19. The latter
6 will have the opportunity to be matched with one of the studied clusters of COVID-19
7 patients with clear, albeit probabilistic, treatment recommendations ready for
8 implementation.

9

10 **2. Methods**

11 We conducted a single-center retrospective cohort study. We included 556 hospitalized
12 patients with confirmed COVID-19 between September 22, 2019, and October 12, 2021. All
13 statistical analyses were performed using R 4.3.1 (The R Foundation for Statistical
14 Computing, Vienna, Austria) [8].

15 2.1. Data collection

16 The data were extracted from the Hôtel Dieu de France hospital electronic database. Only the
17 first data values collected within 24 hours of hospital admission were used for cluster
18 analysis (Table 1). The prognostic value of the different treatments administered was
19 assessed by studying their effect on severity variables such as contraction of nosocomial
20 infections, development of pneumo-mediastinum, composite fatal outcome (i.e admission,
21 intubation and death), day of ICU transfer, date of intubation, occurrence of thromboembolic
22 or hemorrhagic events and the corresponding dates, duration of hospitalization and all-cause
23 death.

1 2.2. Data pre-processing and Clustering

2 The data studied is composed of continuous and categorical/ordinal variables (Table 1).

3 Scaling and normalization were applied to continuous variables to meet the requirements of

4 each algorithm [9,10]. Missing data were handled using the *mice* package [11].

Continuous Variables

• Age	• Lymphocytes	• Weight
• CRP	• Neutrophiles	• Baseline CAT Ground Glass estimation
• D-dimers	• Procalcitonin	• Baseline CAT Pulmonary artery diameter
• Ferritin	• Serum Creatinine	• 1st PCR CT value
• LDH	• Leucocytes	• Day symptoms started

Categorical Variables	Definition
Sex	Biological gender
• 0	Male
• 1	Female
Chronic renal failure	Progressive kidney damage causing a buildup of waste and toxins in the body
• 0	Absent
• 1	Present
Cardiovascular disease	Any cardiac or vascular disease, e.g. coronary disease and heart failure
• 0	Absent

• 1	Present
Diabetes mellitus	Chronic metabolic disorder caused by a buildup of glucose in vessels and organs
• 0	Absent
• 1	Present
Hypertension	Chronic high blood pressure increasing the risk of stroke and cardiac events
• 0	Absent
• 1	Present
Immunosuppression	Patients with low immunity, e.g. cancer, transplantation and autoimmunity
• 0	Absent
• 1	Present
Lung Disease	Chronic respiratory disorder, e.g. COPD, asthma, and pulmonary fibrosis
• 0	Absent
• 1	Present
Smoker	People addicted to smoking
• 0	Absent
• 1	Present
Symptomatic Covid-19	Whether or not a patient is expressing symptoms
• 0	Absent

• 1	Present
Baseline CAT lobar condensation	Increased density of a lobe in the lung due to fluid, inflammation, or infection
• 0	Absent
• 1	Present
O ₂ needs	The volume of additional oxygen needed to keep the blood well saturated
• 0	No O ₂
• 1	O ₂ < 4 L/min
• 2	O ₂ 4-8 L/min
• 3	High Flow

1 *Table 1. Parameters used for patient clustering. The values used in the algorithm for*
2 *categorical parameters are illustrated.*

3 2.3. Cluster Analysis

4 The study used the 26 variables described in Table 1 and three clustering methods (K-
5 prototypes , KAMILA [12] and LCM [13]) to cluster COVID-19 patients. The appropriate
6 number of clusters was determined using silhouette scores [14] and Harrel's concordance
7 index [15]. The clinical relevance of the results was assessed on the basis of differences in
8 survival and hospital stay between clusters, and the identification of signature variables likely
9 to differentiate clusters. Further information on the clustering process can be found in
10 Supplementary file 1.

11

1 2.4. Statistical analysis

2 Kaplan-Meier risk curves, logistic and Cox regressions were applied to 4 subclusters of
3 patients, each defined by whether they were admitted to the ICU and whether they were
4 superinfected or not (procalcitonin ≥ 0.5). Odd ratio (OR) and Hazard Ratio (HR) with 95%
5 Confidence Interval (CI) were used to assess treatment effects. Mann-Whitney U or Kruskal-
6 Wallis tests were used for continuous variables, followed by Dunn's post-test if the latter test
7 was used [16]. The Chi-2 test (and Goodness-of-Fit test) and Fisher's exact test were used to
8 compare categorical variables, followed by a post hoc test using Bonferroni correction [17].

1 3. Results

2 3.1. Comparison of characteristics

3 After performing a systematic and comparative analysis of the three algorithms considered,
4 namely K-prototypes, KAMILA and LCM, KAMILA proved to be the best clustering
5 algorithm because it had the highest silhouette score and C-index, and produced the highest
6 number of signature features, indicating superior clustering quality and differentiation ability
7 compared to the other methods. A comprehensive, meticulously-documented, step-by-step
8 analysis is made available for reference in Supplementary File 1.

9 Table 2 summarizes the demographic data and days since symptom onset of our sample,
10 while Table 3 summarizes that of each cluster. Cluster 1 was the largest (239 patients) in
11 contrast to the 2nd and 3rd clusters, both having almost the same number of patients (153 and
12 156, $p = 0.865$ for pairs 2-3). Roughly, Cluster 1 had the youngest patients with 5 out of 6
13 patients being ≤ 68 years, Cluster 2 had the oldest patients with the same ratio being ≥ 69
14 years, and Cluster 3 had the older half of Cluster 1 and the younger half of Cluster 2. Cluster
15 1 had the lowest weighted patients with Cluster 3 consisting mostly of the heavier. Cluster 1
16 had an almost equal distribution of genders, unlike Cluster 2 and 3 having predominantly
17 men. Cluster 2 patients presented the earliest to the hospital, whereas for Cluster 3 the latest.

18
19
20
21
22
23
24

Patients' Characteristics

Age (years) \pm SD	62.82 \pm 16.61
----------------------	-------------------

Symptom onset (days) ± SD		6.53 ± 5.49
Sex n (%)	Male	373 (67.09)
	Female	183 (32.91)
Outcomes n (%)	Dead	101 (18.17)
	Alive	455 (81.83)

1 *Table 2. Demography and days since symptom onset of 556 COVID-19 patients.*

	Cluster 1	Cluster 2	Cluster 3	P
<i>Age</i>	57 (46 - 68)	76 (69 - 82.5)	63 (54 - 71.75)	< 0.01
<i>Weight</i>	76 (65.75 - 85)	81 (70 - 94)	85 (75 - 95)	< 0.01
<i>Sex: Male</i>	139 (57.2%)	111 (71.61%)	123 (77.85%)	< 0.001
<i>PCR CT value</i>	22.5 (19 - 28)	21 (18 - 25)	22 (18.75 - 26)	0.163
<i>Symptoms (days ago)</i>	-6 (-9 - -3)	-5 (-7 - -2)	-9 (-12 - -7)	< 0.001
<i>PA diameter (1st CAT)</i>	25 (24 - 27)	28 (26 - 31)	27 (24 - 29)	< 0.001
<i>GGO (1st CAT)</i>	15 (5 - 23.5)	20 (10 - 30)	40 (30 - 50)	< 0.001

2 *Table 3. Anthropometric and demographic measurements of three COVID-19 clusters*
3 *identified using KAMILA cluster analysis. Grey cells in a row: 1 = all pairs significantly*
4 *different. Green = lowest, Orange = highest. Abbreviations: PA pulmonary artery, CAT*
5 *computerized axial tomography, GGO: ground-glass opacities*

1 Cluster 1 has the fewest hypertensive and diabetic patients, but percentages comparable to
 2 Cluster 3 for other risk factors, while Cluster 2 has the highest rates for all risk factors.
 3 Roughly, half of Cluster 1 did not require any oxygen, unlike half of Cluster 3 needing higher
 4 flow of oxygen. Cluster 1 had the lowest laboratory values including IL6, except for
 5 lymphocyte count, LDH and Ferritin being comparable to Cluster 2. While the latter
 6 exhibited the highest procalcitonin and creatinine levels, Cluster 3 had the highest leucocyte
 7 (viz. neutrophile) count, CRP, LDH and Ferritin but lowest Lymphocyte count. Cluster 3 had
 8 the most ground-glass opacities on serial CT scans while Cluster 2 had the largest pulmonary
 9 artery diameter. Lastly, Cluster 1 has the lowest rates of ICU admissions and intubations,
 10 Cluster 2 the highest rates of hemorrhagic events and Cluster 3 the highest rates of
 11 thromboembolic events. Relevant patient risk factors are shown in Table 4, lab results in
 12 Table 5 and hospital outcomes in Table 6.

Risk Factors	Cluster 1	Cluster 2	Cluster 3	P
<i>Hypertension</i>	78 (32.1%)	147 (94.84%)	79 (50%)	< 0.001
<i>Diabetes</i>	28 (11.52%)	86 (55.48%)	43 (27.22%)	< 0.001
<i>Coronary Artery Disease</i>	24 (9.88%)	91 (58.71%)	26 (16.46%)	< 0.001
<i>Pulmonary Diseases</i>	32 (13.79%)	49 (32.03%)	23 (14.94%)	< 0.001
<i>Heart Failure</i>	1 (4.55%)	13 (30.95%)†	4 (7.14%)	0.002
<i>Chronic Kidney Disease</i>	12 (4.94%)	66 (42.58%)	5 (3.16%)	< 0.001
<i>Smoking</i>	34 (14.66%)	56 (36.6%)	38 (24.68%)	< 0.001
<i>Immunocompromised</i>	22 (9.05%)	18 (11.61%)	10 (6.33%)	0.272

13
 14 *Table 4. Relevant risk factors in the three COVID-19 clusters obtained using KAMILA. Grey*
 15 *cells in a row: 1 = all pairs significantly different, 2 = those aren't significantly different. †*
 16 *= slightly different. Green = lowest, Orange = highest.*

1

	Cluster 1	Cluster 2	Cluster 3	P
<i>Leukocytes</i>	5850 (4525 - 7700)	7300 (5300 - 9200)	9700 (7000 - 12550)	< 0.001
<i>Serum creatinine</i>	66 (53 - 80)	121 (88 - 201)	74 (62 - 93)	< 0.001
<i>Procalcitonin</i>	0.11 (0.06 - 0.22)	0.32 (0.14 - 0.94)	0.195 (0.1 - 0.4)	< 0.001
<i>Neutrophils</i>	4170 (3060 - 5940)	5380 (3875 - 7505)	8105 (5678 - 11010)	< 0.001
<i>Lymphocytes</i>	1015 (630 - 1440)	830 (580 - 1265)	690 (483 - 1025)	< 0.001
<i>LDH</i>	275.5 (219 - 349)	291.5 (235 - 366)	488.5 (394 - 649)	< 0.001
<i>Ferritin</i>	611 (277 - 1078)	579 (286 - 1044)	1196 (674 - 2152)	< 0.001
<i>D-Dimer</i>	0.58 (0.36 - 1.23)	1.23 (0.63 - 2.77)	0.995 (0.59 - 1.71)	< 0.001
<i>CRP</i>	49 (19 - 93)	94.25 (46 - 169)	136.5 (79 - 189.5)	< 0.001
<i>IL6</i>	24.6 (13.3 - 94)	41 (29.7 - 97.5)	47.35 (19.25 - 117.3)	0.071

1 *Table 5. Laboratory findings in the three distinct COVID-19 clusters. Number of grey cells in*
 2 *a row: 1 = all pair-wise comparisons are significant, 2 = this comparison isn't significant. †*

	Cluster 1	Cluster 2	Cluster 3	P	
ICU Admission	23 (9.47%)	43 (27.74%)	57 (36.08%)	< 0.001	
Intubation	15 (6.17%)	26 (16.77%)	37 (23.42%)	< 0.001	
Thrombo- embolic Events	9 (3.72%)	2 (1.29%)	16 (10.13%)	< 0.001	
Hemorrhagic Events	11 (4.55%)	18 (11.61%)†	11 (6.96%)	0.036	
PA diameter (2nd CAT)	26 (23 – 28)	29 (26 – 31)	27 (25 – 30)	< 0.001	
GGO (2nd CAT)	27.5 (10 – 45)	32.5 (20 – 60)	50 (30 – 75)	< 0.001	
Oxygen Requirement	No need	124 (51.03%)	36 (15.66%)	3 (1.92%)	< 0.001
	<4 L	80 (32.92%)	46 (29.68%)	19 (12.1%)	
	4 - 8 L	30 (12.35%)	60 (38.71%)	61 (38.85%)	
	High Flow	9 (3.7%)	13 (8.39%)	74 (47.13%)	

3 = slightly different. Green = lowest, Orange = highest.

1 *Table 6. Outcomes during hospitalization of the three distinct COVID-19 clusters. Number of*
2 *grey cells in a row: 1 = all pair-wise comparisons are significant, 2 = this comparison isn't*
3 *significant. † = slightly different. Green = lowest, Orange = highest. Abbreviations: PA*
4 *pulmonary artery, CAT computerized axial tomography, GGO: ground-glass opacities.*

5 3.2. Survival analysis

6 Survival analysis is a statistical method used to study the time until an event of interest
7 occurs, like patient deaths. In the classical approach, it examines the probability of an event
8 occurring: an $HR > 1$ indicates an increased probability of the event occurring and $HR < 1$
9 indicates a decreased probability. However, a modified analysis will be featured here, where
10 deaths are considered discharges, hence $HR > 1$ will indicate a positive outcome, i.e. faster
11 discharge.

12 The analysis suggests that cluster 2 had the highest risk of all-cause death, while Clusters 1
13 and 3 were not different in terms of mortality. Moreover, clusters 2 and 3 have significantly
14 higher risks of prolonged stay resulting in faster Cluster 1 patient discharges. The results of
15 the two analyses are summarized in Table 7.

16 Based on all the previous results, the clusters will be hereafter labeled according to the
17 population they describe. Namely, Cluster 1 will be dubbed "Resilient Recoverees", Cluster 2
18 "Vulnerable Veterans," and Cluster 3 "Paradoxical Patients".

19

20 3.3. Regression analysis

21 The association of various treatments with multiple outcomes was evaluated within each
22 cluster to minimize complications and unnecessary interventions. Detailed treatment results,
23 including HR and OR, are available in Supplementary file 2. Subclusters were constructed
24 based on whether PCT at admission ≥ 0.5 and admission location (ICU or regular wards, as

1 shown in Table 7). Mean and SD for treatment initiation, duration, and doses is presented in
2 Table A (Supplementary file 2).

3

4

5

Cluster	N	Mortality	Survival Time	HR Death	Length of Stay	HR Discharge†
1	239	7.11%	49	Reference	6	Reference
2	153	34%	41	2.68 (1.54-4.66)	11	0.51 (0.40-0.65)
3	156	13.4%	41	1.26 (0.70-2.29)	11	0.53 (0.42-0.66)

6 *Table 7. Overview of key outcomes for the three COVID-19 clusters. Durations are*
7 *represented as medians in days. † HR of discharge < 1 is counter intuitively a bad outcome.*

8

9 ***Non-superinfected non-ICU resilient recoverees.*** The use of carbapenem treatment was
10 associated with longer hospital stays. In contrast, Tocilizumab doses of ~750 mg or more
11 were associated with faster patient discharge.

12 ***Superinfected non-ICU resilient recoverees.*** Aminoglycosides, glucocorticoid treatment
13 equal to or greater than ~6 weeks, and azithromycin at ~1.5 weeks from symptom onset or
14 later were correlated with prolonged hospitalization periods and provided no benefit.

15 ***Superinfected vulnerable veterans.*** Non-ICU patients required oxygen therapy ranging from
16 4 to 8 liters which extended their hospital stays. The odds of reaching the composite outcome
17 (ICU admission, intubation, then death) if a patient duration of glucocorticoid therapy

1 averaged ~3 weeks and they did not receive aminoglycoside treatment averaged 0.220 (0.054
2 – 0.619).

3 ***Non-superinfected ICU vulnerable veterans.*** Deferred ICU admissions were associated with
4 a delay of ~1.5 weeks in starting Tocilizumab and/or glucocorticoid therapy from onset of
5 symptoms. This also applied to Hydroxychloroquine treatment if it extended beyond 3 weeks.
6 This suggests that all 3 therapies were started after ICU admission, not preventively before.
7 Conversely, early ICU admissions were associated with the administration of doses of
8 Aspirin \geq 324mg. Regardless administration date, the mentioned therapies did not benefit the
9 patient.

10 ***Non-superinfected non-ICU vulnerable veterans.*** Cephalosporins failed to impact the
11 composite outcome, proving its uselessness. Antibiotic therapy beyond ~2 weeks benefitted
12 patients suggested later superinfection. Glucocorticoid use for ~6 weeks or more offered no
13 advantage other than prolonging stays. Similarly, glycopeptides extended hospital stays and
14 were associated with increased bleeding risk, suggesting their use as poor prognostic
15 indicator.

16 ***Non-superinfected non-ICU paradoxical patients.*** An increase in mortality was significantly
17 linked to the use of prophylactic doses of antiplatelets in comparison to alternative dosage
18 regimens of the same drug. The administration of doxycycline and the implementation of
19 prone positioning were associated with expedited discharges, as opposed to those subjected to
20 glucocorticoid treatment for a duration of ~6 weeks or more and prednisone exceeding 25
21 mg.

22 ***Non-superinfected ICU paradoxical patients.*** Glucocorticoid therapies that lasted beyond ~6
23 weeks also delayed ICU admission. Delayed ICU admission was associated with very early
24 and multiple courses of Tocilizumab or Baricitinib treatment lasting beyond ~2 weeks, but
25 there was no advantage in starting Tocilizumab thereafter. Subsequent superinfections and

1 abstaining from using Tocilizumab were the main promoters of ICU admission. As for
2 antibiotics, expecting to administer glycopeptides or cotrimoxazole at ~2 weeks from
3 symptom onset if a bacterium was elucidated could delay ICU admissions. During ICU stay,
4 neither carbapenems nor azithromycin had a positive impact on patient survival. The use of
5 Remdesivir did not show any benefit.

6

7 **4. Discussion**

8 Over the past three years, SARS-COV-2 has spread globally manifesting itself in different
9 clinical presentations ranging from a fleeting flu to critical illness [18]. Few studies have
10 linked pathophysiologic findings to identify patient patterns for personalized treatments
11 [19,20]. Cluster analysis is a promising approach to categorizing patients, and our study
12 classified patients based on medical history, biochemistry, and radiology. KAMILA
13 algorithm produced the best results [12], identifying three distinct patient clusters with unique
14 characteristics. This approach challenges the most updated treatment guidelines that rely on
15 univariate patient data that has not been interpreted using a multivariate approach, such as
16 clustering. Specifically, it challenges the COVID-19 clinical spectrum outlined in the
17 frequently updated NIH treatment guidelines [18]. As an overview, adults were considered to
18 have asymptomatic or pre-symptomatic infection; mild, symptomatic illness without stigmata
19 of pneumonia; moderate illness, with signs of lower respiratory tract disease but SpO₂ ≥94%
20 on room air; severe illness, including patients with SpO₂ <94% on room air who are not in
21 shock or respiratory and organ failure; and critical illness with system failure. This
22 classification is primarily based on oxygen saturation, a measurement that can be
23 inconclusive, particularly in people aged ≥50 years with high-risk comorbidities and severe
24 outcomes. NIH recommendations on oximetry interpretation favor consideration of the
25 patient's overall clinical presentation and history. That said, it is essential to be proactive and

1 detect and therefore treat those having risky histories so as to avoid admissions to the IC,
2 invasive mechanical ventilation, and death. CDC researchers reviewed the risk factors that
3 favor COVID-19 progression into severe statuses [21], which have been taken into account in
4 the therapeutic management of hospitalized patients as per the NIH recommendations [22].
5 Our study aims to refine these predictive criteria through clustering.

6 To illustrate, the resilient recoverees, the largest cluster having the fastest recovery and
7 lowest mortality, can be considered a moderate-to-severe COVID-19 group. They include
8 middle aged fit patients, with markedly few cardiovascular risk factors, e.g. hypertension.
9 They can benefit from minimal (e.g. < 8 L O₂) to no therapeutic approaches considering the
10 low mortality burden. Laboratory results were roughly normal with no markers of severe
11 COVID-19 [23]. It's worth noting that IL-6 levels in this cluster were the lowest, but the lack
12 of significant difference from other clusters is unreliable due to the infrequency of IL-6
13 measurements. The use of Tocilizumab may accelerate the discharge of non-ICU patients
14 with no bacterial superinfections, but its cost-effectiveness and priority in this cluster should
15 be minded.

16 As for the vulnerable veterans, they mostly include elderly men who have multiple risk
17 factors and multiple comorbidities. They are at higher risk of severe to critical COVID-19
18 with more hemorrhagic events, superinfections and a higher mortality rate, most likely
19 favored by their comorbidities. In addition, these patients had the largest pulmonary artery
20 and the highest serum creatinine and procalcitonin suggesting the presence of pulmonary
21 hypertension and the prevalence of superinfections. That said, superinfected ICU patients
22 appeared to benefit from a ~ 3 week course of glucocorticoid therapy, which aligns with the
23 NIH recommendations [18].

24 The paradoxical patient cluster, predominantly middle to old age men with nearly twice
25 commoner hypertensive and diabetic patients than resilient recoverees, presenting 1 to 2

1 weeks late with more thromboembolic events. Despite severe to critical COVID-19
2 classification by NIH guidelines, their mortality rates mirror resilient recoverees.
3 Paradoxically, if one were to consider merely their history, they would risk misclassifying the
4 cluster as resilient. This cluster experiences prolonged stays akin to vulnerable veterans,
5 similarly demanding ICU admissions and intubations, but requiring noticeably greater high-
6 flow O2 supplementation. Glass opacities on CT scans predominate and are associated with
7 elevated COVID-19 severity biomarkers [23], notably lymphopenia but also CRP, ferritin
8 and LDH alluding to a “cytokine storm”. This excessive, uncontrolled response contribute to
9 tissue damage, organ failure, and heightened mortality [24]. Ancillary should aid in
10 distinguishing resilient from paradoxical patients, emphasizing the impact of the virus itself.
11 For non-superinfected regular wards patients, recent literature echoed our conclusions on the
12 effectiveness of prone positioning [18] and doxycycline therapy [25], particularly in
13 hastening discharge. More, it agrees on the uselessness of anti-platelet therapy in this group
14 [18]. As for non-infected ICU patients, Tocilizumab or Baricitinib early on and for long
15 periods delay ICU admissions, as well as glucocorticoid therapy for more than 6 weeks.
16 Two studies have also attempted to cluster COVID-19 patients into groups. The first study by
17 Han *et al.* used factor analysis for mixed data (FAMD) [26]. What differentiates this study
18 from ours is the addition of patient-experienced symptoms. Their results showed that the
19 patients could be divided into three distinct clusters: Cluster A, the most severe with the
20 longest hospital stays; Cluster B, of intermediate severity COVID-19 with a length of stay as
21 long as Cluster A; and Cluster C, the mildest with the shortest length of stay. Their analysis
22 showed that cluster A had the worst survival rate, whereas cluster B had higher CRP, D-
23 dimer, AST, and LDH levels, indicating a quintessential COVID-19 phenotype. Clusters A
24 and B are thus comparable to our vulnerable and paradoxical patients, respectively. Cluster C
25 had mainly systemic and digestive symptoms and a low frequency of typical symptoms of

1 fever and cough; because of its low severity, it mostly resembles the resilient recoverees. Our
2 study, in comparison, included imaging studies. We were also able to find significant
3 correlation to age. Be that as it may, old age was proven to be associated with adverse
4 outcomes for patients with COVID-19 [27].

5 Arévalo-Lorido *et al.*, the second study similar to ours, analyzed datasets by applying the
6 Random Forest model and the Gaussian mixed model by clustering [28]. The algorithms
7 generated six clusters, the last three of which had high mortality rates from any cause or
8 ended up in intensive care, whereas the first three included patients who did not. The most
9 important comorbidities were heart failure, atrial fibrillation, vascular disease, and
10 neurodegenerative disease, which were mainly present in the last three clusters. The fifth
11 cluster, with the poorest prognosis, included those with liver, kidney, and gastrointestinal
12 diseases, as well as chronic obstructive pulmonary disease. From what has been described,
13 the first three clusters converge on the resilient cluster, and the last three clusters converge on
14 the vulnerable and paradoxical clusters, with cluster 5 in the study being the closest to the
15 paradoxical patients. Contrasted to our study, it did not include data on imaging.
16 Furthermore, KAMILA concisely separated patients into a small number of meaningful
17 clusters, unlike Random Forest and Gaussian mixed model, which resulted in multiple
18 clusters which seem unfathomable.

19 Our study has significant strengths. Specifically, we recognize the effectiveness of model-
20 based algorithms in clustering mixed data, providing a rationale for this choice. Additionally,
21 we incorporated imaging findings and pinpointed vulnerable age groups, all within an
22 optimal small number of clusters. Finally, the article extends its analysis by investigating the
23 impact of various treatments on four subtypes of patients within each cluster.

24 We acknowledge our study has some limitations. Firstly, the number of patients decreased
25 significantly due to multiple stratifications, so further analysis with larger populations is

1 recommended. Secondly, patients' symptoms were not taken into account during data
2 collection, which could have made the classification more clinically friendly. Thirdly, the
3 PCT threshold used to consider patients as infected could have been improved by doing serial
4 measures not only for PCT but also other markers in conjunction (e.g. CRP and imaging).
5 Finally, the data collection was performed before vaccination campaigns and when one
6 COVID-19 variant dominated the cases, so it would be interesting to study the effects of
7 vaccination on the classes and the effect of different variants on patients to find a common
8 classification for all COVID-19 strains.

9

10 **Acknowledgments**

11 We want to acknowledge the support of our institution and the interests of the contributors
12 who have invested their time and expertise in this project. Moreover, this research has not
13 received any external financial assistance or grants.

14

15 **Ethical Approval**

16 The present study was conducted in accordance with the principles outlined in the
17 Declaration of Helsinki. Ethical oversight for this study, specifically on the sole use of
18 anonymized patient data, was obtained from the Institutional Review Board at Saint Joseph
19 University affiliated hospital, Hotel-Dieu de France, in Beirut, Lebanon.

20

21 **References**

22 1. Khoury P, Azar E, Hitti E. COVID-19 Response in Lebanon: Current Experience and
23 Challenges in a Low-Resource Setting. *JAMA*. 2020; 324(6):548.

- 1 2. Baj J, Karakuła-Juchnowicz H, Teresiński G, et al. COVID-19: Specific and Non-Specific
2 Clinical Manifestations and Symptoms: The Current State of Knowledge. *J Clin Med.*
3 **2020**; 9(6):1753.
- 4 3. Ma Q, Liu J, Liu Q, et al. Global Percentage of Asymptomatic SARS-CoV-2 Infections
5 Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis:
6 A Systematic Review and Meta-analysis. *JAMA Netw Open.* **2021**; 4(12):e2137257.
- 7 4. Larose DT, Larose CD. *Clustering. Data Min Predict Anal.* Second edition. Hoboken, New
8 Jersey: John Wiley & Sons Inc; 2015. p. 512.
- 9 5. Islam M, Hasan M, Wang X, Germack H, Noor-E-Alam M. A Systematic Review on
10 Healthcare Analytics: Application and Theoretical Perspective of Data Mining.
11 *Healthcare.* **2018**; 6(2):54.
- 12 6. Pina A, Macedo MP, Henriques R. Clustering Clinical Data in R. In: Matthiesen R, editor.
13 *Mass Spectrom Data Anal Proteomics [Internet].* New York, NY: Springer New York;
14 2020 [cited 2023 Feb 4]. p. 309–343. Available from:
15 http://link.springer.com/10.1007/978-1-4939-9744-2_14
- 16 7. El Hadi C, Ayoub G, Bachir Y, Haykal M, Jalkh N, Kourie HR. Polygenic and Network-
17 based studies in risk identification and demystification of cancer. *Expert Rev Mol*
18 *Diagn.* **2022**; 22(4):427–438.
- 19 8. R: The R Project for Statistical Computing [Internet]. [cited 2021 Sep 17]. Available from:
20 <https://www.r-project.org/>
- 21 9. Peterson R A. Finding Optimal Normalizing Transformations via bestNormalize. *R J.*
22 **2021**; 13(1):310.

- 1 10. Peterson RA, Cavanaugh JE. Ordered quantile normalization: a semiparametric
2 transformation built for the cross-validation era. *J Appl Stat.* **2020**; 47(13–15):2312–
3 2327.
- 4 11. Buuren S van, Groothuis-Oudshoorn K. **mice**: Multivariate Imputation by Chained
5 Equations in R. *J Stat Softw* [Internet]. **2011** [cited 2023 Feb 3]; 45(3). Available from:
6 <http://www.jstatsoft.org/v45/i03/>
- 7 12. Foss AH, Markatou M. kamila: Clustering Mixed-Type Data in R and Hadoop. *J Stat*
8 *Softw* [Internet]. **2018** [cited 2021 Oct 8]; 83(13). Available from:
9 <http://www.jstatsoft.org/v83/i13/>
- 10 13. Marbac M, Sedki M. VarSelLCM: an R/C++ package for variable selection in model-
11 based clustering of mixed-data with missing values. Wren J, editor. *Bioinformatics.*
12 **2019**; 35(7):1255–1257.
- 13 14. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster
14 analysis. *J Comput Appl Math.* **1987**; 20:53–65.
- 15 15. Harrell FE. Evaluating the Yield of Medical Tests. *JAMA J Am Med Assoc.* **1982**;
16 247(18):2543.
- 17 16. Dunn OJ. Multiple Comparisons among Means. *J Am Stat Assoc.* **1961**; 56(293):52–64.
- 18 17. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità / Carlo E.
19 Bonferroni. *Teor. Stat. Delle Classi E Calcolo Delle Probab.* Firenze: Seeber; 1936.
- 20 18. COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19)
21 Treatment Guidelines. [Internet]. National Institutes of Health; Available from:
22 <https://www.covid19treatmentguidelines.nih.gov/>

- 1 19. Öztürk Ş, Özkaya U, Barstuğan M. Classification of Coronavirus (COVID □19) from
2 X□RAY and CT images using shrunken features. *Int J Imaging Syst Technol.* **2021**;
3 31(1):5–15.
- 4 20. Liao D, Zhou F, Luo L, et al. Haematological characteristics and risk factors in the
5 classification and prognosis evaluation of COVID-19: a retrospective cohort study.
6 *Lancet Haematol.* **2020**; 7(9):e671–e678.
- 7 21. CDC. Healthcare Workers [Internet]. *Cent. Dis. Control Prev.* 2020 [cited 2024 Feb 21].
8 Available from: [https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-](https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html)
9 [care/underlyingconditions.html](https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html)
- 10 22. Kompaniyets L, Pennington AF, Goodman AB, et al. Underlying Medical Conditions
11 and Severe Illness Among 540,667 Adults Hospitalized With COVID-19, March 2020–
12 March 2021. *Prev Chronic Dis.* **2021**; 18:210123.
- 13 23. Qin R, He L, Yang Z, et al. Identification of Parameters Representative of Immune
14 Dysfunction in Patients with Severe and Fatal COVID-19 Infection: a Systematic
15 Review and Meta-analysis. *Clin Rev Allergy Immunol.* **2022**; 64(1):33–65.
- 16 24. Jones SA, Hunter CA. Is IL-6 a key cytokine target for therapy in COVID-19? *Nat Rev*
17 *Immunol.* **2021**; 21(6):337–339.
- 18 25. Dhar R, Kirkpatrick J, Gilbert L, et al. Doxycycline for the prevention of progression of
19 COVID-19 to severe disease requiring intensive care unit (ICU) admission: A
20 randomized, controlled, open-label, parallel group trial (DOXPREVENT.ICU). Plavec
21 D, editor. *PLOS ONE.* **2023**; 18(1):e0280745.

- 1 26. Han L, Shen P, Yan J, et al. Exploring the Clinical Characteristics of COVID-19
2 Clusters Identified Using Factor Analysis of Mixed Data-Based Cluster Analysis. *Front*
3 *Med.* **2021**; 8:644724.
- 4 27. Booth A, Reed AB, Ponzo S, et al. Population risk factors for severe disease and
5 mortality in COVID-19: A global systematic review and meta-analysis. Madeddu G,
6 editor. *PLOS ONE.* **2021**; 16(3):e0247461.
- 7 28. Arévalo-Lorido JC, Carretero-Gómez J, Casas-Rojo JM, et al. The importance of
8 association of comorbidities on COVID-19 outcomes: a machine learning approach.
9 *Curr Med Res Opin.* **2022**; 38(4):501–510.

10

11

1 Abbreviations

2 ASMD: Absolute standardized mean	15 ICU: Intensive care unit
3 difference	16 IL-6: Interleukin-6
4 ARDS: Acute respiratory distress	17 KAMILA: K-means of Mixed Large data
5 syndrome	18 k: Number of clusters
6 BIC: Bayesian information criterion	19 LCM: Latent Class Models
7 CI: Confidence interval	20 LDH: Lactate dehydrogenase
8 COVID-19: Coronavirus disease 2019	21 OR: Odds ratio
9 CRN: Creatinine	22 PCT: Procalcitonin
10 CRP: C-reactive protein	23 PMM: Predictive mean matching
11 CT: Computed tomography	24 SARS-CoV-2: severe acute respiratory
12 CT-PCR: Cyclic value - polymerase chain	25 syndrome coronavirus 2
13 reaction	26 VIF: Variance inflation factor
14 HR: Hazard ratio	27 WHO: World Health Organization

28