Heterogeneity in COVID-19 severity patterns among age-gender groups: an analysis of 778 692 Mexican patients through a meta-clustering technique

Lexin Zhou[a], Nekane Romero[a], Juan Martínez-Miranda[c], J Alberto Conejero[b†], Juan M García-Gómez[a†], Carlos Sáez[a†] [carsaesi@upv.es]

[a]Biomedical Data Science Lab, Instituto Universitario de Tecnologías de la Información y Comunicaciones (ITACA), Universitat Politècnica de València (UPV), Camino de Vera s/n, Valencia 46022, España. [b]Instituto Universitario de Matemática Pura y Aplicada (IUMPA), Universitat Politécnica de València, Valencia, Spain. [c]CONACyT - Centro de Investigación Científica y de Educación Superior de Ensenada - CICESE-UT3, Mexico

†Senior authors

Abstract

Objective

To describe COVID-19 subphenotypes regarding severity patterns including prognostic, ICU and morbimortality outcomes, through stratification based on gender and age groups, as described by inter-patient variability patterns in clinical phenotypes and demographic features.

Materials and methods

We used the COVID-19 open data from the Mexican Government including patient-level epidemiological and clinical data from 778 692 SARS-CoV-2 patients from January 13, 2020 to September 30, 2020.

Inter-patient variability was analyzed by combining dimensionality reduction and hierarchical clustering methods. We produced cluster analyses for all combinations of gender and age groups (<18, 18-49, 50-64, and >64). For each group, the optimum number of clusters was selected combining a quantitative approach using the Silhouette coefficient, and a qualitative approach through a subgroup expert inspection via visual analytics. Using the features of the resultant age-gender clusters, we performed a meta-clustering analysis to provide an overall description of the population.

Results

We observed a total of 56 age-gender clusters, grouped in 11 clinically distinguishable meta-clusters with different outcomes. Meta-clusters 1 to 3 showed the highest recovery rates (90.27-95.22%). These clusters include: (1) healthy patients of all ages, (2) children with comorbidities who had priority in medical resources, (3) young patients with obesity and smoking habit. Meta-clusters 4 and 5 showed moderate recovery rates (81.3-82.81%): (4) patients with hypertension or diabetes of all ages, (5) typical obese patients with three highly correlated conditions, namely, pneumonia, hypertension and diabetes. Meta-clusters 6 to 11 had very low recovery rates (53.96-66.94%) which include: (6) immunosuppressed patients with the highest comorbidity rate in many diseases, (7) CKD patients with the worse survival length and recovery, (8) elderly smoker with mild COPD, (9) severe diabetic elderly with hypertension, (10, 11) oldest obese smokers with severe COPD and mild cardiovascular disease with the latter (11) showing a relatively higher age and smoke rate, severe COPD and shorter survival length, reinforcing a high correlation between smoking habit and COPD among elderly. Additionally, the source Mexican state and type of clinical institution proved to be an important factor for heterogeneity in severity.

Discussion

The proposed unsupervised learning approach successfully uncovered discriminative COVID-19 severity patterns for both genders and all age groups from clinical phenotypes and demographic features. A careful read of group outcomes showed consistent results regarding recent literature. Regarding the Mexican population, our results suggest that habits and comorbidities may play a key role in predicting mortality in older patients. Centenarians tended to fall in the groups with better outcomes repeatedly. Additionally, immunosuppression was not found as a relevant factor for severity alone but did when present along with

chronic kidney disease. Further useful correlations could be found by evaluating the duration of unhealthy habits, demographic features, comorbidities, the time since diagnosis, recovery progress, readmission record, and the effect of source variability.

Conclusion

The resultant eleven meta-clusters provide bases to comprehend the classification of patients with COVID-19 based on comorbidities, habits, demographic characteristics, geographic data and type of clinical institutions, as well as revealing the correlations between the above characteristics thereby help to anticipate the possible clinical outcomes for every specifically characterized patient. These subphenotypes can establish target groups for automated stratification or triage systems to provide personalized therapies or treatments.

Code available at: https://github.com/bdslab-upv/covid19-metaclustering

Dynamic results visualization at: http://covid19sdetool.upv.es/?tab=mexicoGov

Keywords: COVID-19, SARS-CoV-2, observational, heterogeneity, epidemiology, clustering, Mexico

1 Introduction

In Mexico, mid-January 2020 reported the first cases of COVID-19. In early March 2020, the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was declared by the World Health Organization as a pandemicsour[1]. As the outbreak goes, the number of infectious individuals increases rapidly. As of November 24, the Mexican nation already surpassed one million cases[2].

While this novel emerging virus is spreading out the world, affecting worldwide economic and restricting people's social interaction, medical groups and researchers have been making a huge effort to discover this novel virus's influence and risk factors. Several studies have suggested potential subphenotypes in COVID-19, mainly within specific comorbidities such as pulmonary diseases or diabetes[3, 4] or related to distinct genetic variants[5]. However, the Mexican population shows a particular high prevalence of comorbidities, such hypertension and diabetes, and obesity, which are directing the population to undesirable risks for severe coronavirus outcomes. For example in 2020, Type 2 diabetes (T2D) is a leading cause of death in Mexico, with a prevalence of 15.9%[6]. Given the potential heterogeneity in Mexican population and COVID-19 severity, mainly in comorbidities and morbimortality, it is crucial to have a pragmatic understanding of how severity patterns vary among these patients to anticipate individuals' prognostic outcomes.

This work shows the results of an unsupervised Machine Learning (ML) meta-clustering approach to identify potential subphenotypes of COVID-19 patients in Mexico from clinical phenotypes and demographic features. Stratification on gender and age groups was included to reduce potential confounding factors since age and gender are highly correlated with comorbidity, habits and mortality. By using a large cohort of more than 700,000 patient-level cases, this is probably the largest cluster analysis about coronavirus patient-level cases so far. Other studies proposed unsupervised ML methods to subgroup aggregated population data[7], CT image analyses[8,9], molecular-level clustering[10], or scientific texts to discover associations among coronavirus and other diseases[11]. However, to our knowledge, few studies provided to date results from unsupervised ML on patient-level epidemiological data[12,13,14]. The resultant subphenotypes can potentially establish target groups for automated stratification or triage systems to provide personalized therapies or treatments for the specific group severity patterns.

This article is organized as follows. Section 'Materials and methods' describes the dataset and the proposed meta-clustering approach. Section 'Results' first presents the general statistics of the studied sample, then describes the results of age-gender subgroups, and finally describes the meta-clusters results. Section 'Discussion' compares our results with the current literature to verify whether our pragmatic classification of the clusters is consistent compared to Mexican and global COVID-19 population, and discusses possible study limitations. Finally, 'Conclusion' summarizes and concludes this work.

2 Materials and methods

2.1 Data

We used the COVID-19 Open Data published by the Mexican Government[15]. The dataset comprises public patient-level data from Mexican nationwide cases tested during COVID-19 outbreak, including demographic, comorbidities, habits, and prognosis data for both positive and non-positive cases. The original dataset was released at 2 November 2020, including totally 2 414 882 individuals from January 13, 2020 to November 2, 2020.

We established severity labels from five derived outcome variables. First, the patient outcome classifies patients into deceased or not from the recording of a date of death. Second, the survival days, calculated from the date of death minus the date of symptoms. Third, the number of days it took from presenting symptoms to hospitalization, calculated from the admission date minus the date of symptoms. Lastly, we created two binary variables for overall survival, to categorize the patients who survived more than 15 and 30 days after presenting symptoms.

The inclusion criteria consisted of patients confirmed as positive of SARS-CoV-2. As part of an initial Data Quality assessment, we excluded invalid and pending result individuals, as follows. Cases who presented missing data or an "ignored" value in at least one of their chronic disease records, to our knowledge being Missing Completely At Random (MCAR), were excluded (n=20094, 0.83% of the total dataset). We also excluded inconsistent, or non-plausible, records for some combinations of variables like between-dates consistency, survival days, from symptom to hospital days (n=476, 0.02% of the total). Since the last update of the dataset was 2nd November, we also excluded patients who presented the symptoms after September 30, since their recovery status is yet unknown. These patients may bias our results when analysing the correlation between patients' mortality and other characteristics because 95.28% of deceased patients died within 31 days in Mexico (section 1, Appendix A).

The temporal variability[16] assessment of the statistical distributions of the data showed a variable transient state in the distributions of some variables from January to April, possibly associated with the smaller sample size at these moths. Thus, we decided to keep the data from all the period of the study. Lastly, the source variability[17] assessment by comparing Mexican states and the type of clinical institutions where patients received medical attention showed a slight variability pattern in the distributions among data sources in some variables. However, we decided to include all sources in the meta-clustering analysis, so that differences on these Mexican states and clinical institutions could be further assessed through the clustering approach. Section 2 of Appendix A provides more details on these DQ results.

The original dataset was provided in Spanish and using coded values. For this work, we coded it into the corresponding English terms. The final sample includes 778 692 positive cases (85.99% of confirmed cases). Figure 1 presents a CONSORT-like flowchart describing the dataset preprocessing. Table 1 shows the list of studied variables. Extra information on the materials is described in the section 3 of Appendix A.

Table 1. List of variables contained in the study case.

Variable	Description	Type (value/format)
Sex	Sex of the person	Discrete (Male, Female)
Age	Age in years at the time of the admission	Numerical Integer
Pregnant	Presence of pregnancy	Discrete (Yes, No)
Obesity	Presence of obesity	Discrete (Yes, No)
Smoke	Presence of smoking habit	Discrete (Yes, No)
Pneumonia	Presence of pneumonia	Discrete (Yes, No)
Diabetes	Presence of diabetes	Discrete (Yes, No)
COPD	Presence of chronic obstructive pulmonary disease	Discrete (Yes, No)
Asthma	Presence of asthma	Discrete (Yes, No)
INMUSUPR	Presence of immunosuppression	Discrete (Yes, No)
Hypertension	Presence of hypertension	Discrete (Yes, No)

CKD	Presence of chronic kidney disease	Discrete (Yes, No)
Cardiovascular	Presence of cardiovascular	Discrete (Yes, No)
Other disease	Presence of other diseases	Discrete (Yes, No)
Hospitalized	Whether a patient was hospitalized	Discrete (Yes, No)
Intubated	Whether a patient was intubated	Discrete (Yes, No)
ICU	Whether a patient had been in an intensive care unit	Discrete (Yes, No)
Other case contact	Whether a patient was detected to have contacted with other coronavirus cases	Discrete (Yes, No)
Result_lab	Coronavirus test result	Discrete (Positive SARS-CoV-2, Non-Positive SARS-CoV-2, Pending, Inadequate result, Not Applied)
Admission_date	The date when a patient was attended by the care unit (not necessarily hospitalized)	Date (dd/mm/yyyy)
Symptoms_date	The date when a patient presented symptoms	Date (dd/mm/yyyy)
Death_date	The date of death	Date (dd/mm/yyyy)
Entity_um	The state where a patient received attention from medical unit	Discrete
Sector	The type of institution of National Health System that provided medical care	Discrete[a]
Outcome[b]	Death result of the patient (we used this to calculate mortality and recovery rate)	Discrete (Deceased, Non-Deceased)
Survival days[b]	The survival period for a patient from presenting symptoms to his/her death	Numerical Integer
Survival>15days[b]	Whether a patient survived more than 15 days from presenting symptoms.	Discrete (Yes, No)
Survival>30days[b]	Whether a patient survived more than 30 days from presenting symptoms.	Discrete (Yes, No)
Survival>15days_deceased[b]	Whether a deceased patient survived more than 15 days from presenting symptoms.	Discrete (Yes, No)
Survival>30days_deceased[b]	Whether a deceased patient survived more than 30 days from presenting symptoms.	Discrete (Yes, No)
From Symptom to Hospital days[b]	The days that took for a patient from presenting symptoms to the hospitalization	Numerical Integer

[a] *IMSS, SSA, ISSSTE, PRIVATE, PEMEX, STATE, SEMAR, SEDENA, IMSS-BIENESTAR, UNIVERSITARY, MUNICIPAL, RED CROSS, DIF.*

[b] *Variables that were created by combining or transform other variables in the original dataset. See explanations in the "materials and method" section. See section 3 of Appendix A for the original dataset description.*

2.2 Methodology

Our methodology consists in applying a two-level subgroup discovery approach. In the first level, we apply multiple subgroup discoveries at stratified groups according to gender and age. In the second level, in a wider perspective, we perform subgroup discovery on the resultant clusters from the first level by aggregating their clinical phenotypes and demographic features. Figure 2 describes the research methodology.

Subgroup discovery was performed through a hierarchical clustering algorithm –using Ward's minimum variance method with Euclidean squared distance[18]– fed by a dimensionality reduction algorithm taking as input variables: obesity, smoking habit, pneumonia, diabetes, COPD, asthma, INMUSUPR, hypertension, CKD, cardiovascular, and other diseases. Dimensionality reduction is known to help in the process of clustering by compressing information into a smaller number of variables, making unsupervised learning less prone to overfitting[19], as well as to facilitate further visual analytics. For each subgroup analyses, we implemented cluster analyses from 2 through 12 clusters. The proper number of subgroups for each

analysis was obtained by combining a quantitative and qualitative cluster accuracy analysis. Quantitative cluster analysis was performed using Silhouette Coefficient[20], which measures the tightness and separation of the objects within clusters, reflecting how similar an object is to its own cluster compared to other clusters.

Qualitative cluster analysis was performed through an exploratory and statistical audit by the authors of this work, including medical, health informatics and machine learning experts from Spain and Mexico. We firstly selected the group of clusters that showed relatively better Silhouette Coefficient values, then chose the number of clusters from these which provided the most reasonable and clinically distinguishable classification regarding clinical phenotypes and demographic features. This process was supported by the COVID-19 pipelines and exploratory tool we developed in previous work[21].

Once obtained the proper number of clusters for each study group, namely for age <18, 18-49, 50-64, and >64 and by male and female, for each cluster among all groups we averaged the values of their clinical and habits features. This vector characterized the pattern of each subgroup independently of their strata, leading to a matrix of N subgroups by the number of variables. This matrix was applied a PCA analysis[22] with two aims. First, to explore the patterns between the clusters through their embedding on the PCA loadings. And second, to embed the clusters into a lower dimension to facilitate the latter meta-cluster analysis through a hierarchical clustering. Again, the appropriate number of meta-clusters (MCs) was selected by combining qualitative Silhouette Coefficient and the expert assessment.

Being aware about the variability among Mexican states and type of clinical institutions found in the initial DQ analysis, we assessed the variability among these factors within each cluster and meta-cluster results. Additionally, we made complementary statistical analysis regarding pregnancy's effect in COVID-19 patients (section 4, Appendix A).

We additionally provided descriptive analyses for the variables and assessed for differences between groups using the following statistics: odds ratio and chi-square ($\chi 2$) test for categorical variables, one sample t-test for two normally distributed numerical variables, and One-way ANOVA test when there are more than two means of samples to compare. The normality of two included variables was defined by both visual plot and Shapiro-Wilk test. A p-value < 0.05 was considered to be significant.

MCA, PCA and clustering analyses were performed using RStudio (version 3.6). Data processing and additional statistical were performed using Python (version 3.8). Temporal and source variability DQ analyses were performed using the EHRtemporalVariability[23] and EHRsourceVariability[17,24,25] packages. The methods developed in this work are available in our GitHub repository https://github.com/bdslab-upv/covid19-metaclustering.

Figure 1. Dataset preprocessing flowchart.

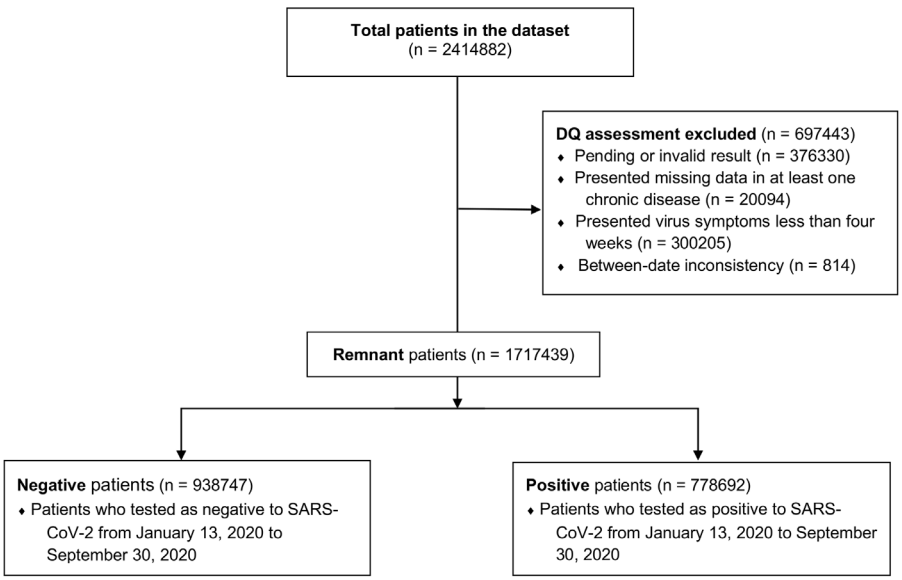
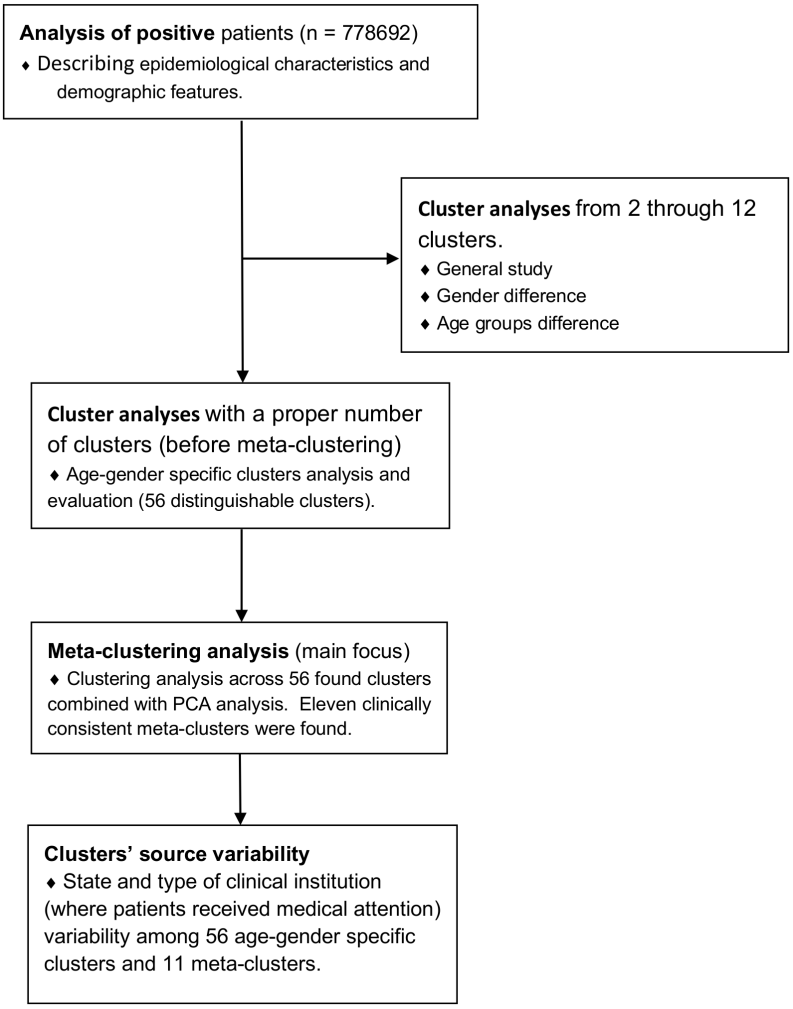


Figure 2. Research methodology flowchart.



3 Results

3.1 Dataset descriptive statistics

Table 2 shows the epidemiological characteristics and clinical outcomes of the dataset patients, and the gender difference.

From the 778 692 COVID-19 patients, 402 655 (51.7%) were male, and 376 037 (48.3%) were female with a male-to-female sex ratio of 1:0.93. The patients who aged 18-49 accounted for the largest proportion (60.4%), and the mean age was 44.5±16.7 years. The age-independent mortality rate was 10.5% (93.3% and 90% survived more than 15 and 30 days respectively); whereas among deceased patients, 36.34% and 5.38% survived more than 15 and 30 days respectively. Furthermore, the mean days for a hospitalized patient from presenting the symptoms to hospitalization was 4.8±3.7 days (23.5% were hospitalized).

There is a baseline significant severity difference between male and female. For example, 21.3% of male presented pneumonia with a mortality of 13.1%, whereas women presented 14.6% pneumonia rate and 7.9% mortality rate (Odds Ratio: 1.58 [95%CI; 1.56-1.60] and 1.76 [95%CI; 1.74-1.79] respectively, male vs female). COPD and hypertension showed no significant difference. Noteworthy, large sample sizes contribute to test positive for statistically significant differences while reducing confidence intervals. The severity difference among age-groups is described in section 5 of Appendix A.

Table 2. Epidemiological characteristics and demographic features of 778692 COVID-19 patients. P-Value was calculated for numerical variables and odds ratio for categorical variables (Jany 13 – Sep 30, 2020).

	Total (n = 778692)	Male (n = 402655)	Female (n = 376037)	P-Value and/or Odds Ratio (95%CI) (Male vs Female)
Age, mean (SD)	44.5 (16.7)	45.2 (16.8)	43.8 (16.5)	P<0.001[a]
Age range, n (%)				
<18	22868 (2.9)	11572 (2.9)	11296 (3.0)	OR=0.96 [0.93-0.98]
18-49	470349 (60.4)	236765 (58.8)	233584 (62.1)	OR=0.87 [0.86-0.88]
50-64	184452 (23.7)	97943 (24.3)	86509 (23.0)	OR=1.07 [1.06-1.08]
>64	101023 (13.0)	56375 (14.0)	44648 (11.9)	OR=1.21 [1.19-1.22]
Features, (%)				
Pregnant	5735 (0.7)	0 (0)	5735 (1.5)	
Smoke	56960 (7.3)	39355 (9.8)	17605 (4.7)	OR=2.21 [2.17-2.25]
Obesity	138929 (17.8)	68858 (17.1)	70071 (18.6)	OR=0.90 [0.89-0.91]
Comorbidities, n (%)				
Diabetes	118867 (15.3)	62495 (15.5)	56372 (15.0)	OR=1.04 [1.03-1.05]
COPD	11119 (1.4)	5677 (1.4)	5442 (1.4)	OR=0.97 [0.94-1.01]
Asthma	20057 (2.6)	7708 (1.9)	12349 (3.3)	OR=0.57 [0.56-0.59]
INMUSUPR	8311 (1.1)	3939 (1.0)	4372 (1.2)	OR=0.84 [0.80-0.88]
Hypertension	149444 (19.2)	76966 (19.1)	72478 (19.3)	OR=0.99 [0.98-1.00]
CKD	14526 (1.9)	8234 (2.0)	6292 (1.7)	OR=1.23 [1.19-1.27]
Cardiovascular	15072 (1.9)	8526 (2.1)	6546 (1.7)	OR=1.22 [1.18-1.26]
Other disease	18525 (2.4)	8232 (2.0)	10293 (2.7)	OR=0.74 [0.72-0.76]
Treatment, n (%)				
Hospitalized	182675 (23.5)	110758 (27.5)	71917 (19.1)	OR=1.60 [1.59-1.62]
Intubated	31978 (4.1)	20680 (5.1)	11298 (3.0)	OR=1.75 [1.71-1.79]
ICU	15916 (2.0)	10156 (2.5)	5760 (1.5)	OR=1.66 [1.61-1.72]
Pneumonia	140720 (18.1)	85692 (21.3)	55028 (14.6)	OR=1.58 [1.56-1.60]
Mortality, n (%)	82077 (10.5)	52553 (13.1)	29524 (7.9)	OR=1.76 [1.74-1.79]
Survival>15days, n (%)	726443 (93.3)	369553 (91.8)	356890 (94.9)	OR=0.60 [0.59-0.61]
Survival>30days, n (%)	701027 (90.0)	352913 (87.6)	348114 (92.6)	OR=0.57 [0.56- 0.58]
Survival>15days_deceased, n (%)	29828 (36.34)	19451 (37.01)	10377 (35.15)	OR=1.08 [1.05-1.12]
Survival>30days_deceased, n (%)	4412 (5.38)	2811 (5.35)	1601 (5.42)	OR=0.99 [0.93-1.05]
From Symptom to Hospital days, mean (SD)	4.8 (3.7)	4.9 (3.7)	4.7 (3.6)	P<0.001[a]
Other case contact, n (%)	338709 (43.5)	161772 (40.2)	176937 (47.1)	OR=0.76 [0.75-0.76]

[a] *One-sample T-test.*

3.2 Age-gender groups and meta-clustering

In the following, for referencing each age-gender subgroup we used the following abbreviation composition: [Age Group][Gender][Cluster ID]. For example: <18F1 means age<18 group, female and cluster ID 1 within that age and gender group results.

After evaluating the clustering results for each age-gender group, we selected the following k –number of clusters– for each specific age-gender group (table 3): <18M: k=5, <18F: k=4, 18-49M: k=7, 18-49F: k=7, 50-64M: k=9, 50-64F: k=8, >64M: k=8, >64F: k=8. This led to a total number of 56 age gender clusters. Table 3 shows the number of individuals (n) for each age-gender cluster. The details for each age-gender group can be fully explored at http://covid19sdetool.upv.es/?tab=mexicoGov.

Table 3. Size of the 56 age-gender specific clusters. The number indicates the number of patients for each age-gender cluster.

Cluster size (n)	0-1000	1001-5000	5001-10000	10001-30000	>=30000

Cluster ID (unrelated among groups)	<18M	<18F	18-49M	18-49F	50-64M	50-64F	>64M	>64F
1	9463	753	138266	149573	17716	29725	16997	14861
2	542	10228	19863	22897	3525	4017	8057	2806
3	410	81	24449	17555	41583	15293	3372	11023
4	969	234	3947	6776	12113	3478	17144	6342
5	188		22905	21131	10231	4042	1192	1886
6			22224	12186	2543	16527	4334	2469
7			5111	3466	484	2570	2866	4001
8					2884	10857	2413	1260
9					6864			

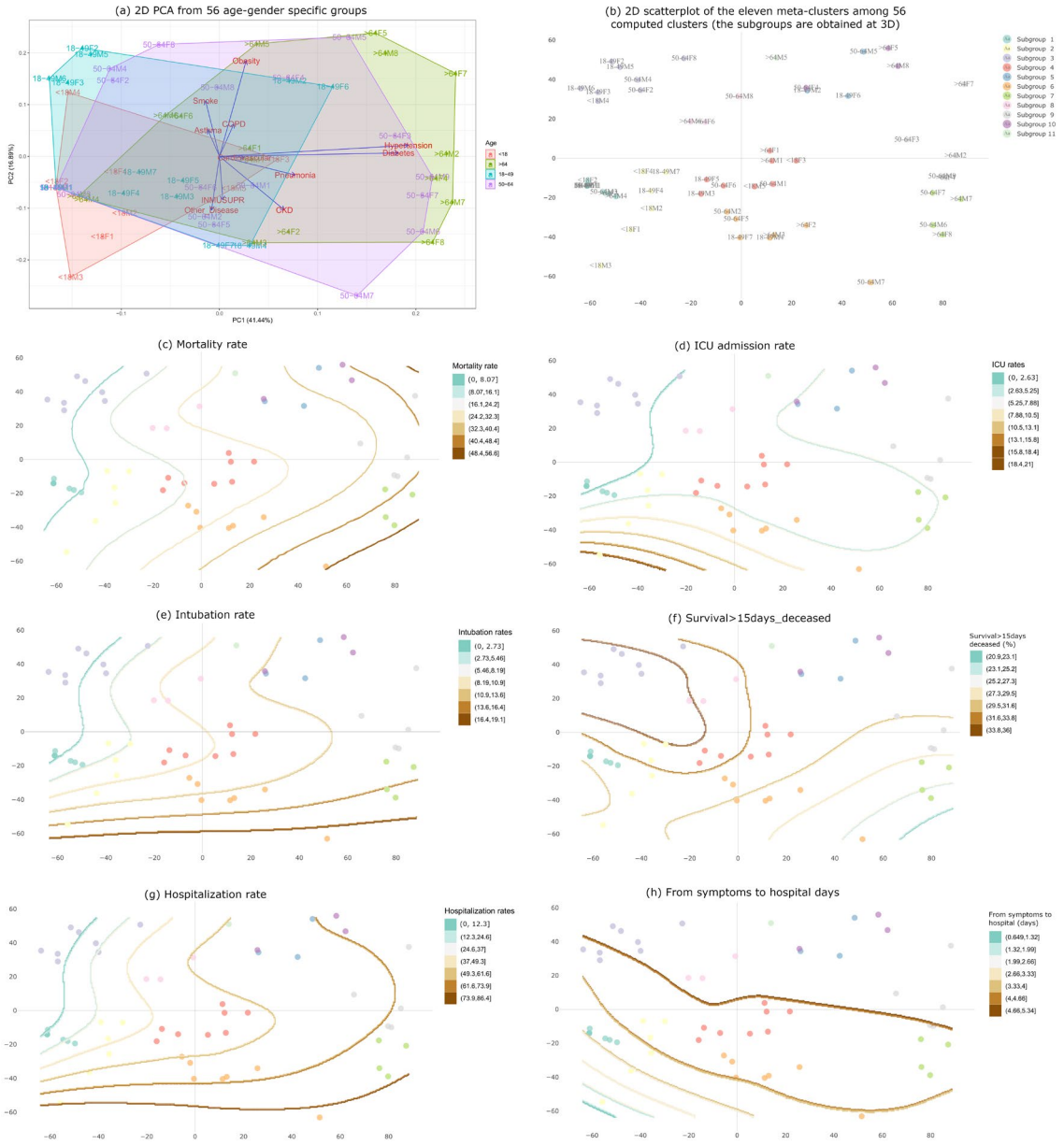
The PCA analysis across the features of the 56 clusters found remarkable patterns and heterogeneity among clusters of different ages in both genders (Figure 3a). We then leveraged the three first PCA components and applied the meta-clustering to these clusters (Figure 3b). Both figures show each cluster with its abbreviation at its corresponding coordinates. We found heterogeneity patterns among the clusters of different age groups: for example, young adults are prone to asthma and smoking habit; whereas elderly are prone to hypertension, diabetes, obesity, COPD, pneumonia, and CKD. The results also show that obesity and smoking habit –both positively correlated– are strongly separated from INMUSUPR and other diseases –both positively correlated–, implying these two pairs of features are negatively correlated in the studied data subgroups. Further characterizations of these patterns are defined next based on the meta-clustering results.

We found eleven MCs across the 56 age-gender clusters that were clinically distinguishable (Figure 3b). Additionally, we applied locally estimated scatterplot smoothing[26] (LOESS) models among the eleven MCs to delineate severity patterns for mortality, ICU, intubation, survival length, hospitalization, and from symptoms to hospital days. For example, children took fewer days from presenting symptoms to hospitalization, and have higher ICU, intubation, and hospitalization rates than adults with similar conditions (Figure 3d, e, g, h; Figure 4). MC3 –young obese cluster with moderate asthma and smoke rates– behaves inversely implicating that children, under similar clinical conditions, may receive priority regarding medical attention.

Furthermore, by inspecting simultaneously the PCA plot (Figure 3a) and LOESS models (Figure 3c-h) it helps to visualize the correlation between the studied severity and comorbidities/habits. For example, CKD decreases survival length significantly among deceased patients and increases intubation rates (Figure 3e, D). Mortality constantly increases from children to elderly, but the most severe zones are inclined toward pneumonia, CKD, and COPD (Figure 3c) independently of the age groups.

The heatmap in Figure 4 describes and quantifies the main characteristics among the 56 age-gender clusters and relates them to the MC they belong, simultaneously ordering rows and columns through a biclustering technique[27]. Figure 4 highlights other relevant patterns among the 56 age-gender clusters. E.G, children clusters, especially the youngest took significantly fewer days to receive medical attention after showing the symptoms, and are prone to ICU admission despite presenting similar clinical condition than adult clusters (e.g., cluster <18M3 compared with 50-64F5). Regarding gender discrepancy, our results reinforced that males have higher risks than females since the females clearly show a better RR despite presenting similar clinical conditions than males (e.g., >64M1 versus >64F1). Complementary, Table 4 quantifies the main characteristics of the eleven MCs, and table 5 summarizes their main features according to age group, habits, comorbidities, pneumonia, and recovery.

Figure 3. Principal component analysis (PCA) plot of the 56 age-gender clusters and Meta-clustering results: (a) 2D principal component analysis (PCA) plot from 56 COVID-19 clusters; (b) 2D scatterplot of the eleven MCs among 56 computed clusters. (c-h) 2D scatterplot regarding the severities of the eleven MCs among 56 computed clusters. The text that each cluster possesses corresponds with its severity rate. We applied LOESS model to detect general severity patterns. 7 severity ranges for each plot, each severity range corresponds with the area inside two curves of the same color scale. The found severity patterns correspond to: (c) mortality; (d) ICU; (e) intubation, (f) Survival>15days_deceased, (g) hospitalization, and (h) from symptoms to hospital days. The coordinates of these eight plots are the same.



Regarding MCs, MC1 includes generally the two healthiest clusters per each age group with very high RR (90%). It is worth noting that most deceased patients in MC1 with pneumonia were mostly older patients (see Figure 4). MC2 includes children and young individuals (mean age 18) with healthy habits and little incidence of relevant diseases (13% INMUSUPR, 17% cardiovascular disease, 4% CKD); albeit RR is very high (91%), MC2 holds the highest ICU admission rate (9%) that is caused by three children clusters whose ICU rate vary from 13.41% to 18.45%. MC3 includes young adults (mean age 40) with significant obesity, smoking and also a little incidence of other diseases; its RR is the highest (95%). The three MCs have similarly high RRs; while 1 and 3 have a little incidence of pneumonia, and therefore good recovery is attributable to mild SARS-CoV-2 disease, MC2 has at least 1/3 of patients with pneumonia. Thus, recovery

is more likely explained by a better immune response, better response to treatments or to priority in health care attention when they suffered from severe disease.

Similar to MC1, MC4 includes individuals of all ages with healthy habits but, unlike MC1, most patients in MC4 have hypertension (41%) or diabetes (39%) but not both together. MC5 includes young adults with obesity (75%), diabetes (57%) and/or hypertension (69%). Both MCs still have high RRs, of approximately 80%. From MC4 onwards, all MCs have from 40 to 50% incidence of SARS-CoV-2019 pneumonia at the moment the case was reported to the health care system; this does not exclude the possibility that some patients developed pneumonia some days after. It is worth noting that in groups 4 to 11 more than 70% of deceased patients were diagnosed with pneumonia at the moment of admission.

RRs from MC6, 8, 9 and 10 are similar (64-67%). MC6 includes older adults with no obesity or smoking, but with some frequent diseases like diabetes, hypertension, INMUSUPR or other. MC8 includes elderly with smoking habit, plus hypertension (34%) and/or COPD (44%), two smoking-related diseases. MC10 is quite similar, including elderly with obesity (50%) or smoking habit (42%), who also suffer from COPD (37%) but with a much higher incidence of diabetes (61%) and hypertension (78%). MC9 contains older adults and elderly with both diabetes (95%) and hypertension (96%).

MC7 and 11 hold the lowest RRs (54% and 56%). MC7 includes older adults and elderly with common diseases –diabetes, hypertension and cardiovascular disease– plus CKD (81%), which stands out as the differential factor with similar MCs with low RRs, such as 6 or 9. MC11 is similar to 8 and 10; the key differences are the prevalence of smoking (78%, which doubles the former), the prevalence of COPD (almost all patients, 91%) and the mean age (76 years old), eight years older than that of MC8 and 10 (68 years old). This implies older obese MCs with smoking habits –MC8, 10, and 11– have significantly higher COPD and cardiovascular incidence, which does not occur with the young smoker –MC3.

Figure 4. Heatmap showing the quantified characteristics among 56 of each age-gender specific cluster of the eleven MCs, the size of each cluster (n) was categorized into 6 categories as shown in Table 3.

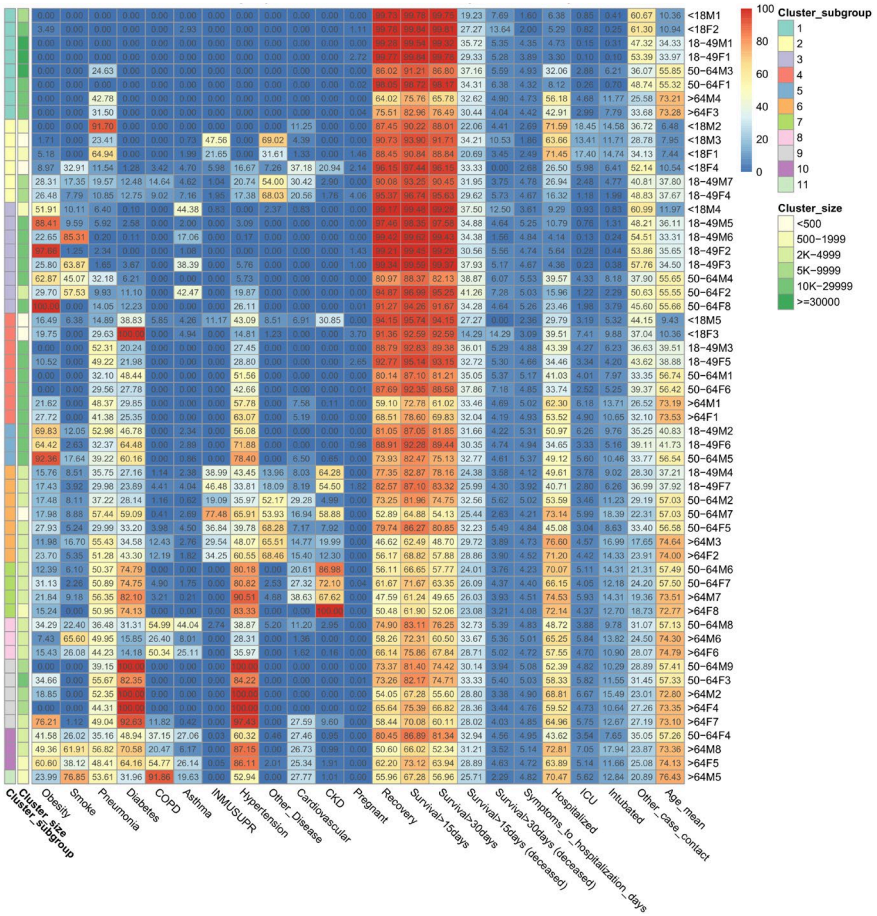


Table 4. The distribution of age, features and comorbidity with the quantitative description of demographic features, treatment, and epidemiological characteristics among eleven MCs with a CI of 95%. In these results, we applied arithmetic mean presuming that each age-gender cluster is representative to its population. Thus, the size (n) of each age-gender cluster was ignored.

	MC1	MC2	MC3	MC4	MC5	MC6	MC7	MC8	MC9	MC10	MC11
No. of age-gender clusters (n total = 56)	8	6	8	8	3	7	4	3	5	3	1
Age ($\bar{x}$)	43.4	18	39.8	44.8	46.4	56.3	65.3	68.7	66.8	68.2	76.4
Sex (%)											
Female	50	50	50	50	33.33	42.86	50	33.33	60	66.67	0
Male	50	50	50	50	66.67	57.14	50	66.67	40	33.33	100
Age range (%)											
<18	25	66.67	12.5	25	0	0	0	0	0	0	0
18-49	25	33.33	50	25	66.67	28.57	0	0	0	0	0
50-64	25	0	37.5	25	33.33	42.86	50	33.33	40	33.33	0
>64	25	0	0	25	0	28.57	50	66.67	60	66.67	100
Features (%)											
Pregnancy	0.49	1.28	0.3	0.8	0.33	0.26	0.01	0	0.01	0	0
Obesity	0.44	11.78	59.88	12.01	75.54	18.89	20.15	19.05	25.94	50.51	23.99
Smoke	0	9.67	34.09	0.8	10.77	8.1	4.38	38.03	0.22	42.02	76.85
Comorbidities (%)											
Diabetes	0	4.42	4.5	39.06	57.14	35.62	76.44	20.45	95	61.23	31.96
COPD	0	4.51	0	0.73	0	5.1	2.03	43.91	2.36	37.46	91.86
Asthma	0.37	3.2	18.17	1.15	2.03	2.69	0.49	25.72	0.08	19.79	19.63
INMUSUPR	0	13.03	0.1	1.4	0	40.38	0	0.91	0	0.03	0
Hypertension	0	9.13	7.59	41.15	68.79	46.79	83.71	34.38	96.33	77.86	52.94
Other Disease	0	38.32	0.3	1.22	0	48.63	1.85	1.73	0	0.82	0
Cardiovascular	0	17.52	0.1	2.46	2.17	14.25	21.64	4.73	5.52	26.51	27.77
CKD	0	4.27	0	3.87	0.22	31.84	81.67	1.04	1.92	1.28	1.01
Treatment (%)											
Hospitalized	19.87	46.08	14.15	42.22	44.91	58.56	70.72	57.17	60.8	60.11	70.47
ICU	1.59	9.82	1.23	4.48	5.06	4.01	4.87	4.81	5.56	5.24	5.62
INTUBATED	3.44	9.03	2.18	7.9	8.46	12.12	13.38	11.5	12.13	12.42	12.84
Pneumonia (%)	12.36	37	9.08	37.18	41.52	42.44	52.14	43.55	48.1	46.8	53.61
Recovery (%)	90.27	91.37	95.22	82.81	81.3	66.94	53.96	66.43	64.95	64.42	55.96
Survival>15days (%)	93.46	93.73	97.01	88.39	87.27	76.34	65.37	77.1	75.26	75.34	67.28
Survival>30days (%)	90.74	91.8	95.5	83.74	82.14	68.26	55.71	68.2	66.33	65.88	56.96
Survival>15days, deceased (%)	30.76	28.64	36.21	31.09	31.59	28.46	24.8	31.7	29.73	31.01	25.71
Survival>30days, deceased (%)	6.61	4.64	5.93	5.79	4.52	4.2	3.82	5.26	4.04	4.24	2.29
From Symptoms to Hospital days ($\bar{x}$)	3.78	3.2	4.87	4.37	5.21	4.48	4.3	4.85	4.92	4.94	4.82
Other case contact (%)	45.84	40.23	51.18	36.6	36.04	27.39	20.9	27.88	27.56	28	20.89

Table 5. Main features of the 11 MCs, sorted by recovery. The thresholds for the different variable categories are displayed with a bullet graph.

Meta-cluster ID	Recovery	ICU	Intubation	Age Group	Habit	Comorbidity	Pneumonia
1	**Very high**	**Low**	**Low**	All	Healthy	Healthy	No
2	**Very high**	**Very high**	**Moderate**	Children & Young	Healthy	*Some w/* INMUSUPR, cardiovascular or other disease.	Yes
3	**Very high**	**Low**	**Low**	Young adults	Obesity Smoke	*Some w/* asthma	No
4	**High**	**Moderate**	**Moderate**	All	Healthy	**Diabetes Hypertension**	Yes
5	**High**	**High**	**Moderate**	Young adults	Obesity	**Diabetes Hypertension**	Yes
6	**Moderate**	**Moderate**	**High**	Older adults	Healthy	**Diabetes Hypertension INMUSUPR Other disease** *Some* w/ CKD	Yes
8	**Moderate**	**Moderate**	**High**	Elderly	Smoke	**Hypertension COPD** *Some w/* diabetes or asthma	Yes
9	**Moderate**	**High**	**High**	Older adults & Elderly	Healthy	*All* **Diabetes** *All* **Hypertension**	Yes
10	**Moderate**	**High**	**High**	Elderly	Obesity Smoke	**COPD Hypertension Diabetes** *Some w/* asthma or cardiovascular.	Yes
7	**Low**	**Moderate**	**Very High**	Older adults & Elderly	Healthy	**Diabetes Hypertension CKD Other disease** *Some w/* cardiovascular.	Yes
11	**Low**	**High**	**High**	Elderly	Smoke	**Hypertension** *All* **COPD** *Some w/* diabetes, asthma or cardiovascular.	Yes

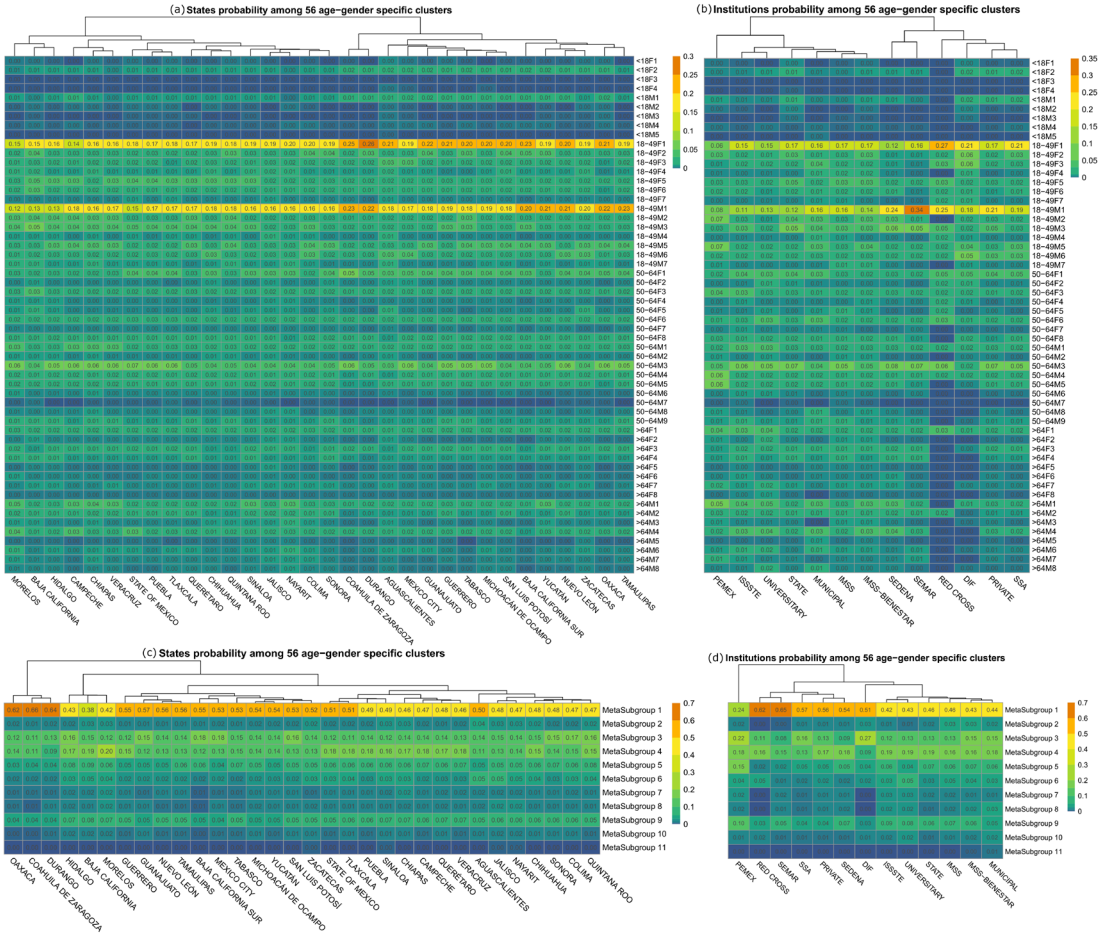
3.3 Variability among states and type of clinical institutions

We found heterogeneity patterns regarding the severity and clusters' distribution among the different Mexican states and types of clinical institution. We note that we compared the clusters' distribution within each age range to circumvent any correlation effect with comorbidities and habits.

Regarding state variability, half of the states tend to a higher probability in healthy clusters with better RR, lower hospitalization, ICU, and Intubation rates among each age-gender group (Figure 6a, e.g., 18F2, 18M1, 18-49F1, 18-49M1, 50-64F1), whereas another half behave inversely; Hidalgo, Baja California and Morelos all together –healthier– compared with Oaxaca, Coahuila de Zaragoza and Durango represent the extremes of these two groups of states. Surprisingly, Mexico City has significantly higher probability in healthier clusters than State of Mexico, albeit the population of their main urban areas are close and both have similar resources and economic development level. Similar patterns occur among MCs (Figure 6c).

Regarding clinical institution variability (Figure 6b, d), SSA, DIF, Private and Red Cross are prone to healthier young patients. This pattern occurred inversely in other institutions, especially the Mexican Petroleum Institution, whose severe cluster probabilities are generally higher. The clinical institutions of the armed forces (SEMAR, SEDENA) were mostly healthy, intuitively with a higher probability of male patients. Interestingly, among the three primary types of clinical institutions in Mexico, the public health system (SSA) is prone to mild-comorbidity, and have relatively higher probabilities in healthy clusters among each age-gender groups, and mostly were MC1 (57%) and 3 (16%); whereas the two main social security systems (IMSS, ISSSTE) occur the opposite.

Figure 6. Heatmaps of probability distribution of the type of clinical institution (TCI) and state where patients received the treatment or medical attention, among 56 age-gender specific clusters (a, b) and eleven MCs (c, d). Rows represent the clusters and columns represent the states and TCI itution. Columns are arranged according to a hierarchical clustering on their values. (a) State probability among age-gender specific clusters; (b) TCI probability among age-gender specific clusters; (c) State probability among eleven MCs; (d) TCI probability among eleven MCs.



4 Discussion

To date, only a few reports have used cluster analysis to describe subgroups heterogeneity in COVID-19 patient-level epidemiological or EHR data[12,13,14] but none included gender and age factors to implement age-gender clustering and meta-clustering analyses on such large dataset (778 692 patients), aiming to find potential patient strata throughout these factors. Thus, it is crucial to comprehend the inter-patient variability patterns to anticipate their risk, susceptibility for viral infection, and morbimortality, based on their clinical phenotypes and demographic characteristics, including age-gender groups analyses.

Our results show 11 clinically distinguishable MCs among 56 age-gender clusters. Each one of the 11 MCs is consistent from a clinical point of view, meaning that the group outcomes can be up to some extent predicted from the proposed input variables according to the literature published up to date. From an outcomes perspective, a dividing line can be clearly drawn between MC1 to 5, with recovery rates (RR) always over 80%, and the rest, whose overall survival never exceeds 70%. Several factors can explain these findings, namely the age distribution, habits and comorbidity. Since all MCs contain 30-60% of women, gender does not seem to be a significant factor among MCs, but in age-gender clusters and statistical analysis showed clearly less severity in female such as pneumonia and mortality rate (Odds Ratio [OR]: 1.58 [95%CI; 1.56-1.60] and 1.76 [95%CI; 1.74-1.79] respectively, male vs female). Thus, we discuss our results among both MCs and age-gender clusters, and then relate them with supporting literature based on the distinct sets of features.

4.1 Age

Notably, MC6 to 11 are exclusively composed of older adults and elderly patients, with the only exception of meta-cluster 6 which contains less than one third (28.57%) of young adults. However, as widely described in literature[28,29], age does not seem to be necessarily linked to higher mortality. MC1 and 4 support this idea since, despite containing the same number of groups (25%) of each age, they show similar RRs (RR MC1 = 90.27%, RR MC4 = 82.81%) to those of groups made only of young adults with little incidence of previous disease (RR MC2 = 91.37%; RR MC3 = 95.22%) and those made of young adults with some frequent diseases, such as diabetes and hypertension (RR MC5 = 81.30%), respectively.

Seemingly, children –MC2– receive priority regarding medical attention since children took fewer days from presenting symptoms to hospitalization, and have significantly higher ICU, intubation, and hospitalization rates than adults with similar clinical conditions. After discussion with Mexican clinicians, a potential reason for this was that in early ages the decompensation or deterioration caused by a pulmonary disease is faster than in adults, and with a higher risk that can result in death. In some cases, in adults there is some margin of time to see how the patient condition evolves before the intubation or ICU admission, but not in children. These results are similar with some recent literature: a study with a small cohort from Madrid, Tagarro et al.[30] found 10% of 41 children with SARS-CoV-2 infection required admission to ICU. As described by Götzinger et al.[31], severe COVID-19 can also happen in small children and adolescents; factors associated with an increased likelihood of requiring ICU admission include age younger than one-month, male sex, presence of lower respiratory tract infection signs and presence of a pre-existing medical condition. Within MC6 to 11, overall survival cannot be explained only by age neither. While MC11 shows the highest mortality and mean age, MC7 shows a similar RR with its mean age being approximately ten years younger, and thus much more similar to the groups with better RRs.

The discussed findings support the idea that, while a young age predisposes to mild disease[29,32], habits and comorbidities may play a key role in predicting mortality rates in older patients with SARS-CoV-2 infection. Interestingly, when we performed the age-gender clustering for the age group of >65 years, we found that centenarians –individuals of over 100 years of age– tended to repeatedly fall in the groups with better outcomes, which is in line with the well-studied good health and low frailty scores[33] of this subpopulation. Therefore, age is a key factor to explain the dividing line between “high” and “moderate” RRs, as well as the low RR in MC11 (56%) compared to MC8 and 10 (64 and 66%), all of which share “hypertension”, “COPD” and “smoke” as only inputs, differing in mean age (76 years for MC11 versus 66-64 years for MC8 and 10).

4.2 Habits

The role of obesity and smoking as risk factors for severe disease are complex, since they are both associated with the development of a number of conditions (e.g. COPD[34] or cardiovascular[35]). In our study, the effect of obesity is more clearly seen on the comparison between MC4 and 5. Both have diabetes and hypertension and moderate RRs (81-82%); however, whereas MC4 includes patients of all ages (25%) without obesity, MC5 contains mostly young adults (66.7%) who suffer from obesity. This suggests that obese young adults may behave as "older", implying higher mortality[29,36]. We found just the opposite in young individuals without previous conditions; MC2 and 3 have similar RRs even though MC3 contains a significant number (59.27%) of obese patients or smokers. These findings suggest the role of habits cannot be considered alone, but always along with age and duration of unhealthy habits. Our results confirm smoking and obesity are simultaneously risk factors for severe SARS-CoV-2 and the development of other diseases, such as cardiovascular disease or COPD, especially in older patients –MC8, 10, 11; it is then feasible that the longer the time as a smoker, the greater the incidence of severe disease. The effect of obesity is not so clear in older groups, since they all have about 20% of obese individuals. Still, in young obese patients without comorbidity (18-49M5 and 18-49F2), obesity seems unrelated to mortality.

Regarding smoking, the evidence of a negative impact is not so straightforward. Some reviews have presented current smoking as a protective factor versus former smoking, while it is clearly a risk factor versus never smoking[37]. Our results show that groups gathering young smokers have RRs which are not inferior to age-matched non-smoking groups, as proven by MC3 (RR = 95.22%, 34% smokers) versus MC2 (RR =91.37%, 9.7% smokers). In older individuals, the effect of tobacco is harder to evaluate since it is inevitably linked to the development of COPD. MC8, 10 and 11 are most representative for older adult and elderly smokers. In conclusion, when evaluating habits, the patient's age and time since diagnosis may help establish useful correlations.

4.3 Comorbidities

Among the recorded comorbidities, diabetes and hypertension hold the highest prevalence. Actually, their prevalence seems to explain the decrease in RRs rates from over 90% in MCs 1-3 to 81% in MCs 4-5, all of which are young adult groups. If we consider older MCs (6-11), both diseases are present in nearly every group, so it doesn't seem to specifically characterize any cluster. While MC9 represents older patients with both diseases simultaneously (>95%). MC10 differs from 8, both having similar characteristics, but the former has double diabetes and hypertension rates. According to current literature, both diabetes and hypertension are independent risk factors for severe disease[29,38,39]. On the other hand, some diseases tend to be descriptive of a certain group. Immunosuppressed patients fall mostly on MC6 –older adults with either diabetes, hypertension, INMUSUPR or other disease. We were surprised not to find INMUSUPR patients within the clusters with the lowest RRs. However, INMUSUPR has not yet been confirmed a relevant factor for disease severity, except for cancer patients[40,41]. Furthermore, MC6 also holds a low amount of CKD patients, a factor which has been widely studied as a key factor for disease progression[42,43] and may be the cause for the INMUSUPR in this group as we computed an odds ratio of 9.65 (95%CI [9.05-10.28]) according to the prevalence of INMUSUPR of CKD patients vs non-CKD patients.

MC7 is characterised by the high prevalence of CKD and other disease. RR falls here almost 10% compared with severe subgroups probably due to prevalence CKD since our result demonstrated it is highly correlated with mortality and shortens survival length among deceased patients; which is definitely in line with a study from Mexico which found that CKD is the factor that best explains mortality[44]. MC8 is similar to 10 and 11, all of which can be explained through COPD, MC11 gathering more than 90% COPD. According to several reviews, COPD patients have increased risk of severe pneumonia and poor outcomes when they develop COVID-19[45,46]. Cardiovascular disease is quite homogeneously distributed among groups, particularly on MC7, 10 and 11. Nowadays, cardiovascular disease may be a double-edged factor, since the disease itself is a proven risk factor for SARS-CoV-2 infection severity, but some of the treatments used, as it is the case of ACE inhibitors, have also been proven to protect against the severe infection from this virus[47,48]. Thus, understanding group outcomes requires a careful read of all single factors individually, the interaction between them and the changes in prevalence with age.

4.4 State and Type of Clinical Institution

To date, the Mexican states and the type of clinical institution variability regarding severity are rarely reported[49,50,51]. Our methods differ with the previous literature by combining MCs and age-gender clusters to counterbalance the effect of age since age and gender are highly correlated with comorbidity and habits. For example, one state (e.g., Morelos) may display higher severity than others if the former includes relatively more elderly and male, but when we only compare age-gender groups the result displays that actually no severity difference exists in terms of probability among age-gender groups of the same age range.

The inclination towards healthy and severe clusters are distinct among different states. This discrepancy may be influenced by many factors such as the number and type –urban or rural– of population, the quantity of medical institutions and availability of resources, and virus transmission level since some states are more industrialized, have the greater cities and have more economical resources (e.g., Mexico City, Jalisco, the State of Mexico) than others (e.g., Oaxaca, Chiapas, Guerrero). Surprisingly, despite having similar resources and development level, Mexico City is prone toward healthier clusters among age-gender groups as well as overall severity through observing the MCs, whereas in the State of Mexico occurs the opposite.

Regarding the type of clinical institutions, two primary social security institutions (IMSS and ISSSTE) that have a national coverage are prone to have more elderly patients and also have a higher probability of severe clusters among each age range in both gender groups; whereas local public hospitals (SSA) behave inversely. One possible explanation is that SSA depends on the local states, and the resources among states often differ. This phenomenon is reflected in these institutions' quality and resources to attend their populations. Another possible explanation we obtained after discuss with several Mexican physicians is that when SSA receives severe patients and have no sufficient medical resources, the patients can be transferred to the IMSS COVID-19 facilities. Consequently, this may saturate IMSS and deplete more resources due to an increasing number of patients, making the distribution of resources harder. These results are in line with a previous study where it was found that the risk of death for an average patient attending IMSS and ISSSTE is 2 times the national average and 3 times higher relative to the private sector[49].

The complex correlation between severity and state/type of clinical institution implies a crucial population and source-inequality. Thus, both considering state and type of clinical institution combined with MCs and age-gender clusters altogether help lead a better classification of patients.

4.5 Limitations

As possible limitations, we excluded patients confirmed after September 30 to avoid possible analysis disturbance about the patient's death result. This approach impeded us to use the most recent data whose variability of epidemiological characteristics could have changed to some degree. The patients' real characteristics comprise many other characteristics such as discharge, cough, fever, and dyspnea which were not available in the data; it would be interesting to include these characteristics in future experiments to explore heterogeneity patterns. Furthermore, the dataset did not include any further information about the patients who were discharged nor readmissions, which is another interesting focus that are rarely reported currently. Thus, further study about the severity patterns discovery among discharged patients who received post-surveillance is highly needed.

5 Conclusion

The analysis of inter-patient variability at COVID-19 heterogenous clusters through an unsupervised ML approach produced compelling models with discriminative severity patterns for all age-gender specific groups. The resultant eleven MCs provide bases to comprehend the classification of patients with COVID-19 based on comorbidities, habits, demographic characteristics, geographic data and type of clinical institutions, as well as revealing the correlations between the above characteristics to anticipate the possible clinical outcomes of each patient with a specific profile. For example, an older obese patient who smokes could be classified into subgroups –MC8, 10, 11– distinguished by pervasive differences in severity

and comorbid patterns. After obtaining further clinical information, preferably, we can extract the age-gender groups within the selected MC to select the age-gender cluster whose characteristics coincide the most with our patient and then evaluate the patient's expected outcomes.

While our findings are informative for designing a novel data-driven model for stratification of COVID-19 patients in Mexico, these may be restricted by limited follow-up systems and other important unregistered geographic, demographic, and epidemiological characteristics such as the duration of the comorbidities and unhealthy habits. We made available the code to replicate the study in other countries or datasets.

Availability of supporting data and materials

The data of epidemiological and clinical patient-level open-source database in Mexico is publicly available at https://www.gob.mx/salud/documentos/datos-abiertos-152127 in Spanish. The English version and the studied sample of this dataset are available in our GitHub Repository https://github.com/bdslab-upv/covid19-metaclustering. The results from 2 through 12 clusters for both gender and age subgroups are available at http://covid19sdetool.upv.es/?tab=mexicoGov.

Abbreviations

COPD: Chronic Obstructive Pulmonary Disease
CKD: Chronic Kidney Disease
INMUSUPR: Immunosuppression
ICU: Intensive Care Unit
EHR: Electronic Health Record
RR: Recovery Rate
MC: Meta-Cluster
DIF: National System for Integral Family Development
IMSS: Mexican Institute of Social Security
ISSSTE: Institute for Social Security and Services for State Workers
PEMEX: Mexican Petroleum Institution
SEDENA: Secretariat of the National Defense
SEMAR: Secretariat of the Navy
SSA: Secretariat of Health

Funding

This work was supported by Universitat Politècnica de València contract no. UPV-SUB.2-1302 and FONDO SUPERA COVID-19 by CRUE-Santander Bank grant "Severity Subgroup Discovery and Classification on COVID-19 Real World Data through Machine Learning and Data Quality assessment (SUBCOVERWD-19).

Acknowledgements

We sincerely thank the different types of clinical institutions and the Mexican government that have made a huge effort to make these data publicly available. We also thank the clinicians and epidemiologists from the Servicios de Salud de Nayarit for the useful discussions on specific aspects of the medical attention to hospitalized patients and the reporting of epidemiological data processes related to COVID-19. Furthermore, we would also like to thank Francisco Tomás García Ruiz for his valuable help in data visualization design.

Authorship Statement

LZ, CS, JMGG, JAC designed the research; LZ, NR, CS, JMGG, JAC, JMM conducted the research; LZ, CS processed and analyzed the data and performed the statistical analysis; all authors assessed the clinical consistency of the cluster analyses. LZ, NR, CS drafted the manuscript; all authors: revised the manuscript critically; all authors approved the final manuscript.

References

1. Organization, W. H. Coronavirus disease 2019 (COVID-19): situation report, 51. (2020).
2. Organization, W. H. COVID-19 weekly epidemiological update, 24 November 2020. (2020).
3. Gattinoni, L., Camporota, L. & Marini, J. J. COVID-19 phenotypes: leading or misleading? *Eur. Respir. J.* **56**, (2020).
4. Gattinoni, L. *et al.* COVID-19 pneumonia: different respiratory treatments for different phenotypes? (2020).
5. Murray, M. F. *et al.* COVID-19 outcomes and the human genome. *Genet. Med.* 1–3 (2020).
6. Whittemore, R. *et al.* ¡ Sí, Yo Puedo Vivir Sano con Diabetes! A Self-Management Randomized Controlled Pilot Trial for Low-Income Adults with Type 2 Diabetes in Mexico City. *Curr. Dev. Nutr.* **4**, nzaa074 (2020).
7. Lai, Y., Charpignon, M.-L., Ebner, D. K. & Celi, L. A. Unsupervised learning for county-level typological classification for COVID-19 research. *Intell. Med.* 100002 (2020).
8. Huang, L. *et al.* Serial quantitative chest ct assessment of covid-19: Deep-learning approach. *Radiol. Cardiothorac. Imaging* **2**, e200075 (2020).
9. Meng, H. *et al.* CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia at admission in Wuhan, China. *J. Infect.* (2020).
10. Barone, S. M. *et al.* Unsupervised machine learning reveals key immune cell subsets in COVID-19, rhinovirus infection, and cancer therapy. *bioRxiv* (2020).
11. Oniani, D., Jiang, G., Liu, H. & Shen, F. Constructing Co-occurrence Network Embeddings to Assist Association Extraction for COVID-19 and Other Coronavirus Infectious Diseases. *J. Am. Med. Informatics Assoc.* (2020).
12. Pung, R. *et al.* Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *Lancet* (2020).
13. Jia, J. *et al.* Epidemiological characteristics on the clustering nature of COVID-19 in Qingdao City, 2020: a descriptive analysis. *Disaster Med. Public Health Prep.* 1–5 (2020).
14. Rubio-Rivas, M. *et al.* Predicting clinical outcome with phenotypic clusters in COVID-19 pneumonia: an analysis of 12,066 hospitalized patients from the spanish registry SEMI-COVID-19. *J. Clin. Med.* **9**, 3488 (2020).
15. de Salud, S. Datos Abiertos-Dirección General de Epidemiologıa. https://www.gob.mx/salud/documentos/datos-abiertos-152127.
16. Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M. & Avillach, P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *Gigascience* **9**, giaa079 (2020).
17. Sáez, C., Robles, M. & García-Gómez, J. M. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat. Methods Med. Res.* **26**, 312–336 (2017).
18. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
19. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
20. Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K. & Kerdprasopb, N. The clustering validity with silhouette and sum of squared errors. *learning* **3**, (2015).
21. Sáez, C., Romero, N., Conejero, J. A. & García-Gómez, J. M. Potential limitations in COVID-19 machine learning due to data source variability: a case study in the nCov2019 dataset. *J. Am. Med. Informatics Assoc.* (2020).

22. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987).

23. Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M. & Avillach, P. EHRtemporalVariability: delineating temporal dataset shifts in electronic health records. *medRxiv* (2020).

24. Sáez, C. *et al.* Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. *J. Am. Med. Informatics Assoc.* **23**, 1085–1095 (2016).

25. Sáez, C. & García-Gómez, J. M. EHRsourceVariability. *GitHub Repos.* (2019).

26. Cleveland, W. S., Grosse, E. & Shyu, W. M. Local regression models. Chapter 8 in Statistical models in S (JM Chambers and TJ Hastie eds.), 608 p. *Wadsworth Brooks/Cole, Pacific Grove, CA* (1992).

27. Cheng, Y. & Church, G. M. Biclustering of expression data. in *Ismb* vol. 8 93–103 (2000).

28. Zhao, X. *et al.* Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. *MedRxiv* (2020).

29. Stawicki, S. P. *et al.* The 2019–2020 novel coronavirus (severe acute respiratory syndrome coronavirus 2) pandemic: A joint american college of academic international medicine-world academic council of emergency medicine multidisciplinary COVID-19 working group consensus paper. *J. Glob. Infect. Dis.* **12**, 47 (2020).

30. Tagarro, A. *et al.* Screening and severity of coronavirus disease 2019 (COVID-19) in children in Madrid, Spain. *JAMA Pediatr.* (2020).

31. Götzinger, F. *et al.* COVID-19 in children and adolescents in Europe: a multinational, multicentre cohort study. *Lancet Child Adolesc. Heal.* **4**, 653–661 (2020).

32. Davies, N. G. *et al.* Age-dependent effects in the transmission and control of COVID-19 epidemics. *MedRxiv* (2020).

33. Borras, C. *et al.* Centenarians: An excellent example of resilience for successful ageing. *Mech. Ageing Dev.* **186**, 111199 (2020).

34. Zamzam, M. A., Azab, N. Y., El Wahsh, R. A., Ragab, A. Z. & Allam, E. M. Quality of life in COPD patients. *Egypt. J. chest Dis. Tuberc.* **61**, 281–289 (2012).

35. Ezzati, M., Henley, S. J., Thun, M. J. & Lopez, A. D. Role of smoking in global and regional cardiovascular mortality. *Circulation* **112**, 489–497 (2005).

36. Farsalinos, K. *et al.* Current smoking, former smoking, and adverse outcome among hospitalized COVID-19 patients: a systematic review and meta-analysis. *Ther. Adv. Chronic Dis.* **11**, 2040622320935765 (2020).

37. Kwok, S. *et al.* Obesity: A critical risk factor in the COVID-19 pandemic. *Clin. Obes.* **10**, e12403 (2020).

38. Zaki, N., Alashwal, H. & Ibrahim, S. Association of hypertension, diabetes, stroke, cancer, kidney disease, and high-cholesterol with COVID-19 disease severity and fatality: A systematic review. *Diabetes Metab. Syndr. Clin. Res. Rev.* **14**, 1133–1142 (2020).

39. Abdi, A., Jalilian, M., Sarbarzeh, P. A. & Vlaisavljevic, Z. Diabetes and COVID-19: A systematic review on the current evidences. *Diabetes Res. Clin. Pract.* **166**, 108347 (2020).

40. Cajamarca-Baron, J. *et al.* SARS-CoV-2 (COVID-19) in Patients with some Degree of Immunosuppression. *Reumatol. Clínica (English Ed.* (2020).

41. Thng, Z. X. *et al.* COVID-19 and immunosuppression: a review of current clinical experiences and implications for ophthalmology patients taking immunosuppressive drugs. *Br. J. Ophthalmol.* (2020).

42. Gansevoort, R. T. & Hilbrands, L. B. CKD is a key risk factor for COVID-19 mortality. *Nat. Rev. Nephrol.* **16**, 705–706 (2020).

43. Wu, C. *et al.* Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. *JAMA Intern. Med.* (2020).

44. Hernández-Galdamez, D. R. *et al.* Increased risk of hospitalization and death in patients with COVID-19 and pre-existing noncommunicable diseases and modifiable risk factors in Mexico. *Arch. Med. Res.* 51, 683–689 (2020).

45. Leung, J. M., Niikura, M., Yang, C. W. T. & Sin, D. D. COVID-19 and COPD. *Eur. Respir. J.* 56, (2020).

46. Zhao, Q. *et al.* The impact of COPD and smoking history on the severity of COVID-19: a systemic review and meta-analysis. *J. Med. Virol.* (2020).

47. Barison, A. *et al.* Cardiovascular disease and COVID-19: les liaisons dangereuses. *Eur. J. Prev. Cardiol.* 2047487320924501 (2020).

48. Guzik, T. J. *et al.* COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options. *Cardiovasc. Res.* (2020).

49. Rivera-Hernandez, M., Ferdows, N. B. & Kumar, A. The Impact of the Covid-19 Epidemic on Older Adults in Rural and Urban Areas in Mexico. *Journals Gerontol. Ser. B* (2020).

50. Salinas-Escudero, G. *et al.* A survival analysis of COVID-19 in the Mexican population. *BMC Public Health* 20, 1–8 (2020).

51. Najera, H. & Ortega-Avila, A. G. Health and Institutional Risk Factors of COVID-19 Mortality in Mexico, 2020. *Am. J. Prev. Med.* (2020).