Identification and validation of clinical phenotypes with prognostic implications in patients admitted to hospital with COVID-19: a multicentre cohort study

*Belén Gutiérrez-Gutiérrez, María Dolores del Toro, Alberto M Borobia, Antonio Carcas, Inmaculada Jarrín, María Yllescas, Pablo Ryan, Jerónimo Pachón, Jordi Carratalà, Juan Berenguer, Jose Ramón Arribas, Jesús Rodríguez-Baño, on behalf of the REIPI-SEIMC COVID-19 group and COVID@HULP groups**

Summary

Background The clinical presentation of COVID-19 in patients admitted to hospital is heterogeneous. We aimed to determine whether clinical phenotypes of patients with COVID-19 can be derived from clinical data, to assess the reproducibility of these phenotypes and correlation with prognosis, and to derive and validate a simplified probabilistic model for phenotype assignment. Phenotype identification was not primarily intended as a predictive tool for mortality.

Methods In this study, we used data from two cohorts: the COVID-19@Spain cohort, a retrospective cohort including 4035 consecutive adult patients admitted to 127 hospitals in Spain with COVID-19 between Feb 2 and March 17, 2020, and the COVID-19@HULP cohort, including 2226 consecutive adult patients admitted to a teaching hospital in Madrid between Feb 25 and April 19, 2020. The COVID-19@Spain cohort was divided into a derivation cohort, comprising 2667 randomly selected patients, and an internal validation cohort, comprising the remaining 1368 patients. The COVID-19@HULP cohort was used as an external validation cohort. A probabilistic model for phenotype assignment was derived in the derivation cohort using multinomial logistic regression and validated in the internal validation cohort. The model was also applied to the external validation cohort. 30-day mortality and other prognostic variables were assessed in the derived phenotypes and in the phenotypes assigned by the probabilistic model.

Findings Three distinct phenotypes were derived in the derivation cohort (n=2667)—phenotype A (516 [19%] patients), phenotype B (1955 [73%]) and phenotype C (196 [7%])—and reproduced in the internal validation cohort (n=1368)—phenotype A (233 [17%] patients), phenotype B (1019 [74%]), and phenotype C (116 [8%]). Patients with phenotype A were younger, were less frequently male, had mild viral symptoms, and had normal inflammatory parameters. Patients with phenotype B included more patients with obesity, lymphocytopenia, and moderately elevated inflammatory parameters. Patients with phenotype C included older patients with more comorbidities and even higher inflammatory parameters than phenotype B. We developed a simplified probabilistic model (validated in the internal validation cohort) for phenotype assignment, including 16 variables. In the derivation cohort, 30-day mortality rates were 2·5% (95% CI 1·4–4·3) for patients with phenotype A, 30·5% (28·5–32·6) for patients with phenotype B, and 60·7% (53·7–67·2) for patients with phenotype C (log-rank test p<0·0001). The predicted phenotypes in the internal validation cohort and external validation cohort showed similar mortality rates to the assigned phenotypes (internal validation cohort: 5·3% [95% CI 3·4–8·1] for phenotype A, 31·3% [28·5–34·2] for phenotype B, and 59·5% [48·8–69·3] for phenotype C; external validation cohort: 3·7% [2·0–6·4] for phenotype A, 23·7% [21·8–25·7] for phenotype B, and 51·4% [41·9–60·7] for phenotype C).

Interpretation Patients admitted to hospital with COVID-19 can be classified into three phenotypes that correlate with mortality. We developed and validated a simplified tool for the probabilistic assignment of patients into phenotypes. These results might help to better classify patients for clinical management, but the pathophysiological mechanisms of the phenotypes must be investigated.

Funding Instituto de Salud Carlos III, Spanish Ministry of Science and Innovation, and Fundación SEIMC/GeSIDA.

Copyright © 2021 Elsevier Ltd. All rights reserved.

Lancet Infect Dis 2021; 21: 783–92

Published Online February 23, 2021 https://doi.org/10.1016/S1473-3099(21)00019-0

See Comment page 744

*Members of the REIPI-SEIMC COVID-19 group and COVID@HULP group are listed in appendix pp 23–29

Unidad Clínica de Enfermedades Infecciosas, Microbiología y Medicina Preventiva, Hospital Universitario Virgen Macarena, Seville, Spain (B Gutiérrez-Gutiérrez PhD, M D del Toro PhD, J Rodríguez-Baño PhD); **Departamento de Medicina, Universidad de Sevilla, Seville, Spain** (B Gutiérrez-Gutiérrez, M D del Toro, Prof J Pachón PhD, J Rodríguez-Baño); **Instituto de Biomedicina de Sevilla, Seville, Spain** (B Gutiérrez-Gutiérrez, M D del Toro, Prof J Pachón, J Rodríguez-Baño); **Departamento de Farmacología Clínica, Hospital Universitario La Paz, Universidad Autónoma de Madrid, Madrid, Spain** (A M Borobia PhD, A Carcas PhD); **Instituto de Investigación La Paz, Madrid, Spain** (A M Borobia, A Carcas, J R Arribas MD); **Centro Nacional de Epidemiología, Instituto de Salud Carlos III, Madrid, Spain** (I Jarrín MSc); **Fundación SEIMC/GeSIDA, Madrid, Spain** (M Yllescas PharmD); **Servicio de Medicina Interna, Hospital Universitario Infanta Leonor, Madrid, Spain** (P Ryan PhD); **Unidad Clínica de Enfermedades Infecciosas, Microbiología y Medicina Preventiva, Hospital Universitario Virgen del Rocío, Seville, Spain** (Prof J Pachón); **Servei de Malalties Infeccioses, Hospital Universitari de Bellvitge, Barcelona, Spain**

Introduction

Patients admitted to hospital with COVID-19 show various clinical signs and symptoms and laboratory abnormalities.[1–5] Some of these features have been found to be predictors of mortality.[3,4] The reasons for this heterogeneous presentation are not fully understood. However, it could be related to factors such as viral load,[6] partial immune protection due to previous infections with other coronaviruses,[7] genetic determinants,[8] and other non-genetic-mediated factors such as age and underlying conditions.[3,4]

We hypothesise that patients admitted to hospital with COVID-19 might be classified into few clinical patterns (phenotypes) according to their demographics, underlying conditions, signs, symptoms, radiological findings,

(Prof J Carratalà PhD); **Instituto de Investigación Biomédica de Bellvitge, Barcelona, Spain** (Prof J Carratalà); **Universitat de Barcelona, Barcelona, Spain** (Prof J Carratalà); **Servicio de Microbiología Clínica y Enfermedades Infecciosas, Hospital General Universitario Gregorio Marañón, Madrid, Spain** (J Berenguer PhD); **Instituto de Investigación Sanitaria Gregorio Marañón, Madrid, Spain** (J Berenguer); **Unidad de Enfermedades Infecciosas, Servicio de Medicina Interna, Hospital Universitario La Paz, Madrid, Spain** (J R Arribas)

Correspondence to:
Dr Jesús Rodríguez-Baño, Servicio de Enfermedades Infecciosas, Hospital Universitario Virgen Macarena, 41009 Seville, Spain
jesusrb@us.es

See **Online** for appendix

For the **online phenotype assignment tool** see http://fen-covid.com/index.html

Research in context

Evidence before this study

We searched PubMed, Scopus, and *medRxiv* from Jan 9 to Sept 30, 2020, using the terms ["COVID-19" OR "SARS-CoV-2"] AND ["phenotypes" OR "clinical features"], with no language restrictions, to detect any published study identifying and characterising phenotypes among patients with COVID-19. We found one study that identified three phenotypes in a cohort of 85 patients admitted to the intensive care unit, which were correlated with mortality, and one preprint study in which phenotypes were investigated in ambulatory patients with self-declaration of symptoms. We also found studies referring to distress syndrome-associated phenotypes or hyperinflammatory phenotypes.

Added value of this study

To our knowledge, this is the first study investigating the existence and characterisation of clinical phenotypes for COVID-19 patients at hospital admission. We identified three distinct clinical phenotypes on the basis of demographics, underlying conditions, clinical and laboratory data, and radiological features at presentation among patients admitted to hospital with COVID-19. The phenotypes were shown to have clinical implications, since they were associated with patient prognosis. Furthermore, we developed and validated a simplified probabilistic model for phenotype assignment. This model is available as a tool online to facilitate the probabilistic classification of patients with COVID-19 who are admitted to hospital into phenotypes.

Implications of all the available evidence

Identification of COVID-19 phenotypes allows investigation of potential differences in their underlying pathophysiological mechanisms, which could allow better pathogenesis-targeted approaches for therapies in the design and selection of participants in clinical trials, depending on the mechanism of action of specific drugs and their use in clinical management. Furthermore, phenotype assignment would be helpful in identifying low-risk patients and patients who might need closer monitoring during admission.

and laboratory data at presentation. If they exist, these phenotypes might denote different pathophysiological routes and outcomes and be useful for better classifying patients for testing and treatment strategies.

The objectives of this study were to determine whether clinical phenotypes of patients with COVID-19 can be derived from clinical data, to assess their reproducibility and correlation with prognosis, and to derive and validate a simplified probabilistic model for phenotype assignment.

Methods

Databases

In this study, we used data from two cohorts: the COVID-19@Spain cohort, a retrospective cohort including 4035 consecutive adult patients admitted to 127 hospitals in Spain with COVID-19 between Feb 2 and March 17, 2020, and the COVID-19@HULP cohort, including 2226 consecutive adult patients admitted to a teaching hospital in Madrid between Feb 25 and April 19, 2020. The cohort designs and patient characteristics were previously reported in detail.[4,5] 41 patients in the COVID-19@HULP cohort who were also included in the COVID-19@Spain cohort were excluded from the COVID-19@HULP cohort for the current study (2185 remaining patients in this cohort). The COVID-19@Spain cohort was divided into a derivation cohort, comprising 2667 randomly selected patients, selected using the SPSS function for selection of random samples from a database, and an internal validation cohort, comprising the remaining 1368 patients. The COVID-19@HULP cohort was used as an external validation cohort. An overview of the analyses done in the derivation and validation cohort is shown in the appendix (p 19). The study was approved by the University Hospitals Virgen Macarena and Virgen del Rocío ethics committee (Seville, Spain), which waived the need to obtain written informed consent because of the observational nature of the study. STROBE recommendations were followed (appendix pp 2–3).

We discussed the objectives of the study, the study design, and results with several health-care workers who had had COVID-19.

Phenotype derivation

We considered 69 variables to derive the clinical phenotypes. The variables were selected based on the available information about the features of patients admitted to hospital[1–3] and the early clinical experience gained at the participating sites. All data were collected at hospital admission and included age, sex, race or ethnicity, comorbidities, drugs previously used for underlying diseases, COVID-19-related signs and symptoms at presentation, laboratory data, and chest radiographical data (table 1). As our objective was to explore the existence of phenotypes, we did not preselect any variables.

The proportion of missing data per variable in the COVID-19@Spain cohort is shown in the appendix (pp 5–6). The Little MCAR test was used to verify that missing data were at random, and imputation was done using the Markov chain Monte Carlo method.

Analyses to identify the phenotypes were first done in the derivation cohort. We assessed the distributions of values and missing data, and correlation among the variables, using the χ^2 test and Pearson's correlation coefficient for categorical and continuous variables, respectively. We excluded highly correlated variables. We did a two-step cluster analysis using both continuous and categorical variables, which provided the optimal number

of clusters. We used silhouette analysis to assess the quality of the cluster derivation. We did a sensitivity analysis excluding variables with more than 50% missing data.

Features of the patients in the phenotypes obtained were compared using χ^2 test and Kruskal-Wallis test for categorical and continuous variables, respectively. We visualised the patterns of distribution of the variables in

	Phenotype A *vs* phenotype C		Phenotype B *vs* phenotype C	
	OR (95% CI)	p value	OR (95% CI)	p value
Demographics				
Age (per year)	0·92 (0·90–0·93)	<0·0001	0·96 (0·95–0·97)	<0·0001
Female sex	1·79 (1·27–2·54)	0·0014	1·32 (0·97–1·82)	0·089
Race or ethnicity				
White	0·48 (0·06–4·16)	0·50	0·48 (0·06–3·62)	0·48
Black	0·80 (0·04–17·20)	0·89	0·10 (0·01–2·29)	0·15
Hispanic	2·08 (0·20–21·48)	0·54	1·08 (0·12–9·74)	0·94
Asian	0·20 (0·01–6·66)	0·37	0·55 (0·03–9·68)	0·68
Arab	0·50 (0·03–7·45)	0·61	0·40 (0·03–4·82)	0·47
Other	1 (ref)	..	1 (ref)	..
Comorbidities				
Chronic heart disease	0·12 (0·08–0·17)	<0·0001	0·23 (0·17–0·31)	<0·0001
Hypertension	0·08 (0·05–0·12)	<0·0001	0·19 (0·12–0·28)	<0·0001
Chronic lung disease	0·19 (0·12–0·29)	<0·0001	0·54 (0·39–0·74)	<0·0001
Asthma	1·75 (0·83–3·66)	0·14	1·73 (0·87–3·44)	0·12
Chronic kidney disease (stage 4)	0·05 (0·03–0·10)	<0·0001	0·06 (0·04–0·09)	<0·0001
Liver cirrhosis	0·76 (0·23–2·56)	0·65	0·75 (0·26–2·13)	0·59
Chronic neurological disease	0·45 (0·27–0·76)	0·0031	0·65 (0·43–1·00)	0·057
Active solid malignancy	0·63 (0·34–1·17)	0·14	0·73 (0·43–1·24)	0·21
Active haematological malignancy	0·83 (0·29–2·43)	0·74	0·90 (0·35–2·29)	0·82
HIV/AIDS	1·52 (0·17–14·29)	0·71	1·92 (0·26–14·29)	0·53
Obesity (body-mass index >30 kg/m^2)	0·28 (0·18–0·45)	<0·0001	0·52 (0·37–0·74)	0·0043
Diabetes	0·12 (0·08–0·18)	<0·0001	0·31 (0·23–0·42)	<0·0001
Chronic inflammatory disease	1·33 (0·62–2·84)	0·47	1·20 (0·60–2·42)	0·60
Dementia	0·19 (0·10–0·35)	<0·0001	0·57 (0·38–0·87)	0·0086
Malnutrition	0·40 (0·21–0·76)	0·012	0·51 (0·31–0·86)	0·016
Smoking status				
Never	2·67 (1·88–3·81)	<0·0001	1·26 (0·93–1·71)	0·14
Current smoker	2·56 (1·37–4·79)	0·0037	1·11 (0·63–1·97)	0·71
Former smoker	1 (ref)	..	1 (ref)	..
Treatments for underlying conditions				
Angiotensin converting enzyme inhibitors	0·39 (0·25–0·59)	<0·0001	0·76 (0·54–1·08)	0·12
Angiotensin receptor blockers	0·41 (0·27–0·62)	<0·0001	0·57 (0·40–0·80)	0·0010
Inhaled corticosteroids	0·40 (0·24–0·67)	<0·0001	0·79 (0·53–1·19)	0·26
Systemic corticosteroids	0·61 (0·31–1·20)	0·15	0·69 (0·39–1·23)	0·22
Cancer chemotherapy	1·15 (0·41–3·23)	0·80	1·12 (0·45–2·86)	0·80
Biological drugs	1·08 (0·42–2·78)	0·87	0·65 (0·27–1·54)	0·32
Infection data at admission				
Non-focal symptoms				
Reported fever	1·85 (1·27–2·63)	0·0014	2·17 (1·59–3·03)	<0·0001
Temperature (per 1°C)	0·93 (0·78–1·11)	0·41	1·25 (1·06–1·46)	0·0063
Myalgia or arthralgia	2·70 (1·69–4·17)	<0·0001	2·27 (1·49–3·45)	<0·0001
Headache	3·03 (1·69–5·56)	<0·0001	1·33 (0·76–2·33)	0·32
Skin rash	0·95 (0·18–5·00)	0·95	1·10 (0·26–4·76)	0·89
Anosmia	3·51 (0·81–15·15)	0·098	1·77 (0·42–7·41)	0·44
Altered mental status	0·25 (0·15–0·43)	<0·0001	0·67 (0·45–0·99)	0·042

(Table 1 continues on next page)

	Phenotype A *vs* phenotype C		Phenotype B *vs* phenotype C	
	OR (95% CI)	p value	OR (95% CI)	p value
(Continued from previous page)				
Inflammation				
White blood cells (per 10^3 cells/μL)	0·79 (0·76–0·83)	<0·0001	0·83 (0·80–0·86)	<0·0001
Lymphocytes (per 10^3 cells/μL)	1·11 (0·97–1·28)	0·14	0·99 (0·86–1·14)	0·87
Neutrophils (per 10^3 cells/μL)	0·74 (0·70–0·78)	<0·0001	0·82 (0·79–0·85)	<0·0001
D-dimer (per 10^3 μg/L)	0·79 (0·66–0·93)	0·0050	0·98 (0·95–1·01)	0·18
Procalcitonin (per 1 ng/mL)	0·09 (0·04–0·17)	<0·0001	0·52 (0·44–0·61)	<0·0001
C-reactive protein (per 10^2 mg/L)	0·92 (0·84–1·02)	0·11	0·97 (0·95–0·99)	0·0090
IL-6 (per 10^2 μg/mL)	0·17 (0·11–0·27)	<0·0001	1·00 (0·92–1·09)	0·97
Ferritin (per 10^3 ng/mL)	0·19 (0·11–0·31)	<0·0001	1·30 (0·89–1·89)	0·17
Cardiovascular				
Heart rate per minute (per unit)	1·00 (0·99–1·01)	0·69	1·01 (1·00–1·02)	0·15
Systolic blood pressure (per 1 mmHg)	1·00 (0·99–1·00)	0·56	1·00 (0·99–1·00)	0·62
Diastolic blood pressure (per 1 mmHg)	1·02 (1·01–1·04)	<0·0001	1·02 (1·01–1·03)	<0·0001
Respiratory tract				
Chest pain	1·45 (0·86–2·38)	0·16	0·98 (0·61–1·59)	0·94
Dyspnoea	0·19 (0·13–0·27)	<0·0001	0·65 (0·48–0·88)	0·0061
Cough	1·22 (0·87–1·72)	0·25	1·61 (1·19–2·22)	0·0021
Expectoration	0·46 (0·32–0·68)	<0·0001	0·74 (0·53–1·01)	0·055
Haemoptysis	0·51 (0·20–1·30)	0·16	0·48 (0·22–1·04)	0·062
Respiratory rate per min (per unit)	0·80 (0·77–0·83)	<0·001	0·94 (0·92–0·97)	<0·0001
Oxygen saturation, room air, pulse oximetry (per 1%)	1·62 (1·55–1·70)	<0·0001	1·09 (1·07–1·11)	<0·0001
Oxygen saturation after oxygen supplementation (per 1%)	1·35 (1·26–1·45)	<0·0001	1·07 (1·03–1·11)	<0·0001
Oxygen saturation, room air, venous blood (per 1%)	1·07 (1·05–1·09)	<0·0001	1·03 (1·02–1·03)	<0·0001
PCO_2, venous blood (per 1 mmHg)	1·01 (0·99–1·02)	0·61	0·98 (0·96–0·99)	0·022
Lung infiltrates on chest radiography				
No infiltrate	3·43 (2·32–5·08)	<0·0001	0·77 (0·54–1·10)	0·15
Unilateral	2·25 (1·44–3·51)	<0·0001	1·16 (0·78–1·72)	0·45
Bilateral	1 (ref)	..	1 (ref)	..
Interstitial lung infiltrate	0·48 (0·34–0·68)	<0·0001	1·18 (0·88–1·59)	0·27
Ground-glass opacity infiltrate	0·74 (0·43–1·25)	0·26	1·19 (0·75–1·85)	0·46
Liver				
Albumin, mean (SD; per 1 g/dL)	9·81 (6·46–14·87)	<0·0001	3·37 (2·37–4·79)	<0·0001
Lactic acid dehydrogenase (per 10^2 U/L)	0·61 (0·55–0·68)	<0·0001	1·00 (0·96–1·04)	1·00
Bilirubin (per 1 mg/dL)	0·93 (0·77–1·13)	0·49	1·00 (0·96–1·04)	0·87
Renal				
Creatinine (per 1 × mg/dL)	0·10 (0·07–0·15)	<0·0001	0·13 (0·10–0·17)	<0·0001
Sodium (per 1 × mEq/L)	1·07 (1·03–1·11)	0·0011	1·00 (0·97–1·04)	0·78
Potassium (per 1 × mEq/L)	0·25 (0·18–0·34)	<0·0001	0·24 (0·18–0·31)	<0·0001
Haematological				
Haemoglobin (per 1 × g/dL)	1·66 (1·53–1·81)	<0·0001	1·60 (1·48–1·72)	<0·0001
Haematocrit (per 1%)	1·19 (1·15–1·23)	<0·0001	1·16 (1·13–1·20)	<0·0001
Platelets (per 10^5/μL)	0·86 (0·76–0·99)	0·031	0·80 (0·71–0·90)	<0·0001
Activated partial thromboplastin time (per 1 × s)	0·99 (0·98–1·00)	0·038	0·99 (0·99–1·00)	0·075
International normalised ratio (per unit)	0·18 (0·12–0·28)	<0·0001	0·32 (0·26–0·40)	<0·0001
Other				
Creatine phosphokinase (per 10^2 × U/L)	1·01 (0·95–1·08)	0·71	1·02 (0·96–1·08)	0·50
Blood glucose (per 1 × mg/dL)	0·98 (0·98–0·98)	<0·0001	0·99 (0·99–0·99)	<0·0001

OR=odds ratio.

Table 1: **Bivariate analysis of variables associated with phenotypes in the derivation cohort**

the different phenotypes using chord diagrams and heatmaps after grouping variables into comorbidities and system-related or organ-related data (appendix p 4). A two-step cluster analysis was also done in the internal validation cohort to check the reproducibility of phenotype identification.

Derivation and validation of a parsimonious probabilistic model for phenotypes

As the number of variables used to derive the phenotypes was very high, assigning patients to phenotypes was neither intuitive nor applicable for clinical practice. Therefore, we developed a simplified probabilistic model to assign patients into the phenotypes. As we identified three phenotypes, we did a multinomial logistic regression analysis in the derivation cohort. First, we analysed the bivariate association of each variable of the phenotypes using the χ^2 and Kruskal-Wallis tests for categorical and continuous variables, respectively. Those with $p<0{\cdot}20$ were included in a multinomial logistic regression model; the variance inflation factor value was used to detect the potential occurrence of collinearity and interactions were tested. The variables were selected using a manual backward selection process. The ability of the final model to predict the phenotypes as identified by the derivation process was checked by calculating the area under the receiver operating characteristic curves (AUROC) with 95% CIs for the three phenotypes. We also tested the predictive ability of the model in 60 randomly chosen subcohorts (using a tool in the SPM software) with 80%, 60%, or 40% of the sample size of the derivation cohort.

The probabilistic model for phenotype assignment was used in two ways. First, we applied the model to the internal validation cohort to check its ability to predict the phenotypes obtained from this cohort. Second, we applied the model to both the internal and external validation cohorts to obtain a probabilistic assignment of patients to the phenotypes (model-derived formulae used for probability calculations are in the appendix p 4). Patients were assigned to the phenotype with the highest belonging probability according to the model-derived formula. We checked the distribution of variables among the assigned phenotypes.

Prognostic assessment of the phenotypes

We compared the 30-day mortality of patients in the different phenotypes in the derivation cohort with Kaplan-Meier curves and log-rank tests, and calculated hazard ratios (HRs) with 95% CIs. We also collected data on complications that occurred during treatment in hospital (listed in the appendix p 9). These variables were also analysed in the validation cohorts, in which patients were assigned to the phenotype with the highest probability according to the probabilistic model-derived formula. Since any association of phenotypes with mortality might be caused by a different distribution of a few strong independent prognostic variables in the phenotypes, such as age and oxygen saturation,[4] we did a stratified analysis to check if any mortality association was maintained in all strata of these variables.

All analyses were done with IBM SPSS Statistics 26, SPM 8.2, and R version 3.6.0.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

The features of the patients in the cohorts used for this study were previously reported in detail.[4,5] A two-step cluster analysis of variables collected at hospital admission identified three clinical phenotypes in the derivation cohort: phenotype A (516 [19%] of 2667 patients), phenotype B (1955 [73%] of 2667 patients), and phenotype C (196 [7%] of 2667 patients). The silhouette score was 0·6, indicating good quality of clustering. Exclusion of variables with a high proportion of missing data did not cause any evident changes (data not shown).

The baseline characteristics of the derivation and internal validation cohorts are present in the appendix (pp 7–12). Overall, patients with phenotype A were younger (mean age 55·2 years [SD 18·4] *vs* 68·7 years [15·9] and 77·2 years [10·9] in phenotypes B and C, respectively), were less frequently male (55% *vs* 63% and 69%), presented more frequently with headache (19% *vs* 9% and 7%), myalgia (29% *vs* 26% and 13%), and chest pain (15% *vs* 11% and 11%), had higher lymphocyte count (mean 1439 cells/µL [SD 1761] *vs* 1094 cells/µL [1424] and 1096 cells/µL [1170]), and had lower levels of inflammatory parameters such as C-reactive protein, IL-6, ferritin, or lactic acid dehydrogenase (appendix pp 7–9). Patients with phenotype B more frequently reported fever (83% *vs* 80% and 69% in phenotypes A and C, respectively) and cough (74% *vs* 68% and 63%), more frequently lacked pulmonary infiltrates in chest radiography (20% *vs* 46% and 25%), more frequently had interstitial infiltrates (45% *vs* 25% and 41%), and had higher levels of ferritin (mean 809·5 ng/mL [SD 588·4] *vs* 616·4 ng/mL [219·7] and 752·8 ng/mL [320·4]) and creatine phosphokinase (mean 164·3 U/L [SD 464·0] *vs* 150·8 U/L [368·2] and 141·4 U/L [199·0]; appendix pp 7–9). Patients with phenotype C more frequently had chronic heart disease (56% *vs* 13% and 23% in phenotypes A and B, respectively), hypertension (86% *vs* 31% and 53%), chronic lung disease (31% *vs* 8% and 19%), stage 4 chronic kidney disease (34% *vs* 3% and 3%), obesity (body-mass index >30 kg/m²; 23% *vs* 8% and 14%), diabetes (48% *vs* 10% and 22%), and acute altered mental status (18% *vs* 5% and 12%); had higher levels of neutrophils (mean 8539 cells/µL [SD 6656] *vs* 4112 cells/µL [2511] and 4892 cells/µL [2844]), D-dimer (mean 1343·1 µg/L [SD 2419·7] *vs* 715·8 µg/L [986·5] and 986·3 µg/L [3290·5]), procalcitonin (mean 0·70 ng/mL [SD 0·96] *vs* 0·17 ng/mL [0·26] and 0·27 ng/mL [0·51]), C-reactive protein (mean

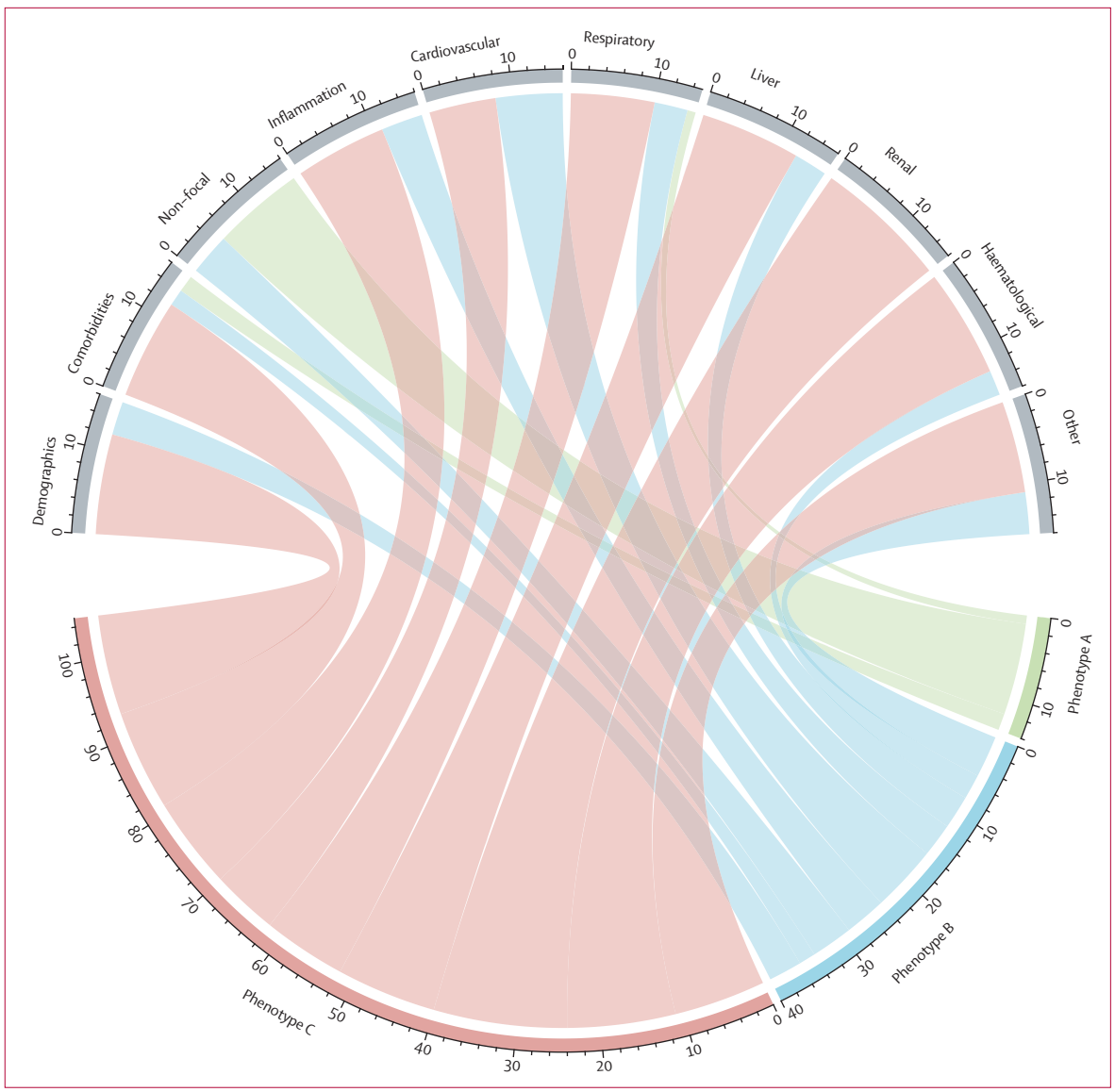

Figure 1: **Chord diagram of the distribution of groups of variables in the phenotypes in the derivation cohort**
Variables are grouped into categories. The phenotypes are shown in different colours: phenotype A is green, phenotype B is blue, and phenotype C is red. For each phenotype, if a variable mean (for continuous variables) or proportion (for categorical variables) is significantly different to the mean or proportion in the full derivation cohort, a ribbon connects the phenotype and the variable group. The width of the ribbons correlates with the number of variables that are significantly different from those in the derivation cohort for that phenotype.

127·1 mg/L [SD 119·8] *vs* 47·4 mg/L [68·3] and 88·8 mg/L [84·2]), creatinine (mean 2·76 mg/dL [SD 2·11] *vs* 0·96 mg/dL [0·56] and 0·99 mg/dL [0·36]), and potassium (mean 4·5 mEq/L [SD 0·7] *vs* 4·0 mEq/L [0·5] and 4·0 mEq/L [0·5]); and had poorer oxygenation parameters (appendix pp 7–9, figures 1, 2).

We repeated the two-step cluster analysis in the internal validation cohort. This analysis also selected three clusters with a very similar distribution of patients to the derivation cohort: phenotype A (233 [17%] of 1368 patients), phenotype B (1019 [74%] of 1368 patients), and phenotype C (116 [8%] of 1368 patients). The silhouette score was also 0·6, and the distribution of variables in the phenotypes was as in the derivation cohort, except for the proportion of patients with liver cirrhosis and active solid malignancies (which were not significantly different in the derivation cohort but were more frequent in phenotype C than in phenotype A or phenotype B in the internal validation cohort), haematological malignancy (no difference in the derivation cohort but less frequent in phenotype A than in phenotypes B and C in the internal validation cohort), and ferritin and creatine phosphokinase concentrations (which were higher in phenotype B than in phenotypes A and C in the derivation cohort and in phenotype C than in

phenotypes A and B in the internal validation cohort; appendix pp 10–12).

To develop a simple way to assign patients to a phenotype, we developed and validated a parsimonious probabilistic model for belonging to phenotypes. We first did a bivariate analysis of the association of the different variables with phenotype A versus phenotype C and phenotype B versus phenotype C in the derivation cohort. We found a significant crude association with phenotype for many variables (table 1). After a variable selection process, we developed a final multinomial logistic regression model with 16 variables, including age, sex, chronic lung disease, obesity, diastolic blood pressure, oxygen saturation (room air), white blood cell count, neutrophils, haematocrit, coagulation international normalised ratio, C-reactive protein, glucose, creatinine, sodium, potassium, and type of lung infiltrate on chest radiograph (table 2). Therefore, we derived a simplified probabilistic model for patient assignment to phenotypes. The AUROC of the model for the observed data in the derivation cohort showed very good predictive ability for the three phenotypes (0·86, 95% CI 0·85–0·88 for phenotype A, 0·88, 0·86–0·89 for phenotype B, and 0·99, 0·99–0·99 for phenotype C). The predictive ability was similar in smaller, randomly selected subcohorts (appendix p 13).

The capacity of the model to correctly assign patients to phenotypes was validated in the internal validation cohort for the phenotypes directly derived from that cohort. The ability of the model to predict the observed phenotypes in the internal validation cohort was also high (AUROC 0·86, 95% CI 0·84–0·89 for phenotype A; 0·86, 0·84–0·88 for phenotype B; and 0·95, 0·93–0·98 for phenotype C; appendix p 22).

The probabilistic model was then applied to the internal and external validation cohorts to obtain the individual probability of being assigned a specific phenotype. The number of patients in the internal validation cohort assigned to phenotypes A, B, and C according to their highest probability were 263 (19%), 1021 (75%), and 84 (6%), respectively (appendix pp 14–15). The corresponding figures for the external validation cohort were 323 (15%), 1757 (80%), and 105 (5%; appendix p 16). In the internal validation cohort, the distribution of all variables in the three predicted phenotypes was similar to that in the derivation cohort (appendix pp 14–15). For the external validation cohort, not all variables collected in the derivation cohort were available. Therefore, we checked the distribution of the variables included in the model, which was similar to that in the derivation cohort (appendix p 16).

In the derivation cohort, 30-day mortality rates were 2·5% (95% CI 1·4–4·3) for patients with phenotype A, 30·5% (28·5–32·6) for patients with phenotype B, and 60·7% (53·7–67·2) for patients with phenotype C (figure 3; appendix p 17). In the internal validation cohort, the mortality in the reproduced phenotypes was 2·6% (95% CI 1·0–5·6) for phenotype A, 31·0% (28·2–33·9) for phenotype B, and 53·4% (44·4–62·2) for phenotype C (appendix p 17). Regarding the phenotypes assigned on the basis of the probabilistic model, the mortality rates in the internal validation cohort were 5·3% (95% CI 3·4–8·1) for phenotype A, 31·3% (28·5–34·2) for phenotype B, and 59·5% (48·8–69·3) for phenotype C (figure 3; appendix p 17) and in the external validation cohort were 3·7% (2·0–6·4) for phenotype A, 23·7% (21·8–25·7) for phenotype B, and 51·4% (41·9–60·7) for phenotype C (the external validation cohort only had in-hospital mortality and not 30-day mortality data; figure 3; appendix p 17). All mortality data are summarised in the appendix (p 17).

The proportion of patients in the derivation cohort who needed intensive care unit care or had transfusion-requiring anaemia, pleural effusion, acute kidney failure,

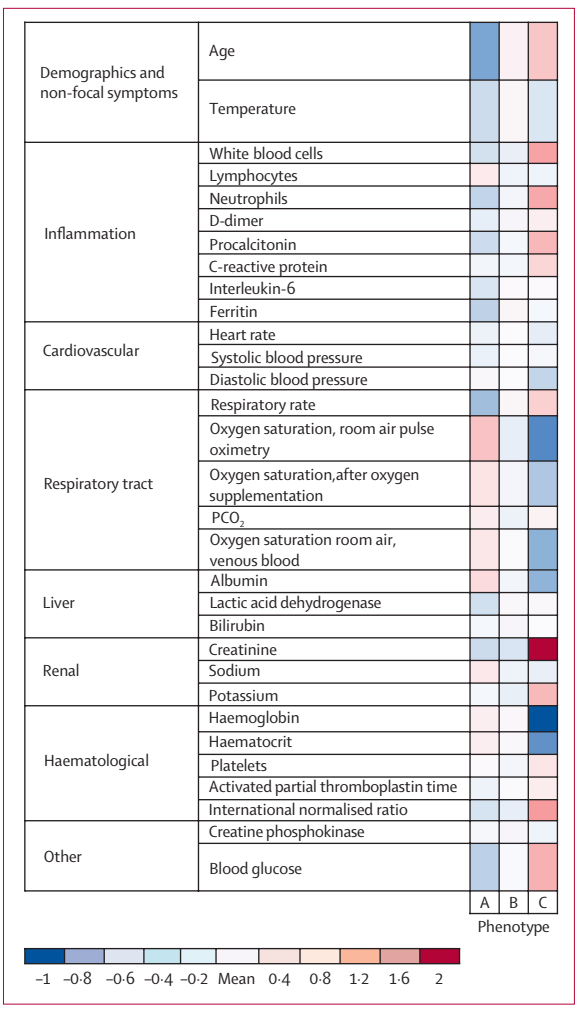

Figure 2: **Heatmap of the distribution of continuous variables in the phenotypes in the derivation cohort**

A colour gradient is used to show differences in mean values in relation to the full derivation cohort, towards red for higher values and blue for lower values. The colour gradient indicates the number of SDs that the mean value in the subcohort of interest is below or above the mean value in the full cohort.

	Phenotype A *vs* phenotype C		Phenotype B *vs* phenotype C	
	OR (95% CI)	p value	OR (95% CI)	p value
Age (per year)	0·93 (0·90–0·96)	<0·0001	0·96 (0·93–0·99)	0·0051
Female sex	0·68 (0·33–1·41)	0·30	0·44 (0·22–0·89)	0·021
Chronic lung disease	0·55 (0·26–1·16)	0·10	0·79 (0·42–1·54)	0·48
Obesity (body-mass index >30 kg/m^2)	0·49 (0·20–1·23)	0·12	0·71 (0·31–1·64)	0·42
White blood cells (per 10^3 cells/μL)	0·80 (0·73–0·87)	<0·0001	0·73 (0·68–0·79)	<0·0011
Neutrophils (per 10^3 cells/μL)	0·89 (0·80–0·99)	0·032	0·99 (0·90–1·08)	0·86
C-reactive protein (per 10^2 mg/L)	0·95 (0·91–1·00)	0·055	0·94 (0·90–0·99)	0·011
Diastolic blood pressure (per 1 mmHg)	1·03 (1·01–1·05)	0·011	1·02 (1·01–1·04)	0·013
Oxygen saturation, room air, pulse oximetry (per 1%)	1·56 (1·46–1·66)	<0·0001	1·11 (1·07–1·16)	<0·0001
Lung infiltrate on chest radiography				
No infiltrate	4·07 (1·83–9·02)	0·00055	1·17 (0·55–2·49)	0·69
Unilateral	3·50 (1·51–8·06)	0·0032	2·05 (0·93–4·51)	0·071
Bilateral	1 (ref)	..	1 (ref)	..
Creatinine (per 1 mg/dL)	0·09 (0·05–0·15)	<0·0001	0·06 (0·04–0·10)	<0·0001
Sodium (per 1 mEq/L)	1·09 (1·02–1·17)	0·010	1·04 (0·98–1·11)	0·14
Potassium (per 1 mEq/L)	0·37 (0·21–0·67)	0·00093	0·26 (0·15–0·45)	<0·0001
Haematocrit (per 1%)	1·29 (1·21–1·38)	<0·0001	1·27 (1·19–1·35)	<0·0001
International normalised ratio (per unit)	0·12 (0·07–0·22)	<0·0001	0·12 (0·08–0·18)	<0·0001
Blood glucose (per 1 mg/dL)	0·99 (0·98–0·99)	<0·0001	0·99 (0·98–0·99)	<0·0001

The variance inflation factor value was less than 2 in all cases. OR=odds ratio.

Table 2: **Multinomial logistic regression model for the prediction of phenotypes in the derivation cohort**

heart failure, bacterial pneumonia, acute respiratory distress syndrome, or cardiorespiratory arrest during admission was significantly increased in phenotype C compared with phenotypes A and B and significantly decreased in phenotype A compared with phenotypes B and C; differences were not significant for stroke, ischaemic coronary event, liver failure, or disseminated intravascular coagulation (appendix pp 7–9). Results were similar in the internal validation cohort, with the exception that liver failure was more frequent in phenotype B (appendix p 15).

To check whether the association of the phenotypes with mortality was maintained after considering different distributions of strong mortality predictors across the phenotypes, such as age and oxygen saturation, we did a stratified analysis per strata of these variables in the derivation cohort. In all strata, phenotypes were significantly associated with mortality (appendix p 18).

Discussion

We identified three phenotypes based on demographics, underlying conditions, clinical and laboratory data, and radiological features at presentation among patients admitted to hospital with COVID-19. The phenotypes, despite not intended to be used for predicting mortality, had clinical implications, as we observed associations with patient prognosis. We also developed a simplified probabilistic model that is potentially applicable to other cohorts.

Clinical presentation of COVID-19 is polymorphic. Clinical phenotypes have been described for patients with severe acute respiratory distress with potential implications for respiratory support therapy.[9] Phenotypes based only on self-declaration of symptoms by non-hospitalised patients with COVID-19 using an app have been reported.[10] Clinical phenotypes have been identified in patients with sepsis,[11] and a so-called hyperinflammatory phenotype has been proposed in patients with COVID-19.[12,13] However, to our knowledge, only one other study[14] has specifically investigated the existence of diverse clinical phenotypes for patients with COVID-19 at hospital admission; three phenotypes were also identified in that study[14] on the basis of clinical and laboratory features, using hierarchical clustering in 85 patients admitted to the intensive care unit, with a small number of variables.

In our study, the phenotypes we identified were associated with patient prognosis. By contrast with studies that generate outcome prediction scores or identify outcome predictors, in which the independent predictive association of each variable with the outcome is assessed, phenotypes provide information about how the population can be classified according to clustering of variables and how such clusters are associated with the outcome. As age and oxygen saturation are strong independent predictors of mortality,[4] we did a stratified analysis of these variables. The results of this analysis suggest that the association of phenotypes with mortality is not only due to the different distribution of these variables in the phenotypes, but that the phenotypes are

consistently associated with different mortality risks. However, the phenotypes are not expected to provide accurate prediction of prognosis, as done by predictive modelling, as the outcome rates in the phenotypes depend on the exact distribution of the strongest outcome predictors in each population to which the phenotypes are applied. In this sense, phenotypes are complementary to predictive scores. Beyond that, the phenotypes might reflect different profiles of pathogen and host interactions, as a consequence of different infecting viral load, natural or acquired humoral and cellular immune response against SARS-CoV-2, or cell–receptor features and expression, alongside host genetic background.[6–8] Since the databases used in this study only included phenotypic profiles and manifestations, we cannot provide information about underlying immunological or virological mechanisms. Future studies could reproduce the phenotypes and investigate their correlations with virological, immunological, and genetic data.

We did not analyse the duration of disease at hospital admission because the start of symptoms can be difficult to assess in many patients and can be confused with manifestations related to chronic conditions; in our experience, this is particularly frequent in older patients with comorbidities. The duration of symptoms could be relevant to differentiate between the viral and inflammatory phases of the disease,[13] but a clear cutoff in the number of days to differentiate between the phases cannot currently be defined.

Classification of patients into phenotypes might be useful to design treatment strategies. Very low-risk patients (eg, those with phenotype A who are younger than 60 years or with oxygen saturation >95%), who would need lower degrees of watchfulness and care, might be identified and discharged for ambulatory follow-up. Patients without initial criteria for being admitted to the intensive care unit but with phenotype B or phenotype C could be closely monitored during admission. As some aspects of the pathophysiology of the infection in patients with different phenotypes might be different, the therapeutic approach might need to be tailored on a patient-by-patient basis.[15] Since phenotype C comprises patients with laboratory parameters suggestive of a hyperinflammatory state, such patients might be selected to investigate the efficacy of anti-inflammatory drugs. This strategy would allow more specific and efficient design of randomised trials. However, whether these phenotypes are useful for clinical purposes requires further investigation of the underlying mechanisms and more specific studies.

Since the phenotypes were identified using a high number of variables, it would be difficult to apply

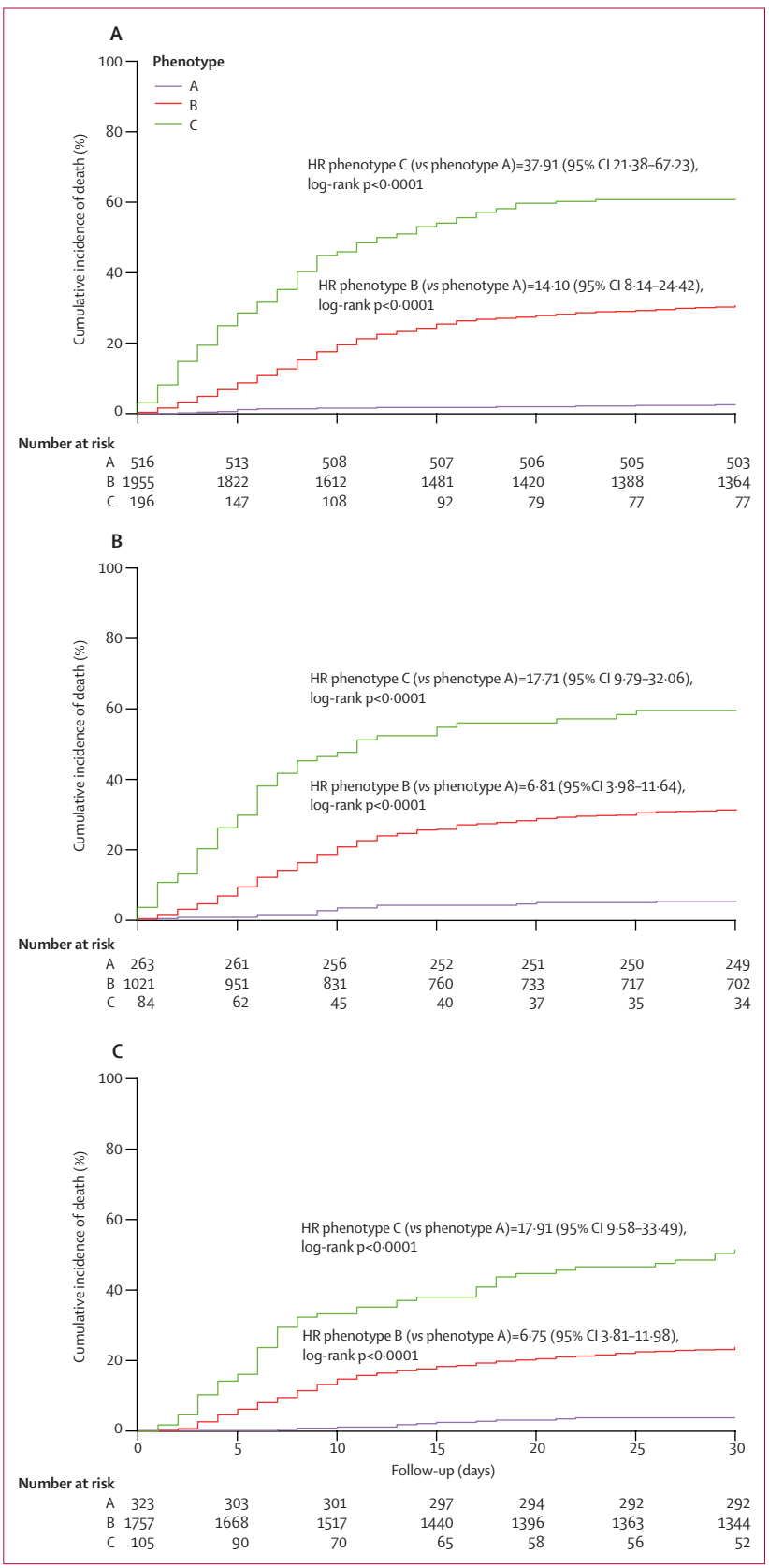

Figure 3: **Probability of death up to day 30 according to phenotypes in the derivation cohort (A), internal validation cohort (B) and external validation cohort (C)**
HR=hazard ratio.

them clinically in the absence of automated big data management. Therefore, we developed and validated a simplified probabilistic prediction model for phenotype assignment. A publicly available calculator and app have been developed to facilitate the classification of patients admitted to hospital with COVID-19 into phenotypes, using the probabilistic model for phenotype assignment.

Limitations of our study are the high proportion of patients classified into phenotype B, reflecting the profile of the patients admitted during the first weeks of the epidemic in saturated hospitals, the exclusive participation of Spanish hospitals, and the high proportion of missing data for several variables. Hospital admission criteria might be different in other countries or at different times during the pandemic; however, the cohorts we used included patients with varying severity of disease. Some symptoms might not have been reported by the most severely ill patients. Finally, the phenotypes were derived and validated at hospital admission and would be useful for decisions at that time; whether changes in evolution due to the natural history of the disease or the influence of treatments modify the phenotype assignment needs further study. Strengths of our study include the use of well characterised cohorts, the inclusion of a high number of variables from different domains, and the validation.

In conclusion, patients admitted to hospital with COVID-19 can be classified into phenotypes that have prognostic implications. We developed a simplified tool for the probabilistic classification of patients into phenotypes. Further studies are needed to elucidate the underlying pathophysiological mechanisms leading to a particular phenotype.

Contributors

JR-B, BG-G, MDdT, JP, JC, PR, IJ, MY, JRA, and JB conceived and designed the study. All members of the REIPI-SEIMC COVID-19 and COVID@HULP groups acquired the study data. JR-B, BG-G, MDdT, AMB, AC, JP, JC, PR, IJ, MY, JRA, and JB analysed and interpreted the data. JRB, BGG, and MDdT drafted the manuscript. Critical revisions to the manuscript were made by all members of the REIPI-SEIMC COVID-19 Group. JR-B, PR, JB, and JRA obtained the funds for the study. BG-G and JR-B verified all data. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

IJ received honoraria for participating in an advisory board for Gilead Sciences and educational activities for ViiV. JB received research grants from AbbVie, Gilead Sciences, Merck, and ViiV, and honoraria for speaker or advisory board participation from AbbVie, Gilead Sciences, Janssen, Merck, and ViiV. JRA received fees for advisory board and speaker participation and research grant support from Viiv, Janssen, Gilead Sciences, Merck, Teva, Alexa, and Serono. PR has received research grants from Gilead Sciences and Merck, and honoraria for speaker or advisory board participation from Gilead Sciences, AbbVie, and ViiV. All other authors declare no competing interests.

Data sharing

Data collected for the study, including deidentified participant data and a data dictionary defining each field in the set, will be made available to other investigators upon request to the corresponding author, after approval of a proposal by the REIPI-SEIMC COVID-19 and COVID@HULP groups boards, with a signed data access agreement.

Acknowledgments

The study was funded by Instituto de Salud Carlos III, Spanish Ministry of Science and Innovation (COV20/01031), and Fundación SEIMC/GeSIDA. Additionally, JR-B, BG-G, JB, IJ, JC, JP, and JRA received funding for research from Plan Nacional de I+D+i 2013–2016 and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación y Universidades (cofinanced by European Development Regional Fund "A way to achieve Europe"), and operative programme Intelligent Growth 2014–2020 through the following networks: Spanish Network for Research in Infectious Diseases (RD16/0016/0001 [JR-B, BG-G, MDdT], RD16/0016/0005 [JC], and RD16/0016/0009 [JP]) and Spanish AIDS Research Network (RD16/0025/0017 [JB], RD16/0025/0018 [JRA], RD16/0025/00XX [IJ]). We thank Alejandro González-Herrero for programming of the web tool and app. This study was presented at the ESCMID Conference on Coronavirus Disease, Sept 23–25, 2020.

References

1 Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med* 2020; **382:** 1708–20.

2 Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *JAMA* 2020; **323:** 2052–59.

3 Docherty AB, Harrison EM, Green CA, et al. Features of 20133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* 2020; **369:** m1985.

4 Berenguer J, Ryan P, Rodríguez-Baño J, et al. Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. *Clin Microbiol Infect* 2020; **26:** 1525–36.

5 Borobia AM, Carcas AJ, Arnalich F, et al. A cohort of patients with COVID-19 in a major teaching hospital in Europe. *J Clin Med* 2020; **9:** 1733.

6 Magleby R, Westblade LF, Trzebucki A, et al. Impact of SARS-CoV-2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019. *Clin Infect Dis* 2020; published online June 30. https://doi.org/10.1093/cid/ciaa851.

7 Mateus J, Grifoni A, Tarke A, et al. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* 2020; **370:** 89–94.

8 Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 2020; **383:** 1522–34.

9 Gattinoni L, Chiumello D, Caironi P, et al. COVID-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive Care Med* 2020; **46:** 1099–102.

10 Sudre CH, Lee KA, Lochlainn MN, et al. Symptom clusters in Covid19: a potential clinical prediction tool from the COVID Symptom study app. *medRxiv* 2020; published online June 16. https://doi.org/10.1101/2020.06.12.20129056 (preprint).

11 Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019; **321:** 2003–17.

12 Sinha P, Calfee CS, Cherian S, et al. Prevalence of phenotypes of acute respiratory distress syndrome in critically ill patients with COVID-19: a prospective observational study. *Lancet Respir Med* 2020; **8:** 1209–18.

13 Manson JJ, Crooks C, Naja M, et al. COVID-19-associated hyperinflammation and escalation of patient care: a retrospective longitudinal cohort study. *Lancet Rheumatol* 2020; **2:** e594–602.

14 Azoulay E, Zafrani L, Mirouse A, Lengliné E, Darmon M, Chevret S. Clinical phenotypes of critically ill COVID-19 patients. *Intensive Care Med* 2020; **46:** 1651–52.

15 Cao W, Li T. COVID-19: towards understanding of pathogenesis. *Cell Res* 2020; **30:** 367–69.