Development of a multivariable model predicting mortality risk from comorbidities in an Italian cohort of 18,286 confirmed COVID-19 cases aged 40 years or older

Sviluppo di un modello predittivo del rischio di decesso sulla base delle comorbidità in una coorte di 18.286 casi confermati di COVID-19 con almeno 40 anni d'età

Anita Andreano,[1] Rossella Murtas,[1] Sara Tunesi,[1] Federico Gervasi,[1] Pietro Magnoni,[1] Antonio Giampiero Russo[1]
[1] UOC Epidemiology Unit, Agency for Health Protection of the Metropolitan Area of Milan (Italy)
Corresponding author: Antonio Giampiero Russo; agrusso@ats-milano.it

WHAT IS ALREADY KNOWN

- **Age and male gender are associated with a higher mortality risk from COVID-19.**
- **Hypertension, cardiovascular diseases, diabetes, cancer, and several other chronic conditions are associated with a higher mortality risk from COVID-19.**

WHAT THIS STUDY ADDS

- **In a large cohort of confirmed COVID-19 cases from the Metropolitan area of Milan (Lombardy Region, Northern Italy), an accurate and well discriminating prediction model based on age, gender, and comorbidities was developed using administrative data.**
- **The most important chronic conditions predicting 30-day mortality were chronic heart failure, tumours, diabetes, and severe kidney disease.**
- **The developed predictive model will allow to identify high-risk subjects to primarily target for prevention and therapy in case of further epidemic waves of COVID-19.**

ABSTRACT

OBJECTIVES: to develop a risk prediction model for 30-day mortality from COVID-19 in an Italian cohort aged 40 years or older.

DESIGN: a population-based retrospective cohort study on prospectively collected data was conducted.

SETTING AND PARTICIPANTS: the cohort included all swab positive cases aged 40 years older (No. 18,286) among residents in the territory of the Milan's Agency for Health Protection (ATS-MI) up to 27.04.2020. Data on comorbidities were obtained from the ATS administrative database of chronic conditions.

MAIN OUTCOME MEASURES: to predict 30-day mortality risk, a multivariable logistic regression model, including age, gender, and the selected conditions, was developed following the TRIPOD guidelines. Discrimination and calibration of the model were assessed.

RESULTS: after age and gender, the most important predictors of 30-day mortality were diabetes, tumour in first-line treatment, chronic heart failure, and complicated diabetes. The bootstrap-validated c-index was 0.78, which suggests that this model is useful in predicting death after COVID-19 infection in swab positive cases. The model had good discrimination (Brier score 0.13) and was well calibrated (Index of prediction accuracy of 14.8%).

CONCLUSIONS: a risk prediction model for 30-day mortality in a large COVID-19 cohort aged 40 years or older was developed. In a new epidemic wave, it would help to define groups at different risk and to identify high-risk subjects to target for specific prevention and therapeutic strategies.

Keywords: COVID-19, chronic conditions and COVID-19, predictors of death from COVID-19, multivariable logistic prediction model

RIASSUNTO

OBIETTIVI: sviluppare un modello predittivo di morte a 30 giorni per COVID-19 in una coorte italiana di età pari o superiore a 40 anni.

DISEGNO: è stato condotto uno studio di coorte retrospettivo basato sui dati raccolti in modo prospettico.

SETTING E PARTECIPANTI: la coorte includeva tutti i casi positivi al tampone nasofaringeo di età uguale o superiore a 40 anni (n. 18.286) tra i residenti nel territorio dell'Agenzia di tutela della salute (ATS) di Milano registrati sino al 27.04.2020. I dati sulle comorbidità sono stati ottenuti dal database delle patologie croniche dell'ATS stessa.

PRINCIPALI MISURE DI OUTCOME: per prevedere il rischio di mortalità a 30 giorni, è stato sviluppato un modello di regressione logistica multipla, comprendente età, genere e condizioni selezionate, seguendo le linee guida TRIPOD. Sono state valutate la discriminazione e la calibrazione del modello.

RISULTATI: dopo l'età e il genere, i fattori predittivi più importanti di mortalità a 30 giorni si sono rivelati il diabete mellito, il tumore in trattamento di prima linea, l'insufficienza cardiaca cronica e il diabete complicato. Il c-index validato mediante bootstrap è di 0,78, il che suggerisce che questo modello è utile per prevedere la morte dopo infezione da COVID-19 nei casi positivi al tampone. Il modello ha una buona discriminazione (Brier score 0,13) ed è ben calibrato (*index of prediction accuracy* 14,8%).

CONCLUSIONI: è stato sviluppato un modello predittivo del rischio di mortalità a 30 giorni in un'ampia coorte di soggetti positivi a COVID-19 di età pari o superiore a 40 anni. In nuove ondate epidemiche, sarà utile nel definire gruppi di rischio ed identificare soggetti ad alto rischio.

Parole chiave: COVID-19, patologie croniche e COVID-19, fattori predittivi del rischio di decesso per COVID-19, modello predittivo di regressione logistica multipla

INTRODUCTION

After starting in China in December 2019 and extending to bordering countries, the first European small clusters of Coronavirus disease 2019 (COVID-19) were detected in France, Germany, and UK in late January 2020.[1] Then, the epidemic outburst in Italy, which has been the first and one of the most hit European countries. 'Case one' was diagnosed on 20th February in the Province of Lodi (Lombardy Region, Northern Italy).[1,2] Reported crude case-fatality rate from COVID-19 in Italy was 9.3% in the first phase of the epidemic, higher than those reported in other European countries, such as France (2.6%) and Germany (0.7%).[3] In that period, Italy had also the second largest percentage of cases older than 60 years after the Netherlands (respectively, 56% and 58%) compared to other EU countries (e.g., Germany 24% and France 36%).[4] COVID-19 is characterized by a reported case-fatality rate critically varying with age, with 94% fatalities occurring in individuals older than 60 years.[4.] This parallels with the increased prevalence of comorbidities in the older population.[5] Increased risk of developing critical illness has been signalled for hospitalized patients with cancer and chronic obstructive pulmonary disease (COPD), and being associated with Charlson's index or number of comorbidities.[6-8] In other studies, history of hypertension, cardiovascular disease, diabetes mellitus (DM), severe asthma, dementia, cancer, chronic kidney, liver, and rheumatic diseases were found to be associated with severe disease or death from COVID-19.[7,9-18] During the course of the epidemic, different prediction models were developed for short-term mortality in different countries.[8,19-23] However, the majority of them predicts mortality risk for either hospitalized patients or for subjects presenting to an emergency room or an outpatient clinic, and includes some clinical or biochemical parameter. For this reason, and also because they were developed on selected sub-populations of severe cases, they cannot be used at a community level to predict the individual risk for an entire population. Therefore, it is relevant to develop a prognostic model on a general population of COVID-19 confirmed cases based only on demographic factors and comorbidities readily available from health databases. All this allows to determine the individual risk of every subject in the population and to give tailored preventive recommendations to those at higher risk, and will help physicians to decide, based on the estimated individual risk, which strategy has to be undertaken (e.g., home monitoring, hospitalization) in case of suspected symptomatology or nasopharyngeal swab positivity.

The aim of this study was to develop a 30-day mortality prediction model in swab-positive cases of COVID-19 aged 40 years or older using demographic and comorbidities data from administrative health databases, in order to score the entire population residing in the Metropolitan area of Milan based on the predicted risk.

METHODS

STUDY DESIGN, DATA SOURCES, AND MEASURES

This was a population-based cohort study on data prospectively collected between February and 27 April 2020. The cohort included all COVID-19 cases in residents in the study area registered with the Regional Health System (RHS). The study area corresponds to the territory covered by the Agency for Health Protection of the Metropolitan Area of Milan (ATS-MI), including 193 municipalities in Lombardy Region (Northern Italy), with a total population of 3.48 million inhabitants. It also comprises the municipality of Codogno, which was at the origin of the first Italian epidemic outbreak. The estimated percentage of people resident but not registered with the RHS is 4.1%. This was estimated on the Milan and Sesto San Giovanni municipalities, for which the civil registry of residents is available to the ATS.

A confirmed case is defined as a person with a real-time polymerase chain reaction (RT-PCR) positive result for COVID-19, irrespective of clinical signs and symptoms. From the beginning of the outbreak, all tracing activities of the ATS-MI were included in a web-based platform, developed by the Epidemiological Unit of ATS-MI, called *Milano COV*, including cases and related contacts (details on the information system are described in Murtas et al.).[24] Also, from the beginning of the epidemic, the Lombardy Region daily sent the lists of swab-positive outpatients and hospitalized COVID-19 cases to each ATS. This information was integrated with *Milano COV* data in the Integrated Datawarehouse for COVID Analysis in Milan, through deterministic record linkage on individual tax code. Cases underwent epidemiological investigation to provide description of the clinical presentation and course of COVID-19. The demographic information in the Integrated Datawarehouse for COVID Analysis in Milan was verified with the Health Service Register of the Lombardy Region (age, gender, place of residence). A random unique id was attributed to every subject. This same id was assigned to each subject in all other administrative databases of the ATS-MI, deterministically linking it on individual tax code. The Integrated Datawarehouse for COVID Analysis in Milan and the other administrative databases were then anonymized prior to analysis. Individual level comorbidities data were derived using the chronic disease administrative database of the ATS-MI, according to the algorithms specified in the Regional Act X/6164[25] and X/7655[26] of 2017, summarized in English in Murtas et al.[24] The algorithms are based on the following databases: hospital discharge, out-

patient visits and exams, exemption from co-payment, and drug prescriptions.

Vital status was derived from the early notification system of the ATS-MI,[27] set-up from the beginning of the epidemic. In this system, deaths are either communicated from each Municipality to the ATS and manually entered in the Health Service Register or directly imputed in *Milano COV* for subjects already in the database. The vital status was assessed on 23rd May 2020. A vital status at 30-day from diagnosis, which was defined for confirmed-cases as the first date between registered symptom onset and the swab positivity result, was determined. The date of symptom onset in the database was derived from the epidemiological interview or from the date of first access to an emergency department or first thorax CT scan, in this order of priority. If none of these data was available and the patient had been hospitalized, the date of hospital admission was used. For a minority of patients, infected in the early phase of the epidemic and for whom no onset dates were available, a univariate random imputation was performed, according to a uniform distribution U(a,b) with parameters a=10Feb2020 and b=17Feb2020. For this analysis, patients with a date of death which exceeded 30 days after diagnosis were considered as alive.

POPULATION

From the *Milano COV* database of the ATS-MI, all subjects with nasopharyngeal swab positive for COVID-19 at 27th April 2020 were extracted. All subjects that were at least 40 years old when diagnosed with COVID-19 were included in the analysis (figure S1, on-line Supplementary material). The choice was made both because deaths were very rare under age 40 (n. 8) and because the majority of comorbidities of interest were very rare. No further exclusion was performed.

DEVELOPMENT OF THE PREDICTIVE MODEL

The TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guidelines were followed,[28] including a 22-item checklist, to report the development and validation of the model predicting 30-day mortality risk in COVID-19 cases aged 40 years or older, using data on 65 comorbidities from an administrative health database. Given the high number of events and the minimal cost represented by the collection of this information, while the a priori clinical knowledge on the associations between comorbidities and death from COVID-19 is limited and in order to maximize the expected discrimination ability based on administrative data only, it was decided to develop a full model without performing model selection using automatic statistical techniques. The number of variables to be firstly introduced in the model on the basis of clinical-epidemiological considerations were reduced. Conditions that were absent or very rare in the cohort were not included and some relatively rare diseases with similar clinical consequences were grouped. No patient was lost to follow-up (i.e., emigrated outside the region) before 30 days and, among alive patients, only 160 (1.1%) had a follow-up time shorter than 30 days at 23rd May 2020 (minimum and median follow-up times: 26 and 29 days, respectively). Consequently, multiple logistic regression was considered adequate to develop the predictive model. Collinearity using Phi correlation index and the Variance Inflation Factor (VIF) were evaluated. Twenty-one pre-specified interaction terms, chosen on epidemiological and clinical considerations (table S2), and kept in the model only the significant ones at p=0.05, were also tested. The heuristic shrinkage estimator (van Houwelingen-le Cessie) was calculated including d.f. for testing interaction.[29] Overfitting is likely to be a concern when this estimator has values below 0.85.[30] The functional form of the relationship between age and death was evaluated and a restricted cubic regression spline with three knots based on graphical evaluation and AIC comparisons was chosen. Knots were chosen according to Harrell's rule.[31] All the other predictors were included as binary variables. Estimated model parameters are reported in table S2. The adjusted effects of predictors on the risk of death were reported as the odds ratio (OR) and the corresponding 95% and 99% Wald confidence intervals (CIs). To graphically present effect estimates, the OR of increasing age from 60 to 80 years was calculated. Also, to present a single global OR for factors having interaction terms, values were set to age 70 (because it was the mean age in the development cohort), gender to female, and absence of the interested comorbidity.[31] A metric of the absolute importance of each model term in predicting 30-day mortality from COVID-19 was calculated as the log+1 transformed Wald chi^2 minus the predictor degrees of freedom, and presented graphically.[31] Internal validation of the model was performed using bootstrap resampling (B 1,000) to evaluate the discrimination and calibration of the model.[32] Ten-fold validation was additionally used. Discrimination was assessed through Harrell's c-index/area under the curve (AUC): a value of 0.5 is equivalent to random prediction, while a value of 1 implies perfect discrimination. Weak model calibration was evaluated assessing calibration intercept and slope, with an intercept of 0 and a slope of 1 indicating no over or underfitting, and no systematic over or underestimation of predicted risk. Model moderate calibration was evaluated constructing a calibration plot, using locally-weighted polynomial regression for smoothing, to assess correspondence between predicted risk and observed event rates among patients with the same pre-

dicted risk.[33-35] Overall prediction accuracy was evaluated through Brier score, which can take values from 0 – for a perfect model – to 0.25 – for a non-informative model. It was also calculated the index of prediction accuracy (IPA), a scaled version of the Brier score, reflecting both discrimination and calibration, and having negative values for models performing worse than the null.[36] As a sensitivity analysis, the same prediction model was fitted on the subsets of never hospitalized and hospitalized patients. All analyses were performed with R software (v. 3.6.3, R Core Team, Vienna, Austria) and R packages rms (v 5.1-4, F. Harrell).

RESULTS

DESCRIPTION OF THE COHORT

On 27th April 2020, the COVID-19 database of the ATS-MI included 20,364 swab positive cases; 10% of them were younger than 40 years old and were therefore excluded, leaving 18,286 subjects. Less than 1% (No. 165) of swab positive cases had a missing diagnosis date, which was imputed as described in the methods. The earliest diagnosis date was 10th February, while the latest was dated 27th April. In the cohort, 3,832 patients deceased by 23rd May 2020; 9% of them (No. 333) died after 30 days and were consequently considered as alive for the analysis, resulting in 3,499 events.

Median age was 71 years and there were slightly more females (52%). Fifty-six percent of the swab positive cases were hospitalized at some point, and 21% got infected in a residential setting. Thirty-five percent of cases were in the City of Milan, while 15% in the province of Lodi, where the epidemic started. Among the most common comorbidities, hypertension had a prevalence of 45%, with a 2019 prevalence in the ≥40-year population of the ATS-MI of 30%. The same figure for ischaemic heart disease was 13%, with an ATS-MI population prevalence of 6%, and 15% for DM (including type 1 and 2 and complicated DM), with an ATS-MI overall prevalence of 8% (table 1).

DEVELOPMENT AND INTERNAL VALIDATION OF THE RISK PREDICTION MODEL FOR 30-DAY MORTALITY

The database of chronic diseases includes 65 conditions (table S1). The number of predictors was reduced to include in the model to 32, based on clinical and epidemiological judgment. The following six variables were excluded, because they were not present or very rare in the confirmed COVID-19 cohort:

- perinatal conditions: no subjects with this condition in the development cohort;
- ill-defined conditions: no subjects;
- optic neuromyelitis: 1 subject;
- diseases of the genitourinary system: 2 subjects;
- infectious and parasitic diseases: no subjects;
- congenital malformations: 10 subjects and no events.

Being an implausible risk factor, also chronic cutaneous diseases (25 subjects) were not included. The following categories of the chronic conditions database were merged into one, according to sparsity or based on similar clinical consequences:

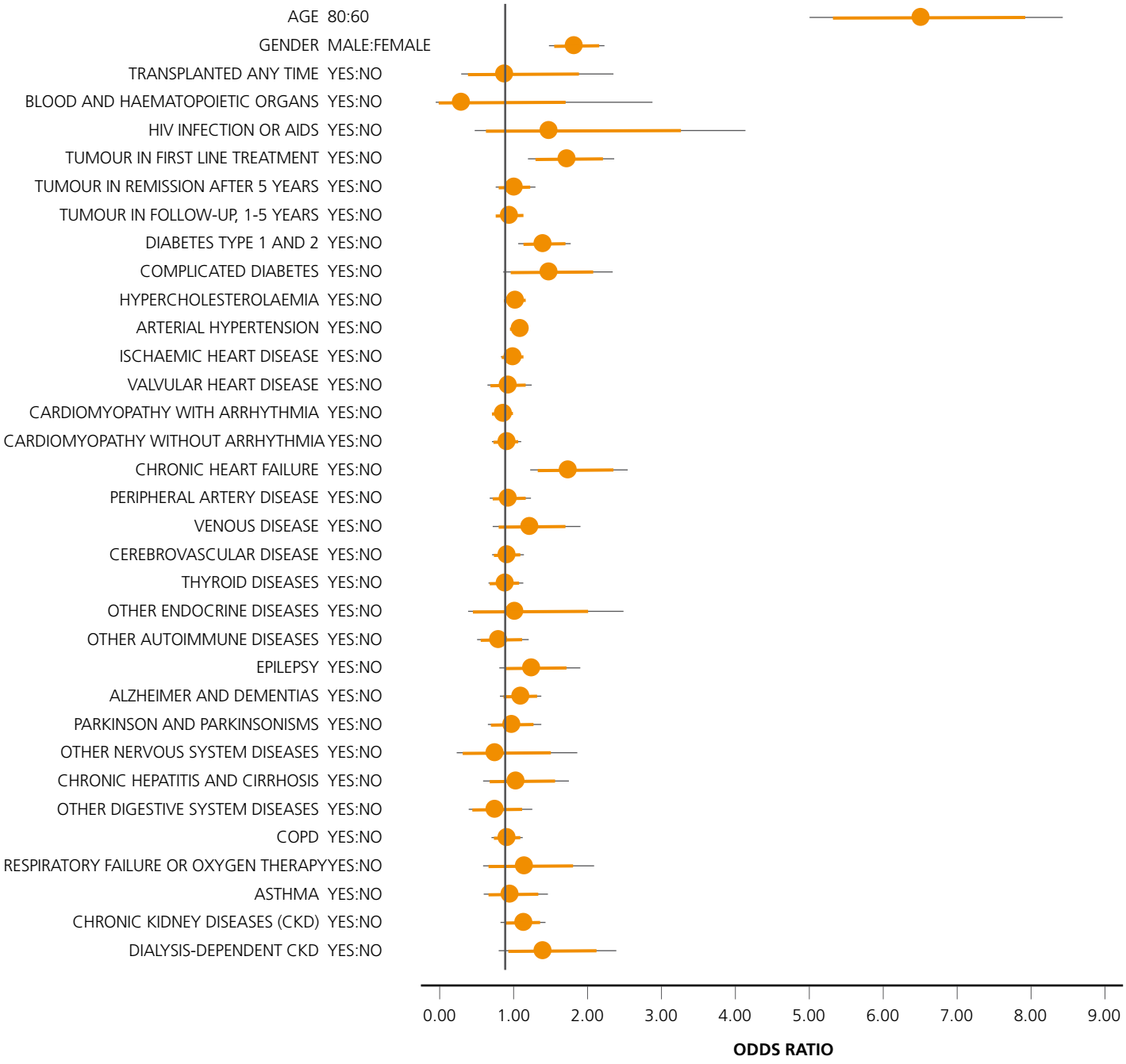
- transplanted within 2 years and transplanted from more than 2 years;
- complicated type 1 and type 2 diabetes;
- thyroid diseases (hypothyroidism, Basedow's disease and hyperthyroidism, Hashimoto's thyroiditis);
- other endocrine diseases (Cushing's syndrome, Addison's disease, hyper and hypoparathyroidism, acromegaly gigantism, diabetes insipidus, pituitary dwarfism, others);
- chronic hepatitis and cirrhosis;
- digestive system diseases (chronic pancreatitis, ulcerative colitis and Crohn's disease, others);
- autoimmune diseases (autoimmune haemolytic anaemias, systemic sclerosis, ankylosing spondylitis, Sjogren's disease, rheumatoid arthritis, psoriasis and psoriatic arthropathy, systemic lupus erythematosus, myasthenia gravis);
- Alzheimer's disease and dementia;
- other nervous system diseases (multiple sclerosis, other diseases of the nervous system and sense organs).

Diabetes and complicated diabetes are mutually exclusive, as well as Chronic Kidney Disease (CKD) and dialysis-dependent CKD, and the three tumour categories. No collinearity problems were detected (largest Phi correlation index =0.36 between complicated diabetes and ischemic heart disease, and between ischemic heart disease and hypercholesterolemia). The variables with the highest variance inflation factor (VIF) were ischemic heart disease (1.30), chronic heart failure (CHF) (1.28), and age (1.27). Among the pre-specified tested interactions (table S2), those between age and five other predictors (gender, CHF, DM, complicated DM, and tumour in first-line treatment) were statistically significant and were kept in the model, after comparing bootstrap validated c-indexes of the model with and without the interaction terms. The van Houwelingen-le Cessie shrinkage estimator considering 66 d.f. (including 21 d.f. for the tested, but not included interactions) was 0.98, implying no concern for overfitting.

The full multivariable logistic regression model included age, gender, the 32 comorbidities, and the 5 interaction terms. Its results are graphically displayed in figure 1, in terms of adjusted OR of the main effect combined with all interactions involving the predictor. Model coefficients for all terms and p-values are reported in table S3. The model-adjusted likelihood of dying within 30 days

CHARACTERISTICS	SWAB POSITIVE CASES		
	OVERALL (No. 18,286)	DECEASED WITHIN 30-DAYS	
		NO (No. 14,787)	YES (No. 3,499)
	No. (%)	No. (%)	No. (%)
GENDER			
Male	8,704 (47.6)	6,664 (45.1)	2,040 (58.3)
AGE CLASS			
40-59 years	5,884 (32.2)	5,740 (38.8)	144 (4.1)
60-79 years	6,188 (33.8)	4,901 (33.1)	1,287 (36.8)
80+ years	6,214 (34.0)	4,146 (28.0)	2,068 (59.1)
SETTING			
Home	4,765 (26.1)	4,401 (29.8)	364 (10.4)
Hospitalized	9,604 (52.5)	7,117 (49.2)	2,487 (64.9)
Residential	3,363 (18.4)	2,755 (18.6)	608 (17.4)
Residential followed by hospitalization	554 (3.0)	283 (1.9)	271 (7.7)
GEOGRAPHIC LOCATION			
Lodi Province (starting place of the outburst)	2,669 (14.6)	2,101 (14.2)	568 (16.2)
City of Milan	6,478 (35.4)	5,188 (35.1)	1,290 (36.9)
Milan Province	9,139 (50.0)	7,498 (50.7)	1,641 (46.9)
NUMBER OF COMORBIDITIES			
None	6,363 (34.8)	5,717 (39.6)	646 (16.9)
1-3	8,709 (47.6)	6,756 (46.7)	1,953 (51.0)
≥4	3,214 (17.6)	1,981 (13.7)	1,233 (32.2)
SPECIFIC COMORBIDITIES			
Transplanted any time	69 (0.4)	57 (0.4)	12 (0.3)
Blood and Hematopoietic organs	26 (0.1)	31 (0.2)	5 (0.1)
HIV infection or AIDS	65 (0.4)	55 (0.4)	10 (0.3)
Tumour in first line treatment	1,005 (5.5)	692 (4.7)	313 (8.9)
Tumour in follow-up 1-5 years	772 (4.2)	553 (3.7)	219 (6.3)
Tumour in remission after 5 years	1,071 (5.9)	783 (5.3)	288 (8.2)
Type 1 Diabetes	23 (0.1)	17 (0.1)	6 (0.2)
Type 2 Diabetes	2,385 (13.0)	1,653 (11.2)	732 (20.9)
Complicated diabetes mellitus Type 1 and 2	422 (2.3)	273 (1.8)	149 (4.3)
Hypercholesterolaemia	2,293 (12.5)	1,572 (10.6)	721 (20.6)
Arterial hypertension	8,156 (44.6)	5,907 (39.9)	2,249 (64.3)
Ischaemic heart disease	2,361 (12.9)	1,534 (10.4)	827 (23.6)
Valvular heart disease	458 (2.5)	319 (2.2)	139 (4.0)
Cardiomyopathy with arrhythmia	2,297 (12.6)	1,541 (10.4)	756 (21.6)
Cardiomyopathy without arrhythmia	1,693 (9.3)	1,145 (7.7)	548 (15.7)
Chronic heart failure	1,398 (7.6)	861 (5.8)	537 (15.3)
Peripheral artery disease	587 (3.2)	387 (2.6)	200 (5.7)
Venous diseases	197 (1.1)	139 (0.9)	58 (1.7)
Cerebrovascular disease	820 (4.5)	543 (3.7)	277 (7.9)
Thyroid diseases	1064 (5.8)	896 (6.1)	168 (4.8)
Other endocrine diseases	63 (0.3)	50 (0.3)	13 (0.4)
Other autoimmune diseases	346 (1.9)	272 (1.8)	64 (1.8)
Epilepsy	253 (1.4)	179 (1.2)	74 (2.1)
Alzheimer and Dementias	669 (3.7)	446 (3.0)	223 (6.4)
Parkinson and Parkinsonism	310 (1.7)	211 (1.4)	99 (2.8)
Other nervous system diseases	87 (0.5)	76 (0.5)	11 (0.3)
Chronic hepatitis and cirrhosis	419 (2.3)	334 (2.3)	85 (2.4)
Other digestive system diseases	232 (1.3)	195 (1.3)	37 (1.1)
Chronic obstructive pulmonary disease	896 (4.9)	603 (4.1)	293 (8.4)
Respiratory failure or Oxygen therapy	97 (0.5)	63 (0.4)	34 (1.0)
Asthma	391 (2.1)	335 (2.3)	56 (1.6)
Chronic kidney disease	584 (3.2)	369 (2.5)	215 (6.1)
Dialysis-dependent chronic kidney disease	147 (0.8)	95 (0.6)	52 (1.5)

Table 1. Characteristics of the cohort of residents in the ATS of Milan with swab positive SARS-CoV-2 infection between February and 27th April 2020, followed-up to 23rd May 2020, surviving or not 30 days after the infection.
Tabella 1. Caratteristiche della coorte di residenti nell'ATS di Milano positivi al tampone per SARS-CoV-2 tra febbraio e il 27 aprile 2020, seguiti fino al 23 maggio 2020, vivi o deceduti entro 30 giorni dall'infezione.



Note: For age, the adjusted OR for increasing from 60 to 80 years is displayed. Global OR for factors having interaction terms in the model were calculated setting age to 70, gender to female, and absence of the involved comorbidity. / ***Nota:*** *Per l'età si visualizza l'OR aggiustato per un incremento da 60 a 80 anni. Gli OR globali per fattori che hanno termini d'interazione nel modello sono calcolati fissando l'età a 70 anni, il genere a femminile e la comorbidità coinvolta come assente.*

Figure 1. Results from the multivariable logistic regression model predicting 30-days mortality risk from COVID-19 in the development cohort of 18,286 swab positive cases, presented as adjusted odds ratio (orange bullets) with 95% (orange line) and 99% confidence intervals (black line).
Figura 1. Risultati dal modello di regressione logistica multivariata predittivo del rischio di mortalità a 30 giorni per COVID-19 nella coorte di sviluppo del modello, composta da 18.286 casi positivi al tampone. I risultati sono presentati mediante odds ratio aggiustati (punti blu scuro) e relativi intervalli di confidenza al 95% (barre arancioni) e al 99% (linee nere).

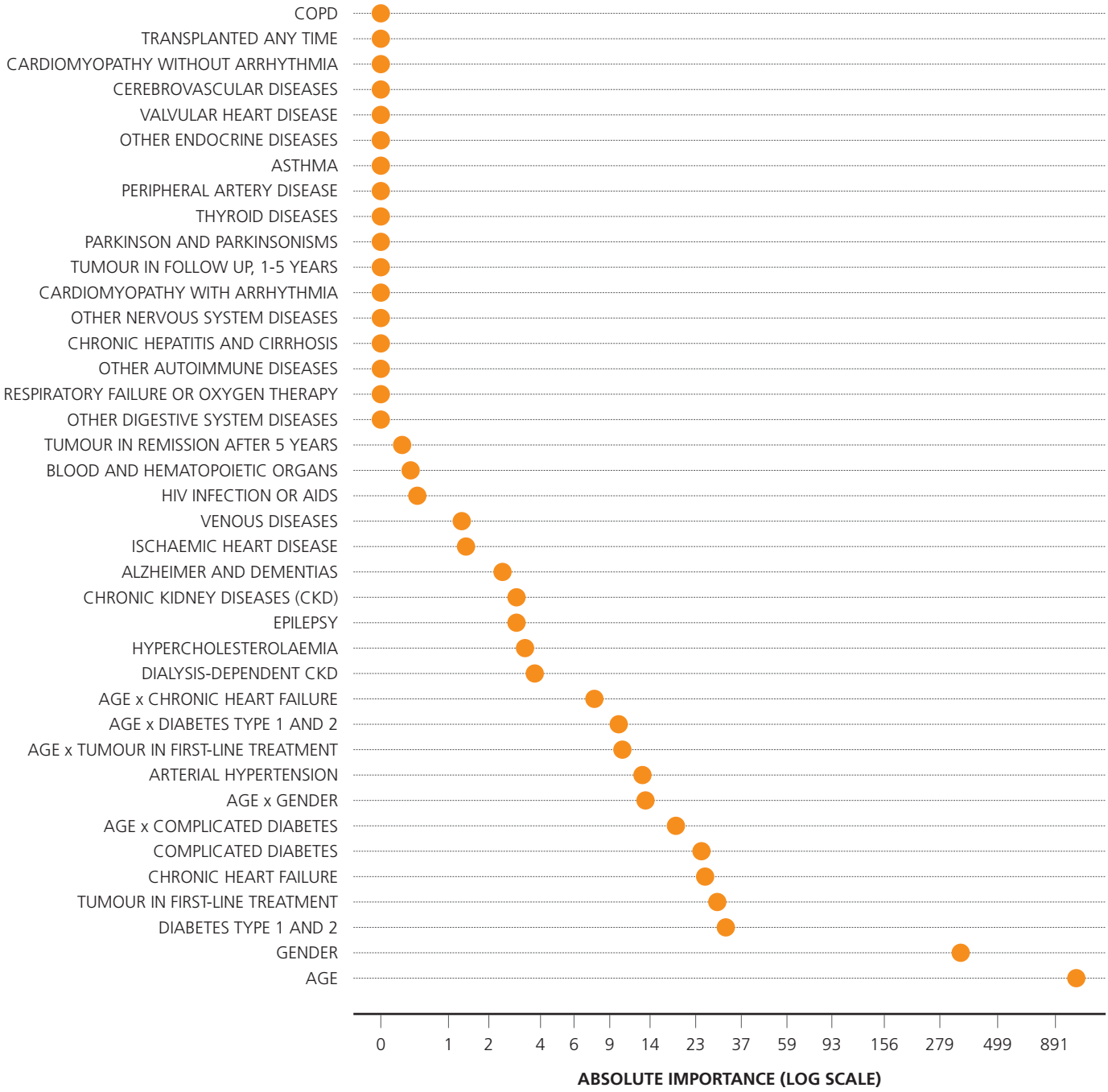


Figure 2. Absolute importance of predictors, measured by Wald chi² value minus the degrees of freedom of the predictor (log-scale), based on multivariable logistic regression in the development cohort of 18,286 COVID-19 swab positive cases.
Figura 2. Importanza assoluta dei predittori, misurati tramite il chi² di Wald meno i gradi di libertà del predittore (scala logaritmica), basata sul modello di regressione logistica multivariata nella coorte di sviluppo costituita da 18.286 persone positive al tampone per COVID-19.

		BOOTSTRAP CORRECTED (B=1000)	10-FOLD VALIDATION
Brier score	0.130	0.1309	0.1308
c-index/AUC	0.788	0.7835	0.7838
CALIBRATION			
intercept	0 (-0.04;0.04)		
slope	1 (0.95;1.05)		

Table 2. Internal validation of the multivariable logistic model predicting 30-day risk of death in the swab positive COVID-19 cohort (No. 18,286).
Tabella 2. Validazione interna del modello logistico multivariato predittivo del rischio di morte a 30 giorni nella coorte di soggetti con tampone positivo a COVID-19 (n. 18.286).

from COVID-19 symptom onset was higher in older patients with OR 6.8 (95%CI 5.6-8.2, for patients with 80 years vs 60 years), males with OR 2.0 (95%CI 1.7-2.3, compared to females), and in patients with chronic heart failure with OR 1.9 (95%CI 1.5-2.5), tumours in first-line treatment OR 1.8 (95%CI 1.4-2.3), diabetes with OR 1.5 (95%CI, 1.3-1.8), complicated diabetes OR 1.6 (95%CI, 1.1-2.2), dialysis-dependent CKD with OR 1.5 (95%CI 1.0-2.2), epilepsy with OR 1.4 (95%CI 1.0-1.8), arterial hypertension with OR 1.2 (95%CI 1.1-1.3), CKD with OR 1.2 (95%CI 1.0-1.5), and hypercholesterolemia with OR 1.2 (95%CI 1.0-1.3). The relative importance of predictors in the model is summarized in figure 2, with diabetes, tumour in first-line treatment, and chronic heart failure being the most important predictor after age and gender. The bootstrap-validated c-index was 0.78, which suggests that the model here presented is useful in predicting death after COVID-19 infection in swab positive cases. This model had good discrimination (Brier score 0.13) and was well calibrated (table 2, figure S2), summarized by an IPA of 14.8%. The results from the stratified models are presented in figures S3 and S4. The AUC in never hospitalized patient was 0.81 and in hospitalized patients 0.78. The first most important predictors were age, gender, and chronic heart failure in both sub-cohorts. Diabetes, with and without complications, the interaction between age and complicated diabetes, and tumour in first line treatment were among the 3rd and 9th most important factors in both sub-cohorts.

DISCUSSION

A risk prediction model for 30-day mortality in a large cohort of confirmed COVID-19 cases was developed using age, gender, and a large number of chronic conditions derived from administrative data. The model has a good discriminative capability, especially considering that predictors derive from administrative data. The prominent role of age and, to a less extent, of male gender on predicting mortality risk was confirmed. The chronic conditions with the greatest ability to predict short-term death were found to be diabetes, tumours in first-line treatment, and chronic heart failure. Although a variety of prediction models has been developed for COVID-19 severity and mortality so far,[6,9,19,21,23,37] the majority are based on hospitalized patients or use clinical parameters at presentation, while few of them allow to predict individual risk before infection from administrative databases.[38] The developed model may be used to calculate the individual risk at population level in subjects aged 40 years or older, and compute the percentage of high-risk subjects corresponding to different thresholds of death not only in the Lombardy RHS, but also in any health system having population-level administrative data on comorbidities. In the event of further waves of COVID-19, this would allow a management of social distancing and quarantine which considers the greater or lesser actual risk afflicting each subject.

Concerning the association between comorbidities used in the model and risk of death from COVID-19, a few studies investigated concurrently the role of several chronic conditions and of demographic factors. The large study on NHS England examining about 11,000 COVID-19-related deaths found that male gender, age, deprivation, diabetes, severe asthma, respiratory disease, chronic heart disease, liver disease, stroke, dementia, other neurological diseases, and reduced kidney function had the greatest association.[17] In addition, a study on 507 hospitalized patients with COVID-19 aged older than 65 years and investigating the relationship between short-term mortality and demographic factor plus several comorbidities found that pre-existing dementia, diabetes, COPD, and depression had the strongest association.[22] Diabetes has been reported to be associated with increased mortality risk in different studies summarized by a metanalysis from Huang et al. that found a pooled relative risk of 2.1 (95%CI 1.4-3.1) varying with age and hypertension.[11] Concerning cardiovascular diseases, many studies did not specify which conditions were actually in-

cluded, so it was maintained separate categories as it was made possible from the large number of subjects. In the developed predictive model, only CHF and hypertension led to an increased risk. A new finding was the predictive role of epilepsy, increasing the odds of death by 40%, that should be further investigated in aetiological studies. Of notice, respiratory failure, COPD, and asthma did not increase 30-day mortality risk in the multivariate model. The most important predictive factors are relatively common in the population aged 40 years or older in the study area, for example, tumour in first line treatment has a 4% prevalence and chronic heart failure 2.6%. Moreover, several important predictors are often present at the same time in this population, such as diabetes and arterial hypertension or either of them with hypercholesterolemia. The consequence is that a large number of subjects would qualify to receive targeted intervention to reduce the risk of COVID-19. Those predictors and their combinations are prevalent in most countries, not only in high-income ones,[39] even if a lower proportion of high-risk subjects may be expected in countries with a younger population. The results presented in this paper also suggest that, even if age is the most important single predictor of short-term death from COVID-19, it is not the only factor to be considered when developing health policies for protecting particular groups of individuals. In this model, being male was a very important predictor of short-term mortality, even accounting for several comorbidities. This was not the object of the study, but further research would be needed to determine which factors put men at higher risk, besides a higher prevalence of comorbidities.

Strengths of this study are the large number of cases in both the development and validation cohort, and the availability of information on a large number of pre-existing chronic conditions assessed in a uniform and inexpensive way. A possible limitation of the study is the exclusion of cases under 40 years which does not allow to estimate the total population at high risk. This choice was made because subjects younger than 40 are very few in the cohort (171 patients in the 0-17 age class and 1,900 patients in the 18-39 age class) and the events so scarce (one death) that the risk in the younger population will be not accurately predicted with the data used for this study. This concern is also due to the fact that, in out of the eight deceased subjects under age 40, six had none of the investigated comorbidities.

Most likely this segment of the population would deserve a different analysis, including data on pre-existing/concomitant acute conditions and information on body mass index (BMI), which has been reported as a relevant factor in younger people, and that at present are not available in the chosen database. The second limitation concerns lack of information on BMI and smoking status, which have been reported as potentially having an effect on mortality.[17,40] Third, it is possible that some of the chronic conditions, including COPD and asthma, are underestimated from administrative databases using the algorithms of the Lombardy region. Unpublished data concerning the validation of 196,472 signalled conditions by general practitioner of the ATS found an overall concordance of 83%. However, BPCO and asthma had a concordance of 70% and 50%, respectively, mainly due to missed diagnosis with the algorithm. Lastly, the baseline risk of the cohort may overestimate the actual population risk in further waves, because in the first phase of the epidemic many patients with a clinically mild presentation were not tested; on the contrary, all hospitalized patients received a nasopharyngeal swab, either before or after hospitalization. However, it is likely that individuating groups with different relative risks is more important for public health policies than precisely estimate absolute individual risk and it is less dependent on the lethality of different variants of the virus. Also, the sensitivity analysis on the sub-cohorts of never-hospitalized and hospitalized patients showed that the most important predictors are the same and that the magnitude of the effects are comparable.

CONCLUSIONS

A predictive model for 30-day mortality risk from COVID-19 in subject aged 40 years or older has been developed and the most important comorbidities predicting early risk of death in a large cohort have been identified. In a new epidemic wave, it would help physicians and health systems to define groups at different risk and to identify high-risk subject in the over 40 population to target for specific prevention and therapeutic strategies. For example, people at the highest risk could be able to remain better isolated until fully immunized if supported in their needs by social assistance or, if infected, they could be selected for intensive monitoring at home or early hospitalization. Now that monoclonal antibodies are available, the choice for which patient to use them could be based on risk models. Lastly, priority in immunization could be based on such a risk model and not only on the age criteria.

Conflicts of interest: none declared.

REFERENCES

1. European Center for Disease Prevention and Control. COVID-19 Situation Dashboard. Available from: https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html (last accessed: 28.05.2020).
2. Riccardo F, Ajelli M, Andrianou X, et al. Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. medRxiv 2020. doi: 10.1101/2020.04.08.20056861. Available from: https://www.medrxiv.org/content/10.1101/2020.04.08.20056861v1
3. Sudharsanan N, Didzun O, Bärnighausen T, Geldsetzer P. The Contribution of the Age Distribution of Cases to COVID-19 Case Fatality Across Countries: A Nine-Country Demographic Study. Ann Intern Med 2020;173(9):714-20.
4. Natale F, Ghio D, Tarchi D, Goujon A, Conte A. COVID-19 Cases and Case Fatality Rate by age. Published on-line: 04.05.2020. Available from: https://ec.europa.eu/knowledge4policy/publication/covid-19-cases-case-fatality-rate-age_en (last accessed: 04.06.2020).
5. Yang Y. Trends in U.S. adult chronic disease mortality, 1960-1999: age, period, and cohort variations. Demography 2008;45(2):387-416.
6. Liang W, Liang H, Ou L, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern Med 2020;180(8):1081-89.
7. Grasselli G, Greco M, Zanella A, et al. Risk Factors Associated With Mortality Among Patients With COVID-19 in Intensive Care Units in Lombardy, Italy. JAMA Intern Med 2020;180(10):1345-55.
8. Chow DS, Glabis-Bloom J, Soun J, et al. Development and external validation of a prognostic tool for COVID-19 critical disease. PLos One 2020;15(12):e0242953.
9. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med 2020;46(5):846-48.
10. Wang L, He W, Yu X, et al. Coronavirus disease 2019 in elderly patients: Characteristics and prognostic factors based on 4-week follow-up. J Infect 2020;80(6):639-45.
11. Huang I, Lim MA, Pranata R. Diabetes mellitus is associated with increased mortality and severity of disease in COVID-19 pneumonia – A systematic review, meta-analysis, and meta-regression. Diabetes Metab Syndr 2020;14(4):395-403.
12. Cheng Y, Luo R, Wang K, et al. Kidney disease is associated with in-hospital death of patients with COVID-19. Kidney Int 2020;97(5):829-38.
13. Du R-H, Liang L-R, Yang C-Q, et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. Eur Respir J. 2020;55(5).
14. Johnston SL. Asthma and COVID-19: is asthma a risk factor for severe outcomes? Allergy 2020;75:1543-45.
15. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol 2020;146(1):110-18.
16. Zheng Z, Peng F, Xu B, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect 2020;81(2):e16-25.
17. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020;584(7821):430-36.
18. Shi Y, Yu X, Zhao H, et al. Host susceptibility to severe COVID-19: a retrospective analysis of 487 case outside Wuhan. Published on-line: 04.03.2020. doi:10.21203/rs.3.rs-16021/v1.
19. Yadaw AS, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health 2020;2(10):e516-25.
20. 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.
21. Hajifathalian K, Sharaiha RZ, Kumar S, et al. Development and external validation of a prediction risk model for short-term mortality among hospitalized U.S. COVID-19 patients: A proposal for the COVID-AID risk tool. PLoS One 2020;15(9):e0239536.
22. Atkins JL, Masoli JAH, Delgado J, et al. Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. J Gerontol A Biol Sci Med Sci 2020;75(11):2224-30.
23. 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.
24. Murtas R, Andreano A, Gervasi F, et al. Association between autoimmune diseases and COVID-19 as assessed in both a test-negative case-control and population case-control design. Auto Immun Highlights 2020;11(1):15.
25. Regione Lombardia. Attivazione della presa in carico di pazienti cronici e fragili: DGR n. X/6164 del 30.01.17 "Governo della domanda". Available from: https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/servizi-e-informazioni/Enti-e-Operatori/sistema-welfare/attuazione-della-riforma-sociosanitaria-lombarda/avvio-presa-carico-pazienti-cronici-fragili/dgr2017-6164-avvio-presa-carico-pazienti-cronici-fragili
26. Regione Lombardia. Avvio del percorso di presa in carico dei pazienti cronici e fragili: DGR n. x/7655 del 28.12.2017. Available from: https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/servizi-e-informazioni/Enti-e-Operatori/sistema-welfare/attuazione-della-riforma-sociosanitaria-lombarda/dgr2017-7655-avvio-presa-carico-cronici/dgr2017-7655-avvio-presa-carico-cronici
27. Sandrini M, Andreano A, Murtas R, et al. Assessment of the Overall Mortality during the COVID-19 Outbreak in the Provinces of Milan and Lodi (Lombardy Region, Northern Italy). Epidemiol Prev 2020;44(5-6) Suppl 2:244-51.
28. The EQUATOR Network. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Available from: https://www.equator-network.org/reporting-guidelines/tripod-statement/ (last accessed: 20.05.2020).
29. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9(11):1303-25.
30. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
31. Harell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer 2015.
32. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54(8):774-81.
33. Fenlon C, O'Grady L, Doherty ML, Dunnion J. A discussion of calibration techniques for evaluating binary and categorical predictive models. Prev Vet Med 2018;149:107-14.
34. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167-76.
35. Cleveland WS. LOWLESS: A program for smoothing scatterplots by robust locally weighted regression. Am Stat 1981;35(1):54.
36. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagn Progn Res 2018;2(1):7.
37. Mendes A, Serratrice C, Herrmann FR, et al. Predictors of In-Hospital Mortality in Older Patients With COVID-19: The COVIDAge Study. J Am Med Dir Assoc 2020;21(11):1546-54.
38. Gude-Sampedro F, Fernández-Merino C, Ferreiro L, et al. Development and validation of a prognostic model based on comorbidities to predict COVID-19 severity: a population-based study. Int J Epidemiol 2020;50(1):64-74.
39. Garin N, Koyanagi A, Chatterji S, et al. Global Multimorbidity Patterns: A Cross-Sectional, Population-Based, Multi-Country Study. J Gerontol A Biol Sci Med Sci 2016;71(2):205-14.
40. Ponsford MJ, Gkatzionis A, Walker VM, et al. Cardiometabolic Traits, Sepsis, and Severe COVID-19: A Mendelian Randomization Investigation. Circulation 2020;142(18):1791-93.