

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

4

5

6 **Jose L. Izquierdo^{1,2}, Julio Ancochea^{3,4,5}, Savana COVID-19 Research**
7 **Group*, and Joan B. Soriano^{3,4,5} (ORCID 0000-0001-9740-2994)**

8

9 ¹ Universidad de Alcalá, Madrid

10 ² Hospital Universitario de Guadalajara, Guadalajara

11 ³ Hospital Universitario de La Princesa, Madrid

12 ⁴ Universidad Autónoma de Madrid, Madrid

13 ⁵ Centro de Investigación en Red de Enfermedades Respiratorias (CIBERES),
14 Instituto de Salud Carlos III (ISCIII), Madrid; all in Spain

15

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

24

25 **Address for correspondence:**

26 Dr. Joan B Soriano, MD, PhD, FERS, FCCP
27 Servicio de Neumología
28 Hospital Universitario de la Princesa, UAM
29 Diego de León 62, 28005-Madrid, Spain

30

31 **Email:** jbsoriano2@gmail.com

32 **Cellular:** +34 618867769

33

34 **Date:** May 5, 2020

35 **File name:** BigCOVIData paper_MedRxiv.docx

36 **Word count:** 3,334 words

37 **Number of references:** 36 references

38 **Number of illustrations:** 4 figures and 2 tables; Supplementary Materials: 1 figure and
39 1 table

40 **Running Title:** Big data and COVID-19

41 **Conflict of Interest:** Individual forms from all co-authors are appended

42 **Keywords:** artificial Intelligence, big data, COVID-19, electronic health records,
43 tachypnoea, SARS-CoV-2

44 **ABSTRACT**

45

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

50

51 We explored the unstructured free text in the EHRs within the SESCAM Healthcare
52 Network (Castilla La-Mancha, Spain) from the entire population with available EHRs
53 (1,364,924 patients) from January 1st to March 29th, 2020. We extracted related clinical
54 information upon diagnosis, progression and outcome for all COVID-19 cases, focusing
55 in those requiring ICU admission.

56

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

67

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

71

72 **Abstract word count:** 249 words

73

74 **Funding:** This study was sponsored by SAVANA (<https://www.savanamed.com/>)

75

76

77 INTRODUCTION

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

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

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

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

104

105 METHODS

106

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

116 *Study design and data source*

117 This was a multicenter, non-interventional, retrospective study using data captured in
118 the EHRs of the participating hospitals within the SESCOAM Healthcare Network in
119 Castilla-La Mancha, Spain (**Figure 1**). Data captured in the EHRs was collected from
120 all available departments, including inpatient hospital, outpatient hospital, and ER, for

121 virtually all types of provided services in each participating hospital. The study period
122 was January 1, 2020 – March 29, 2020.

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

129 ***Study sample***

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

143 ***EHRead[®]***

144 To meet the study objectives, we used *EHRead[®]*[21], a technology developed by
145 SAVANA that applies NLP, machine learning, and deep learning to analyse the
146 unstructured free-text information written in millions of de-identified EHRs. This
147 technology enables the extraction of information from all types of EHRs and the
148 subsequent normalization of extracted clinical entities to a unique terminology. This
149 process allows for further analysis of descriptive or predictive nature. Originally based
150 on SNOMED CT terminology, our unique body of terminology comprises more than
151 400,000 medical concepts, acronyms, and laboratory parameters aggregated over the
152 course of five years of free-text mining, targeting the most common diseases (e.g.
153 respiratory diseases, cardiovascular diseases, and diabetes, among others).

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

163 For particular cases where extra specifications are required (i.e., to differentiate
164 COVID cases from other mentions of the term related to fear of the disease or to
165 potential contact), the detection output was manually reviewed in more than 5000
166 reports to avoid any possible ambiguity associated with free-text reporting. All NLP
167 deep learning models used in this study were validated using the standard

168 training/validation/testing approach; we used a 75/12/13 split ratio in the available
169 annotated data (between 2,000 and 3,000 records, depending on the model) to ensure
170 efficient generalization on unseen cases. For all developed models, we obtained F-
171 scores greater than 0.89.

172 **Data Analyses**

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

184 **Predictive model**

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

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

213

214 **RESULTS**

215

216 From a source general population of 2,035,000 inhabitants, we used NLP and machine
217 learning to analyse the clinical information contained in the EHRs of 1,364,924
218 anonymous patients (**Figure 1**). Among these, we identified a total of 10,504 patients
219 diagnosed with COVID-19 (**Figure 2**). The flowchart of participation in the study up to
220 hospital admission, ICU admission, or discharge is presented in **Figure 2**.

221 COVID-19 patients were 52.5% males, with a mean±SD age of 58.2±19.7 years,
222 (**Table 1**). Most COVID-19 patients were 50 years and older (**Figure 3**). Upon
223 diagnosis, the most common symptoms reported were cough, fever and dyspnoea
224 (**Table 1**); notably, less than half of patients presented with these symptoms, probably
225 due to the fact that most were attended in primary care. Further, respiratory crackles,
226 myalgia, and diarrhoea were identified in 5% or more of cases, while other respiratory
227 and non-respiratory signs and symptoms were less common. Sex-dependent
228 differences in sign and symptom frequencies upon diagnosis are shown in **Table 1**. Of
229 note, we observed subtle increases in frequency of diarrhoea, myalgia, headache,
230 chest pain, and anosmia in female COVID-19 patients, while men showed significant
231 increases in fever, dyspnoea, respiratory crackles, ronchus, lymphopenia, and
232 tachypnoea (all $p < 0.05$).

233 Similarly, the most frequent comorbidities were cardiovascular disease (48.2% of
234 patients) -mainly arterial hypertension (33.6%) and heart disease (25.1%)- and
235 diabetes (15.7%) (**Table 1**). Regarding respiratory diseases, COPD was present in
236 6.4%, asthma in 7.2%, OSA in 2%, and bronchiectasis in 1.2% of patients. Sex-
237 dependent differences in comorbidities upon diagnosis are also shown in **Table 1**;
238 except for asthma, the frequency of all comorbidities was significantly higher in male
239 than female COVID-19 patients (all $p < 0.05$).

240 Next, we explored whether the distribution of comorbidities and sign/symptoms
241 captured in the patients' EHRs upon diagnosis differed between those COVID-19
242 patients who were admitted to the ICU vs. those who were not (**Table 2**). Regarding
243 comorbidities, diabetes, obesity, cardiovascular disease (mainly hypertension), heart
244 disease (mainly ischemic heart disease), and renal dysfunction were more common
245 among those patients who were admitted to the ICU (all $p < 0.01$). As for signs and
246 symptoms, cough, fever, dyspnoea, respiratory crackles, diarrhoea, tachypnoea,
247 lymphopenia, and rhonchus were more frequent among ICU patients (all $p < 0.05$).
248 Interestingly, respiratory diseases were not more frequent among patients who were
249 admitted to the ICU (**Table 2**).

250 Finally, by using a machine-learning, data-driven algorithm, we identified that the
251 combination of three easily available clinical variables, namely age, temperature, and
252 respiratory frequency, was the most parsimonious predictor of ICU admission among
253 COVID-19 patients (**Figure 4**). For this model, age and temperature were captured as
254 continuous variables, whereas tachypnoea (yes/no) was defined as respiratory
255 frequency of more than 20 breaths per minute. With accuracy, recall, and AUC values
256 of 0.68, 0.71, and 0.76, respectively, the presented model reached optimal balance in
257 terms of positive and negative predictive value for ICU admission. On the one hand,
258 those younger than 56 years, without tachypnoea, and with temperature $< 39^{\circ}\text{C}/102^{\circ}\text{F}$
259 (entropy = 0, $n = 145$) (or $> 39^{\circ}\text{C}/102^{\circ}\text{F}$ without respiratory crackles), were free of ICU
260 admission, (entropy = 0, $n = 18$). On the other hand, COVID-19 patients aged 40 to 70
261 years were likely to be admitted in the ICU if they presented with tachypnoea and
262 delayed their visit to the ER after being seen in primary care (entropy = 0, $n = 104$). As
263 stated in the Methods section, we performed an additional sensitivity analysis with

264 different data sets to further validate the results of our predictive model. The
265 independent data set of two provinces (Ciudad Real and Guadalajara, including a total
266 of 753,408 individual patients, or 38% of the entire study sample from Castilla-La
267 Mancha; **Figure 1** and **Supplemental Table S1**), was used to retrain our algorithm to
268 identify ICU admission at onset; validation was performed in the remaining three
269 provinces. As shown in **Supplemental Figure S1**, the new decision tree identified the
270 same relevant clinical variables, that is age, tachypnea, temperature, and respiratory
271 crackles/ronchus with similar (but not identical) thresholds in some of them. This
272 additional model reached values of accuracy, recall, and AUC of 0.85, 0.57, and 0.84,
273 respectively, thus providing additional proof of validity for our main findings.

274

275 **DISCUSSION**

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

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

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

297

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

302

303 Further applied to other national and international healthcare networks, the tools and
304 methodology presented here can potentially characterize and predict the prognosis of
305 COVID-19 in a timely and unprecedented manner. As recently pointed out[29, 30],
306 there might be value in the application of artificial intelligence to the current COVID-19
307 pandemic, not only to predict outbreaks[31] or read chest CT scans[32], but also to
308 disentangle COVID-19's clinical onset and natural history in nearly real-time. Indeed,
309 classical methods would have required months of questionnaire-based data collection
310 and questionnaire validation, along with multiple Ethics Board approvals and other
311 practical hurdles, all saved with our current approach.

312

313 In the race against COVID-19[33], where the goal is to curb the pandemic, it is
314 imperative to leverage big data and intelligent analytics for the betterment of public
315 health. However, it is of the utmost importance not to neglect privacy and public trust,
316 to keep best practices, and to maintain responsible standards for data collection and
317 data processing at a global scale[34].

318

319 ***Strengths and Limitations***

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

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

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

356 The results and conclusions of the present study should be interpreted in light of the
357 following limitations. First, we did not distinguish COVID-19 cases confirmed by
358 laboratory results (i.e., RT-PCR) from those exclusively diagnosed through clinical
359 observation (i.e., symptomatology, imaging and laboratory results). However, it should

360 be noted that PCR and other rapid laboratory tests for the detection of SARS-CoV-2
361 were not routinely administered in Spain during the study period. In addition, this
362 decision is supported by recent discussions on the clinical validity and relatively high
363 sensitivity of symptom- and imaging-based identification of COVID-19 patients,
364 especially in early stages of the disease[17, 19, 20]. Second, independent replications
365 by different research groups in larger patient sets are needed to further support the
366 clinical validity of our results.

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

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

374

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

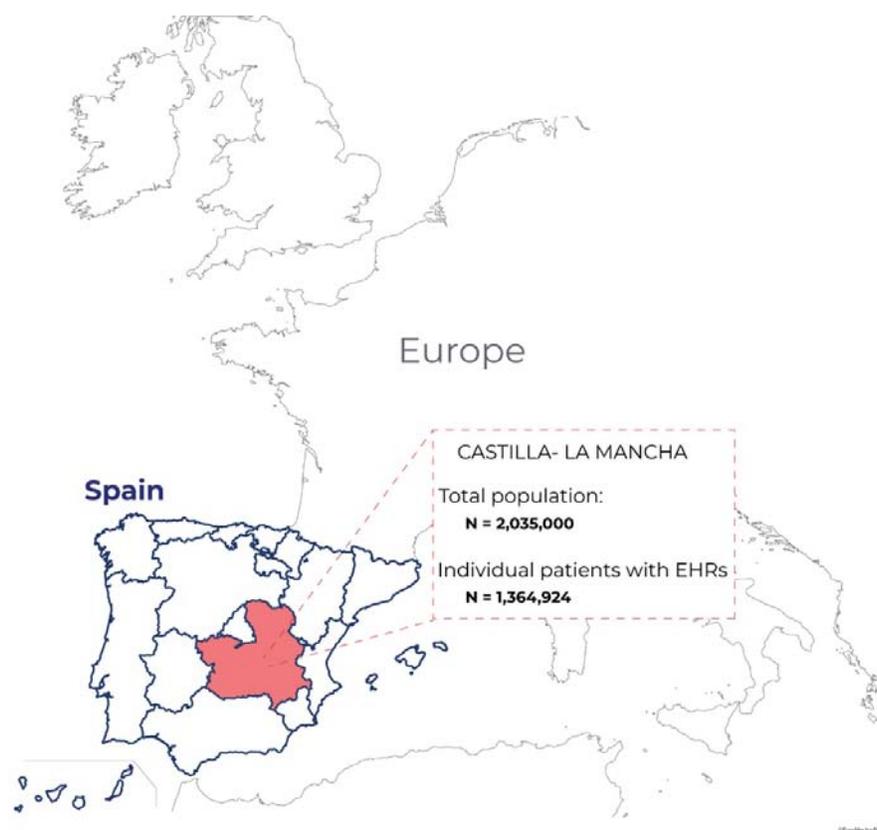
380 REFERENCES

- 381 1. Centers for Disease Control and Prevention (CDC). Coronavirus (COVID-19) at CDC & P.
382 Available from <https://www.cdc.gov/coronavirus/2019-ncov/index.html>. Accessed April 8,
383 2020
- 384 2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P,
385 Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. A Novel Coronavirus from Patients with
386 Pneumonia in China, 2019. *N Engl J Med* 2020: 382(8): 727-733.
- 387 3. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dorner L, Parker M, Bonsall
388 D, Fraser C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital
389 contact tracing. *Science* 2020.
- 390 4. Qin L, Sun Q, Wang Y, Wu KF, Chen M, Shia BC, Wu SY. Prediction of Number of Cases
391 of 2019 Novel Coronavirus (COVID-19) Using Social Media Search Index. *Int J Environ Res Public*
392 *Health* 2020: 17(7).
- 393 5. Topol EJ. High-performance medicine: the convergence of human and artificial
394 intelligence. *Nat Med* 2019: 25(1): 44-56.
- 395 6. Divita G, Carter M, Redd A, Zeng Q, Gupta K, Trautner B, Samore M, Gundlapalli A.
396 Scaling-up NLP Pipelines to Process Large Corpora of Clinical Notes. *Methods Inf Med* 2015:
397 54(6): 548-552.
- 398 7. Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big
399 is beautiful, but will bigger be even better? *Lancet Infect Dis* 2006: 6(10): 653-663.
- 400 8. Liu Y, Mao B, Liang S, Yang JW, Lu HW, Chai YH, Wang L, Zhang L, Li QH, Zhao L, He Y,
401 Gu XL, Ji XB, Li L, Jie ZJ, Li Q, Li XY, Lu HZ, Zhang WH, Song YL, Qu JM, Xu JF. Association
402 between ages and clinical characteristics and outcomes of coronavirus disease 2019. *Eur Respir*
403 *J* 2020.
- 404 9. Horton R. Offline: COVID-19-what countries must do now. *The Lancet* 2020:
405 395(10230): 1100.
- 406 10. Xu B, Gutierrez B, Mearu S, Sewalk K, Goodwin L, Loskill A, Cohn EL, Hsuen Y, Hill SC,
407 Cobo MM, Zarebski AE, Li S, Wu C-H, Hulland E, Morgan JD, Wang L, O'Brien K, Scarpino
408 Samuel V, Brownstein JS, Pybus OG, Pigott DM, Kraemer MUG. Epidemiological data from the
409 COVID-19 outbreak, real-time case information. *Scientific Data* 2020: 7(1): 106.
- 410 11. Murray CJ. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-
411 days and deaths by US state in the next 4 months. *medRxiv* 2020: 2020.2003.2027.20043752.
- 412 12. Sotgiu G, GGCs, Miozzo M, Canonica GW, Virchow JC, Soriano JB. Advanced forecasting
413 of SARS-CoV-2 related deaths in Italy, Germany and Spain. *Allergy* 2020: In Press.
- 414 13. Izquierdo JL, Morena D, Gonzalez Y, Paredero JM, Perez B, Graziani D, Gutierrez M,
415 Rodriguez JM. Clinical Management of COPD in a Real-World Setting. A Big Data Analysis. *Arch*
416 *Bronconeumol* 2020.
- 417 14. Sociedad Española de Neumología y Cirugía Torácica. Chart Review of Patients With
418 COPD, Using Electronic Medical Records and Artificial Intelligence (BigCOPData). Available from
419 <https://clinicaltrials.gov/ct2/show/NCT04206098>. Accessed April 13, 2020.
- 420 15. Koo D, Thacker SB. In snow's footsteps: Commentary on shoe-leather and applied
421 epidemiology. *Am J Epidemiol* 2010: 172(6): 737-739.
- 422 16. STrengthening the Reporting of OBServational studies in Epidemiology (STROBE)
423 guidance for reporting observational research. Available from [http://strobe-](http://strobe-statement.org/index.php?id=strobe-home)
424 [statement.org/index.php?id=strobe-home](http://strobe-statement.org/index.php?id=strobe-home). Accessed April 13, 2020.
- 425 17. Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, Zeng B, Li Z, Li X, Li H. Diagnosis of the
426 Coronavirus disease (COVID-19): rRT-PCR or CT? *Eur J Radiol* 2020: 126: 108961.
- 427 18. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, Tan W. Detection of SARS-CoV-2 in Different
428 Types of Clinical Specimens. *JAMA* 2020.

- 429 19. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L. Correlation of Chest CT
430 and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases.
431 *Radiology*: 0(0): 200642.
- 432 20. Xu J, Wu R, Huang H, Zheng W, Ren X, Wu N, Ji B, Lv Y, Liu Y, Mi R. Computed
433 Tomographic Imaging of 3 Patients With Coronavirus Disease 2019 Pneumonia With Negative
434 Virus Real-time Reverse-Transcription Polymerase Chain Reaction Test. *Clinical Infectious
435 Diseases* 2020.
- 436 21. Hernandez Medrano ITG, J; Belda, C; Urena, A; Salcedo, I; Espinosa-Anke, L; Saggion, H.
437 Savana: Re-using Electronic Health Records with Artificial Intelligence. *International Journal of
438 Interactive Multimedia and Artificial Intelligence* 2017: 4(7): 8-12.
- 439 22. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token
440 selection for patient phenotyping from electronic health records. *Scientific Reports* 2020:
441 10(1): 1432.
- 442 23. The Lancet. The gendered dimensions of COVID-19. *The Lancet* 2020: 395(10231):
443 1168.
- 444 24. Quinlan JR. Induction of decision trees. *Machine Learning* 1986: 1(1): 81-106.
- 445 25. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, Haller MC,
446 Heinze G, Moons KGM, Riley RD, Schuit E, Smits LJM, Snell KIE, Steyerberg EW, Wallisch C, van
447 Smeden M. Prediction models for diagnosis and prognosis of covid-19 infection: systematic
448 review and critical appraisal. *BMJ* 2020: 369: m1328.
- 449 26. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers
450 AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for
451 Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015:
452 162(1): W1-73.
- 453 27. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S, Gaymard A,
454 Bouscambert-Duchamp M, Donati F, Le Hingrat Q, Enouf V, Houhou-Fidouh N, Valette M,
455 Mailles A, Lucet JC, Mentre F, Duval X, Descamps D, Malvy D, Timsit JF, Lina B, van-der-Werf S,
456 Yazdanpanah Y. Clinical and virological data of the first cases of COVID-19 in Europe: a case
457 series. *Lancet Infect Dis* 2020.
- 458 28. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing
459 X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q,
460 Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi
461 G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early Transmission
462 Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* 2020:
463 382(13): 1199-1207.
- 464 29. McCall B. COVID-19 and artificial intelligence: protecting health-care workers and
465 curbing the spread. *The Lancet Digital Health* 2020: 2(4): e166-e167.
- 466 30. Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, Guo GY, Du J, Zheng CL, Zhu Q, Hu M,
467 Li XY, Peng P, Shi HZ. Predictors of Mortality for Patients with COVID-19 Pneumonia Caused by
468 SARS-CoV-2: A Prospective Cohort Study. *Eur Respir J* 2020.
- 469 31. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, S RNK. Predicting COVID-19
470 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot
471 Study. *JMIR Public Health Surveill* 2020: 6(2): e18828.
- 472 32. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang
473 G, Xu Q, Fang X, Zhang S, Xia J, Xia J. Artificial Intelligence Distinguishes COVID-19 from
474 Community Acquired Pneumonia on Chest CT. *Radiology* 2020: 200905.
- 475 33. The race against COVID-19. *Nat Nanotechnol* 2020: 15(4): 239-240.
- 476 34. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19
477 pandemic. *Nat Med* 2020: 26(4): 463-464.
- 478 35. World Health Organization (WHO). Coronavirus disease 2019 (COVID-19) Situation
479 Report –64. 2020. Available from <https://www.who.int/docs/default->

484 **FIGURES and TABLES**

485



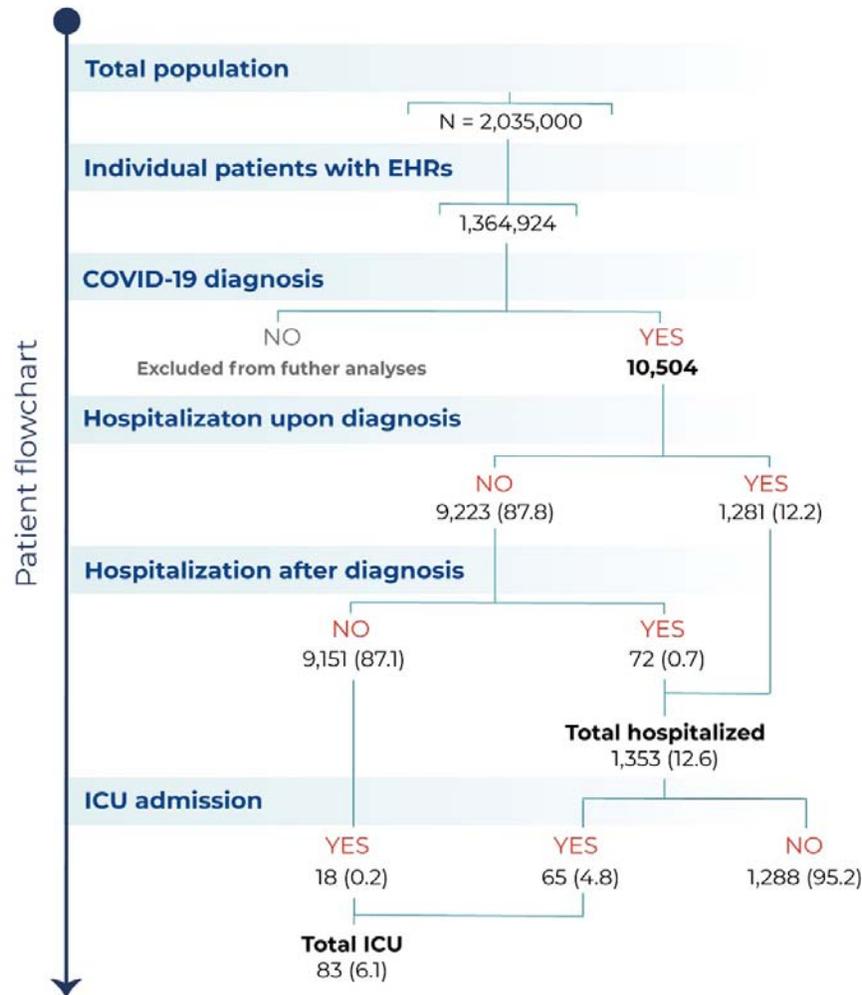
486

487

488

489 **Figure 1. Map of Castilla-La Mancha.**

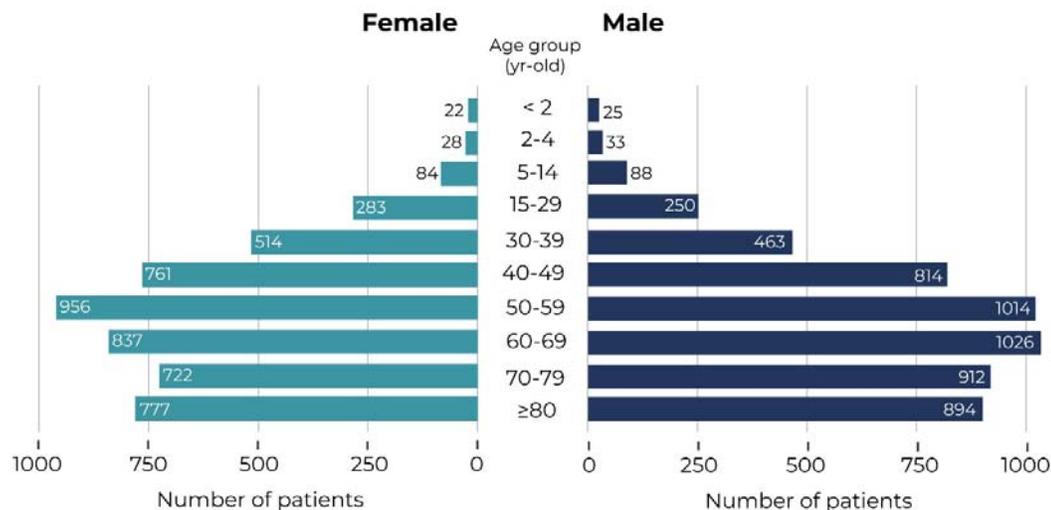
490 **Footnote:** Map of Castilla-La Mancha (red) within the Spanish (blue line) and
491 European territories. From a source general population of 2,035,000 inhabitants, we
492 collected and analyzed the clinical information in the EHRs of 1,364,924 patients within
493 the SESCAM Healthcare Network.



494

495 **Figure 2. Patient flowchart.**

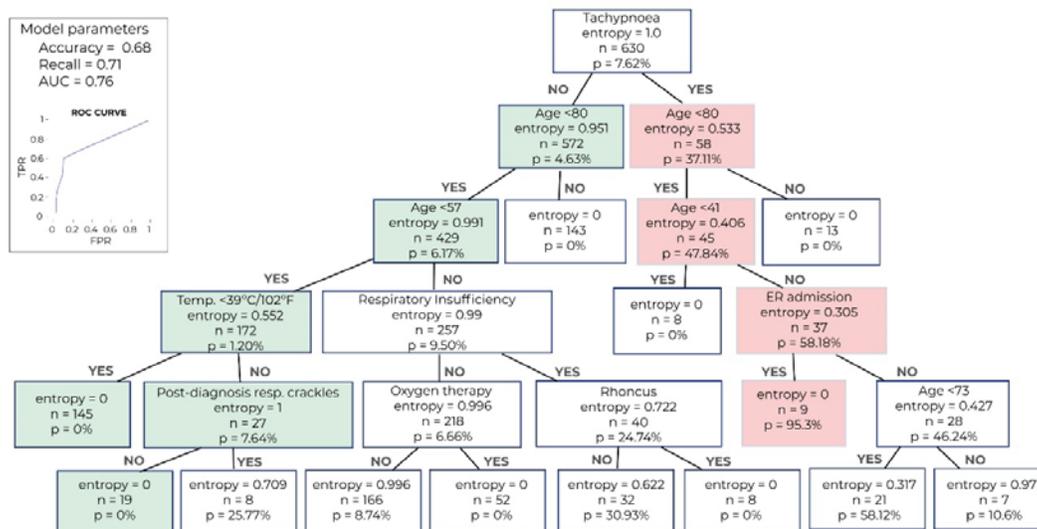
496 **Footnote:** Flowchart depicting the total number of inhabitants in the source population,
 497 the number (%) of patients with available EHRs analyzed, the number of patients
 498 diagnosed with COVID-19, and of those, the number of hospitalizations and ICU
 499 admissions.



500

501 **Figure 3. Age and Sex Distribution of COVID-19 patients.**

502 **Footnote:** Age distribution of incident cases of COVID-19 in females (left) and males
503 (right) in the study population for the period comprised between Jan 1, 2020 and March
504 29, 2020.



505

506 **Figure 4. Decision tree of relevant clinical variables for the prediction of ICU**
507 **admission in COVID-19 patients.**

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

521 **TABLE 1. Baseline demographics and clinical data upon diagnosis.**

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

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

527 **TABLE 2. Association between ICU admission and comorbidities/signs and symptoms upon**
 528 **diagnosis in patients with COVID-19.**

COMORBIDITIES				SIGNS AND SYMPTOMS			
Condition [#]	No ICU n(%)	ICU n(%)	p-value*	Sign or Symptom	No ICU n(%)	ICU n(%)	p-value*
Diabetes	1613(15.5)	33(39.8)	<0.001	Cough	5181(49.7)	62(74.7)	<0.001
Obesity	917(8.8)	19(22.9)	<0.001	Fever	4849(46.5)	55(66.3)	<0.001
Chronic respiratory disease	883(8.5)	5(6)	0.548	Dyspnoea	3246(31.1)	48(57.8)	<0.001
COPD	673(6.5)	2(2.4)	0.2029	Respiratory crackles	1904(18.3)	30(36.1)	<0.001
Asthma	750(7.2)	9(10.8)	0.2868	Myalgia	908(8.7)	11(13.3)	0.2066
OSA	211(2)	1(1.2)	0.8908	Diarrhoea	1084(10.4)	15(18.1)	0.0363
Bronchiectasis	129(1.2)	0(0)	0.6033	Dysphagia	47(0.5)	0(0)	1
Cardiovascular disease	4998(48)	60(72.3)	<0.001	Wheezing	383(3.7)	6(7.2)	0.1568
Hypertension	3487(33.5)	40(48.2)	0.0066	Tachypnoea	311(3)	27(32.5)	<0.001
Ischemic stroke	253(2.4)	1(1.2)	0.716	Chest pain	546(5.2)	8(9.6)	0.1237
Heart Disease	2604(25)	35(42.2)	<0.001	Lymphopenia	524(5)	18(21.7)	<0.001
Ischemic Heart Disease	616(5.9)	11(13.3)	0.0099	Headache	757(7.3)	7(8.4)	0.8442
Heart failure	548(5.3)	4(4.8)	1	Rhonchus	676(6.5)	17(20.5)	<0.001
Renal dysfunction	748(7.2)	16(19.3)	<0.001	Hepatomegaly	8(0.1)	0(0)	1
CKD	488(4.7)	6(7.2)	0.4059	Anosmia	297(2.9)	3(3.6)	0.9317
Chronic Liver Disease	109(1)	2(2.4)	0.502	Ageusia	65(0.6)	0(0)	0.9847
Cirrhosis	51(0.5)	0(0)	1	Neuralgia	41(0.4)	0(0)	1
Depression	699(6.7)	4(4.8)	0.6418	Sore throat	126(1.2)	1(1.2)	1
HIV	33(0.3)	1(1.2)	0.6536	Splenomegaly	21(0.2)	1(1.2)	0.4317

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