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 Politècnica de València, Valencia, Spain. [c]CONACyT - Centro de Investigación Científica y de Educación Superior de Ensenada - CICESE-UT3, Mexico

†Senior authors

Abstract

We describe age-gender unbiased COVID-19 subphenotypes regarding severity patterns including prognostic, ICU and morbimortality outcomes, from patterns in clinical phenotypes, habits and demographic features. We used the Mexican Government COVID-19 open data including 778692 SARS-CoV-2 patient-level data as of September 2020. We applied a two-stage clustering approach combining dimensionality reduction and hierarchical clustering: 56 clusters from independent age-gender analyses supported 11 clinically distinguishable meta-clusters (MCs). MCs 1-3 showed high recovery rates (90.27-95.22%), including healthy patients of all ages, children with comorbidities with priority in medical resources, and young obese, smoker patients. MCs 4-5 showed moderate recovery rates (81.3-82.81%): patients with hypertension or diabetes of all ages, and obese patients with pneumonia, hypertension and diabetes. MCs 6-11 showed low recovery rates (53.96-66.94%): immunosuppressed patients with high comorbidity rate, CKD patients with poor survival and recovery, elderly smokers with COPD, severe diabetic elderly with hypertension, and oldest obese smokers with COPD and mild cardiovascular disease. Group outcomes conformed to the recent literature on dedicated age-gender groups. Combination of unhealthy habits and comorbidities were associated with mortality in older patients. Centenarians tended to better outcomes. Immunosuppression was not found as a relevant factor for severity alone but did when present along with CKD. Mexican states and type of clinical institution revealed relevant heterogeneity in severity, relevant for consideration in further studies. The resultant eleven MCs provide bases for a deep understanding of the epidemiological and phenotypical severity presentation of COVID-19 patients based on comorbidities, habits, demographic characteristics, and on patient provenance 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. These results can establish groups for automated stratification or triage towards personalized treatment enabling a personalized evaluation of the patient's expected outcomes.

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 pandemic[1]. As of November 24, the Mexican nation already surpassed one million COVID-19 cases[2].

Medical institutions and researchers have been making a huge effort to describe specific COVID-19 risk factors and outcomes. Several studies have suggested potential COVID-19 subphenotypes, 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 particularly high prevalence of comorbidities, like hypertension, diabetes –a leading cause of death in 2020[6]– and obesity, which is leading the population to particular and undesirable risks for severe coronavirus outcomes. It remains crucial an understanding of how severity patterns vary among Mexican patients to anticipate individuals' prognostic outcomes.

This work describes 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 to date. Other studies proposed unsupervised ML methods for aggregated population data[7], CT image analyses[8,9], molecular-level clustering[10], or coronavirus-related scientific texts[11]. However, to our knowledge, few studies provided to date results from unsupervised ML on patient-level epidemiological data[12,13,14]. This work aims to describe age-gender unbiased COVID-19 subphenotypes that can potentially establish target groups for automated stratification or triage systems to provide personalized therapies or treatments for the specific group severity patterns.

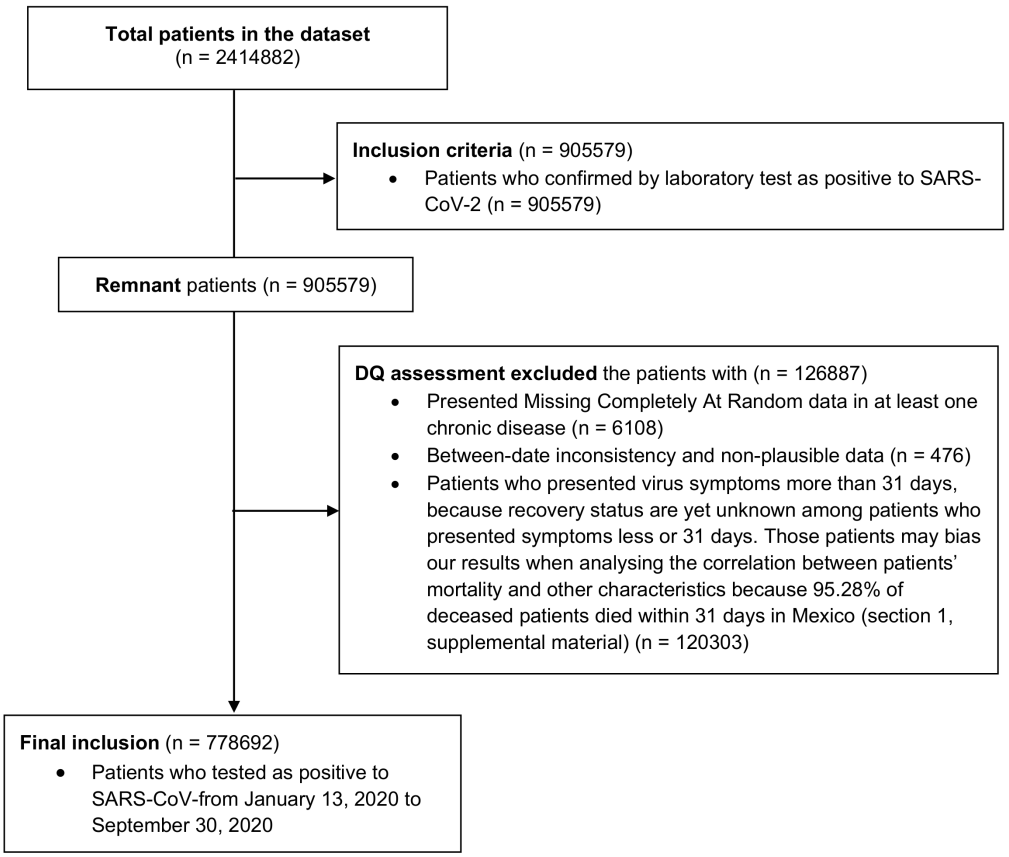
2 Materials and methods

2.1 Data

We used the publicly available COVID-19 Open Data by the Mexican Government[15]. As of 2 November 2020, the dataset comprises a total of 2 414 882 cases including demographic, comorbidities, habits, and prognosis patient-level data, for both positive and non-positive cases.

Figure 1 describes the study inclusion and exclusion criteria, as well as the Data Quality (DQ) assessment process outcomes in a CONSORT-like flowchart. The final sample included 778 692 positive cases.

Figure 1. Dataset preprocessing flowchart.



We derived five outcome variables related with the prospective patient's severity. First, the patient outcome, as deceased or not, from the date of death record. Second, the survival days since the date of symptoms. Third, the days from presenting symptoms to hospital admission. Lastly, we categorized the overall survival at 15 and 30 days after presenting symptoms.

After assessment of potential temporal biases using temporal variability statistical methods[16], and considering not significant temporal changes, we decided keeping the data from all the period of the study. The source variability assessment[17] by comparing differences in data between Mexican states and the type of clinical institutions (TCIs) where patients received medical attention was left as a primary task for this study and is described in section 3.3.

Table 1 shows the list of studied variables. The supplemental material, in sections 2 to 5, describes additional information on the DQ and variability analysis, the original dataset, as well as descriptive statistics of the sample and pregnancy influence in outcomes.

Table 1. List of variables contained in the study case. Originally coded in Spanish, translated into English by the authors for this work.

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 supplemental material for the original dataset description.*

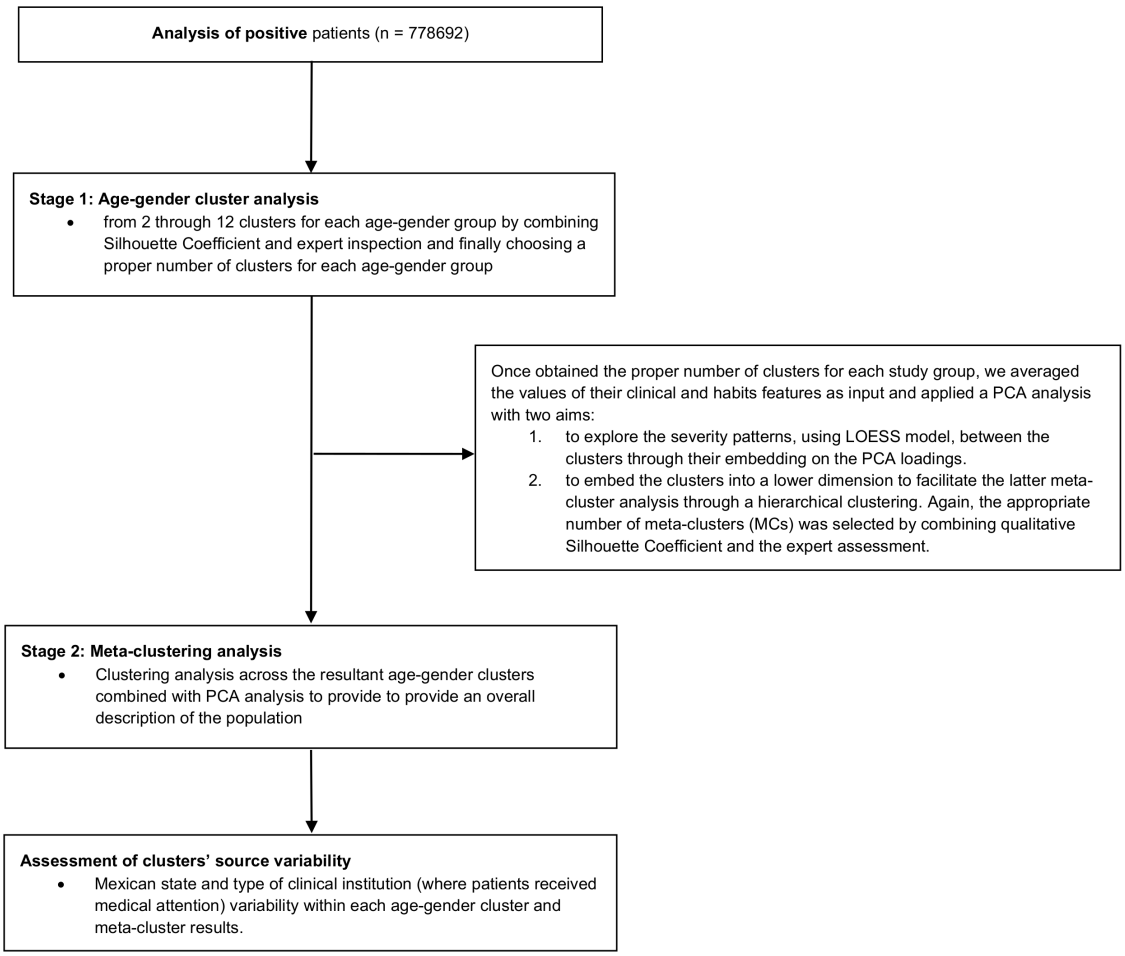
2.2 Methodology

We applied a two-stage subgroup discovery approach. In the first-stage, we applied individual clustering analyses at stratified groups according to gender and age (<18, 18-49, 50-64, and >64 years) to reduce potential confounding factors –since age and gender are highly correlated with comorbidity, habits and mortality– followed by a Principal Component Analysis[18] (PCA) and a locally estimated scatterplot smoothing[19] (LOESS) model on the clusters features to explain their correlations and severity relationships. In the second-stage, in a wider perspective, we performed a clustering analysis on the resultant clusters from the first level by aggregating their clinical phenotypes and demographic features.

Subgroup discovery was performed through a hierarchical clustering algorithm –using Ward's minimum variance method with Euclidean squared distance[20]– 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[21], 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 were obtained by combining a quantitative approach using Silhouette Coefficient[22] –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– and a qualitative cluster analysis audited by the authors of this work, including medical, health informatics and ML experts from Spain and Mexico. We first selected the group of clusters that showed relatively better Silhouette Coefficient values, then adjusted the number for the most reasonable and clinically distinguishable groups regarding clinical phenotypes and demographic features. This process was supported by the pipelines and exploratory tool we developed in previous work[23]. Figure 2 summarizes the full methodology.

The data processing and analyses were performed using RStudio (version 3.6) and Python (version 3.8). Temporal and source variability –DQ analyses– were performed using the EHRtemporalVariability[16] 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 2. Research methodology flowchart.



3 Results

3.1 Meta-clustering analysis

After evaluating the stratified clustering results, we selected the following number of clusters (k) for each specific age-gender group: <18-Male: k=5, <18-Female: k=4, 18-49-Male: k=7, 18-49-Female: k=7, 50-64-Male: k=9, 50-64-Female: k=8, >64-Male: k=8, >64-Female: k=8. This led to a total number of 56 age gender clusters. By taking the features of each group and performing the second-stage meta-clustering analysis we established 11 clinically distinguishable meta-clusters (MCs). Section 6 of supplemental material describes the number of individuals for each age-gender cluster.

Figure 3 describes the relationships between the features of the original 56 age-gender clusters through their principal components (Figure 3A), and provides the correspondence to their assigned meta-clusters (Figure 3B) and their LOESS delineations for the distinct severity outcomes (Figure 3C to 3H).

The 56 clusters PCA analysis uncovered remarkable patterns and heterogeneity among clusters of different ages in both genders. Young adults showed prone to asthma and smoking habit; whereas elderly was 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 immunosuppression and other diseases –both positively correlated–, implying these two pairs of features are negatively correlated in the studied data subgroups.

The LOESS models show that children took fewer days from presenting symptoms to hospitalization, showing higher ICU, intubation, and hospitalization rates than adults with similar conditions (Figure 3D, E, G, H). In contrast, meta-cluster 3 (MC3) –young obese cluster with moderate asthma and smoke rates– behaved inversely, showing that children, under similar clinical conditions, may receive priority regarding medical attention.

Inspecting the relationship between the PCA and LOESS models shows that CKD decreases survival length significantly among deceased patients and increases intubation rates (Figure 3E, D). Mortality constantly increases from children to the elderly, but the most severe zones are inclined toward pneumonia, CKD, and COPD (Figure 3C) independently of the age groups.

Figure 4 describes and quantifies the features of the 56 age-gender clusters and relates them to their MC, and highlights relevant patterns through simultaneously ordering rows and columns through a biclustering technique[26]. We confirm the children have a faster time from symptoms to hospitalization and are prone to ICU admission despite presenting similar clinical condition than adults (e.g., cluster <18M3 versus 50-64F5). Regarding gender discrepancy, females showed a better Recovery Rate (RR) despite presenting similar clinical conditions than males (e.g., >64M1 versus >64F1).

The phenotypes, demographic features and outcomes of each cluster group can be fully explored at http://covid19sdetool.upv.es/?tab=mexicoGov.

Figure 3. Principal component analysis (PCA) of the 56 age-gender clusters, meta-clustering results and LOESS-based severity delineations. (A) PCA from 56 age-gender stratified clusters; (B) scatterplot of the eleven MCs defined from the 56 clusters. (C-H) LOESS scatterplots regarding the severity of the eleven MCs among the 56 computed clusters. The LOESS models delineate seven severity ranges for each outcome including (C) mortality, (D) ICU admission, (E) intubation, (F) survival at 15 days among deceased patients, (G) hospitalization, and (H) days from symptoms to hospitalization. All the scatter plots share coordinates. Each subgroup is denoted using the following abbreviation: [AgeGroup][Gender][ClusterID].

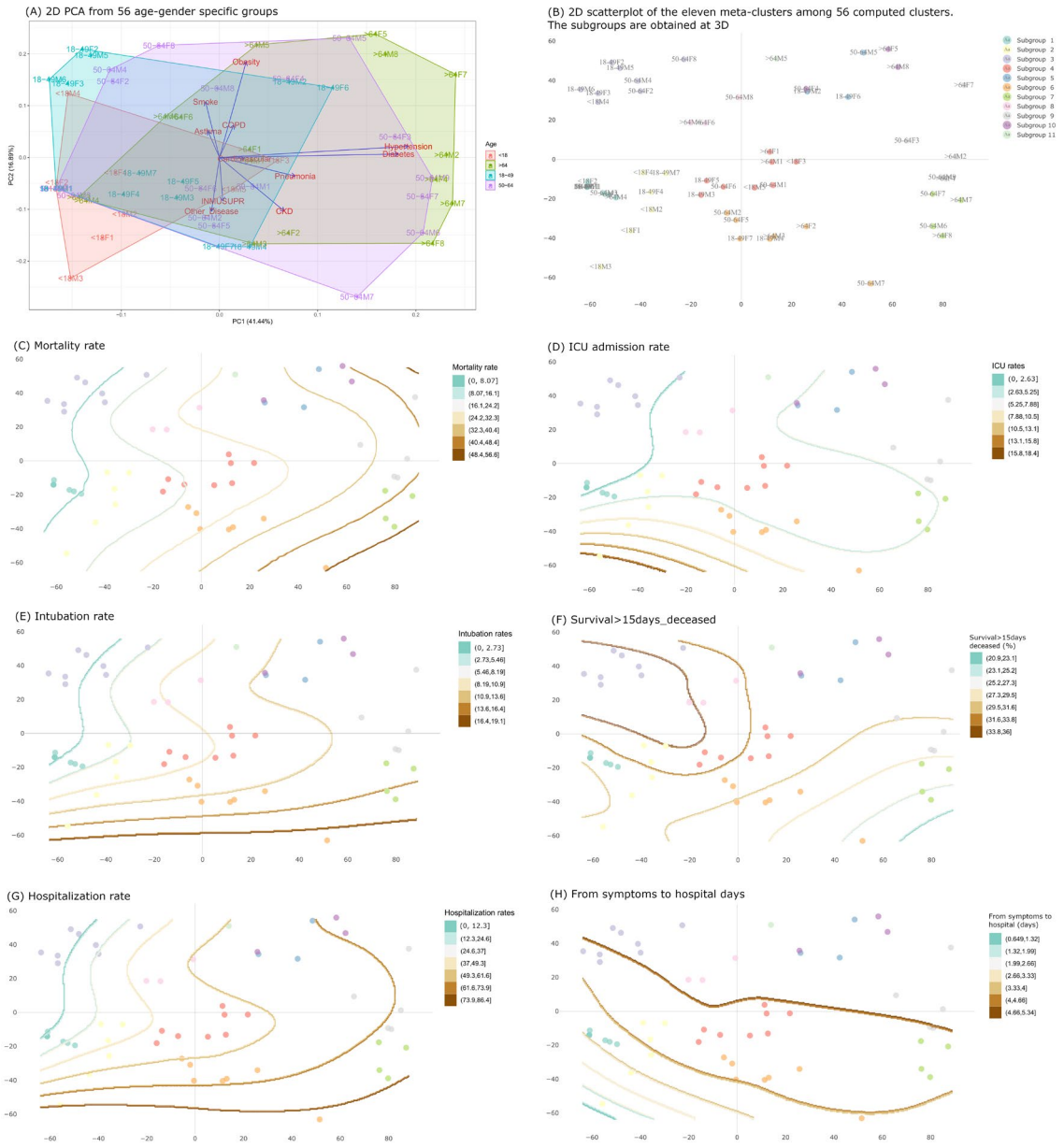
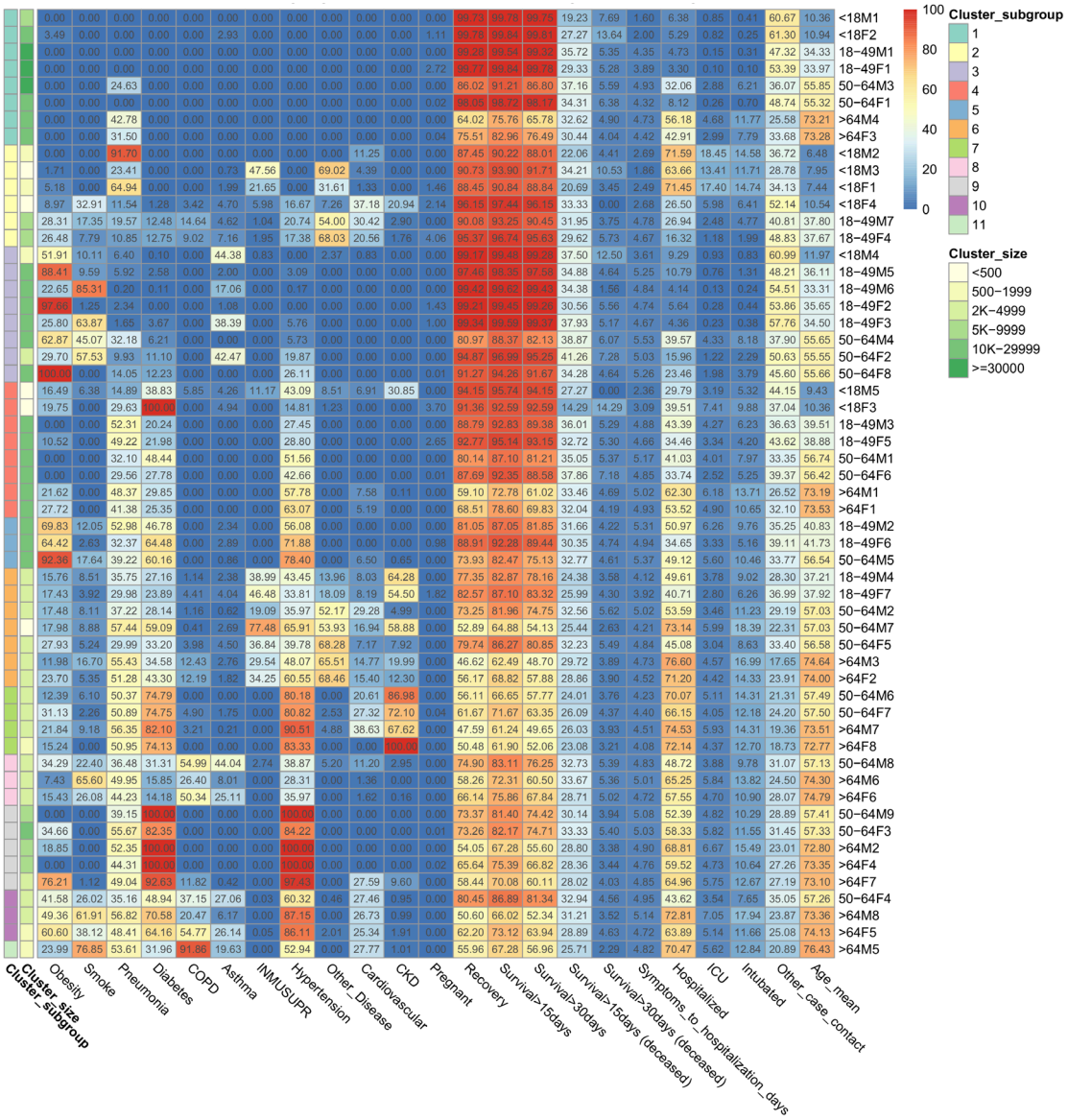


Figure 4. Heatmap showing the quantified characteristics among 56 of each age-gender specific cluster of the eleven MCs, the size of each cluster (n) was categorized into 6 categories as shown in Table 3.

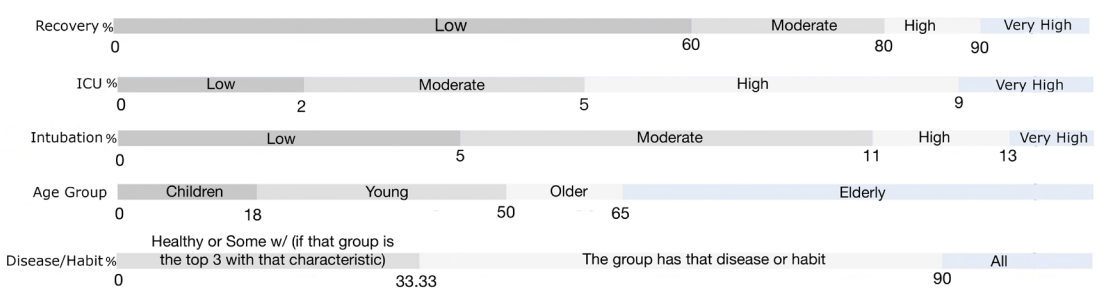


3.2 Epidemiological description of the 11 meta-clusters

Table 3 summarizes their main features of the 11 resultant meta-clusters. Next, we describe the clinically distinguishable main epidemiological findings for each group. All the results and supporting descriptive statistics for each group are available in the supplemental material, section 7.

Table 3. Main features of the 11 age-gender unbiased resultant meta-clusters, sorted by recovery. The thresholds for the different severity outcomes and input variable categories are displayed below.

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



MC1 includes the two healthiest clusters per each age-gender group, with a very high RR (90%). Most deceased patients in MC1 with pneumonia are older patients (Figure 4). MC2 includes children and young individuals (mean age 18) with healthy habits and little incidence of relevant diseases (13% immunosuppression, 17% cardiovascular disease, 4% CKD), albeit RR is very high (91%). MC2 holds the highest ICU admission rate (9%), driven 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, a little incidence of other diseases; and very high RR (95%). Despite the similarly high RRs in MC1 to 3, while MC1 and MC3 show a low incidence of pneumonia, MC2 has 1/3 of pneumonia patients.

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 simultaneously. 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 pneumonia as of case reporting; what does not exclude the possibility that some patients developed pneumonia days after. Noteworthy, in groups 4 to 11 more than 70% of deceased patients were diagnosed with pneumonia.

The RRs from MC6 and 8-10 are similar (64-67%). MC6 includes older adults with no obesity nor smoking, but with frequent diseases including diabetes, hypertension, immunosuppression or other. MC8 includes elderly with smoking habit, plus hypertension (34%) and/or COPD (44%), two smoking-related diseases. Similarly, MC10 includes 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%).

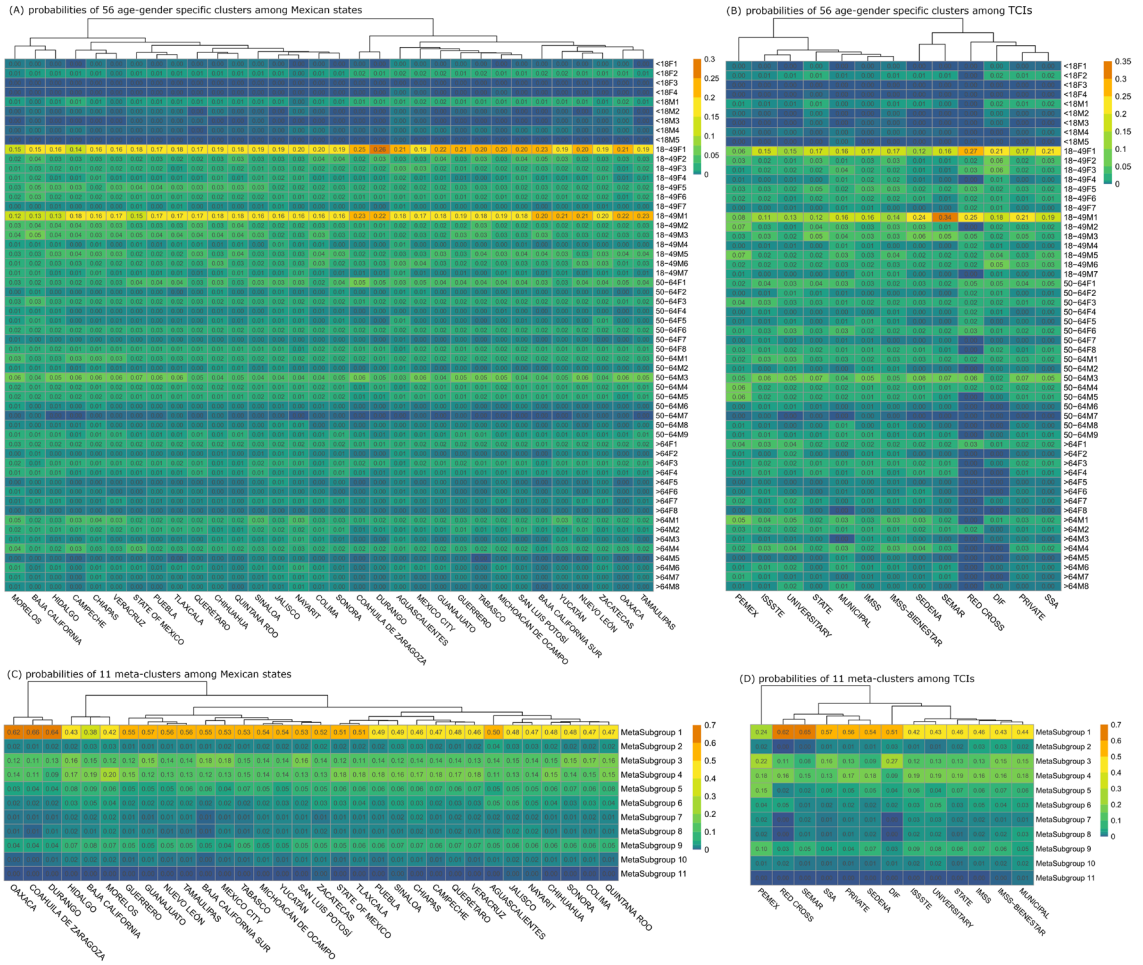
MC7 and 11 hold the lowest RRs (54% and 56%). MC7 includes older adults and elderly with common diseases –diabetes, hypertension and cardiovascular disease– plus CKD (81%). CKD stands out as the differential factor with similar MCs with low RRs, such as 6 or 9. MC11 is similar to 8 and 10; the key differences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

3.3 Variability among states and type of clinical institutions

Regarding state variability, half of the states tended to a higher probability in healthy clusters with better RR, lower hospitalization, ICU, and Intubation rates among each age-gender group (Figure 5A, e.g., 18F2, 18M1, 18-49F1, 18-49M1, 50-64F1) and meta-clusters (Figure 5C), whereas another half behave inversely. Hidalgo, Baja California and Morelos represent healthier groups compared with Oaxaca, Coahuila de Zaragoza and Durango, representing the less healthy. Surprisingly, Mexico City showed 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.

Regarding TCI variability (Figure 5B, D), SSA, DIF, Private and Red Cross are prone to healthier young patients. This pattern occurred inversely in other TCIs, 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 TCIs 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, mostly in MC1 (57%) and 3 (16%); whereas in the two main social security systems (IMSS, ISSSTE) the situation is opposite.

Figure 5. Heatmaps of the probability distribution of the 56 age-gender specific clusters (A, B) and eleven MCs (C, D) for each Mexican state where patients received the treatment or medical attention (A, C) and each type of clinical institution (TCI) (B, D), among. Rows represent the clusters and columns represent the states and TCI. Columns are arranged according to a hierarchical clustering on their values. Note that we compared the clusters' distribution within each age range to circumvent any correlation or association with comorbidities and habits.



4 Discussion

To date, only few reports have used cluster analysis to describe heterogeneity in COVID-19 patient-level epidemiological or EHR data[12,13,14]. To our knowledge, none used such a large dataset (778 692 patients) to find potential patient subphenotypes through cluster analysis on age-gender controlled patient strata. Age-gender unbiased COVID-19 characterization is crucial for a deeper understanding of the inter-patient disease patterns to anticipate their risk, susceptibility for viral infection, and morbimortality.

Our results uncovered 11 clinically distinguishable MCs among 56 age-gender clusters. Each of the 11 MCs shows clinical consistency: their group outcomes can be potentially predicted from the proposed input variables, according to the literature published up to date. From an outcome perspective, a dividing line can be clearly drawn between MCs 1-5 and 6-11. While the former group has RRs always over 80%, the overall survival of the latter never exceeds 70%. Several factors can explain these findings, mainly the age, habits and comorbidity. Since all MCs contain 30-60% of women within their input age-gender clusters, gender does not seem to be a significant factor among MCs. However, statistical analysis and age-gender cluster analysis showed less severity in females including in 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 according to both MCs and age-gender clusters and relate them with supporting literature.

4.1 Age

Two groups with very high RRs are MC2 and MC3, which contain children and young adults. Age may play a protective role against the disease for two reasons. First, as proven by MC3 versus all single-aged groups (MC6-11), pneumonia incidence is lower in young healthy groups; hence, good recovery could be attributable to mild SARS-CoV-2 disease. Second, as shown by good RRs in MC2 –children with severe disease, response to treatment is probably also better at younger ages.

Notably, MC6 to 11 are exclusively composed of older adults and elderly; only MC6 contains less than one third (28.57%) of young adults. However, as described in literature[27,28], older age is not necessarily linked to higher mortality. MC1 and 4 support this fact since, despite containing the same number of groups of each age, they show similar RR (MC1 90.27%, MC4 82.81%) to those RR of groups composed only of young adults with little incidence of previous disease (MC2 91.37%, MC3 95.22%) and those groups made of young adults with some frequent diseases, such as diabetes and hypertension (MC5 81.30%).

Children –MC2– showed to receive priority regarding medical attention, taking fewer days from symptoms to hospitalization, and with significantly higher ICU admission, intubation, and hospitalization rates than adults with similar clinical conditions. After discussion with Mexican clinicians, a potential reason behind this fact seems to be 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. While in adults there is often some time margin to evaluate the patient condition evolution before intubation or ICU admission, it is not the case for children. Supported by recent literature, a study with a small cohort from Madrid,[29] found 10% of 41 children with SARS-CoV-2 infection required ICU admission. Other study[30] showed that severe COVID-19 can also happen in small children and adolescents, where risk factors for ICU admission included 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 a mean age approximately ten years younger, similarly to those groups with better RRs.

These findings support the idea that, while a young age predisposes to mild disease[28,31], habits and comorbidities may play a key role in predicting mortality in older patients with SARS-CoV-2 infection. Interestingly, the clustering for the individual age-gender groups with age >65 years, revealed that centenarians –individuals of over 100 years of age– repeatedly fell in the groups with better outcomes. This fact conforms with the well-studied good health and low frailty scores[32] 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 many conditions (e.g. COPD[33] or cardiovascular[34]). In our study, the influence of obesity is clearly seen by comparing MC4 and 5. Both MCs show diabetes, 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[28,35]. However, we found the opposite in young individuals without previous conditions: MC2 and 3 have similar RRs albeit 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 that smoking and obesity are simultaneously risk factors for severe COPD and cardiovascular, especially in older patients –MC8, 10, 11. It results reasonable that the longer the time as a smoker, the greater the incidence of severe disease. However, the evidence of smoking's negative influence 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[36]. Our results show that groups gathering young smokers have RRs which are not inferior to age-matched non-smoking groups, as proven by the RR of MC3 (95.22%, 34% smokers) versus MC2 (91.37%, 9.7% smokers). In older individuals, the influence of tobacco, inevitably linked to the development of COPD, results harder to evaluate.

Regarding obesity, its influence is not so clear in older groups, having these 20% of obese individuals. Still, in young obese patients without comorbidity (18-49M5 and 18-49F2), obesity seems unrelated to mortality. In conclusion, when evaluating habits, the patient's age and time since diagnosis may help establish more useful correlations.

4.3 Comorbidities

Diabetes and hypertension showed the highest prevalence among the recorded comorbidities. 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. In older MCs (6-11), both diseases are present in nearly every group, not specifically characterizing any cluster; in particular, MC9 represents older patients with both diseases simultaneously (>95%). According to current literature, both diabetes and hypertension are independent risk factors for severe disease[28,37,38].

Immunosuppressed patients fall mostly on MC6 –older adults with diabetes, hypertension, immunosuppression and other disease. Surprisingly, immunosuppressed patients were not the clusters with the lowest RRs. Yet, immunosuppression has not been confirmed as a relevant factor for disease severity, except for cancer patients[39,40]. MC6 also holds few CKD patients, a factor which has been widely studied as a key factor for disease progression[41,42] and it may be the cause for the immunosuppression in this group (OR 9.65 95%CI [9.05-10.28]) according to the prevalence of immunosuppression of CKD patients vs non-CKD patients.

MC7 is characterised by the high prevalence of CKD and other disease. RR decreases here on almost 10% compared with other severe subgroups. Our data showed CKD is highly correlated with mortality and shortens survival length among deceased patients. This conforms with a study from Mexico at which CKD was the factor that best explained mortality[43].

MC8 is similar to 10 and 11, and these 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[44,45].

Cardiovascular disease is 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 COVID-19 severity, but some of the treatments used, such as ACE inhibitors, have also proved to protect against severe infections from SARS-CoV-2[46,47].

4.4 State and Type of Clinical Institution

To date, variability between Mexican states and TCIs regarding severity are rarely reported[48,49,50], nor assessed for variability independently from age and gender. As an example, one state (e.g., Morelos) may show higher severity if it includes more elderly and male patients, but when we compare age-gender groups the results show that no severity difference exists in terms of probability within age-gender groups of the same age range.

The inter-state and TCI variability we found may be influenced from many factors such as the number and type –urban/rural– of population, the quantity of medical institutions and availability of resources and virus transmission level. 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). The differences found between Mexico City and State of Mexico regarding healthy clusters distribution hare hard to explain due their proximity and similarities in the type of population and availability to medical resources.

One possible explanation for the differences in severity between two main social security institutions (IMSS and ISSSTE) and local public hospitals (SSA) is that SSA are administrated by the local states, and the resources among states often differ. This phenomenon could influence these institutions' quality and resources to attend their populations. Another possible explanation is that when SSA receives severe patients and have insufficient medical resources, these 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 conform with previous studies showing that the risk of death for an average patient attending IMSS and ISSSTE is twice the national average and 3 times higher relative to the private sector[48].

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

4.5 Limitations

As a possible limitation, we excluded patients confirmed after September 30 to avoid possible analysis disturbance on survival outcomes, what impeded us using the most recent data whose epidemiological characteristics could have changed to some degree. Furthermore, the dataset did not include additional relevant information about the patients who were discharged, readmissions or the duration of comorbidities and unhealthy habits. Further studies about severity patterns among discharged patients who received post-surveillance or were readmitted is highly needed.

5 Conclusion

The analysis of COVID-19 subphenotypes from the proposed two-stage cluster analysis produced compelling models with discriminative severity patterns and explainability over age and gender. The resultant eleven MCs provide bases for a deep understanding of the epidemiological and phenotypical severity presentation of COVID-19 patients based on comorbidities, habits, demographic characteristics, and on patient provenance 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. These subphenotypes can establish target groups for automated stratification or triage systems to provide personalized therapies or treatments. 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, and then compared with their inner age-gender groups whose characteristics coincide the most with our patient, enabling a personalized evaluation 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 the availability of additional

relevant data including the duration of the comorbidities and unhealthy habits. We facilitate further replicability of the study and generalization to other countries data by making available our experiments code.

Availability of supporting data and materials

The used data is publicly available from the Mexico Government at https://www.gob.mx/salud/documentos/datos-abiertos-152127. The English version of the studied data sample, and the experiments code are available in our GitHub Repository https://github.com/bdslab-upv/covid19-metaclustering. The cluster analysis results can be dynamically explored 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
ML: Machine Learning
DQ: Data Quality
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
TIC: Type of Clinical Institution

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.

Conflict of interest: none declared.

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

19. 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).

20. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).

21. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).

22. Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K. & Kerdprasopb, N. The clustering validity with silhouette and sum of squared errors. *learning* **3**, (2015).

23. Sáez, C., Romero, N., Conejero, J. A. & García-Gómez, J. M. Potential limitations in COVID-19 Mlearning due to data source variability: a case study in the nCov2019 dataset. *J. Am. Med.*

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

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

28. 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).

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

30. 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).

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

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

33. 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).

34. 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).

35. 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).

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

37. 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).

38. 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).

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

40. 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).

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

42. 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).

43. 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).

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

45. 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).

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

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

48. 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).

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

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