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

Anita Andreano
Epidemiology Unit, Agency for Health Protection of Milan
Rossella Murtas
Agency for Health Protection of Milan
Sara Tunesi
Agency for Health Protection of Milan
Maria Teresa Greco
Agency for Health Protection of Milan
David Consolazio
Agency for Health Protection of Milan
Davide Guido
Agency for Health Protection of Milan
Federico Gervasi
Agency for Health Protection of Milan
Maria Elena Gattoni
Agency for Health Protection of Milan
Monica Sandrini
Agency for Health Protection of Milan
Antonio Riussi
Agency for Health Protection of Milan
Antonio Giampiero Russo (agrusso@ats-milano.it)
Agency for Health Protection of Milan

Research Article

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

Posted Date: November 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-117108/v1

Abstract

Background: large studies on the predictive role of chronic conditions on mortality from COVID-19 are scarce. We developed a predictive model of death from COVID-19 in an Italian cohort aged 40 years or older.

Methods: we conducted a cohort study on prospectively collected data. The cohort included all (n=18,286) swab positive cases ≥40 year-old in patients registered with the Agency for Health Protection (AHP) of Milan up to 27/04/2020. Data on comorbidities were obtained from the chronic condition administrative database of the AHP. A multivariable logistic regression model, including age and gender and the selected conditions, was fitted to predict 30-day mortality risk and internally validated. External validation and recalibration were performed in a cohort of untested subjects with COVID-19 like symptoms. R software was used for the analysis.

Results: chronic conditions having the largest model-adjusted odds ratio (OR) of dying within 30 days from COVID-19 infection were chronic heart failure (OR=1.9, 95%CI 1.5-2.5), tumors (OR=1.8, 95%CI 1.4-2.3), complicated diabetes (OR=1.6, 95%CI 1.1-2.2) and dialysis-dependent chronic kidney disease (OR=1.5, 95%CI 1.0-2.2). Bootstrap-validated c-index was 0.78. The model fitted on the validation cohort had a c-index of 0.93, but required recalibration. With this latter model, *at a 10% risk of death threshold, 11% of the AHP population aged 40 years or older is considered at high risk.*

Conclusion: we identified a selected number of comorbidities predicting early risk of death in a large COVID-19 cohort aged 40 years or older. In a new epidemic wave, our results will help physicians and health systems to identify high-risk subject to target for prevention and therapy in this specific age group.

Background

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 20st February in the Lodi Province, Lombardy region, Italy[1, 2]. COVID-19 ongoing pandemic crude case-fatality rate in Italy in the first phase of the epidemic was 9.3%, higher than those reported in other European countries such as France (2.6%) and Germany (0.7%)[3]. In this 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 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 signaled for hospitalized patients with cancer and Chronic Obstructive Pulmonary Disease (COPD), and being associated with Charlson's index.[6] Number of comorbidities was also found to be predictor of severe COVID-19 disease in patients presenting to two large hospitals in the United States[7]. In other studies,

some of which reporting preliminary results or concerning a limited number of patients, history of hypertension, diabetes, cardiovascular disease, coronary artery disease, asthma, chronic kidney disease were found to be associated with severe disease or death from COVID-19[8–16]. These findings came from hospitalized cases or, in a few studies, from subjects presenting to a hospital and may not apply at community-level. Moreover, questions have been raised, for the so-far accumulated evidences, about lack of adjustment for important confounders in etiologic research, and concerning poor methodology in developing predictive models[17, 18]. Therefore, it is relevant to study, at population level, which demographic factors and comorbidities put a subject at higher risk of mortality, if infected, and to estimate the individual risk using information on a large number of chronic conditions. This will allow to give preventive recommendations to high risk subjects, to early protect themselves with social distancing in case of further waves of infection, and will help physicians to decide, based on the estimated individual risk, which strategy to undertake (e.g. home monitoring, hospitalization) in case of suspected symptomatology or nasopharyngeal swab positivity.

Our aim was to develop, and internally validate, a multivariable regression model to predict 30-day mortality risk in nasopharyngeal swab positive cases of COVID-19 aged 40 years or older, using data on comorbidities from administrative health databases. Secondarily, we externally validated the predictive model on cases with COVID-19 like symptoms reported by primary care physicians during the epidemic and who did not receive a nasopharyngeal swab.

Methods

Study design, data sources and measures

This was a population-based cohort study on data prospectively collected, partly ad hoc and partly by routine. The cohort included all COVID-19 cases in the study area, covered by the Agency for Health Protection (AHP) of Milan, corresponding to 193 municipalities in the northern Italian region of Lombardy, with a total population of 3,48 million inhabitants. The study area includes the municipality of Codogno that was at the origin of the first Italian epidemic outbreak. From the beginning of the outbreak, all tracing activities were included in a web-based platform, developed by the Epidemiologic Unit of the AHP, called *Milano COV*, including cases and related contacts (details on the information system are described in the Supplementary Methods). 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. Contacts are defined as all individuals who are associated with a case's sphere of activity, thus potentially exposed to the same source of contagion. Cases and close contacts underwent epidemiological investigation to provide description of the clinical presentation of COVID-19 and its clinical course. Furthermore, data were collected to estimate the serial interval, the symptomatic proportion of COVID-19 cases, and to identify possible routes of transmission.

In order to expand the outbreak reporting system, general practitioners could add symptomatic cases that did not undergo a nasopharyngeal swab, and their close contacts. In the AHP of Milan 95% of the

residents are registered with a GP affiliated with the Lombardy Regional Health System (RHS). From the beginning of the epidemic, the Lombardy Region daily sent the list of hospitalized COVID cases to each AHP, including the date of entry and the name of the hospital where the patient was admitted, and each variation (transfer, home discharge, death) was communicated in the following daily dataflow. Additional information relating to hospitalizations of patients in the ATS-Milano COR was derived from the regular administrative data discharge flow (SDO), which is consolidated for hospital admissions up to the end of April 2020.

We integrated between them through deterministic record linkage on individual tax code, and verified with the demographic information in the Health Service Register of the Lombardy Region (age, gender, place of residence), the described data sources in the Integrated Datawarehouse for COVID Analysis in Milan and attributed a random unique id to every subject. This same id was assigned to each subject in all other administrative databases of the AHP, 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 AHP of Milan, according to the algorithms specified in the Regional Act X/6164*[19]*and X/7655*[20]*of 2017, and* summarized in English in the supplementary material of the article by Murtas et al[21], in which part of the present cohort was analyzed.

Vital status was derived from the early notification system of the AHP of Milan, set-up from the beginning of the epidemic, in which deaths are communicated from the Civil Registry of each Municipality to the AHP and manually introduced in the *Health Service Register,* or directly from the GP and Mayor's offices for the subjects already in the *Milano COV* database through the web-based information system. We determined 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. 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 dates 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, we performed a univariate random imputation according to a uniform distribution U(a,b) with parameters a=10Feb2020 and b=17Feb2020. For symptomatic cases, date of diagnosis was the date of symptoms onset reported by the GP or, if missing, the date in which the subject was introduced in the web-system by the GP. For this analysis, we *considered as alive patients with a date of death more than 30 days after the date of diagnosis. The* vital status was *assessed on May 23rd 2020.*

Population

From the Milano COV database of the Milan AHP we extracted, on April 27, 2020, all subjects with nasopharyngeal swab-positive COVID-19 and all symptomatic cases reported by the GPs affiliated with the RHS.

All subjects that were at least 40 years old when diagnosed with COVID-19 were included in the analysis Supplementary Figure S1*. The choice was made both because deaths were very rare under age 40 (n=8 in the development and n=1 in the external validation cohort) and the majority of comorbidities of interest very rare. No further exclusions were performed.*

Development of the predictive models

We followed the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guidelines[22], including a 22-item checklist, to report the development and validation of the model predicting 30-day mortality risk in nasopharyngeal swab positive cases of COVID-19 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, we decided to develop a full model without performing model selection using automatic statistical techniques. We reduced the number of variables to be firstly introduced in the model on the basis of clinical-epidemiological considerations. We did not include conditions that were absent or very rare in the cohort, and grouped some relatively rare diseases with similar clinical consequences. The predictive model was developed using multiple logistic regression We evaluated collinearity using Phi correlation index and the Variance Inflaction Factor (VIF). We also tested 21 pre-specified interaction terms, chosen on epidemiological and clinical considerations (Supplementary Table S1) and kept in the model only the significant ones at p=0.05. The heuristic shrinkage estimator (van Houwelingen-le Cessie) was calculated including d.f. for testing interaction[23]. Overfitting is likely to be a concern when this estimator has values below 0.85[24]. We evaluated the functional form of the relationship between age and death and chose a restricted cubic regression spline with three knots based on graphical evaluation and AIC comparisons. All the other predictors were included as binary variables. 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). For age, we estimated the OR of increasing age from 60 to 80 years. We adjusted to age 70 (mean and median age in the development cohort 70 and 71 years), gender female and no comorbidities when estimating effects of factors having interaction terms. A metric of the absolute importance of each model term in predicting 30-day mortality from COVID-19 was calculated as the Wald chi-square minus the predictor degrees of freedom, and presented graphically[25]. Internal validation of the model was performed using bootstrap resampling (B=1000) to evaluate the discrimination and calibration of the model[26]. 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. Week model calibration was evaluated assessing calibration intercept and slope, with an intercept of 0 and a slope of 1 indicating no over/underfitting and no systematic over/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 predicted risk[27–29]. 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. We 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[30].

After development on swabs positive cases, the predictive model was externally validated and recalibrated on symptomatic cases, for which the distribution of chronic diseases is closer to that of the AHP population. External validation was evaluating the same discrimination and calibration indexes. Re-calibration of the intercept and slope was performed[31]. To evaluate the usefulness of the re-calibrated model when applied to the population, we used the recalibrated model to predict individual risk of death from COVID-19 in the whole population aged 40 years or older residing in the Milan AHP. We then constructed the predictiveness curve[32] and calculated the percentage of high risk subjects corresponding to different estimated risk of death thresholds.

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) and PredictABEL (v. 1.2-4, S. Kundu et al.)

Results

Description of the cohorts

On April 27, 2020 the COVID-19 database of the AHP of Milan included 20,364 swab positive cases and 12,224 symptomatic but not tested cases. Of those, 10% and 30% respectively were younger than forty years old and were therefore excluded, leaving 18,286 swab positive and 8,611 symptomatic COVID-19 patients (Table 1). Less than 1% (n=165) of swab positive cases had a missing diagnosis date, which was imputed as described in the methods, while all the symptomatic but not tested subjects reported date of onset. *The earliest diagnosis date was Feb 10th for both cohorts, while the latest was April 27th. Among alive patients, 160 (1.1%, minimum and median follow-up times, 26 and 29 days) in the development cohort and 293 (3.4%, minimum and median follow-up times, 19 and 27 days) in the validation cohort had a follow-up time shorter than 30 days at May 23rd 2020. In the development cohort 9% (n=333) of deceased patients at May 23rd 2020 (n=3832), and 10% (n=11) in the validation cohort (total deaths at May 23rd 2020, n=106), died after 30 days and were consequently considered as alive for the analysis, leaving 3499 and 95 events respectively.*

Swab positive cases were considerably older than symptomatic but not tested patients (median age 71 v. 54 years). In both cohorts there were slightly more females (52% and 57%). All symptomatic but not tested cases were treated at home, while 56% of the swab positive cases were hospitalized at some point and 21% got infected in a residential setting. Thirty-five percent of swab positive and 45% of symptomatic cases were in the City of Milan, while 15% and 7% in the province of Lodi, where the epidemic started. Among the most common comorbidities, hypertension had a prevalence of 45% in swab positive and of 25% in symptomatic cases, with a 2019 prevalence in the population 40 year-old or older of the AHP of 30%. The same figures for ischemic heart disease where12% and 4%, with a AHP

population prevalence of 6%, and for DM (including type 1 and 2 and complicated DM) of 15% and 6%, with a AHP population prevalence of 8%.

Table 1. Characteristics of the cohorts of patients aged 40 years or older with either swab positive COVID-19 infection or with COVID-19 like symptoms but not tested, surviving or not after the infection

	Swab positive cases			Symptomatic cases not swab tested			X^2 test for difference p-value
	Overall	Deceased at 30 days		Overall	Deceased at 30 days		
		No	Yes		No	Yes	
	n=18286	*n=14787*	*n=3499*	*n=8611*	*n=8516*	*n=95*	
Gender = Male (%)	8704 (47.6)	6664 (45.1)	2040 (58.3)	3749 (43.5)	3694 (43.4)	55 (57.9)	<0.001
Age class (years) (%)							<0.001
40-59	5884 (32.2)	5740 (38.8)	144 (4.1)	5891 (68.4)	5885 (69.1)	6 (6.3)	
60-79	6188 (33.8)	4901 (33.1)	1287 (36.8)	2124 (24.7)	2106 (24.7)	18 (18.9)	
80+	6214 (34.0)	4146 (28.0)	2068 (59.1)	596 (6.9)	525 (6.2)	71 (74.7)	
Setting (%)							<0.001
Home	4765 (26.1)	4401 (29.8)	364 (10.4)	8610 (100)	8515 (100)	95(100)	
Hospitalized	9604 (52.5)	7117 (49.2)	2487 (64.9)	0	0	0	
Residential	3363 (18.4)	2755 (18.6)	608 (17.4)	1 (0.0)	1 (0.0)	0 (0.0)	
Residential followed by hospitalization	554 (3.0)	283 (1.9)	271 (7.7)	0	0	0	
Geographic location							<0.001
Lodi Province (start of the outburst)	2669 (14.6)	2101 (14.2)	568 (16.2)	628 (7.3)	624 (7.3)	4 (4.2)	
City of Milan	6478 (35.4)	5188 (35.1)	1290 (36.9)	3911 (45.4)	3868 (45.4)	43 (45.3)	
Milan Province	9139 (50.0)	7498 (50.7)	1641 (46.9)	4072 (47.3)	4024 (47.3)	48 (50.5)	
Number of comorbidities							<0.001
None	6363 (34.8)	5717 (39.6)	646 (16.9)	4863 (56.5)	4852 (57.0)	11 (11.6)	
1-3	8709 (47.6)	6756 (46.7)	1953 (51.0)	3261 (37.9)	3220 (37.8)	41 (43.2)	
≥ 4	3214 (17.6)	1981 (13.7)	1233 (32.2)	487 (5.7)	444 (5.2)	43 (45.3)	
Specific comorbidities							
Transplanted any time = Yes (%)	69 (0.4)	57 (0.4)	12 (0.3)	13 (0.2)	13 (0.2)	0 (0.0)	0.003
Blood and Hematopoietic organs = Yes (%)	26 (0.1)	31 (0.2)	5 (0.1)	14 (0.2)	14 (0.2)	0 (0.0)	0.814
HIV infection or AIDS = Yes (%)	65 (0.4)	55 (0.4)	10 (0.3)	47 (0.5)	47 (0.6)	0 (0.0)	0.031
Tumor in first line treatment = Yes (%)	1005 (5.5)	692 (4.7)	313 (8.9)	243 (2.8)	233 (2.7)	10 (10.5)	<0.001
Tumor in follow-up, 1-5years = Yes (%)	772 (4.2)	553 (3.7)	219 (6.3)	221 (2.6)	210 (2.5)	11 (11.6)	<0.001
Tumor in remission after 5 years = Yes (%)	1071 (5.9)	783 (5.3)	288 (8.2)	285 (3.3)	283 (3.3)	2 (2.1)	<0.001
Type 1 Diabetes = Yes (%)	23 (0.1)	17 (0.1)	6 (0.2)	9 (0.1)	9 (0.1)	0 (0.0)	0.778
Type 2 Diabetes = Yes (%)	2385 (13.0)	1653 (11.2)	732 (20.9)	502 (5.8)	487 (5.7)	15 (15.8)	<0.001
Complicated DM Type 1 and 2 = Yes (%)	422 (2.3)	273 (1.8)	149 (4.3)	42 (0.5)	38 (0.4)	4 (4.2)	<0.001
Hypercholesterolaemia = Yes (%)	2293 (12.5)	1572 (10.6)	721 (20.6)	617 (7.2)	598 (7.0)	19 (20.0)	<0.001
Arterial hypertension = Yes (%)	8156 (44.6)	5907 (39.9)	2249 (64.3)	2142 (24.9)	2071 (24.3)	71 (74.7)	<0.001
Ischemic heart disease = Yes (%)	2361 (12.9)	1534 (10.4)	827 (23.6)	374 (4.3)	352 (4.1)	22 (23.2)	<0.001
Valvular heart disease = Yes (%)	458 (2.5)	319 (2.2)	139 (4.0)	85 (1.0)	80 (0.9)	5 (5.3)	<0.001

Cardiomyopathy with arrhythmia = Yes (%)	2297 (12.6)	1541 (10.4)	756 (21.6)	318 (3.7)	289 (3.4)	29 (30.5)	<0.001
Cardiomyopathy without arrhythmia = Yes (%)	1693 (9.3)	1145 (7.7)	548 (15.7)	289 (3.4)	267 (3.1)	22 (23.2)	<0.001
Chronic Heart failure = Yes (%)	1398 (7.6)	861 (5.8)	537 (15.3)	159 (1.8)	134 (1.6)	25 (26.3)	<0.001
Peripheral Artery Disease = Yes (%)	587 (3.2)	387 (2.6)	200 (5.7)	73 (0.8)	65 (0.8)	8 (8.4)	<0.001
Venous diseases = Yes (%)	197 (1.1)	139 (0.9)	58 (1.7)	60 (0.7)	53 (0.6)	7 (7.4)	0.003
Cerebrovascular disease = Yes (%)	820 (4.5)	543 (3.7)	277 (7.9)	74 (0.9)	64 (0.8)	10 (10.5)	<0.001
Thyroid diseases = Yes (%)	1064 (5.8)	896 (6.1)	168 (4.8)	580 (6.7)	574 (6.7)	6 (6.3)	0.004
Other endocrine diseases = Yes (%)	63 (0.3)	50 (0.3)	13 (0.4)	26 (0.3)	25 (0.3)	1 (1.1)	0.65
Other autoimmune diseases= Yes (%)	346 (1.9)	272 (1.8)	64 (1.8)	155 (1.8)	149 (1.7)	6 (6.3)	0.636
Epilepsy = Yes (%)	253 (1.4)	179 (1.2)	74 (2.1)	51 (0.6)	49 (0.6)	2 (2.1)	<0.001
Alzheimer and Dementias = Yes (%)	669 (3.7)	446 (3.0)	223 (6.4)	35 (0.4)	30 (0.4)	5 (5.3)	<0.001
Parkinson and Parkinsonisms = Yes (%)	310 (1.7)	211 (1.4)	99 (2.8)	24 (0.3)	22 (0.3)	2 (2.1)	<0.001
Nervous system diseases, others = Yes (%)	87 (0.5)	76 (0.5)	11 (0.3)	37 (0.4)	37 (0.4)	0 (0.0)	0.671
Chronic hepatitis and cirrhosis = Yes (%)	419 (2.3)	334 (2.3)	85 (2.4)	135 (1.6)	133 (1.6)	2 (2.1)	<0.001
Digestive system diseases, others = Yes (%)	232 (1.3)	195 (1.3)	37 (1.1)	109 (1.3)	109 (1.3)	0 (0.0)	1
COPD = Yes (%)	896 (4.9)	603 (4.1)	293 (8.4)	166 (1.9)	159 (1.9)	7 (7.4)	<0.001
RF or Oxygen therapy = Yes (%)	97 (0.5)	63 (0.4)	34 (1.0)	12 (0.1)	11 (0.1)	1 (1.1)	<0.001
Asthma = Yes (%)	391 (2.1)	335 (2.3)	56 (1.6)	233 (2.7)	232 (2.7)	1 (1.1)	0.004
CKD = Yes (%)	584 (3.2)	369 (2.5)	215 (6.1)	98 (1.1)	86 (1.0)	12 (12.6)	<0.001
Dialysis dependent CKD = Yes (%)	147 (0.8)	95 (0.6)	52 (1.5)	3 (0.0)	3 (0.0)	0 (0.0)	<0.001

Abbreviations: CKD=Chronic kidney disease, COPD=Chronic obstructive pulmonary disease, DM=diabetes mellitus, RF=Respiratory Failure

Development and internal validation of the predictive model on swab positive cases

We reduced the number of chronic conditions to include as predictors in the model from 65, as in the *chronic disease database (Supplementary material of the article by Murtas et al.*[21]*),* to 32 as follows. The following six variables were excluded because they were not present or very rare in the development 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), and congenital malformations (10 subjects and no events). Chronic cutaneous diseases (25 subjects) were not included as predictors as an implausible

risk factor. The following categories of the chronic conditions database were merged into one according to rarity or similar clinical consequences: transplanted within 2 and 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 hemolytic anemias, 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 CKD and dialysis dependent CKD, and the three tumor categories. No collinearity problems were detected (largest Phi correlation index=0.36 between complicated diabetes and ischemic heart disease, and the latter and hypercholesterolemia). The variables with the highest VIF were ischemic heart disease (1.30), chronic heart failure (1.28), and age (1.27). Among the pre-specified tested interactions (Supplementary 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, while model coefficients for all terms and p-values are reported in Supplementary Table S2. The model-adjusted likelihood of dying within 30 days 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), tumors 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, tumor 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 our 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 (Table 2, Supplementary Figure S2), summarized by an IPA of 14.8%.

External validation on symptomatic cases

Results of the external validation of the predictive model in the 8,611 symptomatic subjects are summarized in Table 2, right column. The Brier-score (prediction accuracy) and the c-index (discrimination ability) improved from 0.13 to 0.02 and from 0.79 to 0.93, respectively. However, the model was miscalibrated leading to an overall IPA of -85.4% and required a re-calibration of the parameters leading a new the intercept coefficient of -1.64, and a new slope of 1.57 with a IPA of 10.3%. If we use this recalibrated model to predict 30-days risk of death from COVID-19 in the whole AHP population aged 40 years or older (Supplementary Table S3), 11% (n=249,387) is considered as high risk if we are willing to accept a 10% risk of death, while if we set at 5% the acceptable risk of death the percentage of high risk subject is 18% (n=407,443) (Figure 3).

Table 2 Internal validation of the multivariable logistic model predicting 30-days risk of death in the swab positive COVID-19 cohort aged 40 years or older, and external validation of the same model in cases with COVID-19 like symptoms aged 40 years or older not undergone nasopharyngeal swab.

	Internal validation (n=18286 swab positive subjects)		External validation (n=8611 symptomatic subjects)		
			Before re-calibration	After intercept and slope re-calibration*	
		Bootstrap-corrected (B=1000)			Bootstrap-corrected (B=1000)
Brier score	0.13	0.13	0.02		0.01
c-index/AUC	0.79	0.78	0.93		0.93
IPA	14.9		-85.4	10.3	
Calibration					
intercept	0 (-0.04;0.04)		-2.19(-2.40;-1.98)	0(-021;0.21)	
slope	1 (0.95;1.05)		1.57 (1.32;1.82)	1 (0.84-1.16)	

*Coefficients of the re-calibration models Intercept =-1.64, slope =1.5. AUC=area under the curve, IPA=index of prediction accuracy

Discussion

We developed the first multivariable predictive model of short-term mortality in a large cohort of confirmed COVID-19 cases using age, gender and a large number of chronic conditions derived from administrative data. The model has a good discriminative capability, especially considering predictors derives from administrative data. We confirmed the prominent role of age and, to a less extent, of male gender on mortality risk. We also found that the chronic conditions with the greatest ability to predict short-term death were diabetes, tumor in first-line treatment, and chronic heart failure. These diseases had also the highest estimated model-adjusted OR, together with dialysis-dependent CKD. Particularly, tumor in active treatment and chronic heart failure similarly almost doubled the odds of dying, while dialysis-dependent CKD and both uncomplicated and complicated DM approximately increased the risk of one-half. Model discriminative performance at external validation in symptomatic cases was better than in the development sample, with increasing AUC/c-index and decreasing Brier's score. However, intercept

and slope recalibration was necessary, probably due to the lower mortality risk in a population that includes younger subjects with less comorbidities and un unknown proportion of false positives. Also, even after re-calibration and regardless the good AUC/c-index, the IPA in the external validation cohort was lower than in the development cohort (10.3 vs 14.9). For this reasons, and because of the limited number of events (n=95), results of the prediction model on the external validation cohort may be non-conclusive and further validation on fully independent cohorts will be necessary.

Diabetes has been reported to be associated with increased mortality risk in different studies summarized by a meta-analysis from Huang et al. that found a pooled relative risk of 2.1 (95%CI 1.4-3.1) varying with age and hypertension. Concerning cardiovascular diseases many study did not specify which conditions were actually included, so we 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 etiological studies. Of notice, respiratory failure, COPD and asthma did not increase 30-day mortality risk in the multivariate model, even if RF and COPD are associated with death at univariate analysis. The most important predictive factors are relatively common in the population aged 40 years or older in the study area, for example tumor 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. This implies that, at the *5% thirty-day mortality risk* threshold*, the percentage of high risk subjects in the population aged 40 years or older in the study area was 18% (n=407,443). The consequence is that a large number of subjects would qualify to receive targeted intervention to reduce the risk of COVID-19, including priority access to vaccine once available, and that this number would be even greater if we consider as acceptable lower threshold risk. Those* predictors and their combinations are prevalent in most countries, not only in the high-income ones[33], even if a lower proportion of high-risk subjects may be expected in countries with a younger population. Our results 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 taken into account when developing health policies for protecting particular groups of individuals. In our 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. The most important limitation of the study is the exclusion of cases under 40 years which does not allow to estimate the total population at high risk (the model including all ages is reported as Table S5). This choice was made because subjects younger than 40 are very few in the development 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 and seven deaths, respectively) that the risk in the younger population will be not accurately predicted with our data. 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.

We believe that this segment of the population would deserve a different analysis, including also data on pre-existing/concomitant acute conditions and information on BMI, which has been reported as a relevant factor in younger people and that are not available in our database at present. The second limitation concerns lack of information on body mass index and smoking status, which have been reported as potentially having a predictive effect on mortality[34, 35]. Third, the development cohort is non-representative of the actual population risk because in the first phase of the epidemic many patients with a clinically mild presentation were not tested while all hospitalized patients received a nasopharyngeal swab, either before or after hospitalization. Also patients in the swab positive cohort had a greater number of comorbidities. However, in an epidemic phase, is virtually impossible to perform swabs to the entire population to individuate all diseased subject regardless of symptom intensity, which constitute an important issue in a disease like COVID-19 with a large proportion of asymptomatic or paucisymptomatic cases. Moreover, the symptomatic but not tested cohort in which the model was validated and re-calibrated, has a distribution of chronic diseases which is similar to the population of the studied area.

The recalibrated model may be used to calculate the individual risk at population level in subject aged 40 years or older and consequently identify the segments of the general population in this age class whose risk of COVID-19-related death is higher, not only in the Lombardy RHS, but also in any health system having population-level administrative data on comorbidities. This would allow, in the event of one or more further waves of COVID-19, a management of social distancing and quarantine which takes into account the greater or lesser actual risk afflicting each subject, and which therefore may prevent (at least in part) the already experienced serious economic and social effects of the general lockdown, especially if still no vaccines nor effective medical treatment will be available). It will also provide evidence to define immunization strategies if an effective vaccine will be approved.

Conclusion

We developed a predictive model for 30-day mortality risk from COVID-19 in subject aged 40 years or older and identified the most important comorbidities predicting early risk of death in a large cohort. In a new epidemic wave, these results will help physicians and health systems to suggest preventive recommendations to high-risk subject, information based on measured data and not only on general recommendations.

Declarations

Ethics approval and consent to participate

Ethics approval and consent to participate were not required, as this is an observational study based on data routinely collected by the Agency for Health Protection (ATS) of Milan, a public body of the Regional Health Service – Lombardy Region. The ATS has among its institutional functions, established by the Lombardy Region legislation (R.L. 23/2015), the government of the care pathway at the individual level in

the regional social and healthcare system, the evaluation of the services provided to, and the outcomes of, patients residing in the covered area. This study is also ethically compliant with the National Law (D.Lgs. 101/2018) and the "General Authorisation to Process Personal Data for Scientific Research Purposes" (n.8 and 9/2016, referred to in the Data Protection Authority action of 13/12/2018). Data were anonymized with a unique identifier in the different datasets before being used for the analyses.

Availability of data and materials

The dataset from this study is held securely at the ATS of Milan, Epidemiology Unit. Data sharing agreements prohibit the ATS of Milan from making the dataset publicly available. The full dataset creation plan and underlying analytic code are available from the authors upon request.

Competing interests

The authors declare that they have no competing interests.

Funding

None

Authors' contributions

Concept and design: AGR, AA. Acquisition, analysis, or interpretation of data: AA, RM, ST, DG, MS, AR, MTG, DC, FG, MEG. Drafting of the manuscript: AA, RM, ST, DG, AGR. Critical revision of the manuscript for important intellectual content: MTG, DC, FG, MEG, MS, AR, AGR. Statistical analysis: AA. Administrative, technical, or material support: All authors. AA had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. AGR, the guarantor, accepts full responsibility for the work, had access to the data, and controlled the decision to publish. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

References

1. European Center for Disease Prevention and Control. COVID-19. https://qap.ecdc.europa.eu/public/extensions/COVID-19/COVID-19.html. Accessed 28 May 2020.
2. Riccardo F, Ajelli M, Andrianou X, Bella A, Del Manso M, Fabiani M, et al. Epidemiological characteristics of COVID-19 cases in Italy and estimates of the reproductive numbers one month into the epidemic. preprint. Infectious Diseases (except HIV/AIDS); 2020. doi:10.1101/2020.04.08.20056861.
3. Sudharsanan N, Didzun O, Bärnighausen T, Geldsetzer P. The Contribution of the Age Distribution of Cases to COVID-19 Case Fatality Across Countries. Ann Intern Med. 2020. doi:10.7326/M20-2973.

4. Natale. COVID-19 Cases and Case Fatality Rate by age. 2020. https://ec.europa.eu/knowledge4policy/publication/covid-19-cases-case-fatality-rate-age_en. Accessed 4 Jun 2020.
5. Yang Y. Trends in U.S. adult chronic disease mortality, 1960-1999: age, period, and cohort variations. Demography. 2008;45:387–416.
6. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, 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.
7. Chow DS, Glabis-Bloom J, Soun J, Weinberg B, Berens-Loveless T, Xie X, et al. Development and External Validation of a Prognostic Tool for COVID-19 Critical Disease. medRxiv. 2020;:2020.05.06.20093435.
8. 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. doi:10.1007/s00134-020-05991-x.
9. Wang L, He W, Yu X, Hu D, Bao M, Liu H, et al. Coronavirus disease 2019 in elderly patients: Characteristics and prognostic factors based on 4-week follow-up. J Infect. 2020;80:639–45.
10. 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 Clin Res Rev. 2020;14:395–403.
11. Cheng Y, Luo R, Wang K, Zhang M, Wang Z, Dong L, et al. Kidney disease is associated with in-hospital death of patients with COVID-19. Kidney Int. 2020;97:829–38.
12. Du R-H, Liang L-R, Yang C-Q, Wang W, Cao T-Z, Li M, 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. doi:10.1183/13993003.00524-2020.
13. Johnston SL. Asthma and COVID-19: is asthma a risk factor for severe outcomes? Allergy. n/a n/a. doi:10.1111/all.14348.
14. Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol. 2020. doi:10.1016/j.jaci.2020.04.006.
15. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect. 2020. doi:10.1016/j.jinf.2020.04.021.
16. Host susceptibility to severe COVID-19: a retrospective analysis of 487 case outside Wuhan. 2020. doi:10.21203/rs.3.rs-16021/v1.
17. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. BMJ. 2020;:m1198.
18. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;:m1328.

19. Regione Lombardia. Governo della domanda: avvio della presa in carico di pazienti cronici e fragili. determinazioni in attuazione dell'art. 9 della legge n. 23/2015. 2017. 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.
20. Regione Lombardia. Modalità di avvio del percorso di presa in carico del paziente cronico e/o fragile in attuazione della dgr n. x/6551 del 04/05/2017. 2017. 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.
21. Murtas R, Andreano A, Gervasi F, Guido D, Consolazio D, Tunesi S, 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;11:15.
22. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement | The EQUATOR Network. https://www.equator-network.org/reporting-guidelines/tripod-statement/. Accessed 20 May 2020.
23. Houwelingen JCV, Cessie SL. Predictive value of statistical models. Stat Med. 1990;9:1303–25.
24. Harrell FE, Lee KL, Mark DB. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Stat Med. 1996;15:361–87.
25. Jr FEH. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer; 2015.
26. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
27. 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.
28. 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.
29. Cleveland WS. LOWESS : A program for smoothing scatterplots by robust locally weighted regression. Am Stat. 1981;35:54.
30. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagn Progn Res. 2018;2:7.
31. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61:1085–94.
32. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the Predictiveness of a Marker with Its Performance as a Classifier. Am J Epidemiol. 2008;167:362–8.

33. Garin N, Koyanagi A, Chatterji S, Tyrovolas S, Olaya B, Leonardi M, et al. Global Multimorbidity Patterns: A Cross-Sectional, Population-Based, Multi-Country Study. J Gerontol A Biol Sci Med Sci. 2016;71:205–14.
34. Huang R, Zhu L, Xue L, Liu L, Yan X, Wang J, et al. Clinical findings of patients with coronavirus disease 2019 in Jiangsu province, China: A retrospective, multi-center study. PLoS Negl Trop Dis. 2020;14:e0008280.
35. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020;323:1061.

Figures

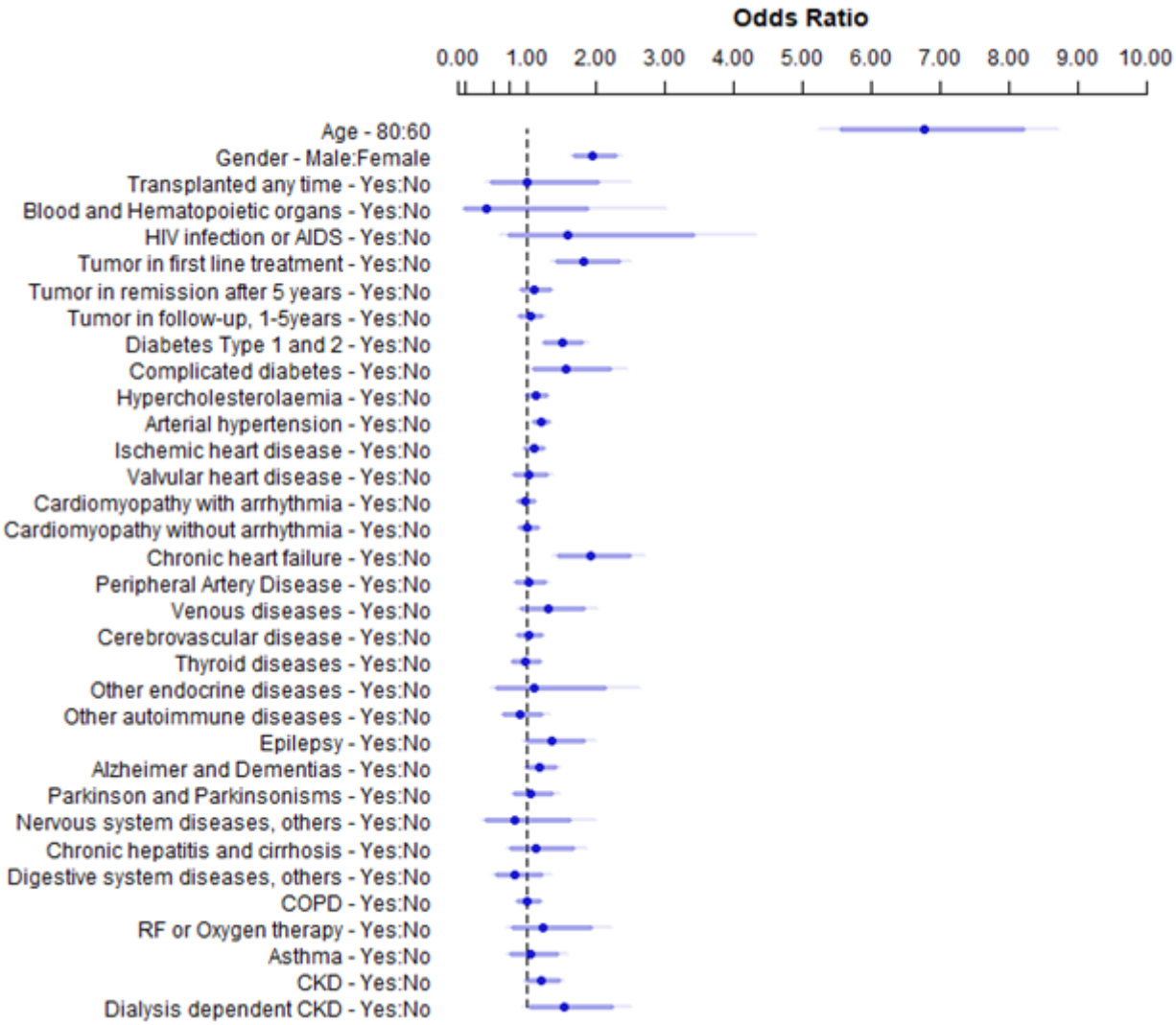

Figure 1

Results from the multivariable logistic regression model predicting 30-days mortality risk from COVID-19 in the development cohort of swab positive cases, presented as adjusted odds ratio (dark blue bullets)

with 95% (blue bar) and 99% confidence intervals (light blue line).

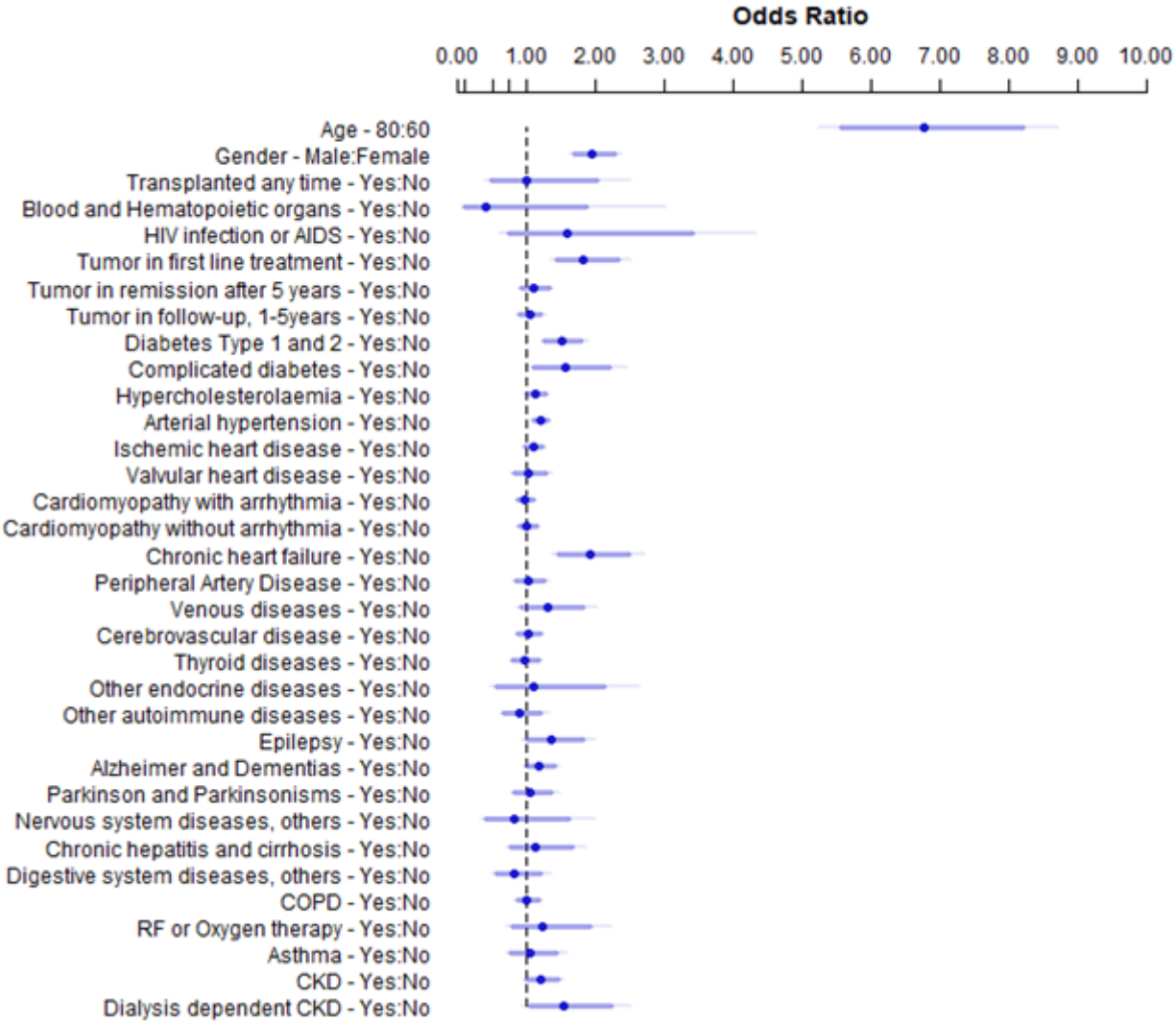

Figure 1

Results from the multivariable logistic regression model predicting 30-days mortality risk from COVID-19 in the development cohort of swab positive cases, presented as adjusted odds ratio (dark blue bullets) with 95% (blue bar) and 99% confidence intervals (light blue line).

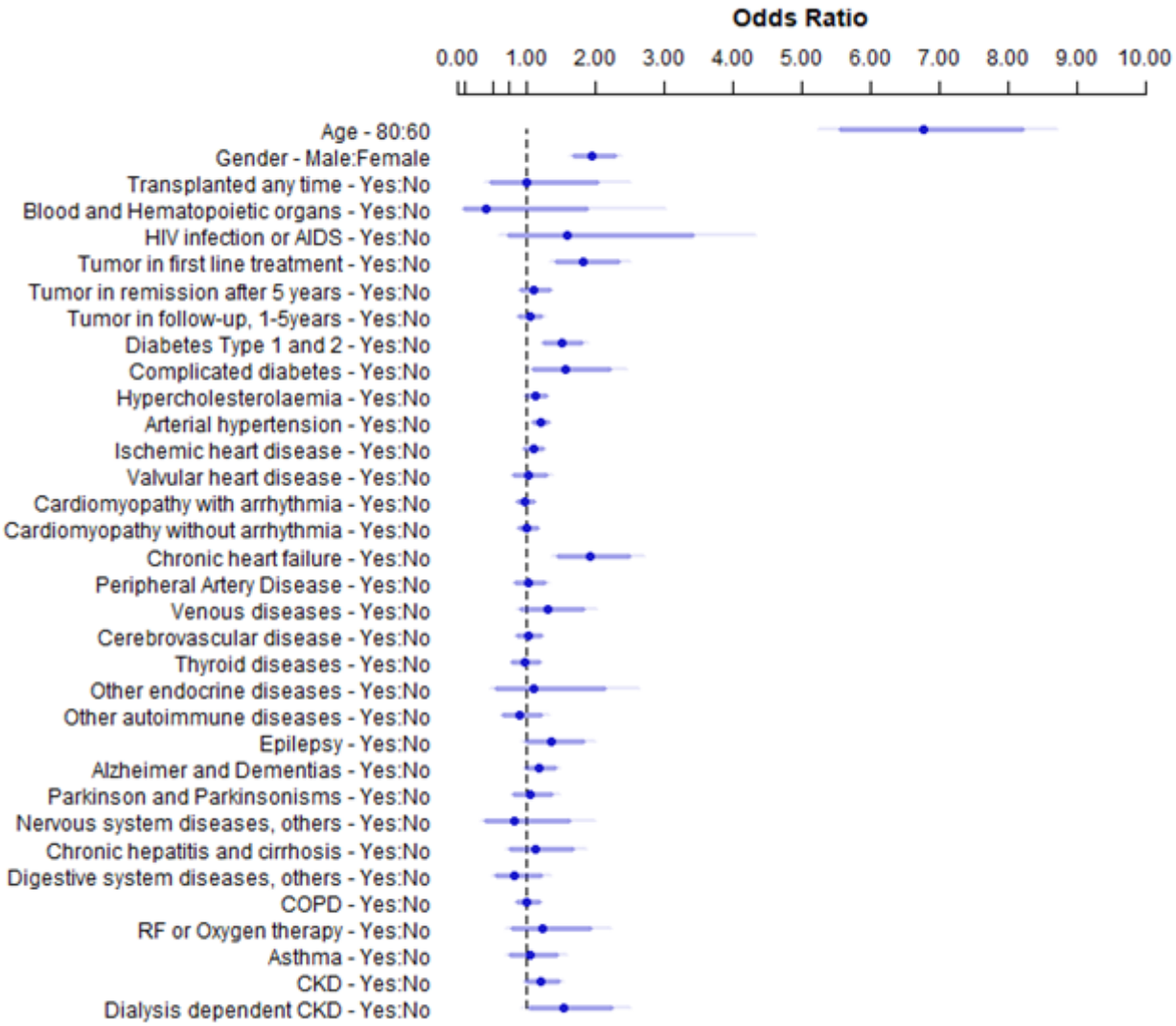

Figure 1

Results from the multivariable logistic regression model predicting 30-days mortality risk from COVID-19 in the development cohort of swab positive cases, presented as adjusted odds ratio (dark blue bullets) with 95% (blue bar) and 99% confidence intervals (light blue line).

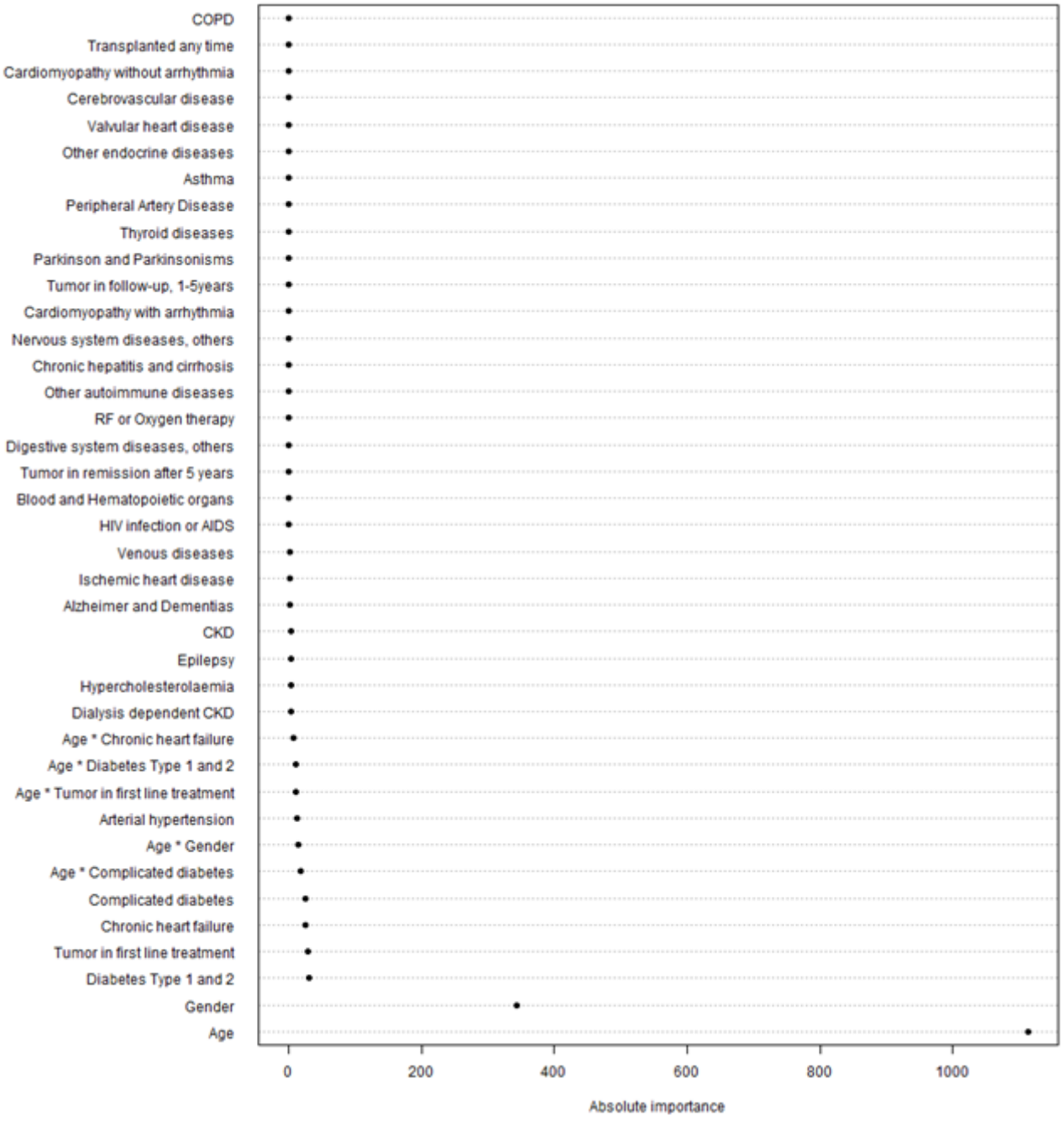

Figure 2

Absolute importance of predictors, measured by Wald chi-square value minus the degrees of freedom of the predictor, based on multivariable logistic regression in the development cohort of COVID-19 swab positive cases.

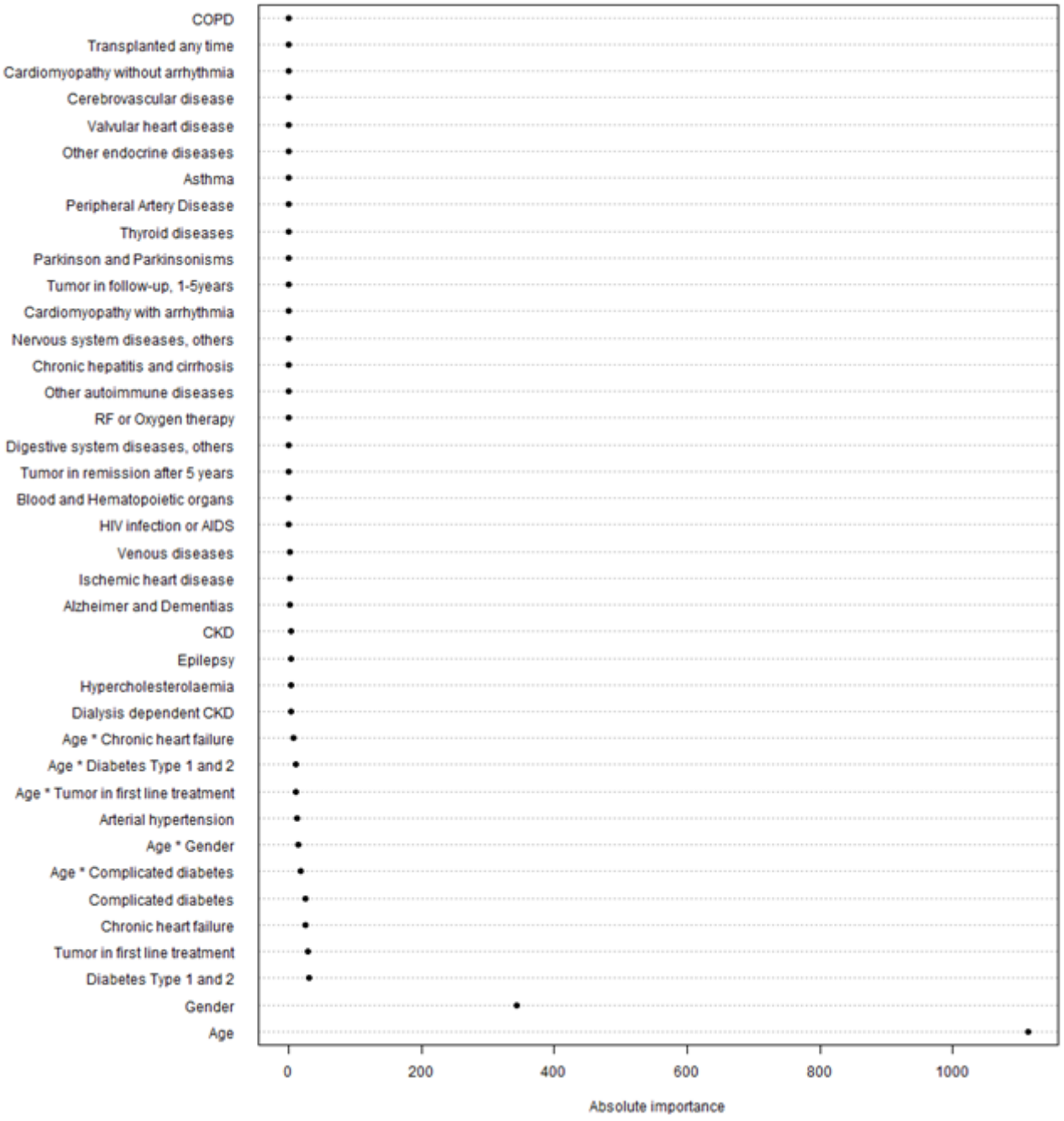

Figure 2

Absolute importance of predictors, measured by Wald chi-square value minus the degrees of freedom of the predictor, based on multivariable logistic regression in the development cohort of COVID-19 swab positive cases.

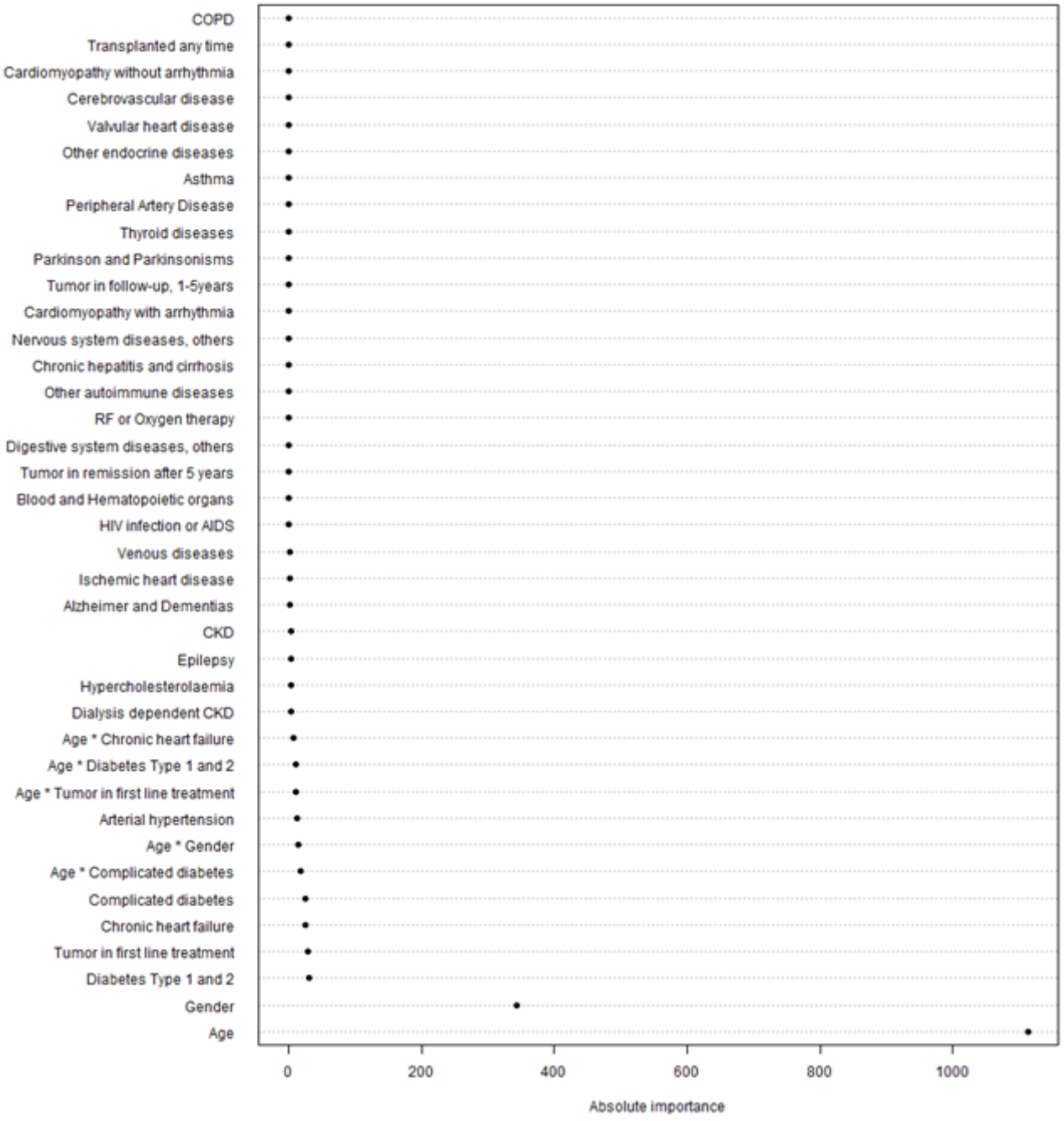

Figure 2

Absolute importance of predictors, measured by Wald chi-square value minus the degrees of freedom of the predictor, based on multivariable logistic regression in the development cohort of COVID-19 swab positive cases.

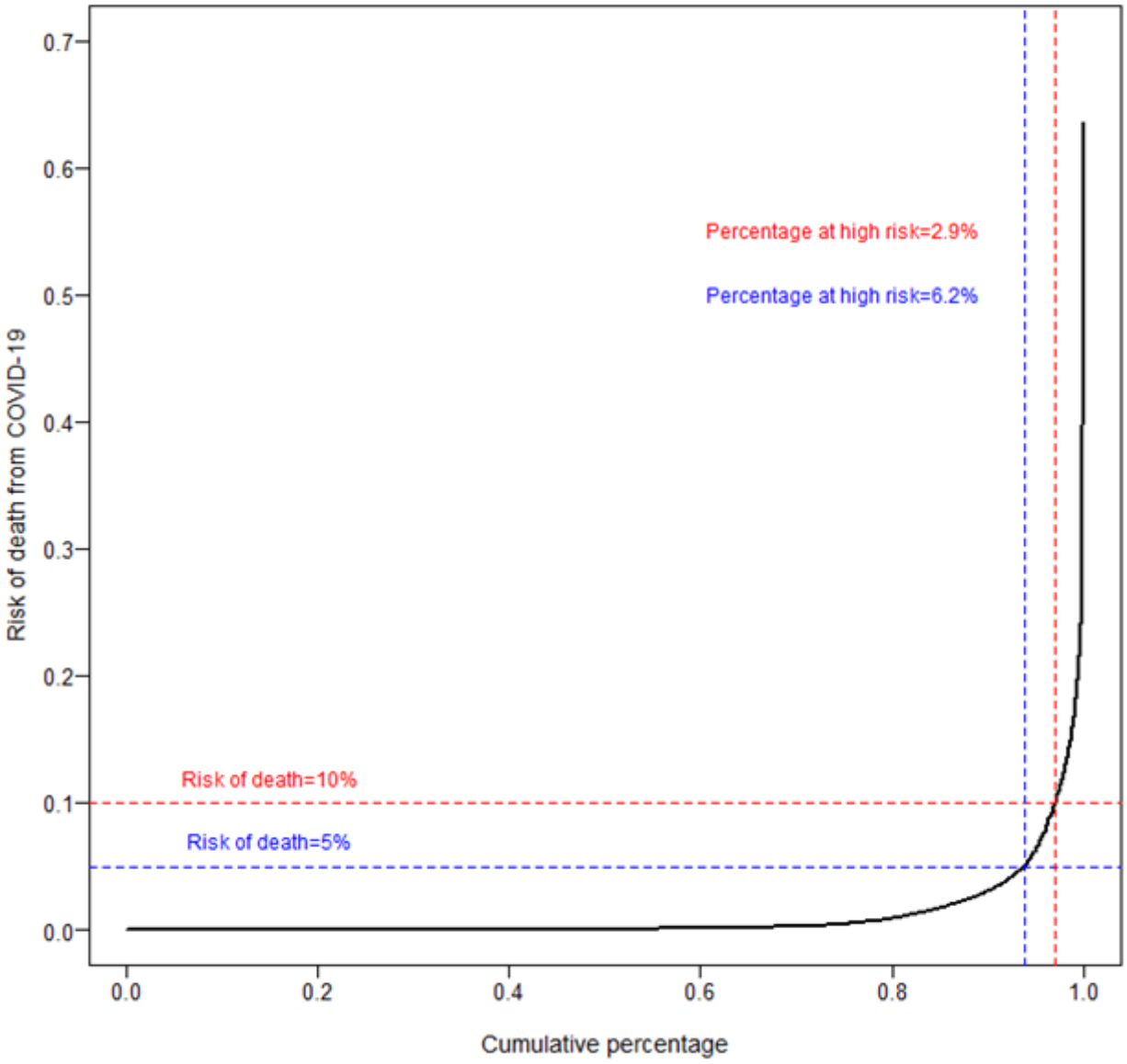

Figure 3

Predictiveness curve of the 30-days mortality risk model from COVID-19, re-calibrated in the symptomatic but not tested cohort, in the whole population aged 40 years or older residing in the territory of competence of the Agency for Health Protection of Milan.

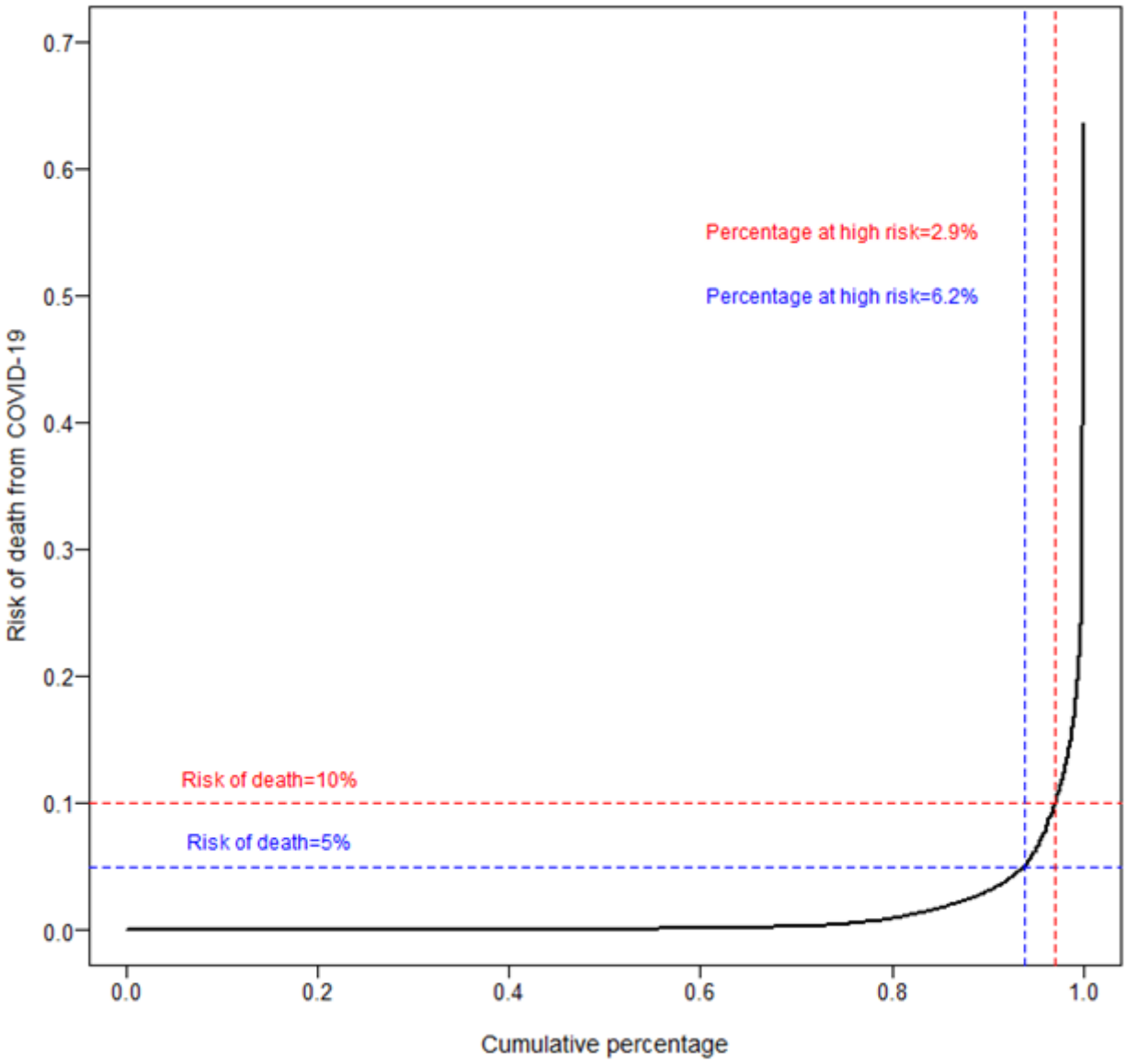

Figure 3

Predictiveness curve of the 30-days mortality risk model from COVID-19, re-calibrated in the symptomatic but not tested cohort, in the whole population aged 40 years or older residing in the territory of competence of the Agency for Health Protection of Milan.

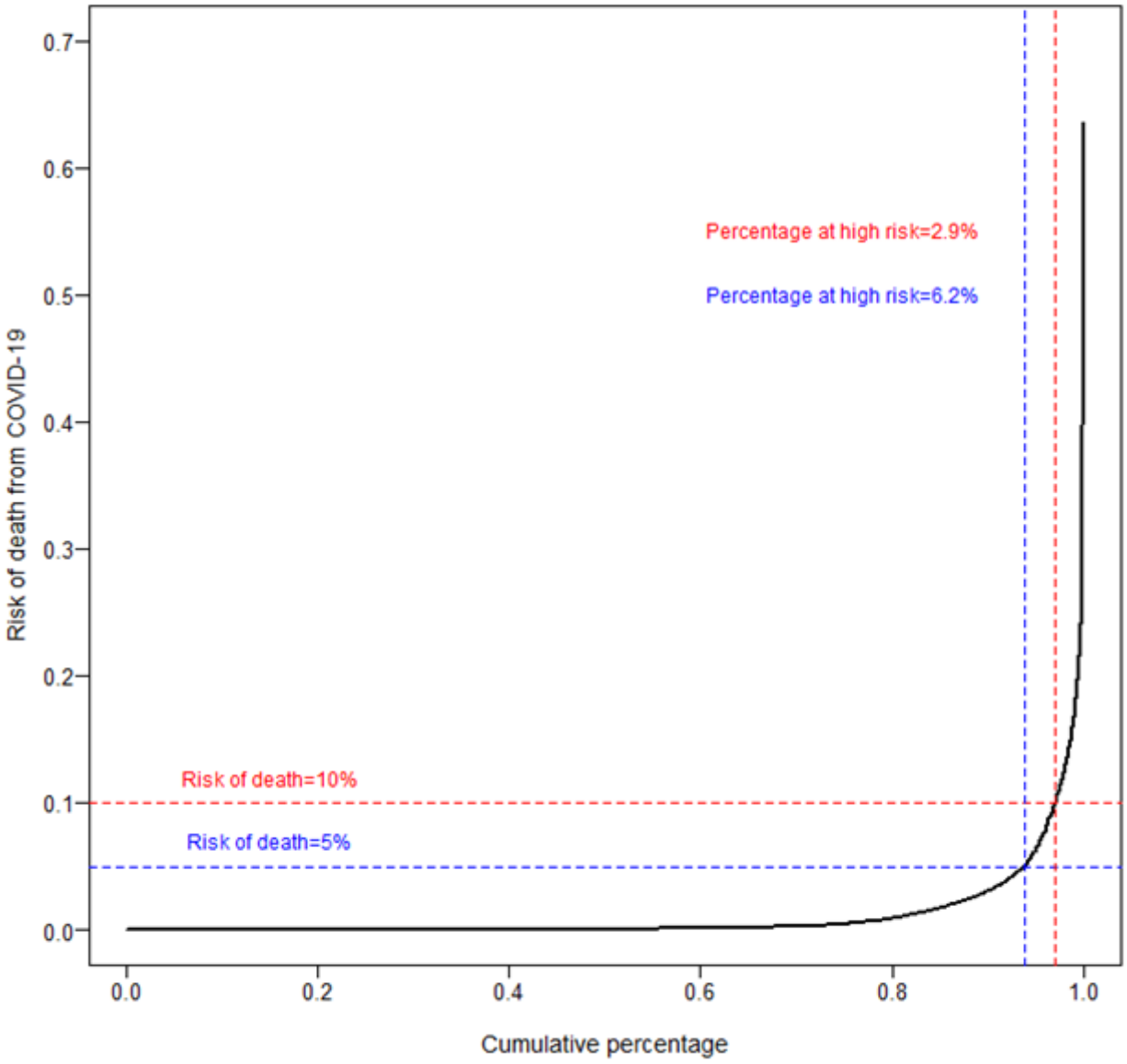

Figure 3

Predictiveness curve of the 30-days mortality risk model from COVID-19, re-calibrated in the symptomatic but not tested cohort, in the whole population aged 40 years or older residing in the territory of competence of the Agency for Health Protection of Milan.