CLINICAL CHARACTERISTICS AND PROGNOSTIC FACTORS
FOR ICU ADMISSION OF PATIENTS WITH COVID-19 USING
MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

Jose L. Izquierdo¹,², Julio Ancochea³,⁴,⁵, Savana COVID-19 Research
Group*, and Joan B. Soriano³,⁴,⁵ (ORCID 0000-0001-9740-2994)

¹ Universidad de Alcalá, Madrid
² Hospital Universitario de Guadalajara, Guadalajara
³ Hospital Universitario de La Princesa, Madrid
⁴ Universidad Autónoma de Madrid, Madrid
⁵ Centro de Investigación en Red de Enfermedades Respiratorias (CIBERES),
Instituto de Salud Carlos III (ISCIII), Madrid; all in Spain

*Savana COVID-19 Research Group are: Ignacio H. Medrano, MD; Jorge Tello;
Alberto Porras, MD, PhD; Marisa Serrano, PhD; Sara Lumbreras, PhD;
Universidad Pontificia Comillas (ORCID: 0000-0002-5506-9027); Carlos Del
Río-Bermudez, PhD (ORCID: 0000-0002-1036-1673); Stephanie Marchesseau,
PhD; Ignacio Salcedo; Andrea Martínez; Claudia Maté, MD; Sergio Collazo,
MD; Jesús Barea, MD; María Villamayor, MD; Antonio Urda, MD, PhD; Carolina
de la Pinta, MD; Imanol Zubizarreta; Yolanda González, PhD; Sebastian
Menke, PhD (ORCID: 0000-0002-2588-6405).

Address for correspondence:
Dr. Joan B Soriano, MD, PhD, FERS, FCCP
Servicio de Neumología
Hospital Universitario de la Princesa, UAM
Diego de León 62, 28005-Madrid, Spain

Email: jbsoriano2@gmail.com
Cellular: +34 618867769

Date: May 5, 2020
File name: BigCOVIData paper_MedRxiv.docx
Word count: 3,334 words
Number of references: 36 references
Number of illustrations: 4 figures and 2 tables; Supplementary Materials: 1 figure and
1 table
Running Title: Big data and COVID-19
Conflict of Interest: Individual forms from all co-authors are appended
Keywords: artificial Intelligence, big data, COVID-19, electronic health records,
tachypnoea, SARS-CoV-2
ABSTRACT

There remain many unknowns regarding the onset and clinical course of the ongoing COVID-19 pandemic. We used a combination of classic epidemiological methods, natural language processing (NLP), and machine learning (for predictive modeling), to analyse the electronic health records (EHRs) of patients with COVID-19.

We explored the unstructured free text in the EHRs within the SESCAM Healthcare Network (Castilla La-Mancha, Spain) from the entire population with available EHRs (1,364,924 patients) from January 1st to March 29th, 2020. We extracted related clinical information upon diagnosis, progression and outcome for all COVID-19 cases, focusing in those requiring ICU admission.

A total of 10,504 patients with a clinical or PCR-confirmed diagnosis of COVID-19 were identified, 52.5% males, with age of 58.2±19.7 years. Upon admission, the most common symptoms were cough, fever, and dyspnoea, but all in less than half of cases. Overall, 6% of hospitalized patients required ICU admission. Using a machine-learning, data-driven algorithm we identified that a combination of age, fever, and tachypnoea was the most parsimonious predictor of ICU admission: those younger than 56 years, without tachypnoea, and temperature <39°C, (or >39°C without respiratory crackles), were free of ICU admission. On the contrary, COVID-19 patients aged 40 to 79 years were likely to be admitted to the ICU if they had tachypnoea and delayed their visit to the ER after being seen in primary care.

Our results show that a combination of easily obtainable clinical variables (age, fever, and tachypnoea with/without respiratory crackles) predicts which COVID-19 patients require ICU admission.

Abstract word count: 249 words

Funding: This study was sponsored by SAVANA (https://www.savanamed.com/)
INTRODUCTION

The unprecedented, global spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes coronavirus disease 2019 (COVID-19) requires innovative approaches that deliver immediate, real-time results[1, 2]. To date, big data technologies have only been used to estimate SARS-CoV-2 transmission[3], and to indirectly estimate COVID-19 cases in China by using social media[4]. However, there remain many unknowns regarding the onset and temporal distribution of the ongoing COVID-19 pandemic. Similarly, both the individual and population burden of COVID-19 are just beginning to be unravelled. To the best of our knowledge, such tools[5-7] have not been used to explore the clinical characteristics and prognostic factors of COVID-19[8].

Considering the unprecedented spread and severity of the ongoing COVID-19 outbreak, focus has been set on hospital’s unmet need, and in particular ICU requirements[8, 9]. Indeed, health systems have been/are near collapse and independent modelling efforts have aimed to forecast a number of epidemiological estimators, including ICU use [10-12].

Previously, our team reported that the combination of big data analytics and machine learning techniques helped to better determine quality of diagnosis and treatment of chronic obstructive pulmonary disease (COPD) via an analysis of hospital electronic health records (EHRs) using natural language processing (NLP) and validated algorithms[13, 14].

By means of The BigCOVIData study, we aimed to better determine the real-world epidemiology of COVID-19 infection in a well-defined population. Using a combination of classic epidemiological methods [15], NLP, and machine learning (for predictive modeling), we analysed the clinical information contained in the EHRs of patients with COVID-19 to advance our understanding of the disease and its associated outcomes, most notably ICU admission.

METHODS

The BigCOVIData study was conducted in compliance with legal and regulatory requirements and followed generally accepted research practices described in the ICH Guideline for Good Clinical Practice, the Helsinki Declaration in its latest edition, Good Pharmacoepidemiology Practices, and applicable local regulations. This study was classified as a ‘non-post-authorization study’ (EPA) by the Spanish Agency of Medicines and Health Products (AEMPS), and it was approved by the Research Ethics Committee at the University Hospital of Guadalajara (Spain). We have followed and endorsed the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidance for reporting observational research[16].

Study design and data source

This was a multicenter, non-interventional, retrospective study using data captured in the EHRs of the participating hospitals within the SESCAM Healthcare Network in Castilla-La Mancha, Spain (Figure 1). Data captured in the EHRs was collected from all available departments, including inpatient hospital, outpatient hospital, and ER, for
virtually all types of provided services in each participating hospital. The study period was January 1, 2020 – March 29, 2020.

The study database was fully anonymized in a structured format and contained no personal information from patients. Likewise, personal information was not accessed during either the application of automated and algorithmic methods (i.e., NLP) or during the conversion of unstructured data into the structured database. Importantly, given that clinical information was handled in an aggregate, anonymized, and irreversibly dissociated manner, patient consent regulations do not apply to the present study.

Study sample

The study sample included all patients in the source population diagnosed with COVID-19. Patients were identified on the basis of clinical diagnosis (i.e., COVID-19 cases determined by observed symptomatology, imaging (mostly chest X-ray) and laboratory results, as captured in the unstructured, free-text information in the EHRs) and/or microbiological test results (i.e., COVID-19 cases confirmed by RT-PCR or similar available tests). Our decision to consider both PCR- and clinically confirmed cases is justified by the limited availability of routinely administered RT-PCR tests in the region during the study period and supported by recent discussions on the far-from-optimal sensitivity of RT-PCR for COVID19 (i.e., a single negative result from a single specimen cannot exclude the disease in suspected cases)[17, 18]. Indeed, recent reports highlight the clinical validity and relatively high sensitivity of symptom- and imaging-based identification of COVID-19 patients, especially in early stages of the disease[17, 19, 20].

EHRead®

To meet the study objectives, we used EHRead®, a technology developed by SAVANA that applies NLP, machine learning, and deep learning to analyse the unstructured free-text information written in millions of de-identified EHRs. This technology enables the extraction of information from all types of EHRs and the subsequent normalization of extracted clinical entities to a unique terminology. This process allows for further analysis of descriptive or predictive nature. Originally based on SNOMED CT terminology, our unique body of terminology comprises more than 400,000 medical concepts, acronyms, and laboratory parameters aggregated over the course of five years of free-text mining, targeting the most common diseases (e.g. respiratory diseases, cardiovascular diseases, and diabetes, among others).

Using a combination of regular expression (regex) rules and machine learning models, the terminology entities are detected in the unstructured text and later classified based on sections typically contained in the EHRs, hospital services, and other clinical specifications. Importantly, each detected term is described in terms of negative, speculative, or affirmative clinical statements; this is achieved by using deep learning CNN classification methods that rely on word embeddings and context information (for a similar methodological approach, see [22]). Limitations in a case by case detection are also overcome with a similar approach to ensure that the detected concepts are used within the appropriate context for the descriptive and predictive analysis.

For particular cases where extra specifications are required (i.e., to differentiate COVID cases from other mentions of the term related to fear of the disease or to potential contact), the detection output was manually reviewed in more than 5000 reports to avoid any possible ambiguity associated with free-text reporting. All NLP deep learning models used in this study were validated using the standard
training/validation/testing approach; we used a 75/12/13 split ratio in the available
annotated data (between 2,000 and 3,000 records, depending on the model) to ensure
efficient generalization on unseen cases. For all developed models, we obtained F-
scores greater than 0.89.

Data Analyses

All categorical variables (e.g., comorbidities, symptoms) are shown in frequency tables,
whereas continuous variables (e.g., age) are described via summary tables that include
the mean, standard deviation, median, minimum, maximum, and quartiles of each
variable. The number of missing data points for each variable is provided, if any. To
test for possible statistically significant differences in the distribution of categorical
variables between study groups (i.e., male vs. female, ICU admission vs. No ICU
admission), we used Yates-corrected chi² tests. For continuous variables, mean
differences were tested using t-tests. Given our general population approach, and our
larger than usual sample size, we were interested in exploring sex-related differences
in COVID-19 patients, so most results are stratified by sex[23]. All statistical inferences
were performed at the 5% significance level using 2-sided tests or 2-sided CIs.

Predictive model

We developed a decision tree to classify COVID-19 patients according to their risk of
being admitted to the ICU. The two types of patients or classes considered in the
model were therefore "admitted to the ICU" and "not admitted to the ICU". The model
maps the characteristics of patients (the variables) to their class in the shape of a tree.
From a clinical perspective, this model contemplates all patient variables upon
admission, meaning that its predictive value is so from symptom debut until outcome.
The tree is composed of nodes that branch to subsequent children nodes depending
on the patient’s variables. The tree is built in such a way that each branch separates
the two classes as much as possible. This separation is measured as Shannon
entropy, where a node with an entropy of zero means that the classification is perfect
(either all or none of the patients were admitted to the ICU) and an entropy of one is
the worst possible mix (50%/50%)[24].

Model training and validation. The model was developed and tested on the available
data from hospitalized patients that had either been admitted to the ICU or not; the
latter were either discharged from the hospital or died in the course of the disease. This
amounted to a total of 900 patients. Our algorithm was validated in a split of our
COVID-19 sample, in a 70% training set and a 30% validation set. This means that the
model was trained with 630 patients (582 who did not require intensive care, vs 48 who
did) and validated over the remaining 270 patients. Because the two classes were
unbalanced (far fewer patients require ICU), we used the standard technique of
oversampling the lower class to guarantee a balance of accuracy and recall (in other
words, the tradeoff between false positives vs. false negatives). Further, we sought to
replicate the results from this validation in a posteriori sensitivity analysis, as per recent
recommendations for predictive modeling in COVID-19[25] and TRIPOD guidance[26].
For this second validation, we trained the model with data from the provinces of Ciudad
Real and Guadalajara (38% of the study sample from Castilla La-Mancha), and used
an independent set with combined data set from the other three provinces, namely
Toledo, Cuenca, and Albacete for validation.

RESULTS
From a source general population of 2,035,000 inhabitants, we used NLP and machine
to analyse the clinical information contained in the EHRs of 1,364,924
anonymous patients (Figure 1). Among these, we identified a total of 10,504 patients
diagnosed with COVID-19 (Figure 2). The flowchart of participation in the study up to
hospital admission, ICU admission, or discharge is presented in Figure 2.
COVID-19 patients were 52.5% males, with a mean±SD age of 58.2±19.7 years,
(Table 1). Most COVID-19 patients were 50 years and older (Figure 3). Upon
diagnosis, the most common symptoms reported were cough, fever and dyspnoea
(Table 1); notably, less than half of patients presented with these symptoms, probably
due to the fact that most were attended in primary care. Further, respiratory crackles,
myalgia, and diarrhoea were identified in 5% or more of cases, while other respiratory
and non-respiratory signs and symptoms were less common. Sex-dependent
differences in sign and symptom frequencies upon diagnosis are shown in Table 1. Of
note, we observed subtle increases in frequency of diarrhoea, myalgia, headache,
chest pain, and anosmia in female COVID-19 patients, while men showed significant
increases in fever, dyspnoea, respiratory crackles, rhonchus, lymphopenia, and
tachypnoea (all p<0.05).
Similarly, the most frequent comorbidities were cardiovascular disease (48.2% of
patients) -mainly arterial hypertension (33.6%) and heart disease (25.1%)- and
diabetes (15.7%) (Table 1). Regarding respiratory diseases, COPD was present in
6.4%, asthma in 7.2%, OSA in 2%, and bronchiectasis in 1.2% of patients. Sex-
dependent differences in comorbidities upon diagnosis are also shown in Table 1;
except for asthma, the frequency of all comorbidities was significantly higher in male
than female COVID-19 patients (all p<0.05).
Next, we explored whether the distribution of comorbidities and sign/symptoms
captured in the patients' EHRs upon diagnosis differed between those COVID-19
patients who were admitted to the ICU vs. those who were not (Table 2). Regarding
comorbidities, diabetes, obesity, cardiovascular disease (mainly hypertension), heart
disease (mainly ischemic heart disease), and renal dysfunction were more common
among those patients who were admitted to the ICU (all p < 0.01). As for signs and
symptoms, cough, fever, dyspnoea, respiratory crackles, diarrhoea, tachypnoea,
lymphopenia, and rhonchus were more frequent among ICU patients (all p < 0.05).
Interestingly, respiratory diseases were not more frequent among patients who were
admitted to the ICU (Table 2).
Finally, by using a machine-learning, data-driven algorithm, we identified that the
combination of three easily available clinical variables, namely age, temperature, and
respiratory frequency, was the most parsimonious predictor of ICU admission among
COVID-19 patients (Figure 4). For this model, age and temperature were captured as
continuous variables, whereas tachypnoea (yes/no) was defined as respiratory
frequency of more than 20 breaths per minute. With accuracy, recall, and AUC values
of 0.68, 0.71, and 0.76, respectively, the presented model reached optimal balance in
terms of positive and negative predictive value for ICU admission. On the one hand,
those younger than 56 years, without tachypnoea, and with temperature <39°C/102°F
(entropy = 0, n = 145) (or >39°C/102°F without respiratory crackles), were free of ICU
admission, (entropy = 0, n = 18). On the other hand, COVID-19 patients aged 40 to 70
years were likely to be admitted in the ICU if they presented with tachypnoea and
delayed their visit to the ER after being seen in primary care (entropy = 0, n = 104). As
stated in the Methods section, we performed an additional sensitivity analysis with
different data sets to further validate the results of our predictive model. The independent data set of two provinces (Ciudad Real and Guadalajara, including a total of 753,408 individual patients, or 38% of the entire study sample from Castilla-La Mancha; Figure 1 and Supplemental Table S1), was used to retrain our algorithm to identify ICU admission at onset; validation was performed in the remaining three provinces. As shown in Supplemental Figure S1, the new decision tree identified the same relevant clinical variables, that is age, tachypnea, temperature, and respiratory crackles/ronchus with similar (but not identical) thresholds in some of them. This additional model reached values of accuracy, recall, and AUC of 0.85, 0.57, and 0.84, respectively, thus providing additional proof of validity for our main findings.

**DISCUSSION**

Recent technological advances allow for the optimal and rapid extraction, integration, and analysis of the unique and massive amount of untapped medical knowledge captured in the EHRs. This possibility is particularly meaningful when the clinical question at hand requires collecting data from a large number of patients in a very limited amount of time, as is the case with the newly described COVID-19 pandemic.

By anonymously accessing the clinical information of more than 10,000 anonymous patients with COVID-19 (a number that largely surpasses samples included in recent reports about the disease[27, 28]), we were able to describe their demographic and clinical characteristics, their clinical journey, and the statistical relationship between the most common symptoms and comorbidities on admission, and COVID-19 prognosis (i.e., ICU admission). There were subtle differences in clinical symptoms at onset by sex, while all comorbidities (but asthma) were significantly higher in male than female COVID-19 patients; these and other findings should be replicated in clinical series elsewhere.

The variables identified in our ICU admission model (i.e., age, temperature, and tachypnoea) are clinically relevant as they are readily available and easily observable in the everyday practice with COVID-19 patients. Although tachypnea is not an exclusive manifestation of COVID-19 and can be present in patients suffering from other respiratory diseases (i.e., pneumonia), our model suggests that this sign (in combination with age and temperature) is the most reliable predictor of ICU use over other common symptoms and signs such as cough, dyspnea, or respiratory crackles.

In addition, given that the stability and capacity of ICUs worldwide is threatened by the rapid spread of the disease, the identification of individual factors that predict ICU admission may not only improve patient management but also optimize healthcare resource use and planning.

Further applied to other national and international healthcare networks, the tools and methodology presented here can potentially characterize and predict the prognosis of COVID-19 in a timely and unprecedented manner. As recently pointed out[29, 30], there might be value in the application of artificial intelligence to the current COVID-19 pandemic, not only to predict outbreaks[31] or read chest CT scans[32], but also to disentangle COVID-19’s clinical onset and natural history in nearly real-time. Indeed, classical methods would have required months of questionnaire-based data collection and questionnaire validation, along with multiple Ethics Board approvals and other practical hurdles, all saved with our current approach.
In the race against COVID-19[33], where the goal is to curb the pandemic, it is imperative to leverage big data and intelligent analytics for the betterment of public health. However, it is of the utmost importance not to neglect privacy and public trust, to keep best practices, and to maintain responsible standards for data collection and data processing at a global scale[34].

**Strengths and Limitations**

To our knowledge, this is the first study using NLP and machine learning to access real-world data in such a large COVID-19 population. Indeed, our state-of-the-art methodology allowed for the rapid analysis of the unstructured free-text narratives in the EHRs of one million patients from the general population of the region of Castilla La-Mancha (Spain).

Our methodology combined modules for sentence segmentation, tokenization, text normalization, acronym disambiguation, negation detection, and a multi-dimensional ranking scheme; the latter involved linguistic knowledge, statistical evidence, and continuous vector representations of words and documents learned via shallow neural networks. When applied to EHRs, NLP enables a) access to entire track records for all patients in the target population, and b) the implementation of exploratory analysis to unravel associations between variables that have remained undetected with traditional research methods. By considering all possible patients with the target disease, the information and analyses used here (i.e., RWD and free-scale statistics) remained unbiased by the research question or the observers. Unlike classical statistical methods (e.g., logistic regression), the main advantage associated with the use of ML in this context is that it allows for the automatic detection of meaningful relationships between variables. For instance, if a given symptom (i.e., fever) is only relevant for certain patients (i.e., older than 50), techniques such as the classification trees used here are suitable to uncover this relationship. In this context, although the total number of patients that required ICU use in the training set was somewhat low (48 patients), the number of variables considered in the model was also very limited. In addition, the inclusion of a validation stage reduces the likelihood of overfitting. Ultimately, the use of classifications trees in this study (as opposed to other models such as Artificial Neural Networks) is especially appropriate in the clinical context because they are easily interpretable.

Regarding the geographical location of our participating hospital sites, it is worth mentioning that with a total of 1,364,924 patients from the region of Castilla La-Mancha (SESCAM Healthcare Network), our sample is representative of the Spanish population; Spain has been among the hardest hit countries by the pandemic, in terms of both total cases and mortality rates [35, 36], and this region in particular is the third most affected in the country, just behind Madrid and Catalonia. For this reason, we anticipate that the clinical conclusions drawn here are relevant for clinicians worldwide. Of note, ICU capacity in the region during the study period was not compromised yet, which protects against possible bias in our training data (all patients requiring intensive care were indeed admitted to the ICU).

The results and conclusions of the present study should be interpreted in light of the following limitations. First, we did not distinguish COVID-19 cases confirmed by laboratory results (i.e, RT-PCR) from those exclusively diagnosed through clinical observation (i.e., symptomatology, imaging and laboratory results). However, it should
be noted that PCR and other rapid laboratory tests for the detection of SARS-CoV-2 were not routinely administered in Spain during the study period. In addition, this decision is supported by recent discussions on the clinical validity and relatively high sensitivity of symptom- and imaging-based identification of COVID-19 patients, especially in early stages of the disease[17, 19, 20]. Second, independent replications by different research groups in larger patient sets are needed to further support the clinical validity of our results.

Finally, future reports from the BIGCOVIDat study may incorporate laboratory results and treatments, and contextualize the results presented here in a larger clinical picture[25].

We conclude that, in the largest series of COVID-19 patients attended during the first three months of the pandemic in Spain, 6% of all hospitalized patients required ICU; and that a combination of easily obtained clinical variables, namely age, fever, and tachypnoea predicts which COVID-19 patients require ICU admission.

**Acknowledgments.** We thank all the Savaners for helping accelerate health science with their daily work. This would have not been possible without every single team member. We also thank SESCAM (Healthcare Network in Castilla-La Mancha, Spain) for its participation in the study and for supporting the development of cutting-edge technology in real time.
REFERENCES


Figure 1. Map of Castilla-La Mancha.

Footnote: Map of Castilla-La Mancha (red) within the Spanish (blue line) and European territories. From a source general population of 2,035,000 inhabitants, we collected and analyzed the clinical information in the EHRs of 1,364,924 patients within the SESCAM Healthcare Network.
Figure 2. Patient flowchart.

Footnote: Flowchart depicting the total number of inhabitants in the source population, the number (%) of patients with available EHRs analyzed, the number of patients diagnosed with COVID-19, and of those, the number of hospitalizations and ICU admissions.
Figure 3. Age and Sex Distribution of COVID-19 patients.

Footnote: Age distribution of incident cases of COVID-19 in females (left) and males (right) in the study population for the period comprised between Jan 1, 2020 and March 29, 2020.
Figure 4. Decision tree of relevant clinical variables for the prediction of ICU admission in COVID-19 patients.

Footnote: The combination of three easily available clinical variables, namely age, temperature, and respiratory frequency, was the most parsimonious predictor of ICU admission among COVID-19 patients. The number of patients, probability (p) of ICU admission predicted by the model, and level of entropy (a measure indicating how mixed or pure the classification is, where 0 indicates optimal separation of classes) are indicated in each box. The green pathway indicates that those patients with no tachypnoea, younger than 56 years old, and with temperature less than 39°C/102°F (OR more than 39°C/102°F without respiratory crackles), did not require ICU admission. On the contrary, the red pathway indicates that patients aged 40-79 years, who presented with tachypnoea, and delayed their visit to the ER after being seen in primary care, were likely to be admitted in the ICU. For this model, we obtained accuracy, recall, and AUC values of 0.68, 0.71, and 0.76, respectively (top right panel). See text for further details.
# TABLE 1. Baseline demographics and clinical data upon diagnosis.

<table>
<thead>
<tr>
<th></th>
<th>Female n = 4,984</th>
<th>Male n = 5,519</th>
<th>TOTAL n = 10,504</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td>4,984(47.4)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td>5,519(52.5)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
<td>1(0.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Age (in years)</strong></td>
<td></td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mean(SD)</td>
<td>57.4(20.0)</td>
<td>59.0(19.5)</td>
<td>58.2(19.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Median(Min-Max)</td>
<td>58.0(0.0-100.0)</td>
<td>60.0(0.0-102.0)</td>
<td>59.0(0.0-102.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>(Q1-Q3)</td>
<td>(44.0-73.0)</td>
<td>(46.0-74.0)</td>
<td>(45.0-73.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Signs and Symptoms n(%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cough</td>
<td>2,482(49.8)</td>
<td>2,760(50.0)</td>
<td>5,243(49.9)</td>
<td>0.8453</td>
</tr>
<tr>
<td>Fever</td>
<td>2,120(42.5)</td>
<td>2,783(50.4)</td>
<td>4,904(46.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dyspnoea</td>
<td>1,476(29.6)</td>
<td>1,818(32.9)</td>
<td>3,294(31.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Respiratory crackles</td>
<td>849(17.0)</td>
<td>1,085(19.7)</td>
<td>1,934(18.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>556(11.2)</td>
<td>543(9.8)</td>
<td>1,099(10.5)</td>
<td>0.03</td>
</tr>
<tr>
<td>Myalgia</td>
<td>467(9.4)</td>
<td>451(8.2)</td>
<td>919(8.7)</td>
<td>0.0326</td>
</tr>
<tr>
<td>Headache</td>
<td>462(9.3)</td>
<td>302(5.5)</td>
<td>764(7.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Rhonchus</td>
<td>279(5.6)</td>
<td>414(7.5)</td>
<td>693(6.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chest pain</td>
<td>287(5.8)</td>
<td>267(4.8)</td>
<td>554(5.3)</td>
<td>0.039</td>
</tr>
<tr>
<td>Lymphopenia</td>
<td>196(3.9)</td>
<td>346(6.3)</td>
<td>542(5.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Wheezing</td>
<td>194(3.9)</td>
<td>195(3.5)</td>
<td>389(3.7)</td>
<td>0.3567</td>
</tr>
<tr>
<td>Tachypnoea</td>
<td>135(2.7)</td>
<td>203(3.7)</td>
<td>338(3.2)</td>
<td>0.0059</td>
</tr>
<tr>
<td>Anosmia</td>
<td>166(3.3)</td>
<td>134(2.4)</td>
<td>300(2.9)</td>
<td>0.0066</td>
</tr>
<tr>
<td>Sore throat</td>
<td>69(1.4)</td>
<td>57(1.0)</td>
<td>127(1.2)</td>
<td>0.118</td>
</tr>
<tr>
<td>Ageusia</td>
<td>33(0.7)</td>
<td>32(0.6)</td>
<td>65(0.6)</td>
<td>0.68</td>
</tr>
<tr>
<td>Dysphagia</td>
<td>19(0.4)</td>
<td>28(0.5)</td>
<td>47(0.4)</td>
<td>0.4119</td>
</tr>
<tr>
<td>Neuralgia</td>
<td>19(0.4)</td>
<td>22(0.4)</td>
<td>41(0.4)</td>
<td>1</td>
</tr>
<tr>
<td>Splenomegaly</td>
<td>8(0.2)</td>
<td>14(0.3)</td>
<td>22(0.2)</td>
<td>0.4071</td>
</tr>
<tr>
<td>Hepatomegaly</td>
<td>2(0.0)</td>
<td>6(0.1)</td>
<td>8(0.1)</td>
<td>0.3586</td>
</tr>
<tr>
<td><strong>Comorbidities n(%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>2,253(45.2)</td>
<td>2,805(50.8)</td>
<td>5,058(48.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1,552(31.1)</td>
<td>1,975(35.8)</td>
<td>3,527(33.6)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>91(1.8)</td>
<td>163(3.0)</td>
<td>254(2.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>1100(22.1)</td>
<td>1539(27.9)</td>
<td>2639(25.1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ischemic heart disease</td>
<td>152(3.0)</td>
<td>475(8.6)</td>
<td>627(6.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart failure</td>
<td>243(4.9)</td>
<td>309(5.6)</td>
<td>552(5.3)</td>
<td>0.1063</td>
</tr>
<tr>
<td>Diabetes</td>
<td>689(13.8)</td>
<td>957(17.3)</td>
<td>1646(15.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Obesity</td>
<td>479(9.6)</td>
<td>457(8.3)</td>
<td>936(8.9)</td>
<td>0.0185</td>
</tr>
<tr>
<td>Renal dysfunction</td>
<td>271(5.4)</td>
<td>493(8.9)</td>
<td>764(7.3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CKD</td>
<td>171(3.4)</td>
<td>323(5.9)</td>
<td>494(4.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Depression</td>
<td>484(9.7)</td>
<td>219(4.0)</td>
<td>703(6.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic respiratory disease</td>
<td>242(4.9)</td>
<td>646(11.7)</td>
<td>888(8.5)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Asthma</td>
<td>496(10.0)</td>
<td>263(4.8)</td>
<td>759(7.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>COPD</td>
<td>126(2.5)</td>
<td>549(9.9)</td>
<td>675(6.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Obstructive sleep apnea syndrome (OSA)</td>
<td>69(1.4)</td>
<td>143(2.6)</td>
<td>212(2.0)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bronchiectasis</td>
<td>42(0.8)</td>
<td>87(1.6)</td>
<td>129(1.2)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic Liver Disease</td>
<td>36(0.7)</td>
<td>75(1.4)</td>
<td>111(1.1)</td>
<td>0.002</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>16(0.3)</td>
<td>35(0.6)</td>
<td>51(0.5)</td>
<td>0.0304</td>
</tr>
<tr>
<td>HIV</td>
<td>12(0.2)</td>
<td>22(0.4)</td>
<td>34(0.3)</td>
<td>0.2113</td>
</tr>
</tbody>
</table>

**Footnote:** *p-values from Yates-corrected chi² test on percentage difference of female vs. male COVID-19 patients. All tests were performed individually for each variable (comorbidity or sign/symptom, where applicable). For numerical values (i.e., age), t-tests of difference between means were used. List of medical conditions according to SNOMED CT terminology.

<table>
<thead>
<tr>
<th>COMORBIDITIES</th>
<th>SIGNS AND SYMPTOMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition*</td>
<td>No ICU n(%)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1613(15.5)</td>
</tr>
<tr>
<td>Obesity</td>
<td>917(8.8)</td>
</tr>
<tr>
<td>Chronic respiratory disease</td>
<td>883(8.5)</td>
</tr>
<tr>
<td>COPD</td>
<td>673(6.5)</td>
</tr>
<tr>
<td>Asthma</td>
<td>750(7.2)</td>
</tr>
<tr>
<td>OSA</td>
<td>211(2)</td>
</tr>
<tr>
<td>Bronchiectasis</td>
<td>129(1.2)</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>4998(48)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>3487(33.5)</td>
</tr>
<tr>
<td>Ischemic stroke</td>
<td>253(2.4)</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>2604(25)</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>616(5.9)</td>
</tr>
<tr>
<td>Heart failure</td>
<td>548(5.3)</td>
</tr>
<tr>
<td>Renal dysfunction</td>
<td>748(7.2)</td>
</tr>
<tr>
<td>CKD</td>
<td>488(4.7)</td>
</tr>
<tr>
<td>Chronic Liver Disease</td>
<td>109(1)</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>51(0.5)</td>
</tr>
<tr>
<td>Depression</td>
<td>699(6.7)</td>
</tr>
<tr>
<td>HIV</td>
<td>33(0.3)</td>
</tr>
</tbody>
</table>

Footnote: *p-values from Yates-corrected chi^2 test of difference between percentage of patients in either outcome group. All tests were performed individually for each variable (comorbidity or sign/symptom, where applicable). *List of medical conditions according to SNOMED CT terminology.