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] [lexinzhouds@gmail.com], 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

Appendix A-Supplemental Material

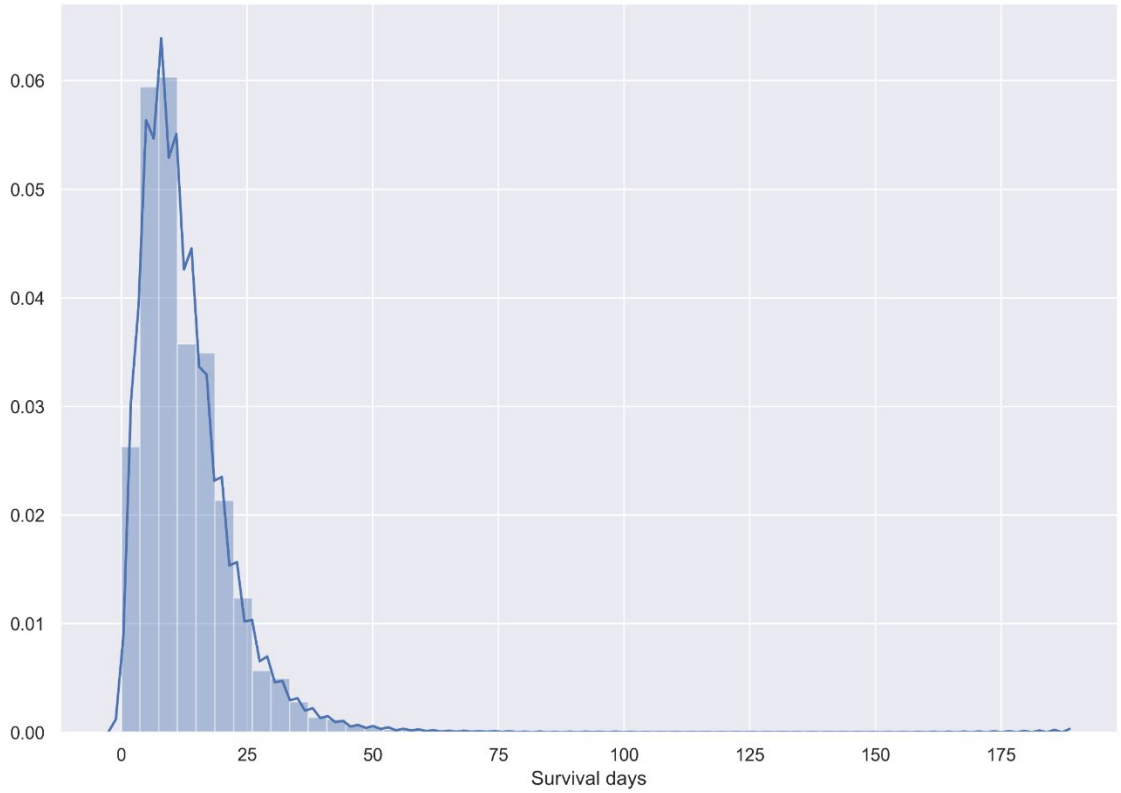
Table of contents

1. Survival days distribution

For completeness, we excluded patients who presented symptoms after September 30 because 95.532% of deceased patients died within 31 days (Figure 1). Some abnormal values are likely to be artificial errors.

Figure 1. Distribution of COVID-19 patients' survival days according to the difference between symptoms date and death date.

Survival weeks	<=4 weeks	<=31 days	<=6 weeks	<= 8 weeks
% individuals	93.111%	95.284	98.7679%	99.615%

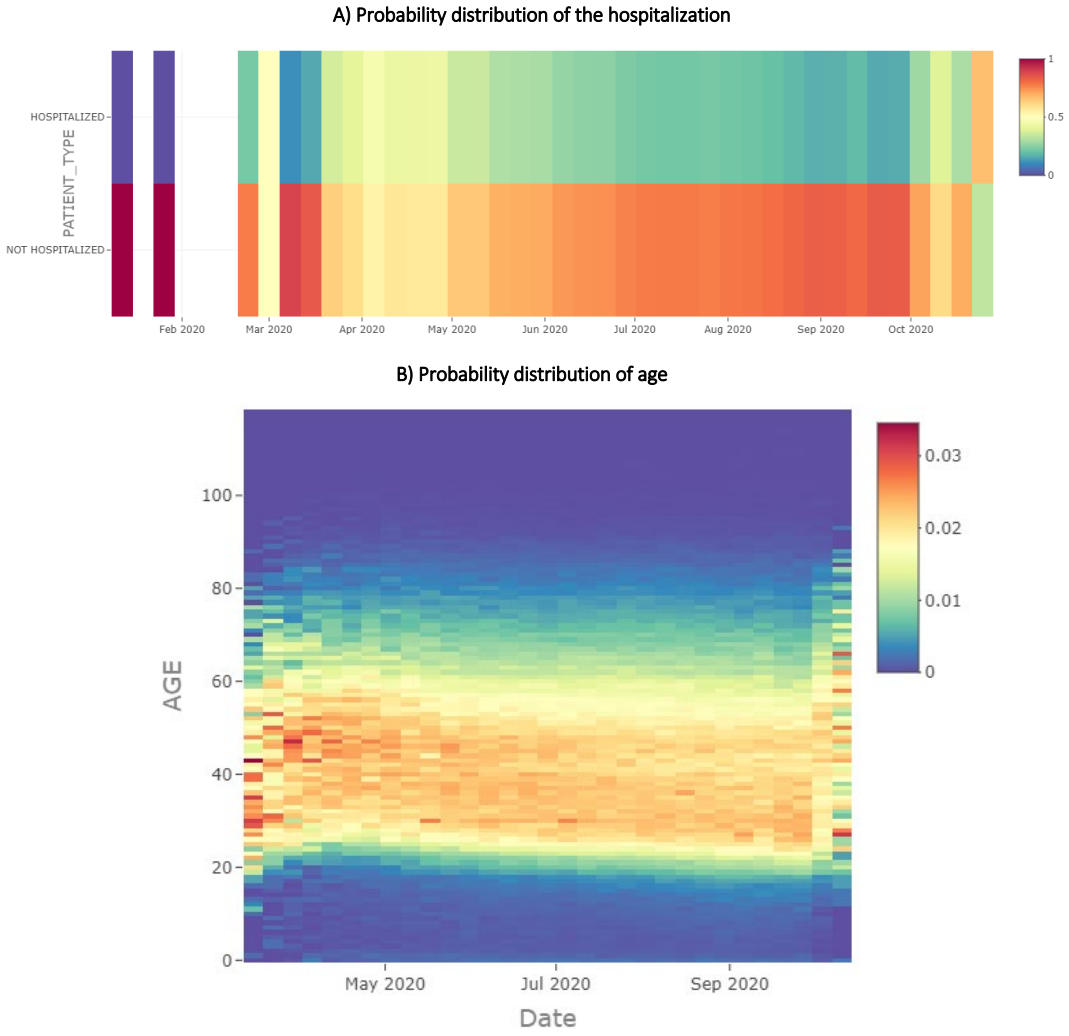


2. DQ results regarding temporal and multi-source variability

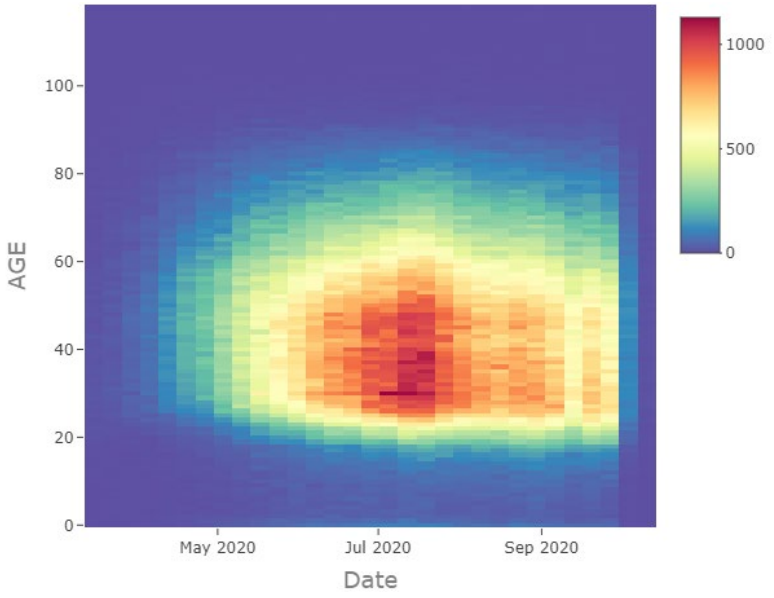
This section illustrates in more detail about DQ analyses (EHR temporal and source variability[1,2]) taken before the clustering analysis. The dataset's temporal variability assessment displays a variable transient state in some variables' distributions from January to April (Figure 1A), possibly associated to the smaller sample size at these moths (Figure 1B). After April, the variation started to stabilize since hospitalization's probability starts to decrease slightly until October. Afterwards, the number of patients was extremely low in the dataset due to a delayed update of patients' information. By excluding those periods in the temporal analyses, we can observe some changes in the age of patients over time (Figure 1B), as well as in the absolute number of patients regarding age (Figure 1C), and patterns in the entity of medical unit. Nevertheless, for the meta-clustering analysis all the period was included, given a flat behavior over time in clinical variables.

The temporal variability of all studied variables is available in http://ehrtemporalvariability.upv.es/ and in a specific tutorial notebook at http://personales.upv.es/carsaesi/covid19-metaclustering/EHRTemporalVariability_tutorial.html

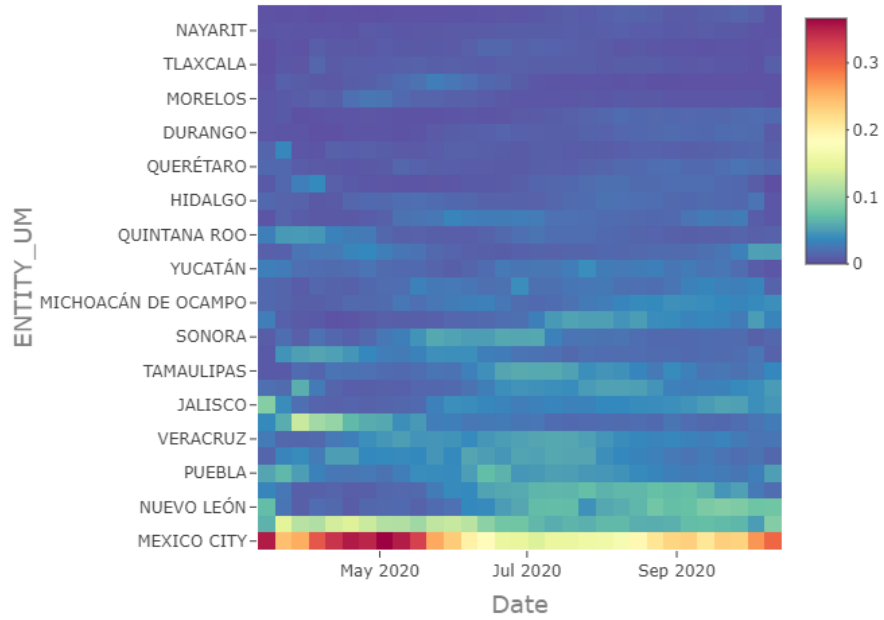
Figure 1. Temporal heatmap regarding the hospitalization and diabetes. Mexico, January 13-November 2, 2020.



C) Absolute frequencies of age



D) Probability distribution of entity of medical unit



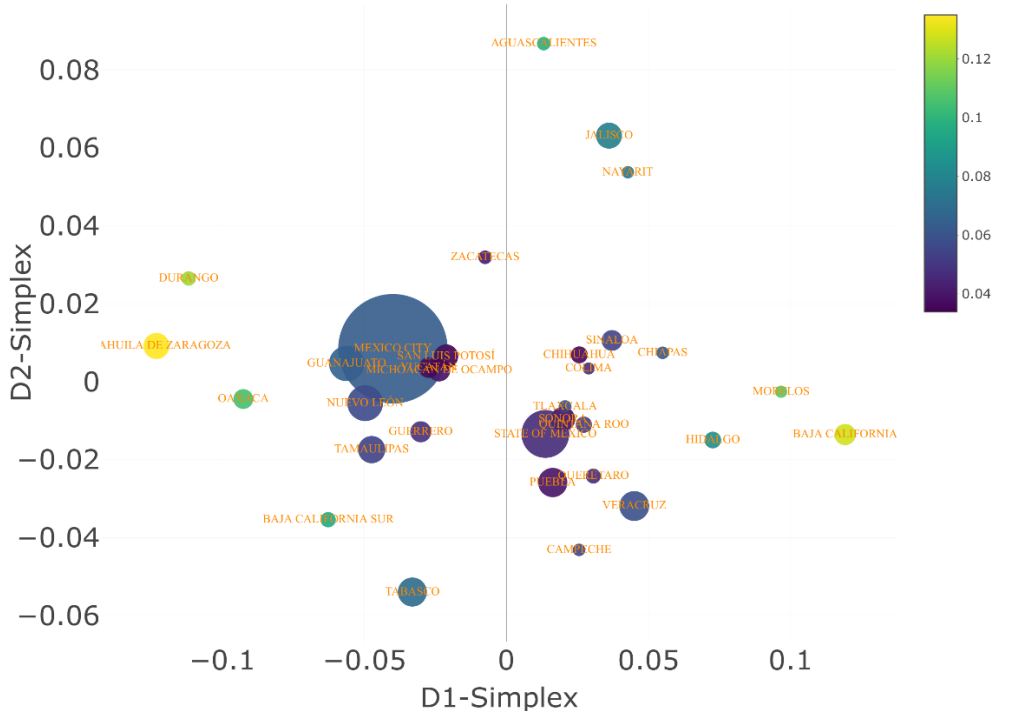
The dataset's multi-source variability assessment found interesting variability patterns in the distributions among data sources in some variables by comparing Mexico's states and the type of clinical institutions where patients received medical attention. Figure 2A shows a striking discrepancy among the states, for example, Hidalgo, Baja California and Morelos all together with Oaxaca, Coahuila de Zaragoza and Durango are located at the left and right extreme, implying these two groups of stats are significantly different upon inclination toward different types of the cluster (in terms of severity, age, and gender) albeit they both are most developed states with similar resources level.

Figure 2B displays the variability between different clinical institution. Surprisingly, we can observe two groups. For example, STATE, SEMAR, UNIVERSITARY, SSA, RED CROSS and SEDENA altogether form one single group and are far located compared with PEMEX, IMSS, ISSSTE, Municipal and the private institution form another group. Surprisingly, most patients were received by the public health system (SSA) and the two main social security systems (IMSS and ISSSTE) whose SPO values differ notably. Thus,

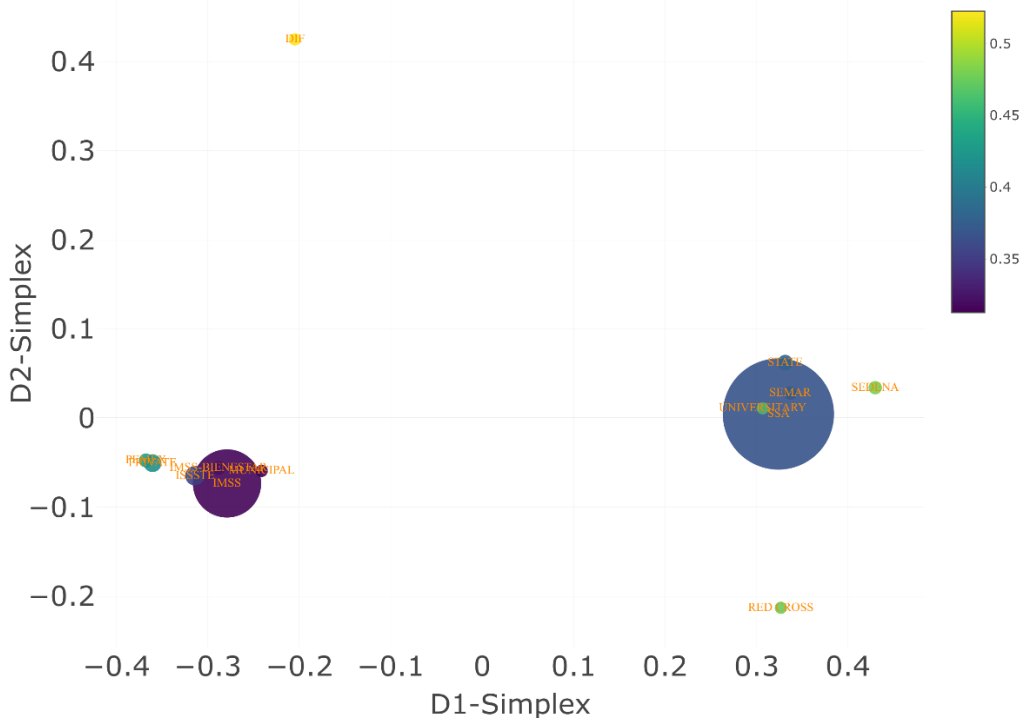
after finding these source variabilities, we decided to include all data in the meta-clustering analysis to further evaluate the effect of these sources' differences.

Figure 2. MSV plot. Each circle represents a data source, where the distances among them represent the distances among their distributions. The color indicates the SPO of each source, and the circle size the source sample size. (A) State variability. (B) Type of clinical institution variability. The 11 meta-clusters were used as input which means that the larger the distance between two states or type of clinical institution the more different their inclination toward clusters of different severity, age, and gender.

A) State variability



B) Type of clinical institution variability



3. List of variables contained in the original dataset

Table 1 displays the variables in the original dataset. The dataset is available in our GitHub repository https://github.com/bdslab-upv/covid19-metaclustering

Table 1. Variables in the original dataset. English version translated by the authors (original in Spanish).

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	Possession of pregnancy	Discrete (Yes, No)
Obesity	Possession of obesity	Discrete (Yes, No)
Smoke	Possession of smoking habit	Discrete (Yes, No)
Pneumonia	Possession of pneumonia	Discrete (Yes, No)
Diabetes	Possession of diabetes	Discrete (Yes, No)
COPD	Possession of chronic obstructive pulmonary disease	Discrete (Yes, No)
Asthma	Possession of asthma	Discrete (Yes, No)
INMUSUPR	Possession of immunosuppression	Discrete (Yes, No)
Hypertension	Possession of hypertension	Discrete (Yes, No)
CKD	Possession of chronic kidney disease	Discrete (Yes, No)
Cardiovascular	Possession of cardiovascular	Discrete (Yes, No)
Other disease	Possession of other diseases	Discrete (Yes, No)
Patient type	Whether a patient was hospitalized	Discrete (Hospitalized, Non-hospitalized)
Intubated	Whether a patient was intubated	Discrete (Yes, No, Not applied)
ICU	Whether a patient had been in an intensive care unit	Discrete (Yes, No, Not applied)
Other case contact	Whether a patient was detected to have contacted with other coronavirus cases	Discrete (Yes, No)
Sector	Type of institution of the National Health System that provided medical care	Discrete[a]
Last_update	The last update of a patient's condition	Date (dd/mm/yyyy)
ID_registration	A random ID number of the case	Numerical Integer
Entity_um	The state where a patient received attention from medical unit	Discrete
Entity_nat	The state where a patient was born	Discrete
Entity_res	The resident state of a patient	Discrete
Municipality_res	The resident municipality of a patient	Discrete
Origin	whether the patient's surveillance was carried out through the system of respiratory disease monitoring units (USMER)	Discrete (USMER, Non-USMER)
Country_Origin	The country where a patient came from	Discrete
Country_nationality	Patient's nationality	Discrete
Nationality	Whether a patient has Mexican nationality	Discrete (Mexican, Foreigner)
Migrant	Whether a patient is emigrant	Discrete (Yes, No)
Speak_indigenous_language	whether a patient speaks an indigenous language	Discrete (Yes, No)
Indigenous	Whether a patient is an indigenous citizen	Discrete (Yes, No)
Tested	Whether the patient was tested to coronavirus	Discrete (Yes, No)
Result_lab	Coronavirus test result	Discrete (Positive SARS-CoV-2, Non-Positive SARS-CoV-2, Pending, Inadequate result, Not Applied)
Final_clasification	Identify which sector confirmed the patient as a coronavirus case	Discrete[b]
Admission_date	The date when a patient attended by the care unit (not necessarily hospitalized)	Date (dd/mm/yyyy)
Symptoms_data	The date when a patient presented symptoms	Date (dd/mm/yyyy)
Death_date	The date when a patient defunct	Date (dd/mm/yyyy)

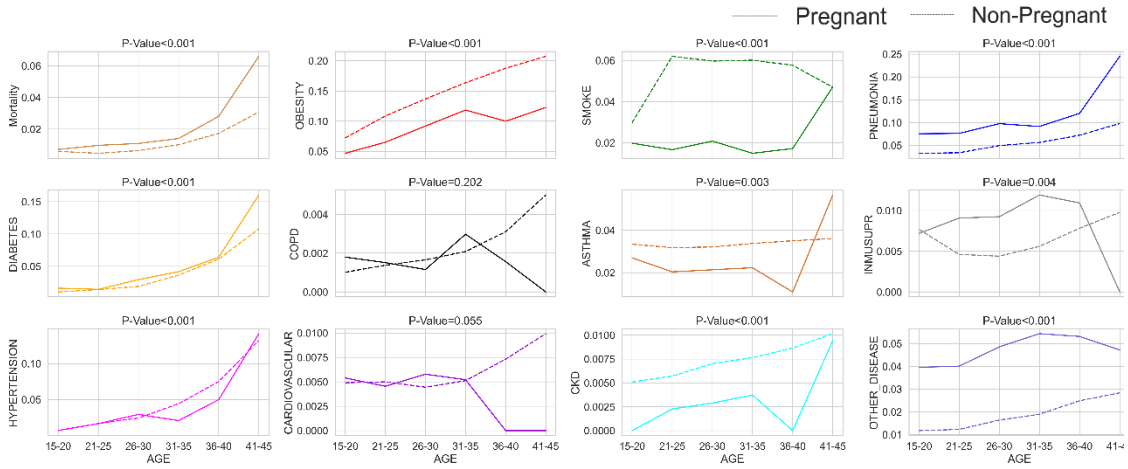
[a] *IMSS, SSA, ISSSTE, PRIVATE, PEMEX, STATE, SEMAR, SEDENA, IMSS-BIENESTAR, UNIVERSITARY, MUNICIPAL, RED CROSS, DIF.*

[b] *Confirmed by laboratory test, Negative tested by laboratory, Not tested by laboratory, Invalid by laboratory, confirmed by epidemiol ogical clinical association, confirmed by ruling committee, Suspected case.*

4. A brief analysis about the effect of pregnancy in female patients

A total of 5735 pregnant patients (0.74% of total cases) were included in our studied sample. Pregnant women were primarily aged between 15-45 (n = 5697, 99.34% of total); so we decided to quantify the mortality and severity difference between pregnant and non-pregnant women (n = 410643) in this age range. Figure 1 demonstrates that pregnancy affects the mortality significantly ($p<0.001$) and clearly an increase in the incidence of pneumonia, INMUSUPR, and other diseases rate ($p<0.001$). Since pregnant women experience immunologic and physiologic alteration that may increase their risk for more severe illness from respiratory infections[3,4,5]. Conversely, the incidence of CKD, asthma, smoke and obesity were lower ($p\leq0.003$). Interestingly, pregnancy did not increase significantly the prevalence of cardiovascular and only a slight increase in diabetes; this is contrary to a recent report in USA[3] that mentioned a double rate of diabetes and cardiovascular in pregnant women.

Figure 1. Probability between pregnant and non-pregnant COVID-19 patients (who aged between 15 to 45; grouped bins in groups of 5 years) in obesity, smoking habit, comorbidities and other case contact. Chi-Square test was used.



5. Age differences among patients

Table 5 shows the discrepancy in studied characteristics among different age-groups (<18, 18-49, 50-64, and >64). It shows asthma and pregnancy are prone to young people, whereas the smoke rates are similar between adults of different ages. Older patients have clearly higher morbimortality, comorbidity and clinical care rate since age and gender are highly correlated with comorbidity and habit (e.g., smoke and obesity).

Table 1. Age differences in regard to epidemiological characteristics and features by COVID-19 subgroups. Mexico, January 13–September 30, 2020.

	≤17 (n=22868)	18-49 (n=470349)	50-64 (n=184452)	≥65 (n=101023)	P-Value
Age, mean (SD)	10.5 (5.5)	35.4 (8.3)	56.2 (4.2)	73.4 (7.0)	<0.001[a]
Sex, n (%)					<0.001[b]
Female	11296 (49.4)	233584 (49.7)	86509 (46.9)	44648 (44.2)	
Male	11572 (50.6)	236765 (50.3)	97943 (53.1)	56375 (55.8)	
Features (%)					
Pregnant	133 (0.6)	5587 (1.2)	10 (0.0)	5 (0.0)	<0.001[b]
Smoke	187 (0.8)	37256 (7.9)	11880 (6.4)	7637 (7.6)	<0.001[b]
Obesity	974 (4.3)	80585 (17.1)	39797 (21.6)	17573 (17.4)	<0.001[b]
Comorbidities, n (%)					
Pneumonia	1223 (5.3)	44070 (9.4)	49514 (26.8)	45913 (45.4)	<0.001[b]
Diabetes	158 (0.7)	31405 (6.7)	50360 (27.3)	36944 (36.6)	<0.001[b]
COPD	19 (0.1)	1557 (0.3)	3208 (1.7)	6335 (6.3)	<0.001[b]
Asthma	771 (3.4)	13009 (2.8)	4267 (2.3)	2010 (2.0)	<0.001[b]
INMUSUPR	401 (1.8)	3335 (0.7)	2617 (1.4)	1958 (1.9)	<0.001[b]
Hypertension	132 (0.6)	39577 (8.4)	58807 (31.9)	50928 (50.4)	<0.001[b]
CKD	107 (0.5)	4693 (1.0)	5031 (2.7)	4695 (4.6)	<0.001[b]
Cardiovascular	197 (0.9)	3549 (0.8)	4573 (2.5)	6753 (6.7)	<0.001[b]
Other disease	578 (2.5)	8548 (1.8)	5091 (2.8)	4308 (4.3)	<0.001[b]
Treatment, n (%)					
Hospitalized	2572 (11.2)	55022 (11.7)	64583 (35.0)	60498 (59.9)	<0.001[b]
Intubated	344 (1.5)	7029 (1.5)	12167 (6.6)	12438 (12.3)	<0.001[b]
ICU	485 (2.1)	4521 (1.0)	5727 (3.1)	5183 (5.1)	<0.001[b]
Mortality, n (%)	276 (1.2)	14042 (3.0)	29485 (16.0)	38274 (37.9)	<0.001[b]
Survival days, mean (SD)	11.9 (11.1)	13.4 (9.4)	14.0 (9.5)	13.1 (8.9)	<0.001[a]
Survival<15days, n (%)	204 (0.9)	8941 (1.9)	18151 (9.8)	24953 (24.7)	<0.001[b]
Survival<30days, n (%)	256 (1.1)	13303 (2.8)	27692 (15.0)	36414 (36.0)	<0.001[b]
From Symptom to Hospital days, mean (SD)	2.2 (2.9)	4.7 (3.6)	5.0 (3.7)	4.8 (3.7)	<0.001[a]
Other case contact, n (%)	13411 (58.6)	228647 (48.6)	69495 (37.7)	27156 (26.9)	<0.001[b]

[a] *One-way ANOVA.*

[b] *Pearson Chi-square Goodness of Fit test.*

References

1. Sáez, C., Gutiérrez-Sacristán, A., Kohane, I., García-Gómez, J. M. & Avillach, P. EHRtemporalVariability: delineating temporal dataset shifts in electronic health records. *medRxiv* (2020).
2. 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).
3. Ellington, S. *et al.* Characteristics of women of reproductive age with laboratory-confirmed SARS-CoV-2 infection by pregnancy status—United States, January 22–June 7, 2020. *Morb. Mortal. Wkly. Rep.* **69**, 769 (2020).
4. Ramsey, P. S. & Ramin, K. D. Pneumonia in pregnancy. *Obstet. Gynecol. Clin. North Am.* **28**, 553–569 (2001).
5. Rasmussen, S. A. *et al.* Preparing for influenza after 2009 H1N1: special considerations for pregnant women and newborns. *Am. J. Obstet. Gynecol.* **204**, S13–S20 (2011).