


# BMJ Open Variables associated with COVID-19 severity: an observational study of non-paediatric confirmed cases from the general population of the Basque Country, Spain

Kalliopi Vrotsou <sup>1,2</sup>, Rafael Rotaecche,<sup>3</sup> Maider Mateo-Abad,<sup>2</sup> Mónica Machón,<sup>1,2</sup> Itziar Vergara<sup>1,2</sup>

**To cite:** Vrotsou K, Rotaecche R, Mateo-Abad M, *et al.* Variables associated with COVID-19 severity: an observational study of non-paediatric confirmed cases from the general population of the Basque Country, Spain. *BMJ Open* 2021;**11**:e049066. doi:10.1136/bmjopen-2021-049066

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-049066>).

Received 14 January 2021  
Revised 04 March 2021  
Accepted 17 March 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

<sup>1</sup>Primary Care Group, Biodonostia Institute for Health Research, Donostia-San Sebastián, Spain

<sup>2</sup>Research Network in Health Services in Chronic Diseases (REDISSEC), Kronikgune Health Services Research Institute, Baracaldo, Spain

<sup>3</sup>Alza Health Center, Osakidetza-Basque Health Service, Donostia-San Sebastian, Spain

## Correspondence to

Kalliopi Vrotsou;  
kalliopi.vrotsoukanari@osakidetza.eus

## ABSTRACT

**Objectives** To investigate which were the most relevant sociodemographic and clinical variables associated with COVID-19 severity, and uncover how their inter-relations may have affected such severity.

**Design** A retrospective observational study based on electronic health record data.

**Participants** Individuals  $\geq 14$  years old with a positive PCR or serology test, between 28 February and 31 May 2020, belonging to the Basque Country (Spain) public health system. Institutionalised and individuals admitted to a hospital at home unit were excluded from the study.

**Main outcome measure** Three severity categories were established: primary care, hospital/intensive care unit admission and death.

**Results** A total of  $n=14\,197$  cases fulfilled the inclusion criteria. Most variables presented statistically significant associations with the outcome ( $p<0.0001$ ). The Classification and Regression Trees recursive partitioning methodology (based on  $n=13\,792$ ) suggested that among all associations, those with, age, sex, stratification of patient healthcare complexity, chronic consumption of blood and blood-forming organ, and nervous system drugs, as well as the total number of chronic Anatomical Therapeutic Chemical types were the most relevant. Psychosis also emerged as a potential factor.

**Conclusions** Older cases are more likely to experience more severe outcomes. However, the sex, underlying health status and chronic drug consumption may interfere and alter the ageing effect. Understanding the factors related to the outcome severity is of key importance when designing and promoting public health intervention plans for the COVID-19 pandemic.

## INTRODUCTION

In December 2019, the new coronavirus SARS-CoV-2 initiated the COVID-19 disease in China, which soon afterwards, on March 12, was declared a pandemic by the WHO. The rapid expansion of the virus, along with its high death toll and the serious health aftermaths, has rendered the COVID-19 outbreak

## Strengths and limitations of this study

- Over 13 000 confirmed COVID-19, non-institutionalised, cases aged  $\geq 14$  years old were explored.
- Electronic health records data were a valuable source of information in this study.
- The three-category outcome severity—primary care only, hospitalised/intensive care unit care and death—was studied in a joint manner.
- The Classification and Regression Trees methodology allowed exploring the big sample and the numerous variables of interest in a flexible way.
- Information on COVID-19 symptoms was not properly registered during the first pandemic wave.

as one of the worst health crises in almost a century worldwide.

Since the first infections were detected in Spain, the statistics have situated this country among the most affected in Europe, both in terms of total cases and in deaths per million people.<sup>1</sup> International literature on COVID-19 is rapidly growing.<sup>2–7</sup> The research conducted so far in Spain has focused mainly on predicting the evolution of the pandemic<sup>8</sup> or the factors associated with mortality.<sup>9</sup> Hospitalised individuals have also been described,<sup>10</sup> and the variables related to severe outcomes in these populations have been explored.<sup>11 12</sup> But so far, none of the previous works has considered the gradient of the COVID-19 severity by studying a multiple category outcome.

The autonomous community of the Basque Country, situated in the north of Spain, has its own public health system (Osakidetza), which offers sanitary coverage to some 2.3 million people. Since 2009, Osakidetza has promoted an integrated healthcare model, by

coordinating its different care levels and offering a more holistic approach on patient care.<sup>13</sup> It counts with an extensive electronic health records infrastructure, where information on patient health data and episodes of care is stored. The objective of the present observational study was to describe a big series of COVID-19-infected individuals during the first pandemic wave; establish their infection severity level, based on electronic health record data and explore what characteristics may be associated with that severity. To this end, the Classification and Regression Trees (CART) methodology was applied. This statistical technique splits the sample into mutually exclusive subgroups that share the same characteristics and can be particularly useful when analysing big data sets.<sup>14</sup>

## METHODS

### Data source and variables

All information was extracted from the electronic health records of the Basque Country Public Health System-Osakidetza, via the Osakidetza Business Intelligence tools. Data extraction covered the period between 28 February 2020 and 31 May 2020, corresponding to the first detected case in the Basque Country and the end of the first pandemic wave in Spain. Only the health records of individuals  $\geq 14$  years old with a COVID-19 positive PCR or antibody test were included, as no antigen tests were performed at that time. Data of cases living in residential homes or those admitted to a hospital at home unit were excluded.

The following variables were studied: age, sex and income level derived by the pharmaceutical co-payment scheme ( $<18000\text{€}$ ,  $18000\text{--}100\,000\text{€}$ ,  $>100\,000\text{€}$ ). Chronic medication consumption was explored using the Anatomical Therapeutic Chemical (ATC) system at the first level (<https://www.who.int/tools/atc-ddd-toolkit>). Polypharmacy, defined as the consumption of five or more chronic drugs, and the number of ATC types consumed were derived. Chronic pathologies based on the International Classification of Diseases (ICD)-9 codes, COVID-19 symptoms registered during consultations and influenza vaccination in the year 2019 were also considered. The Osakidetza stratification according to patient healthcare complexity was studied. Based on a series of health data and the use of health services during the previous year, this variable classifies individuals into four categories, ranging from less to more severe: prevention and promotion of healthy population, self-management support, disease management and case management. Pluripathological individuals belong to the last category. This classification is renewed at the beginning of every calendar year, for all individuals  $\geq 14$  years registered in the Osakidetza system at least during the previous 6 months. A detailed description can be found elsewhere.<sup>15</sup>

Given that the data were anonymous and clinical analyses could not be conducted, it was assumed that the severity of a case would be indicated by the most demanding level of medical attention received, within

the study period. Four severity levels were initially identified: primary care attention only (PC), hospitalisation without intensive care unit (ICU) admission (hospital), ICU admission (ICU) and death. During the pandemic, several emergency ICU units were set up within hospitals across the Basque Country. Nevertheless, this information was not reflected in the electronic health records. As a result, cases admitted to such ICUs were registered as hospital admissions. This fact imposed the necessity to merge hospital and ICU admissions into one category in the current work. Cases meeting the inclusion criteria were considered only once in the current analyses.

### Patient and public involvement

Due to the study design, no patient and public involvement was considered. Nonetheless, two of the authors are medical doctors, who have offered valuable support during this work.

### Statistical analysis

Continuous variables are presented as means with SDs, while medians and IQRs (Q1–Q3) are given for discrete variables. Categorical variables are presented with frequencies and percentages (%). Three-group unadjusted comparisons were performed with the one-way analysis of variance, Kruskal-Wallis and  $\chi^2$  test, respectively. The Jonckheere-Terpstra and Mantel-Haenszel  $\chi^2$ , both testing for a trend along the three severity groups, were additionally tested.<sup>16</sup>

### Classification and Regression Trees

The CART methodology is a non-parametric statistical tool, which can be very useful when handling big data sets with many variables. This statistical technique partitions the sample into smaller homogeneous groups that share the same characteristics. The splitting process starts considering the whole sample that is then recursively partitioned into mutually exclusive subsamples according to the most important variables, selected among all candidate variables. Important variables in CART are those that minimise the variability of the outcome within each subsample. This process results in a tree-like structure with multiple levels, which offers a visual representation of which variables affect the outcome the most. At the same time, it allows understanding the inter-relations the indicated factors may have with one another. CART analysis is a flexible option for data sets with correlated variables, as in our case.<sup>14 17</sup>

The starting point of the tree structure is the root node and each split is an offspring node. Offsprings that do not split any further are called terminal. In the current analyses, splitting was based on the entropy criterion and each variable was allowed only once per tree branch. For a stopping rule, the number of terminal nodes and the observations included in each of them were considered. A tree with 10 terminal nodes, each including at least 1% of the valid sample data was selected. Cost-complexity pruning was applied. Variables with significance levels

**Table 1** Baseline information of the COVID-19 cases during the first wave of the pandemic

Variables	Total (n=14 197)	Primary care (n=9722)	Hospital/ICU (n=3710)	Death (n=765)	P value
Age; mean (SD)	53.7 (17.4)	48.0 (14.4)	62.8 (16.1)	82.3 (10.5)	<0.0001
Sex					
Male	5520 (38.9)	3073 (31.6)	2031 (54.7)	416 (54.4)	<0.0001
Female	8677 (61.1)	6649 (68.4)	1679 (45.3)	349 (45.6)	
Healthcare complexity					
Missing information	405 (2.9)	307 (3.2)	86 (2.3)	12 (1.6)	
Prevention and promotion	3878 (27.3)	3399 (36.1)	470 (12.9)	9 (1.2)	<0.0001
Self-management support	6821 (48.0)	4989 (52.9)	1675 (46.2)	157 (20.8)	
Disease management	2252 (15.9)	891 (9.4)	1050 (28.9)	311 (41.3)	
Case management	841 (5.9)	136 (1.4)	429 (11.8)	276 (36.6)	
Income level					
Missing information	854 (6.0)	251 (2.6)	130 (3.5)	473 (61.8)	
<18 000€	6536 (46.0)	4297 (45.3)	2038 (56.9)	201 (26.8)	<0.0001
18 000–100 000€	6670 (47.0)	5074 (53.5)	1507 (42.0)	89 (11.6)	
>100 000€	137 (1.0)	100 (1.0)	35 (0.9)	2 (0.3)	
Flu vaccination in 2019: yes					
All vaccinated cases	3336 (23.5)	1322 (13.6)	1446 (39.0)	568 (74.2)	<0.0001
Vaccinated cases <65 years old	1103 (10.1)	814 (9.2)	265 (13.6)	24 (3.1)	<0.0001
Vaccinated cases ≥65 years old	2233 (66.5)	508 (57.7)	1181 (66.9)	544 (76.7)	<0.0001

Data are frequency (percentage), unless otherwise stated. For variables with missing information, percentages and statistical comparisons are based on valid data only. Presented p values are based on one-way ANOVA for the variable of age and the  $\chi^2$  test for the categorical variables. Cases <65 years and ≥65 years were n=10843 and n=3354, respectively. Jonckheere-Terpstra and Mantel-Haenszel  $\chi^2$  test for trend also resulted in p<0.0001 in all comparisons.

ANOVA, analysis of variance; ICU, intensive care unit.

p>0.010 in the three-group comparisons and those with a total frequency <1% of the valid sample were excluded from the CART stage, while missing data were omitted.<sup>14</sup> Analyses were performed with the SAS software V.9.4 (Copyright 2016 by SAS Institute). The SAS proc hpsplit function was used for tree construction.

## RESULTS

A total of n=14 197 COVID-19 cases fulfilled the inclusion criteria. Of these, n=9722 (68.5%) received PC attention only, n=3710 (26.1%) had a hospital or ICU admission (n=3630 and 80, respectively), and n=765 died (5.4%). Most cases were detected via PCR (n=8933), and this detection method was the most prevalent in all three outcome groups (PC: 51.0%, hospital/ICU: 87.7%, death: 93.3%). **Table 1** presents the baseline information of the sample. Overall, mean age was 53.7 (SD: 17.4) years, and it increased with outcome severity. Most infected cases were women, but at the same time this sex group presented lower infection severity. In particular, women were more prevalent in PC (68.4%), whereas more men were observed in the hospital/ICU and death groups. As far as the healthcare complexity stratification variable

was concerned, the PC outcome group presented the highest percentage of healthy individuals (36.1%), while case management was most prevalent in the death outcome group (36.6%). Based on the available information, individuals with an annual income <18 000€ were more prevalent in the hospital/ICU and death groups, and those with higher income received mostly PC attention. Finally, the death group had the highest percentage of individuals with an influenza vaccination in the previous year. This observation was consistent for cases <65 and ≥65 years of age, even though the corresponding percentages of the older cases were higher. All comparisons were statistically significant.

Chronic medication consumption data are presented in **table 2**. Overall, the most consumed medications were those for the nervous system (38.7%), alimentary tract and metabolism (33.0%), and cardiovascular system (30.2%). With the exception of musculoskeletal system and antiparasitic products, insecticides and repellents, an increasing consumption trend with severity was observed in all other ATC types. The consumption of alimentary tract and metabolism disorders (A), blood and blood-forming organs (B), cardiovascular system (C) and nervous system diseases drugs (N) exceeded 60% in the

**Table 2** Chronic medication consumption of the COVID-19 sample

	Total (n=14 197)	Primary care (n=9722)	Hospital/ICU (n=3710)	Death (n=765)	P value
Medication (ATC type)					
Alimentary tract and metabolism (A)	4685 (33.0)	2234 (23.0)	1837 (49.5)	614 (80.3)	<0.0001
Blood and blood-forming organs (B)	2414 (17.0)	889 (9.1)	1057 (28.5)	468 (61.2)	<0.0001
Cardiovascular system (C)	4294 (30.2)	1813 (18.6)	1893 (51.0)	588 (76.9)	<0.0001
Dermatologicals (D)	1765 (12.4)	1032 (10.6)	581 (15.7)	152 (19.9)	<0.0001
Genitourinary system and sex hormones (G)	1690 (11.9)	1050 (10.8)	505 (13.6)	135 (17.6)	<0.0001
Systemic hormonal preparations, excluding sex hormones and insulins (H)	1504 (10.6)	876 (9.0)	492 (13.3)	136 (17.8)	<0.0001
Anti-infectives for systemic use (J)	223 (1.6)	122 (1.3)	73 (2.0)	28 (3.7)	<0.0001
Antineoplastic and immunomodulating agents (L)	360 (2.5)	165 (1.7)	141 (3.8)	54 (7.1)	<0.0001
Musculoskeletal system (M)	3137 (22.1)	2010 (20.7)	952 (25.7)	175 (22.9)	<0.0001
Nervous system (N)	5494 (38.7)	2906 (29.9)	1931 (52.0)	657 (85.9)	<0.0001
Antiparasitic products, insecticides and repellents (P)	42 (0.3)	24 (0.2)	15 (0.4)	3 (0.4)	0.284
Respiratory system (R)	2603 (18.3)	1517 (15.6)	864 (23.3)	222 (29.0)	<0.0001
Sensory organs (S)	863 (6.1)	443 (4.6)	297 (8.0)	123 (16.1)	<0.0001
Various (V)	188 (1.3)	45 (0.5)	58 (1.6)	85 (11.1)	<0.0001
Polypharmacy: yes	2921 (20.5)	935 (9.6)	1357 (36.5)	629 (82.2)	<0.0001
Number of ATC types consumed: median (Q1–Q3)	2 (0–3)	1 (0–3)	3 (1–4)	5 (3–6)	<0.0001

Data are frequency (percentage), unless otherwise stated. Q1–Q3: interquartile range values.

Presented p values are based on the  $\chi^2$  test. The Jonckheere-Terpstra and Mantel-Haenszel  $\chi^2$  test for trend resulted in very similar results in all comparisons.

ATC, Anatomical Therapeutic Chemical; ICU, intensive care unit.

death group. Both polypharmacy and the number of ATC types consumed were associated with infection severity.

Regarding the chronic diseases, the most prevalent condition was related to mental pathologies (table 3). In particular, 30% of the sample had received a diagnosis corresponding to the ICD-9 neurotic, personality or other non-psychotic mental disorders. Hypertension was the next more prevalent condition (21%), followed by diseases of the blood and blood-forming organs (11.3%), diseases of the oesophagus, stomach and duodenum (10.4%). Diabetes mellitus was present in 8.5% of the sample. With the exception of neurotic, personality or other non-psychotic conditions that presented the same distribution along the three outcome groups, the prevalence of the most frequent pathologies increased with COVID-19 severity. A similar trend was seen in the total number of chronic diseases. Non-infectious enteritis and colitis, and allergic asthma were the only chronic conditions presenting a descending prevalence with outcome severity, but percentage differences were low.

### Classification and Regression Trees

The CART process indicated that age, sex, healthcare complexity stratification, the ATC categories of blood and blood-forming organ medication (B), as well as nervous

system drugs (N) along with the frequency of ATC types consumed would be the most relevant variables in understanding the main case characteristics associated to the outcome. During this process, the variable of psychosis was also flagged as important. In spite of its low prevalence (2.9%), psychosis was given a lot of weight in the older section of the population. The inclusion of this pathology resulted in a less parsimonious model; with ATC-N drugs placed in an additional tree level. Nonetheless, given that psychosis was the single variable resulting in a node with a death majority, and that other authors have already suggested an association between antipsychotic drugs and mortality in COVID-19 cases,<sup>9</sup> presenting the corresponding findings was considered of relevance. Therefore, the CART process was repeated twice, first excluding and afterwards including psychosis.

### Excluding psychosis

The tree generated by the CART process is depicted in figure 1. Most cases <64.7 years of age (81.4%, node 1) received mainly PC attention. In this tree branch, men presented 15.3% more hospital/ICU compared with women. Among men, those with worse health (node 8) had 19.2% more hospital/ICU admissions, compared with the rest (node 7). The majority of men with worse

**Table 3** Chronic diseases of the COVID-19 cases in the three outcome groups

Disease	Total (n=14 197)	Primary care (n=9722)	Hospital/ICU (n=3710)	Death (n=765)	P value
<b>Infectious disease</b>					
HIV infection	23 (0.2)	7 (0.1)	12 (0.3)	4 (0.5)	0.0002
Liver disease and cirrhosis	133 (0.9)	49 (0.5)	72 (1.9)	12 (1.6)	<0.0001
Malignant neoplasm	918 (6.4)	364 (3.7)	410 (11.0)	144 (18.8)	<0.0001
<b>Endocrine diseases</b>					
Subclinical hypothyroidism without treatment	1101 (7.8)	747 (7.7)	294 (7.9)	60 (7.8)	0.892
Diabetes mellitus	1213 (8.5)	395 (4.1)	606 (16.3)	212 (27.7)	<0.0001
Diseases of the blood and blood-forming organs	1602 (11.3)	930 (9.6)	492 (13.3)	180 (23.5)	<0.0001
<b>Mental disorders</b>					
Psychosis	412 (2.9)	138 (1.4)	143 (3.9)	131 (17.1)	<0.0001
Neurotic disorders, personality disorders and other non-psychotic mental disorders	4258 (30.0)	2926 (30.1)	1096 (29.5)	236 (30.8)	0.712
Mental retardation	39 (0.3)	24 (0.2)	14 (0.4)	1 (0.1)	0.319
<b>Nervous system diseases</b>					
Dementia	126 (0.8)	20 (0.2)	32 (0.8)	74 (9.6)	<0.0001
Other hereditary and degenerative diseases of the central nervous system	307 (2.1)	127 (1.3)	122 (3.2)	58 (7.5)	<0.0001
<b>Diseases of the circulatory system</b>					
Hypertensive disease	2988 (21.0)	1177 (12.1)	1364 (36.8)	447 (58.4)	<0.0001
Ischaemic heart disease	448 (3.2)	111 (1.1)	237 (6.4)	100 (13.1)	<0.0001
Cerebrovascular disease	611 (4.3)	189 (1.9)	266 (7.2)	156 (20.4)	<0.0001
Heart failure and atrial fibrillation and flutter	709 (5.0)	132 (1.4)	361 (9.7)	216 (28.2)	<0.0001
Acute pulmonary heart disease and other venous embolism and thrombosis	150 (1.1)	49 (0.5)	73 (2.0)	28 (3.7)	<0.0001
Arterial embolism and thrombosis	39 (0.3)	17 (0.2)	17 (0.5)	5 (0.7)	0.002
<b>Respiratory disease</b>					
Allergic asthma	354 (2.4)	258 (2.6)	88 (2.3)	8 (1.0)	0.019
Chronic obstructive pulmonary disease and allied conditions (excl. allergic asthma)	1190 (8.3)	630 (6.4)	432 (11.6)	128 (16.7)	<0.0001
Pneumoconioses and other lung diseases due to external agents	20 (0.1)	9 (0.1)	8 (0.2)	3 (0.4)	0.038
<b>Diseases of the digestive system</b>					
Diseases of oesophagus, stomach and duodenum	1481 (10.4)	907 (9.3)	468 (12.6)	106 (13.9)	<0.0001
Non-infectious enteritis and colitis	643 (4.5)	500 (5.1)	121 (3.3)	22 (2.9)	<0.0001
Regional enteritis and ulcerative colitis	73 (0.5)	51 (0.5)	16 (0.4)	6 (0.8)	0.447
<b>Disease of the genitourinary system</b>					
Chronic kidney disease	398 (2.8)	87 (0.9)	188 (5.1)	123 (16.1)	<0.0001
<b>Diseases of the skin and subcutaneous tissue</b>					
Psoriasis	315 (2.2)	180 (1.9)	113 (3.0)	22 (2.9)	<0.0001
<b>Diseases of the musculoskeletal system and connective tissue</b>					
Systemic lupus erythematosus	36 (0.3)	24 (0.2)	10 (0.3)	2 (0.3)	0.972
Rheumatoid arthritis and other inflammatory polyarthropathies	125 (0.9)	59 (0.6)	55 (1.5)	11 (1.4)	<0.0001
Arthropathy associated with other disorders classified elsewhere	8 (0.1)	5 (0.1)	1 (0.0)	2 (0.3)	0.042

Continued

Table 3 Continued

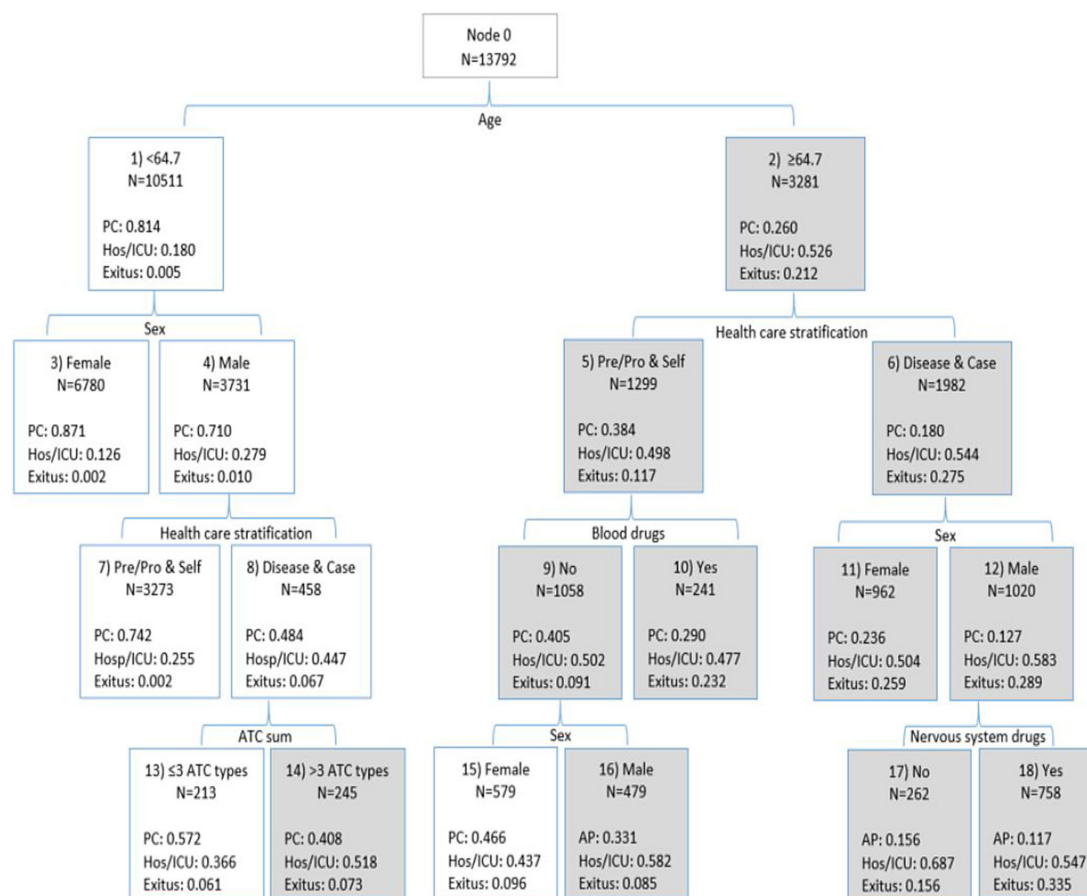
Disease	Total (n=14 197)	Primary care (n=9722)	Hospital/ICU (n=3710)	Death (n=765)	P value
Multimorbidity: $\geq 2$ chronic diseases	5326 (37.5)	2715 (27.9)	1975 (53.2)	636 (83.1)	<0.0001
Total number of chronic diseases					
Median (Q1–Q3)	1 (0–2)	1 (0–2)	2 (1–3)	3 (2–4)	<0.0001

Data are frequency (percentage), unless otherwise stated; Q1–Q3: interquartile range values. Presented p values are based on the  $\chi^2$  test. The Jonckheere-Terpstra and Mantel-Haenszel  $\chi^2$  test for trend resulted in very similar results in all comparisons. ICU, intensive care unit.

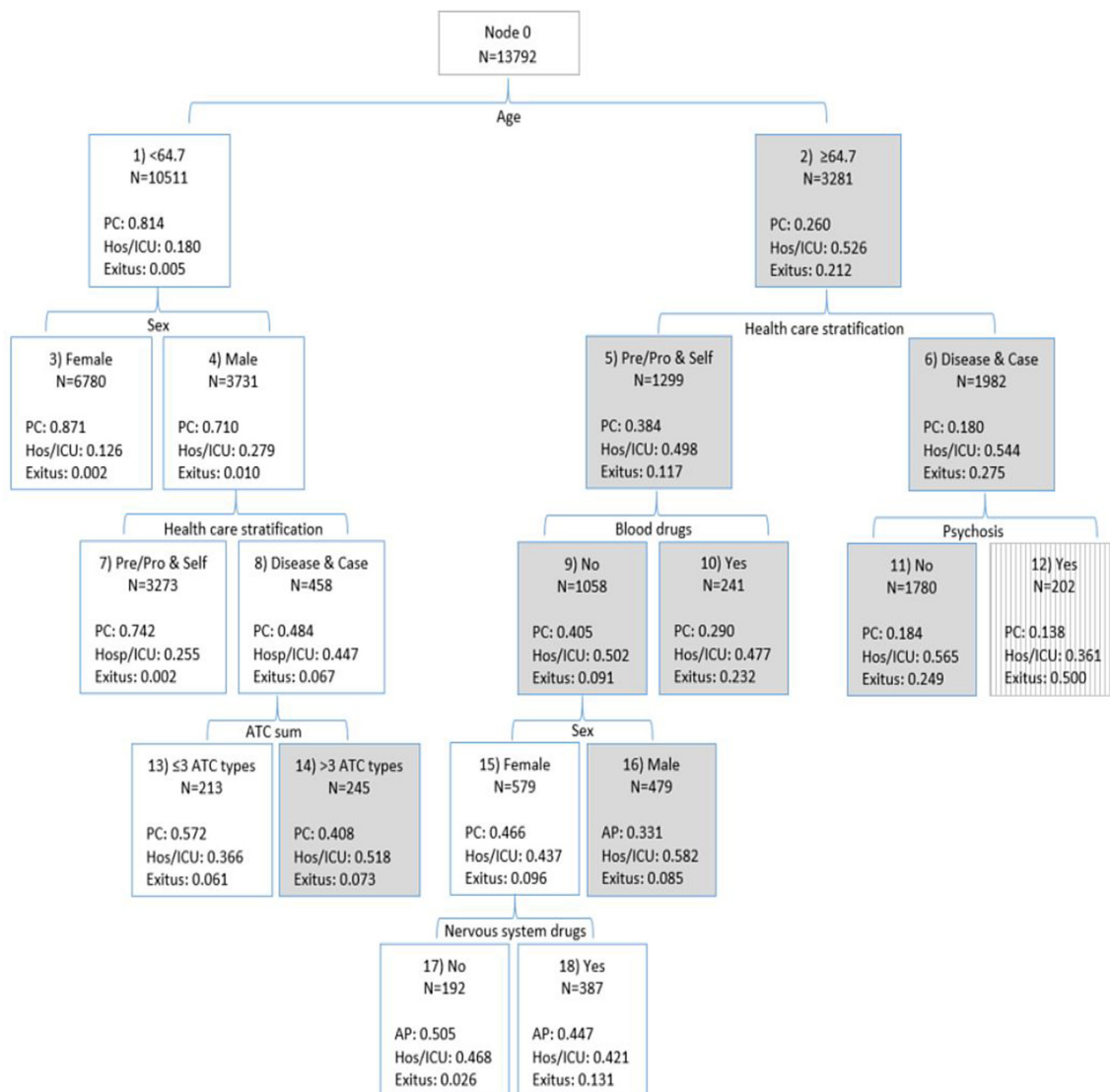
baseline health status who consumed  $\geq 3$  ATC types experienced a hospital/ICU admission.

Cases  $\geq 64.7$  years of age had mainly a hospital/ICU outcome (52.6%), with a considerable death prevalence (21.2%). Those with worse baseline health (node 6) had 4.6% more hospital/ICU admissions and 15.8% more deaths, compared with the rest. Cases with better baseline health status (node 5) were further split according to ATC-B consumption. Death for blood and blood-forming organ drug consumers was experienced in 23.2% of the cases, with the same outcome being 9.1% in non-consumers. Within this last group (node 9),

the majority of women received PC attention, while hospital/ICU was the most prevalent outcome in men. In a similar way, among node 6 cases, men presented worse evolution than women. Finally, men consuming chronic medications for the nervous system (node 18) had 17.9% more deaths compared with non-consumers. The 10 terminal nodes can easily be ordered from less severe (ie,  $<64.7$  year-old women, node 3) to most severe outcome groups (ie,  $\geq 64.7$  year-old disease and case management men who consume nervous system drugs, node 18).



**Figure 1** CART model without psychosis. Percentages up to three decimal places are given. Due to rounding, node percentages may not add to 100%. Pre/Pro & Self: Prevention and promotion, and Self-management support. Disease & Case: Disease management and Case management. White nodes represent groups with a higher percentage of PC care. Grey nodes represent groups with a higher percentage of hospital/ICU admission. ATC, Anatomical Therapeutic Chemical; CART, Classification and Regression Trees; Hos/ICU, hospital and intensive care unit; PC, primary care.



**Figure 2** CART model with psychosis. Percentages up to three decimal places are given. Due to rounding, node percentages may not add to 100%. Pre/Pro & Self: Prevention and promotion, and Self-management support. Disease & Case: Disease management and Case management. White nodes represent groups with a higher percentage of PC care. Grey nodes represent groups with a higher percentage of hospital/ICU admission, and grey-striped node groups with a higher percentage of death. ATC, Anatomical Therapeutic Chemical; CART, Classification and Regression Trees; Hos/ICU, hospital and intensive care unit; PC, primary care.

### Including psychosis

The resulting CART model when the variable of psychosis was included in the recursive process is presented in [figure 2](#). Psychosis was one of the main variables of this model and the single split variable for node 6. Inclusion of this pathology added one more level to the CART tree, with chronic nervous system drugs being a split variable for node 15. Cases with psychosis had a 50% death. The ATC-N consumers presented less PC and higher death compared with non-consumers. No other changes were observed compared with the [figure 1](#) model. In this case, the most severe outcome group was ≥64.7 year-old disease and case management cases with psychosis (node 12).

### DISCUSSION

The present work has studied the sociodemographic and clinical characteristics of a big number of Spanish COVID-19 cases of the first pandemic wave. According to the information extracted from electronic health record data, the variables of age, sex, previous pathologies and chronic drug consumption may be decisive in understanding infection severity.

Both age and male sex have been flagged as important risk factors by previous COVID-19 research.<sup>2 3 7 9 11 12 18</sup> The importance of age is probably undisputable, given the deterioration of the body's immunity mechanisms and the loss of its capacity to adapt to the environment.<sup>19</sup> The present data appear to reflect this known ageing effect. In relation to the variable of sex, women presented

consistently higher PC and lower hospital/ICU in the splits where sex was present. Data from various countries are suggesting that women have better COVID-19 infection outcomes than men.<sup>7 20</sup> Women are considered to have stronger immunity systems.<sup>21</sup> Even though the exact mechanisms responsible for these differences in the COVID-19 context are still unclear and probably multifactorial,<sup>20</sup> certain works are hypothesising that low androgen levels can have a protective role against this disease.<sup>22</sup> The current data, in conjunction with previous evidence, call for a better understanding of the role of sex in the current pandemic. Sex-specific analyses of future wave data should be planned. But more importantly, high-quality prospective studies collecting sex-disaggregated data are needed.<sup>23</sup>

The healthcare complexity stratification variable was present in both main tree arms. It should be mentioned that the way CART divided this four-category variable into a binary one, by merging the two less severe versus the two more severe groups, was imposed by the data, not the investigators. Worse health status at the time of the infection was associated with more hospitalisations for younger cases, and mainly to more deaths among older individuals. The inclusion of this stratification variable in the CART model is a relevant finding. Tools that stratify the general population, identifying those at greater risk, can be an asset for public health prevention programmes. In the COVID-19 literature, the stratification approach has so far mainly focused on hospitalised patients.<sup>12 24 25</sup> While one meta-analysis of in-hospital cases claimed that in COVID-19 infections, underlying health conditions are even more important than age.<sup>26</sup> Our data suggest that, at least at the local level, this very stratification variable can offer valuable information and its implementation may worth be considered when setting up public health action plans. The study of similar indicators used in other health systems would be encouraged.

As far as the drug consumption was concerned, chronic blood and blood-forming organ drugs (B) and drugs for the nervous system (N), both appeared as important variables for cases  $\geq 64.7$  years of age. Cases consuming those drugs presented higher severity levels. ATC-N was the most frequent medication across all three outcome groups. ATC-B had the steepest raising in consumption from one severity level to the next. Several neurological manifestations after a COVID-19 infection have been described in the literature, with the virus perceived by certain authors as a threat for the whole nervous system.<sup>27</sup> It is probable that individuals already suffering from chronic neurological conditions may be indeed more likely to present worse outcomes once infected.<sup>28 29</sup> Blood-related parameters like systolic and diastolic pressure, red and white cell counts, platelets, lymphocytes, among others, have been highlighted as significant predictors in different COVID-19 diagnostic models.<sup>7</sup> An association between certain ATC-B drugs and higher odds of death in infected cases has also been observed.<sup>9</sup> Chronic anticoagulation treatment is referenced as protective against COVID-19

mortality by some,<sup>30</sup> and ineffective by others.<sup>31</sup> COVID-19 cases present a high frequency of thrombotic events, which is leading to an expansion of anticoagulation drug use when treating the disease.<sup>32</sup> But in patients already receiving such drugs prior to infection, drug–drug interactions and infection severity should be carefully assessed before any antiviral therapy is given, or switching from oral to parenteral antithrombotic administration.<sup>33</sup> Worse severity seen among ATC-B consumers in the current data may reflect also an increased risk for patients already under anticoagulation therapy. Poor outcomes due to therapeutic decisions and drug–drug interactions cannot be excluded either. Our continuing COVID-19 work will refine future data explorations. Obtaining, for example, ATC data at the second or third level, as well as information of inpatient treatments, will offer more insight into these associations.

Psychosis was a relevant variable in the CART process. Antipsychotic drugs belong to the ATC-N medication type, which is probably why allowing for the inclusion of psychosis relocated this drug group further down in the tree structure. Older patients with worse baseline health and psychosis had the highest death rate among all CART nodes. We can only hypothesise over the mechanisms that could explain such a finding. On one hand, individuals with psychotic disorders present excess mortality compared with the general population, mainly due to lifestyle choices, associated comorbidities and medication side effects.<sup>34</sup> On the other hand, the treatment management of these cases is challenging as alteration or abrupt cessation of their current medication could potentially lead to a sudden health deterioration or even death.<sup>35</sup> This could happen, for example, during hospital and ICU admissions. In the present sample, 75% of the deaths seen in the psychosis node had been admitted to a hospital during the study period. The available information does not allow knowing whether death took place during the admissions, neither the inpatient treatment regime. An observational US study of >60 000 cases claimed that psychiatric disorders are a risk factor associated with higher COVID-19 diagnosis; with psychosis presenting greater risk ratios versus mood and anxiety disorders. The same study also reported an increased risk of first-time psychiatric disorders for survivors.<sup>36</sup> Others have suggested that antipsychotic use<sup>9</sup> and schizophrenia spectrum disorders<sup>37</sup> are associated with higher COVID-19 mortality. Even though more research in this direction is required, the available data seem to highlight the need for a close monitoring of cases with psychiatric disorders.

The total number of chronically consumed ATC types was an important variable among cases <64.7 years of age. This variable, which could also be perceived as an indicator of the associated comorbidities, stresses even more the importance that underlying pathologies may have in determining the severity of the infection outcome.<sup>26</sup>

In this work, a surrogate outcome variable has been used. Assuming that more intensive care levels represented worse COVID-19 status is a decision also taken by



previous authors.<sup>11 38–40</sup> The available data do not allow studying if admissions and deaths may have been due to other health problems. The female prevalence of this sample was greater than that seen in other COVID-19 publications,<sup>3 4 7</sup> but nonetheless similar to previous studies performed in this country.<sup>9 11</sup> In the Spanish reality, women traditionally assume the caretaker's role for younger and older members of their families, while they also occupy more home-assisting jobs<sup>41</sup> and health-related professions.<sup>42</sup> All these conditions may imply higher exposure rates to the virus, which may offer a possible explanation for the sample's sex distribution.

The current study has certain limitations. The implemented information is based exclusively on electronic health record data within the previously defined dates. After that period, the severity of certain cases may have worsened. Nonetheless, the end study date corresponds to the end of the first COVID-19 wave in our area, where new infections and deaths were very low. This, in combination with the big study sample, should have minimised the effect of possible outcome variations. No COVID-19 symptoms are presented. An attempt to register these symptoms was incorporated at the Osakidetza electronic records early on after the outbreak. But the number of symptoms and registration format evolved over the study period; PC and hospital registrations differed; the medical staff mostly annotated symptoms in text format; while most importantly such registration was totally missing in many cases. During analysis, an effort to recode text annotations and homogenise information from primary care and hospital data was made. In spite of that, and due to the frequency of missing values, the representativeness of the corresponding data could not be assumed. Symptoms are probably more relevant for algorithms discriminating cases from non-cases.<sup>43</sup> During the first pandemic wave, no massive population testings were performed in Spain; but at the end of that wave, serology tests were administered to the health professionals and allied services of our geographical area. Thus, identified cases were either symptomatic, close contacts of cases or individuals working in the health sector. However, the profession of the cases was not an available piece of information in this sample. Working with health records makes recovering missing data or refining variable information a very difficult task. This was also the case with the income level. Its broad categories may have obscured a more appropriate exploration. On the other hand, the high frequency of missing income level data seen in the death group is due to the 'unsubscriptions' of the dead cases from the medication dispensing registry. It is important to note that the target of the Basque public health system is a health coverage based on the health needs and not the earnings of the individuals.

One of the main strengths of this study is its big sample size. The consideration of three outcome groups is another advantage, which allows for a better visualisation of the different severity levels of the disease. Finally, implementing the CART methodology assisted in

translating a complex and multifactorial reality into an easy-to-follow picture. Our findings make clinical sense and are supported by previous evidence. They appear to endorse the need for public health prevention plans that consider population characteristics. At the same time, they highlight that for a multifactorial problem to be properly treated, not only the factors affecting it, but also the inter-relations between the latter should be thoroughly studied. The COVID-19 pandemic may be a new starting point in the public health paradigm. The necessity for public health promoters to work hand in hand with investigators and data analysts has become indisputable under the current circumstances. Prevention plans should be based on rigorous data and understanding of the latter. This is the only way to assure that possible reorganisation and estimation of future resources can reach optimal results.

**Contributors** IV, RR and MM planned this study and obtained the permission for exploring the corresponding data by the Osakidetza central services. MM-A set the filters and performed the data extraction of the electronic health record data. KV and MM-A were both responsible for data cleaning and recoding. KV and MM-A performed all statistical analyses. The input of IV and RR has assured a clinically meaningful perspective of all presented analyses and results. MM performed literature searches. KV drafted the first manuscript version. All authors read and contributed to the consecutive manuscript versions.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not required.

**Ethics approval** The project has been approved by the ethics committee CEIm de Euskadi on 22 July 2020 (reference code: PI2020087).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data may be obtained from a third party and are not publicly available. The data of the current study are stored in a server of our institution. Sharing them with external investigators will be evaluated on an individual basis and will require an approval by the Osakidetza central services. The corresponding author should be contacted.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iD

Kalliopi Vrotsou <http://orcid.org/0000-0002-3296-3923>

#### REFERENCES

- 1 Worldometer. Coronavirus update (live): 109,735,851 cases and 2,420,401 deaths from COVID-19 virus pandemic, 2016. Available: <https://www.worldometers.info/coronavirus/>
- 2 Docherty AB, Harrison EM, Green CA. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC who clinical characterisation protocol: prospective observational cohort study. Available: <https://isaric4c.net>
- 3 Nachtigall I, Lenga P, Józwiak K, *et al*. Clinical course and factors associated with outcomes among 1904 patients hospitalized with COVID-19 in Germany: an observational study. *Clin Microbiol Infect* 2020;26:1663–9. doi:10.1016/j.cmi.2020.08.011
- 4 Hewitt J, Carter B, Vilches-Moraga A, *et al*. The effect of frailty on survival in patients with COVID-19 (COPE): a multicentre, European,

- observational cohort study. *Lancet Public Health* 2020;5:e444–51 [www.thelancet.com/](http://www.thelancet.com/)
- 5 Baqui P, Bica I, Marra V, *et al.* Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. *Lancet Glob Health* 2020;8:e1018–26. doi:10.1016/S2214-109X(20)30285-0
  - 6 Jones RC, Ho JC, Kearney H, *et al.* Evaluating trends in COVID-19 research activity in early 2020: the creation and utilization of a novel open-access database. *Cureus* 2020;12:e9943.
  - 7 Wynants L, Van Calster B, Collins GS, *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
  - 8 Aguiar M, Ortuondo EM, Bidaurrezaga Van-Dierdonck J, *et al.* Modelling COVID 19 in the Basque country from introduction to control measure response. *Sci Rep* 2020;10:17306.
  - 9 Poblador-Plou B, Carmona-Pírez J, Ioakeim-Skoufa I, *et al.* Baseline chronic comorbidity and mortality in laboratory-confirmed COVID-19 cases: results from the PRECOVID study in Spain. *Int J Environ Res Public Health* 2020;17:5171–14.
  - 10 Casas-Rojo JM, Antón-Santos JM, Millán-Núñez-Cortés J, *et al.* Clinical characteristics of patients hospitalized with COVID-19 in Spain: results from the SEMI-COVID-19 registry. *Rev Clin Esp* 2020;220:480–94.
  - 11 Working group for the surveillance and control of COVID-19 in Spain. The first wave of the COVID-19 pandemic in Spain : characterisation of cases and risk factors for severe outcomes, as at 27 April 2020. *Eurosurveillance* 2020;25.
  - 12 Rubio-Rivas M, Corbella X, Mora-Luján JM, *et al.* Predicting clinical outcome with phenotypic clusters in COVID-19 pneumonia: an analysis of 12,066 hospitalized patients from the Spanish registry SEMI-COVID-19. *J Clin Med* 2020;9:3488.
  - 13 Bengoa R. Transforming health care: an approach to system-wide implementation. *Int J Integr Care* 2013;13:e039.
  - 14 Zhang H, Singer BH. *Recursive partitioning and applications. second ed.* New York: Springer Series in Statistics, 2010.
  - 15 Orueta JF, Nuño-Solinis R, Mateos M, *et al.* Predictive risk modelling in the Spanish population: a cross-sectional study. *BMC Health Serv Res* 2013;13:1.
  - 16 Walker GA. *Common statistical methods for clinical research with SAS examples.* Second Edition. Cary, NC: SAS Institute, 2002.
  - 17 Lemon SC, Roy J, Clark MA, *et al.* Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Ann Behav Med* 2003;26:172–81.
  - 18 Wollenstein-Betech S, Cassandras CG, Paschalidis IC. Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: hospitalizations, mortality, and the need for an ICU or ventilator. *Int J Med Inform* 2020;142:104258.
  - 19 WHO Team. Social determinants of health. world report on ageing and health, 2015. Available: <https://www.who.int/publications/i/item/world-report-on-ageing-and-health>
  - 20 Gebhard C, Regitz-Zagrosek V, Neuhauser HK, *et al.* Impact of sex and gender on COVID-19 outcomes in Europe. *Biol Sex Differ* 2020;11.
  - 21 Moulton VR, Rider V, Martocchia A. Sex hormones in acquired immunity and autoimmune disease. *Front Immunol* 2018;9:2279.
  - 22 Mohamed MS, Moulin TC, Schiöth HB. Sex differences in COVID-19: the role of androgens in disease severity and progression. *Endocrine* 2021;71:3–8. doi:10.1007/s12020-020-02536-6
  - 23 Dehingia N, Raj A. Sex differences in COVID-19 case fatality: do we know enough? *Lancet Glob Health* 2021;9:e14–15. doi:10.1016/S2214-109X(20)30464-2
  - 24 Knight SR, Ho A, Pius R, *et al.* Risk stratification of patients admitted to hospital with covid-19 using the ISARIC who clinical characterisation protocol: development and validation of the 4C mortality score. *BMJ* 2020;370:m3339. doi:10.1136/bmj.m3339
  - 25 Mei Y, Weinberg SE, Zhao L, *et al.* Risk stratification of hospitalized COVID-19 patients through comparative studies of laboratory results with influenza. *EClinicalMedicine* 2020;26:100475. doi:10.1016/j.eclinm.2020.100475
  - 26 Romero Starke K, Petereit-Haack G, Schubert M, *et al.* The age-related risk of severe outcomes due to COVID-19 infection: a rapid review, meta-analysis, and meta-regression. *Int J Environ Res Public Health* 2020;17:5974. doi:10.3390/ijerph17165974
  - 27 Koralnik IJ, Tyler KL. COVID-19: a global threat to the nervous system. *Ann Neurol* 2020;88:1–11.
  - 28 Herman C, Mayer K, Sarwal A. Scoping review of prevalence of neurological comorbidities in patients hospitalized for COVID-19. *Neurology* 2020;95:77–84.
  - 29 Romagnolo A, Balestrino R, Imbalzano G, *et al.* Neurological comorbidity and severity of COVID-19. *J Neurol* 2021;268:762–9.
  - 30 Rossi R, Coppi F, Talarico M, *et al.* Protective role of chronic treatment with direct oral anticoagulants in elderly patients affected by interstitial pneumonia in COVID-19 era. *Eur J Intern Med* 2020;77:158–60.
  - 31 Schiavone M, Gasperetti A, Mancone M, *et al.* Oral anticoagulation and clinical outcomes in COVID-19: an Italian multicenter experience. *Int J Cardiol* 2021;323:276–80.
  - 32 Hanff TC, Mohareb AM, Giri J, *et al.* Thrombosis in COVID-19. *Am J Hematol* 2020;95:1578–89.
  - 33 Vivas D, Roldán V, Esteve-Pastor MA. [Recommendations on antithrombotic treatment during the COVID-19 pandemic. Position statement of the Working Group on Cardiovascular Thrombosis of the Spanish Society of Cardiology]. *Rev Esp Cardiol* 2020;73:749–57.
  - 34 Oakley P, Kisely S, Baxter A, *et al.* Increased mortality among people with schizophrenia and other non-affective psychotic disorders in the community: a systematic review and meta-analysis. *J Psychiatr Res* 2018;102:245–53. doi:10.1016/j.jpsychires.2018.04.019
  - 35 InSimon L, Hashmi M, Callahan A. *Neuroleptic malignant syndrome.* Treasure Island, FL: StatPearls Publishing, 2020. <https://www.ncbi.nlm.nih.gov/books/NBK482282/>
  - 36 Taquet M, Luciano S, Geddes JR, *et al.* Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62 354 COVID-19 cases in the USA. *The Lancet Psychiatry [Internet]* 2020 (cited 2020 Dec 10).
  - 37 Nemani K, Li C, Olfson M, *et al.* Association of psychiatric disorders with mortality among patients with COVID-19 multimedia supplemental content. *JAMA Psychiatry [Internet]* 2021 <https://jamanetwork.com/>
  - 38 Clark A, Jit M, Warren-Gash C, *et al.* Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *The Lancet Global Health* 2020;8:e1003–17.
  - 39 Ghahramani S, Tabrizi R, Lankarani KB, *et al.* Laboratory features of severe vs. non-severe COVID-19 patients in Asian populations: a systematic review and meta-analysis. Vol. 25, *European Journal of medical research. BioMed Central* 2020.
  - 40 Ou M, Zhu J, Ji P, *et al.* Risk factors of severe cases with COVID-19: a meta-analysis. *Epidemiol Infect* 2020;148:e175.
  - 41 Hellgren Z, Serrano I. Transnationalism and financial crisis: the hampered migration projects of female domestic workers in Spain. Available: [www.mdpi.com/journal/socsci](http://www.mdpi.com/journal/socsci)
  - 42 Lázaro-Pérez C, Martínez-López Jose Ángel, Gómez-Galán J, *et al.* Anxiety about the risk of death of their patients in health professionals in Spain: analysis at the peak of the COVID-19 pandemic. *Int J Environ Res Public Health* 2020;17:5938.
  - 43 Zimmerman R, Nowalk MP, Bear T. Proposed clinical indicators for efficient screening and testing for COVID-19 infection from classification and regression trees (CART) analysis. *medRxiv : the preprint server for health sciences* 2020.