Supplementary material for "Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK"

Supplementary Methods

Study design and participants

The overall design is the derivation and external validation of a multivariable risk prediction model for poor outcome and death in inpatients with COVID-19, reported in accordance with the TRIPOD statement.1

Derivation and internal validation cohort

The study population comprised two cohorts of adults (≥18 years old) hospitalised with COVID-19 in Wuhan Sixth Hospital and Taikang Tongji Hospital (Wuhan, China). Both hospitals were designated hospitals for admitting patients with confirmed symptomatic COVID-19 who had more severe symptoms (**Figure S3**; patients with less severe symptoms were admitted to Fangcang hospitals). The derivation cohort consisted of adult patients who were admitted between 01/02/2020 and 23/02/2020, and who died or were discharged on or before 29/03/2020 (**Figure S4**). Patients who were readmitted for recurrent COVID-19 were excluded. A confirmed case of COVID-19 was defined by a positive result in a real-time reverse-transcription polymerase chain reaction (RT-PCR) assay (Ct<37) of nasopharyngeal swab (NPS) specimens.

External validation cohort

The external validation cohort consisted of adults (≥18 years old) hospitalized with COVID-19 in King's College Hospital (KCH) NHS Foundation Trust (London, United Kingdom) between 01/03/2020 and 02/04/2020, who have been followed up until 08/04/2020 (KCH cohort) (**Figure S5**). In total, there were 432 patients. However, complete data for risk prediction were only available for 226 patients, which were used in this study. A confirmed case of COVID-19 was defined by a positive result on a real-time RT-PCR assay of NPS specimens.

Data collection

Derivation and internal validation cohort. Demographic, premorbid conditions, clinical symptoms or signs at presentation, laboratory data, treatment and outcome data were extracted from electronic medical records using a standardised data collection form by a team of experienced respiratory clinicians, with double data checking and involvement of a third reviewer where there was disagreement. Anonymised data was entered into a password-protected computerised database.

External validation cohort. Demographic, laboratory and outcome data were extracted from electronic health records using a combination of SQL queries for structured data and elastic queries for unstructured free text data. All data extractions were checked and assessed through a two-stage manual verification.

Outcomes

Derivation and internal validation cohort. Two outcomes were modelled in our study: poor outcome and death. In the derivation cohort, poor outcome was defined by reaching any of the following endpoints: developing acute respiratory distress syndrome (ARDS), receiving intubation or extracorporeal membrane oxygenation (ECMO) treatment, ICU admission (WHO Ordinal Scale 6-7) and death (WHO ordinal scale 8).[2] ARDS was defined according to the Berlin criteria.[3] Metrics corresponding to death and poor outcome were denoted 'death' and 'poor', respectively (eg. OR_{death} and OR_{poor}).

External validation cohort. Due to the ongoing COVID-19 pandemic in the UK, not all patients from the KCH cohort had been discharged or died by the time of external validation analysis (12/04/2020). The primary use of a risk prediction tool for prognosis is to inform early clinical decisions (**Figure 1**), and we therefore externally validated the risk prediction equation in relation to two outcomes (death or poor outcome) occurring within the follow-up period. Poor outcome was defined as ICU admission (WHO Ordinal Scale 6-7) or death (WHO ordinal scale 8).[2]

Predictors

Derivation and internal validation cohort. Predictors were chosen for multivariable modelling with consideration of workflow in clinical practice and their availability in routine hospital care. In total, four groups of predictors were modelled: basic demographic predictors (age and sex), premorbid conditions documented in the medical record (diabetes mellitus, chronic lung disease, immunocompromised, malignancy, hypertension, heart disease and chronic renal disease), symptoms (fever defined as an axillary temperature of ≥37.3°C, coughing, fatigue, dyspnoea and diarrhoea) and laboratory tests obtained at or immediately after inpatient admission (neutrophil count, lymphocyte count, CRP, creatinine). Age and laboratory tests were used as continuous variables, while other variables were used as categorical variables.

External validation cohort. Data were only extracted for variables included in the derived risk prediction model, namely age, sex, lymphocyte count, neutrophil count, platelet count, creatinine and CRP.

Univariate statistical analysis

Continuous variables were presented as medians and interquartile ranges (IQR). Univariate odds ratios (ORs) for continuous variables were computed by univariate logistic regression with intercept fitted. Differences between the mean of continuous variables in the two outcome groups were tested by Student and Welsh t-test, respectively where variance was equal or heterogeneous. Categorical variables were summarised as counts (n) and percentages (%). Univariate ORs of categorical variables and their p-values were derived from contingency table statistics. All analyses were performed using Scipy library (version 1.4.1) in Python (version 3.7.7).[4]

Statistical modelling

Derivation and internal validation cohort

An overview of modeling workflow is given in **Figure S6**. The model development pipeline was developed using the *Scikit-learn* (version 0.22.1) library in *Python*.[6] The code was published in Github repository (https://github.com/huayu-zhang/COVID-19-Risk).

Handling missing data. Premorbid conditions, symptoms and ARDS level were imputed as negative if at least one field in the same group is positive. For example, if a patient has value '1' in the 'ards_severe' field and 'ards_mild' or 'ards_moderate' fields are missing, these 2 fields were naturally imputed 0. No other types of imputation were performed.

For statistical modeling with multiple variables, only complete cases were selected. For comparison of different models, complete cases were determined on the union set of predictors of all models. Numbers of sample size in each analysis were summarised in **Figure S4** and **Table S4**. We used χ^2 test of 2-way contingency table to investigate whether the outcomes were disproportionately affected by missing data (**Table S5**).

Model development. For each outcome, three models were developed with combinations of predictor groups:

1. Demographic + premorbid Conditions + Symptoms (Group DCS)
2. Demographic + premorbid Conditions + Symptoms + Laboratory results (Group DCSL)
3. Demographic + Laboratory results (Group DL)

Group DCS was designated to reflect the information available at community triage and low income countries where blood tests may not be immediately available (Decision 2 in **Figure 1**) level, while Group DCSL was designed to reflect the full information available in the Wuhan datasets at the point of hospital admission (Decision 3 in **Figure 1**). Since prediction tools based on large numbers of predictors cannot be easily implemented in clinical care during an epidemic (due to information incompleteness and time pressures), two parsimonious risk prediction models with Group DL as predictors and death or poor outcome as the outcome were developed. These two models were externally validated. Logistic regression model with L1 regularization (LASSO) was used for model development to account for the large number of predictors relative to outcome. Models were referred to in the manner of 'predictor group-outcome' (eg. DCS-Death).

Cross-validation. For all models, internal assessment of model performance and stability of predictor coefficients was done by repeated (10 times) 5-fold cross-validation. The splits in cross-validation were randomly generated. Splits where either partition of the data contained no positive outcome cases were excluded. Model performance metrics measured were: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and C-index. Predictor coefficients in each iteration were also recorded and summarized in identical manner to model performance metrics.

Model tuning to optimize L1 regularization strength. To estimate the optimal strength of L1 regularization in each analysis (predictor selection or development of final models), the parameter was tuned in the range 2^{-5} to 2^{4} with cross-validation. Optimal L1 regularization strength was chosen by the level which gave rise to the best model performance (judged by C-index).

Predictor selection. Since L1 regularization shrinks coefficients of weak predictors to zero, it also served as the method for predictor selection. Demographic predictors were included in all models. For predictor selection of one model, all predictors to be selected were added to the model. The model was then tuned to find the optimal L1 regularization strength. Predictors which did not have coefficients significantly different to zero in cross-validation at the chosen L1 regularization strength were eliminated. The significance of coefficients was tested by one-mean Student t-test with the null hypothesis that the coefficient is not different to zero. The above process was done recursively until no predictors could be eliminated.

Fitting final models. Predictors that survived predictor selection were used to fit the final model. To obtain the optimal L1 regularization for the final model, the L1 regularization strength was optimized again by model tuning. Using the optimal L1 regularization strength and selected predictors, a final model was fitted using the entire data with complete cases. The coefficients of the final model were reported and used for external validation. Some coefficients of selected predictors were zero in this step, because the entire data was used (instead of cross-validation in predictor selection). In this case, these predictors with coefficients of zeros were also excluded from the final model. Calibration metrics (calibration-in-the-large, calibration slope and calibration plots) were calculated using predicted probabilities from the final models.

Risk stratification based on predicted probability. To explore prediction cut-offs to use to inform clinical decision-making, we grouped patients into low-, high- and very high-risk groups based on predicted probabilities obtained from the final models. The target outcome percentages in low and high-risk groups were arbitrarily assigned to be approximately 10-fold lower and 8-fold higher than each outcome percentage in the whole dataset.

External validation cohort

We validated two outcome models (death and poor outcome) using two groups of predictors: demographics and laboratory results. The KCH cohort did not directly provide laboratory tests at the admission date and not all admission dates are reliably associated with COVID-19 due to the possibility of hospital-acquired COVID-19. To identify eligible laboratory results for validation, we used the most recent laboratory results within 4 days after the recorded date of COVID-19 symptom onset. We fitted the death and poor outcome models using the aforementioned predictors (group DL). Model validity was evaluated in relation to discrimination (C-index), calibration (calibration-in-the-large, calibration slope and calibration plots) and clinical usefulness (sensitivity, specificity, and positive and negative predictive values).[5] We additionally examined the performance of the risk prediction model at the selected cut-offs for the three category risk prediction model for both outcomes.

Comparison of informativeness of predictor groups

L1-regularized logistic regression models with fixed regularization strength (alpha=1) were fitted with demographic predictors together with one or combinations of the rest groups of predictors. Fixed regularizations strength (alpha=1) was chosen here for comparability of model performance and coefficients of same predictors in different models.

Model performance and predictor coefficients were assessed by cross-validation as described above. The multivariable informativeness of each group of predictors was quantified by the gain (or loss) of model

performance (ΔC-index) when the group was added. The model performance metrics here served only for comparing informativeness of predictor groups, because they were not tuned to optimal level. Given predictor coefficients in logistic regression model correspond to per unit influence on log odds ratio of the outcome, the relative informativeness of each predictor was assessed by the change of the predictor coefficient after another group of predictors was added to the model.

To improve the comparability between coefficients of categorical and continuous predictors, coefficients of continuous predictors were normalized by multiplying 2 * IQR to show their effect across an estimator full range of value spread. The normalization was done only for illustrative purposes.

Ethics and governance

The derivation study was approved by the Research Ethics Committee of Shanghai Dongfang Hospital and Taikang Tongji Hospital. The external validation study operated under London South East Research Ethics Committee (reference 18/LO/2048) approval granted to the King's Electronic Records Research Interface (KERRI); specific work on COVID-19 research was reviewed with expert patient input on a virtual committee with Caldicott Guardian oversight. All individual-level data at Kings College Hospital was processed and analysed locally.

Supplementary results

Missing data distribution in non-outcome and outcome groups in derivation cohort

In **Table S5**, distribution of missing data in non-outcome and outcome (either death or poor outcomes) groups is summarized. For all predictor combinations, incomplete data affected the group who died more than the surviving group, primarily because some patients died very soon after hospital admission. Since the healthcare system in Wuhan was operating with challenges to its capacity, records of those patients were not completed afterwards. However, risk prediction tools to inform clinical decision making will not be used in such patients, so this should not affect prediction tool performance in the intended target population. For poor outcome, the imbalance of missing data was only found in DCS predictors. The number of missing data was small (1 and 3 missing in not-poor-outcome and poor outcome), nonetheless. The influence of such missing numbers was minimal.

Analysis of relative informativeness of predictor groups

We next tried to understand relative informativeness of each predictor group by adding the groups alone or in combinations (**Table S4**). The baseline model with only demographic predictors (age and sex) made the largest contributed to risk prediction. Adding underlying conditions as predictors decreased coefficients of age and sex without significantly affecting model performance (ΔC-index$_{death}$=-0.034, ΔC-index$_{poor}$= 0.015) (**Table S6**). Adding symptoms as predictors affected neither model performance (ΔC-index$_{death}$=0.036, ΔC-index$_{poor}$= 0.023) nor coefficients of age and sex. Combining underlying conditions and symptoms in the model did not affect coefficients of either group (**Figure S7A-B**).

On the other hand, adding laboratory results as predictors substantially improved model performance (ΔC-index$_{death}$=0.148, ΔC-index$_{poor}$= 0.153) (**Table S6**). When laboratory results were modeled in combination with underlying conditions or symptoms, their coefficients were stable (**Figure S7C-F**). With death as the outcome, coefficients of underlying conditions or symptoms were generally shrunk to or towards zero (**Figure S7C and S7E**); with poor outcome as the outcome, the trend was less clear (**Figure S7D and S7F**).

Supplementary Table 1: Univariate analysis of factors associated with Death and Pool Outcome of COVID-19 patients in the derivation cohort

	Death (n=33/773)		**Poor Outcome (n=75/772)**	
Variable	**Odds ratio (95%CI)**	**p-value**	**Odds ratio (95%CI)**	**p-value**
Demographic				
Age*	1·06 (1·03-1·10)	<0·001	1·07 (1·04-1·09)	<0·001
Sex (male vs female)	1·88 (0·91-3·87)	0·12	1·54 (0·95-2·50)	0·099
Premorbid Conditions				
Chronic lung disease	5·32 (2·15-13·13)	<0·001	6·13 (3·16-11·88)	<0·001
Diabetes Mellitus	2·38 (1·03-5·49)	0·07	1·58 (0·85-2·96)	0·2
Immunocompromised	5·21 (1·09-24·89)	0·12	7·35 (2·27-23·78)	<0.001
Malignancy	2·33 (0·52-10·39)	0·55	1·99 (0·66-5·99)	0·37
Hypertension	2·68 (1·29-5·58)	0·01	3·37 (2·06-5·53)	<0·001
Heart disease	4·43 (2·00-9·81)	<0.001	4·42 (2·51-7·77)	<0·001
Chronic renal disease	5·26 (1·68-16·42)	0·01	4·99 (2·07-12·02)	<0.001
Symptoms				
Fever	1·16 (0·53-2·55)	0·86	1·08 (0·64-1·82)	0·88
Cough	3·35 (1·16-9·65)	0·03	1·49 (0·85-2·59)	0·2
Fatigue	1·42 (0·68-2·95)	0·44	1·52 (0·93-2·50)	0·12
Dyspnoea	5·43 (2·21-13·35)	<0·0001	3·33 (1·97-5·64)	<0·001
Diarrhoea	0·65 (0·09-4·89)	0.996	0·26 (0·03-1·92)	0·26
Admission clinical features				
Neutrophil count (10^9/L)*	1·08 (1·02-1·14)	<0.001	1·33 (1·22-1·44)	<0.001
Lymphocyte (10^9/L)*	0·02 (0·01-0·08)	0·002	0·17 (0·10-0·29)	0·001
Platelet (10^9/L)*	0·99 (0·98-1·00)	0·04	0·99 (0·99-1·00)	0·02
C-reactive protein (mg/L)*	1·03 (1·02-1·04)	<0·001	1·03 (1·02-1·03)	<0·001
Creatinine (μmol/L)*	1·01 (1·00-1·01)	0·003	1·01 (1·01-1·02)	0·009

Two patients had missing data for death and three patients had missing data for poor outcome. * For continuous variables, OR is per unit increase in value so should be interpreted in relation to the actual range of values (e.g. OR for a 60 year old vs a 30 year old is 1·06^30 = 5·74)

Supplementary Table 2: Multivariable logistic regression analysis of risk factors associated with Death and Poor outcome in COVID-19 patients in the derivation cohort: predictor coefficients

	Death coefficient (95%CI)			**Poor outcome coefficient (95%CI)**		
Predictor	**DCS* 30/769 (3·9%)**	**DCSL 19/651 (2·9%)**	**DL 20/653 (3·1%)**	**DCS 72/768 (9·4%)**	**DCSL 57/651 (8·8%)**	**DL 58/653 (8·9%)**
Demographic						
Age	0·0461 (0·0122:0·0799)	-0·0020 (-0·0531:0·0491)	0·0142 (-0·0370:0·0654)	0·0414 (0·0187:0·0640)	0·0182 (-0·0145:0·0509)	0·0281 (-0·0004:0·0565)
Sex (male vs female)	0·3177 (-0·4743:1·1097)	-0·8563 (-2·1144:0·4019)	-0·8035 (-2·0102:0·4031)	0·2050 (-0·3270:0·7370)	-0·5136 (-1·2856:0·2583)	-0·4723 (-1·1882:0·2436)
Premorbid Conditions						
Chronic lung disease	1·0975 (0·0706:2·1243)	0·4054 (-1·3315:2·1424)	..	1·2284 (0·4757:1·9812)	0·9657 (-0·2093:2·1408)	..
Diabetes Mellitus	0·4448 (-0·5191:1·4087)	0·3168 (-1·0493:1·6830)	..	..	-0·2920 (-1·2810:0·6971)	..
Immunocompromised	1·3798 (-0·4451:3·2047)	..	..	1·4338 (0·0494:2·8182)	1·5738 (-0·4570:3·6045)	..
Malignancy	0·0185 (-1·9155:1·9526)	-0·0019 (-2·0489:2·0452)	..	..	-0·7171 (-2·4933:1·0591)	..
Hypertension	0·3457 (-0·4809:1·1722)	..	..	0·6631 (0·1047:1·2215)	0·4819 (-0·3166:1·2804)	..
Heart disease	0·6761 (-0·2613:1·6135)	..	..	0·7408 (0·0745:1·4070)	0·7360 (-0·2202:1·6922)	..
Chronic renal disease	1·3210 (-0·0856:2·7276)	..	..	0·9920 (-0·1025:2·0865)	-0·3503 (-2·2309:1·5304)	..
Symptoms						
Fever	..	..	..	0·0280 (-0·5615:0·6175)	0·1678 (-0·6413:0·9769)	..
Cough	0·7409 (-0·3614:1·8431)	0·9446 (-0·6341:2·5233)	..	..	-0·6005 (-1·4021:0·2012)	..
Fatigue	..	..	..	0·0996 (-0·4731:0·6724)	0·2024 (-0·5748:0·9796)	..
Dyspnoea	1·5245 (0·5816:2·4675)	0·8559 (-0·4246:2·1365)	..	1·0902 (0·4967:1·6837)	1·2438 (0·4404:2·0472)	..
Diarrhoea	..	..	..	-0·2074 (-1·7452:1·3304)	-2·6571 (-8·2778:2·9636)	..
Admission clinical features						

Neutrophil count (10^9/L)	··	0·1793 (0·0294:0·3292)	0·2249 (0·0844:0·3654)	··	0·2489 (0·1114:0·3864)	0·2575 (0·1340:0·3810)
Lymphocyte (10^9/L)	··	-1·4449 (-2·9297:0·0400)	-1·3255 (-2·7882:0·1372)	··	-1·2256 (-2·1206:-0·3306)	-1·1842 (-1·9857:-0·3827)
Platelet (10^9/L)	··	-0·0059 (-0·0129:0·0011)	-0·0052 (-0·0118:0·0015)	··	-0·0036 (-0·0081:0·0009)	-0·0027 (-0·0068:0·0013)
C-reactive protein (mg/L)	··	0·0140 (0·0028:0·0251)	0·0138 (0·0035:0·0240)	··	0·0078 (-0·0020:0·0176)	0·0082 (-0·0003:0·0167)
Creatinine (μmol/L)	··	0·0012 (-0·0023:0·0047)	0·0015 (-0·0020:0·0049)	··	0·0125 (0·0045:0·0204)	0·0099 (0·0036:0·0161)
Intercept	-8·5016	-3·2048	-3·5172	-6·2993	-4·2863	-4·1865

Data in the table head are number (%). Number is the number patients with complete data of the corresponding risk group. % is the percentage of patients with the outcome. Data in the table cells are coefficient (odds ratio) or number (%). ·· indicates the corresponding predictor is not used. *Initials for specifying predictor groups: D – Demographic, C - Underlying Conditions, S – Symptoms, L - Laboratory Results.

Supplementary Table 3: Risk stratification of patients in derivation and validation cohorts in DL models

	Death		**Poor outcome**	
	Derivation % who die (no. who die/no. in group)	External validation % who die (no. who die/no. in group)	Derivation % with poor outcome (no. with poor outcome/no. in group)	External validation % with poor outcome (no. with poor outcome/no. in group)
Low risk*	0·34% (2/580)	16·5% (20/121)	0·63% (2/320)	26·3% (5/19)
High risk*	15·0% (3/20)	42·9% (12/28)	8·9% (25/280)	28·4% (33/116)
Very high risk*	28·3% (15/53)	58·4% (45/77)	58·5% (31/53)	64·8% (59/91)

* Probability cut offs for Low-High risk are 0·039 for death and 0·030 for poor outcome. Probability cut-offs for High-Very High risk are 0·069 for death and 0·275 for poor outcome.

Supplementary Table 4: Summary of parameters in all modelling analyses

Analysis	Outcome	Predictor	Sample size (n)	Outcome (n)	Non-outcome (n)	Model	Solver	L1-reg. Strength	Cross-Validation	Tuning
Predictor selection										
	Death	DL	653	20	633	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Death	DCS	769	30	739	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Death	DCSL	651	19	632	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DL	653	58	595	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DCS	768	72	696	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DCSL	651	57	594	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
Parameter tuning										
	Death	DL	653	20	633	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Death	DCS	769	30	739	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Death	DCSL	651	19	632	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DL	653	58	595	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DCS	768	72	696	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
	Poor outcome	DCSL	651	57	594	logistic regression	liblinear	range [2^-4, 2^5]	yes	yes
Final model fitting										
	Death	DL	653	20	633	logistic regression	liblinear	2^-4	no	no
	Death	DCS	769	30	739	logistic regression	liblinear	2^-1	no	no
	Death	DCSL	651	19	632	logistic regression	liblinear	2^-2	no	no

	Poor outcome	DL	653	58	595	logistic regression	liblinear	2^-4	no	no
	Poor outcome	DCS	768	72	696	logistic regression	liblinear	2^0	no	no
	Poor outcome	DCSL	651	57	594	logistic regression	liblinear	2^-2	no	no
Comparison of relative informativeness										
	Death	D	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DC	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DS	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DCS	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DCL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DSL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Death	DCSL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	D	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DC	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DS	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DCS	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DCL	651	19	632	logistic regression	liblinear	2^0	yes	no
	Poor outcome	DSL	651	19	632	logistic regression	liblinear	2^0	yes	no

	Poor outcome	DCSL	651	19	632	logistic regression	liblinear	2^0	yes	no

Supplementary Table 5: Missing data distribution in non-outcome and outcome groups

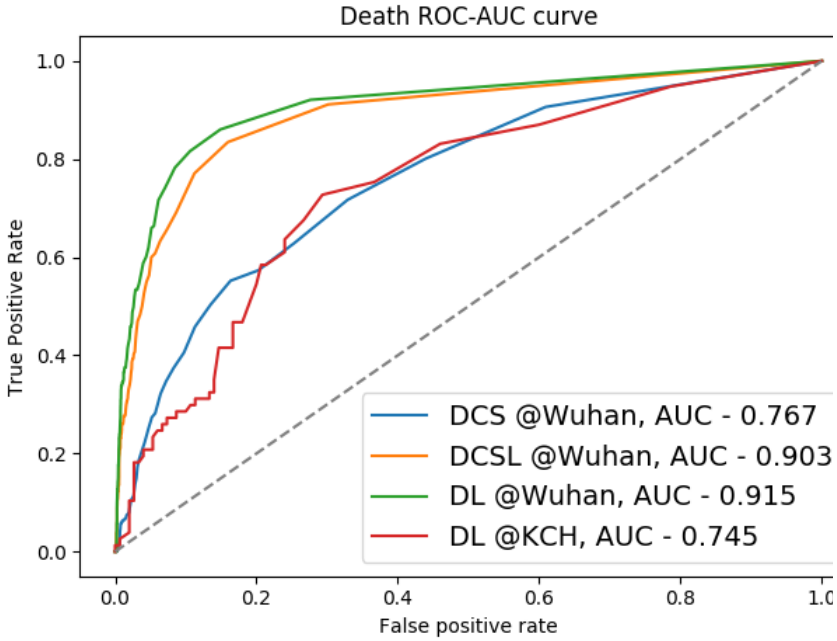
Outcome	Predictor	Sample size (n)	Complete (n)	Missing (n)	Complete in non-outcome (n)	Complete in outcome (n)	Missing in non-outcome (n)	Missing in outcome (n)	Odds ratio	p_value
Death	DCS	773	740	33	738	30	2	3	36·9	p<0·001
Death	DCSL	773	740	33	632	19	108	14	4·3	p<0·001
Death	DL	773	740	33	633	20	107	13	3·8	p<0·001
Poor outcome	DCS	772	697	75	696	72	1	3	29·0	p<0·001
Poor outcome	DCSL	772	697	75	594	57	103	18	1·8	0·055
Poor outcome	DL	772	697	75	595	58	102	17	1·7	0·096

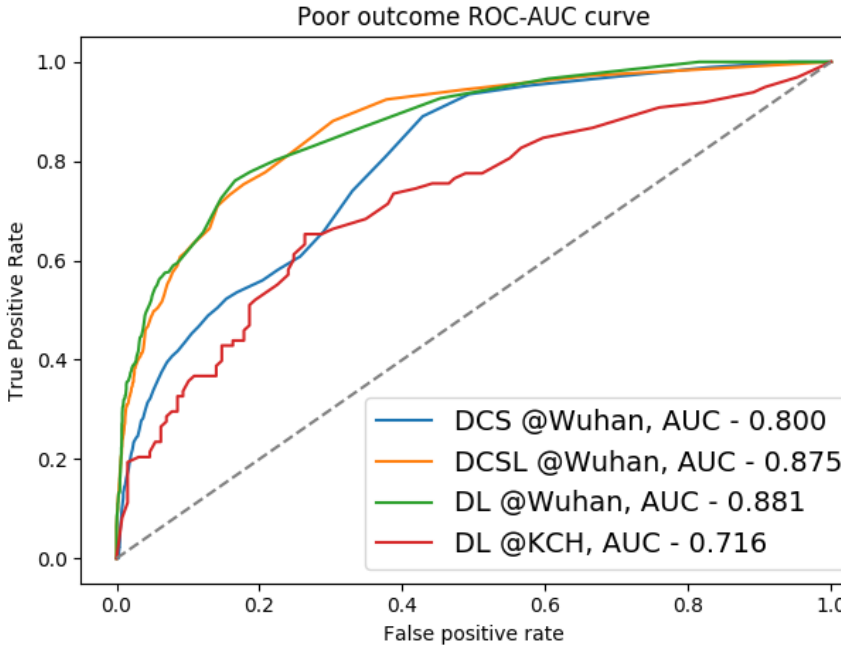
Supplementary Table 6: Relative informativeness of predictor groups

Outcome	**Death**								**Poor outcome**							
Predictor	**D***	**DC**	**DS**	**DL**	**DCS**	**DCL**	**DSL**	**DCSL**	**D**	**DC**	**DS**	**DL**	**DCS**	**DCL**	**DSL**	**DCSL**
sensitivity	0.000	0.000	0.000	0.267	0.000	0.268	0.256	0.250	0.000	0.013	0.000	0.340	0.029	0.335	0.351	0.344
specificity	1.000	1.000	1.000	0.993	1.000	0.992	0.992	0.991	1.000	0.999	1.000	0.987	0.998	0.985	0.985	0.983
PPV	0.000	0.000	0.000	0.472	0.000	0.480	0.441	0.407	0.000	0.160	0.000	0.713	0.270	0.663	0.685	0.643
NPV	0.970	0.970	0.970	0.978	0.970	0.978	0.978	0.978	0.912	0.913	0.912	0.940	0.915	0.940	0.941	0.941
C-index	0.723	0.688	0.759	0.871	0.728	0.866	0.882	0.878	0.718	0.734	0.741	0.871	0.760	0.866	0.879	0.876
ΔC-index**	0.000	-0.035	0.036	0.148	0.005	0.143	0.159	0.155	0.000	0.016	0.023	0.153	0.042	0.148	0.161	0.158

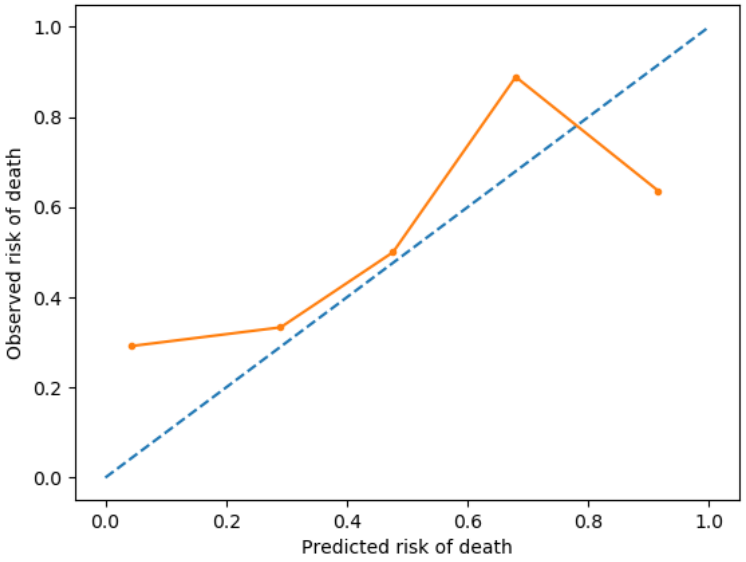
* Initials specifying predictor categories: D – Demographic, C - Premorbid Conditions, S – Symptoms, L - Laboratory Results. ** ΔC-index compared to the baseline models with D predictors.

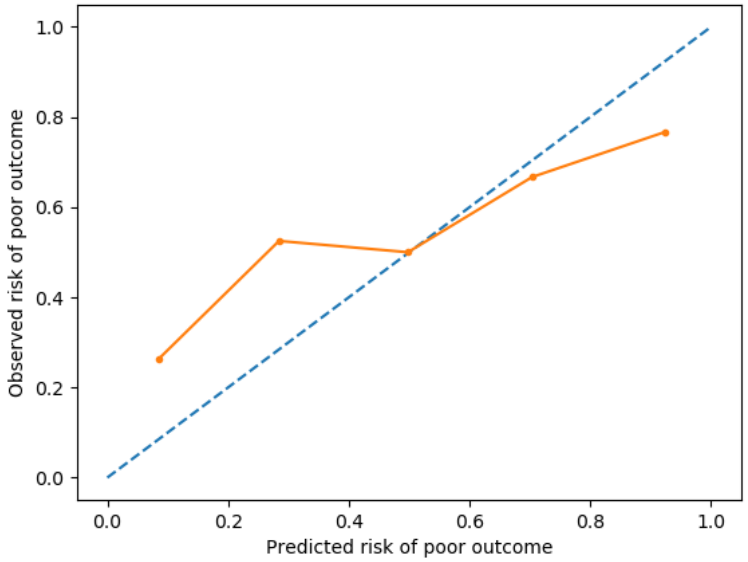
Supplementary Figure 1: ROC-AUC curves of different models in predicting death and poor outcome in derivation and external validation cohorts: A) Death as outcome B) Poor outcome as outcome



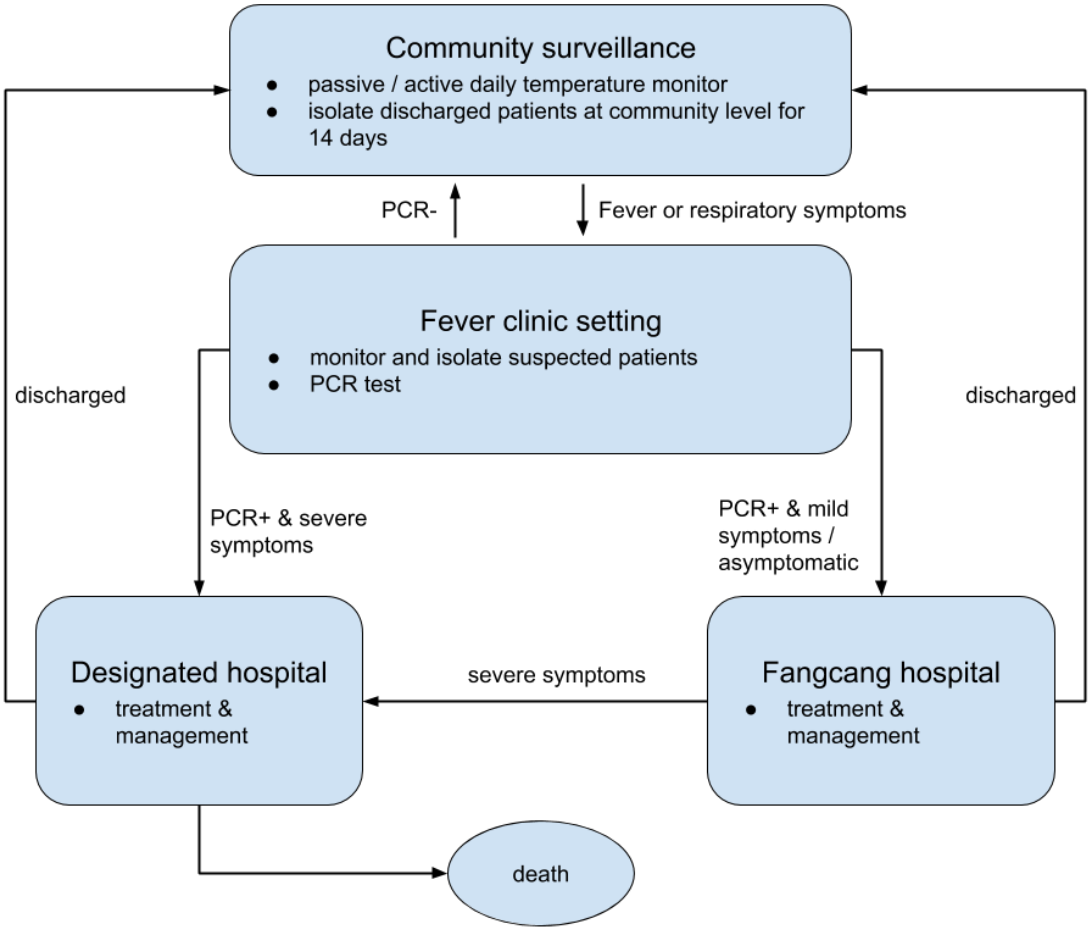


Supplementary Figure 2: Calibration plots of prediction models on external validation cohort: A) Calibration plot of DL-Death model in external validation, B) Calibration plot of DL-Poor model in external validation

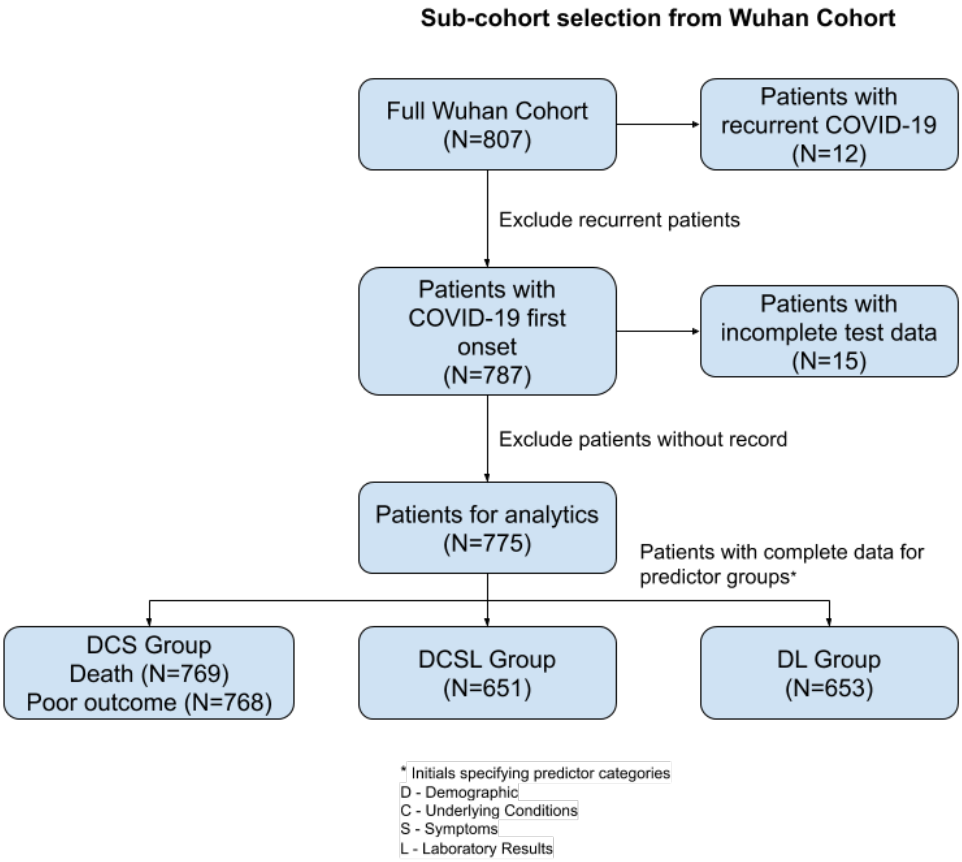




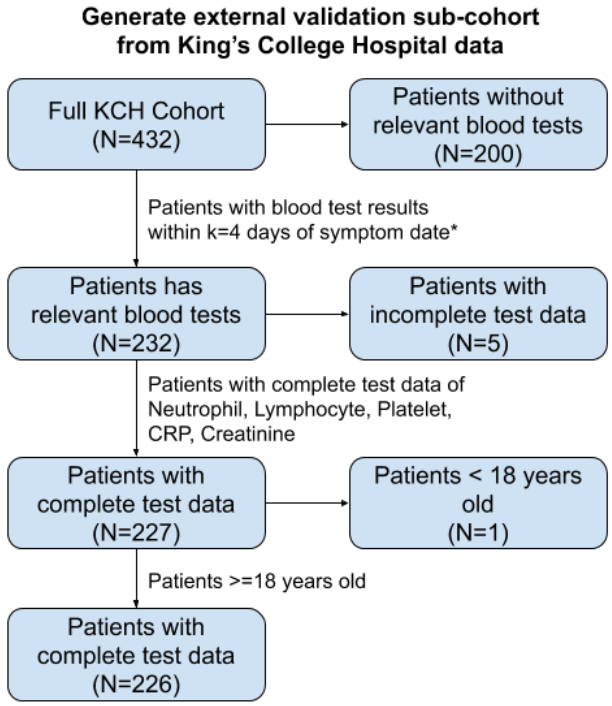
Supplementary Figure 3: Flow chart of COVID-19 patient pathway in Wuhan



Supplementary Figure 4: Wuhan cohort patient selection process



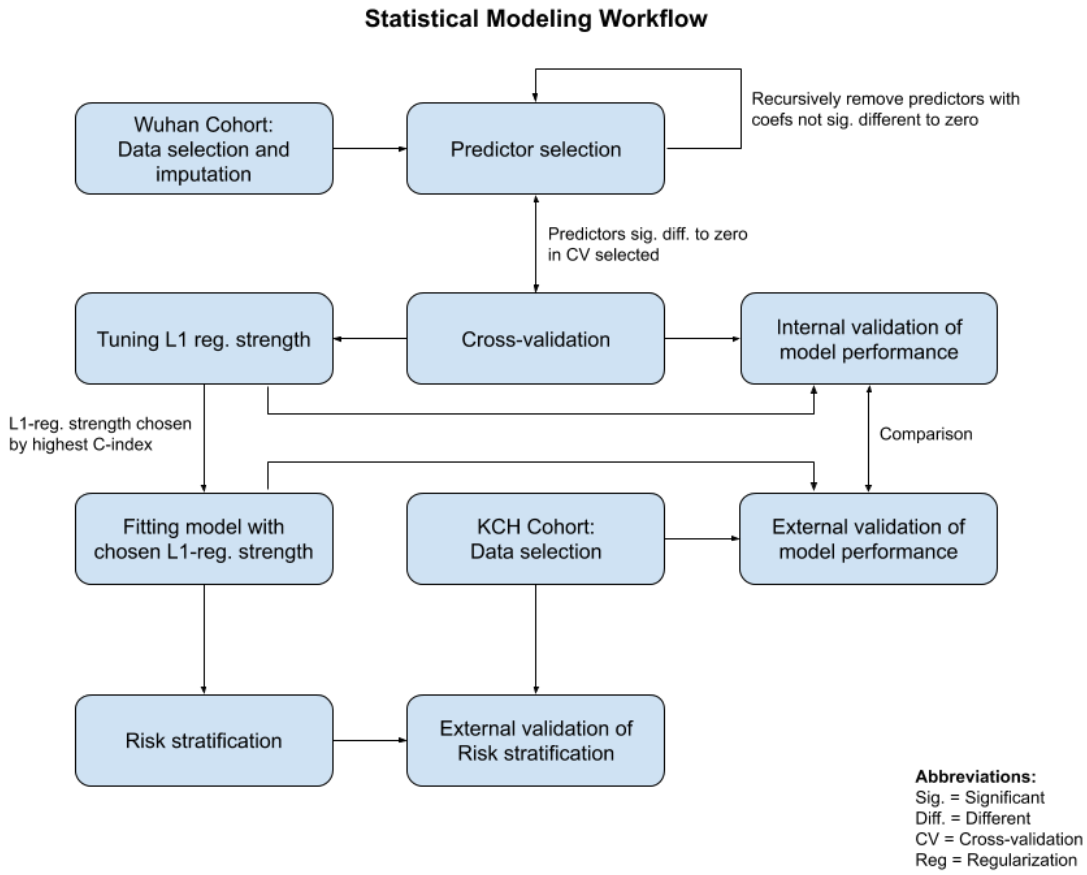
Supplementary Figure 5: King's College Hospital cohort patient selection process



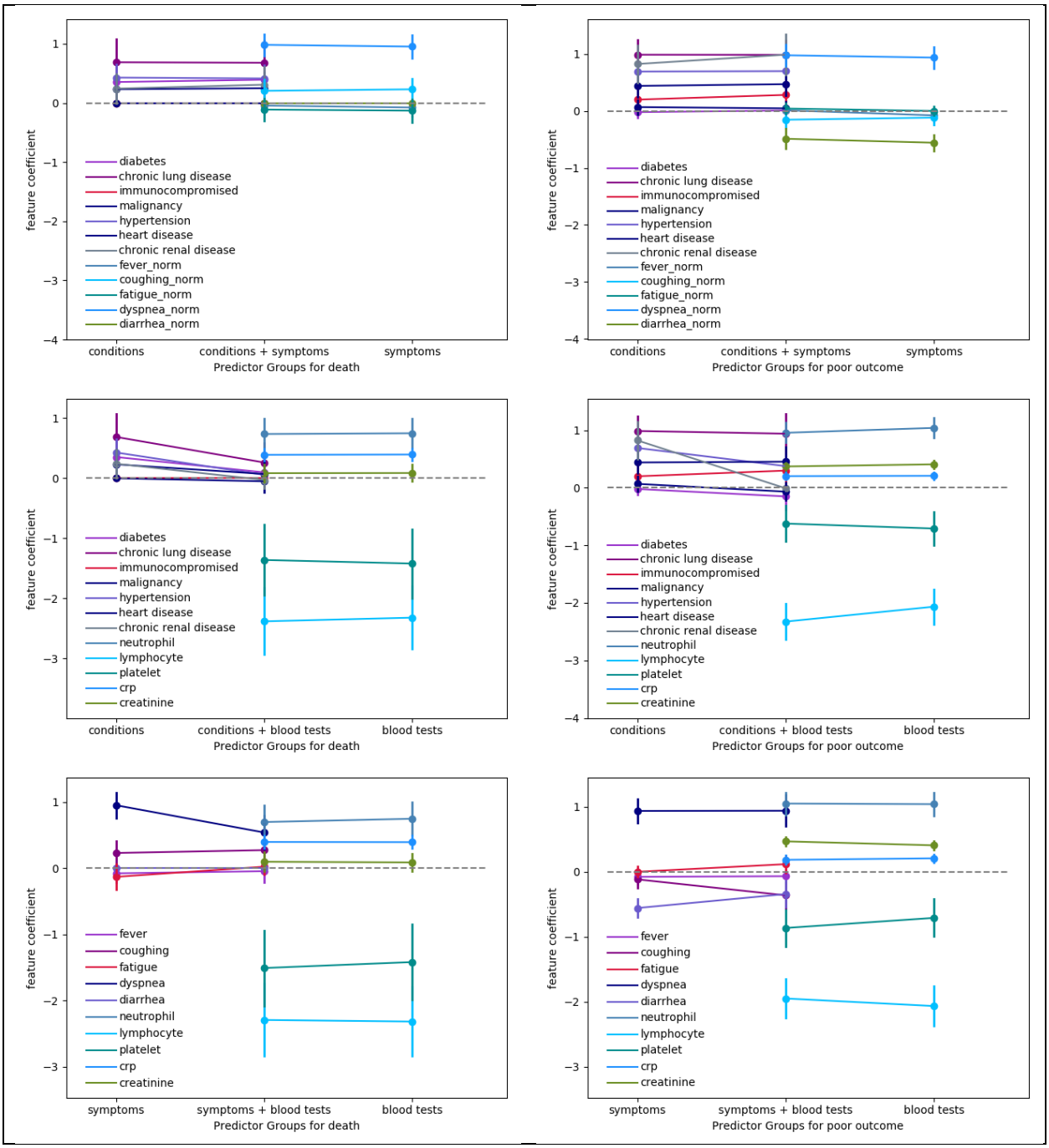
* symptom date is used because it is the only reliable COVID-19 related date for KCH cohort. This date is the patient reported date when COVID-19 symptoms first appeared. When patients could not report such a date (e.g., dementia or dysphasic), the test date was used.

If there are multiple blood tests, only the test closest to the symptom date is used.

Supplementary Figure 6: Statistical Modelling Workflow



Supplementary Figure 7: Relative informativeness of predictor groups in multivariate analysis



References

1. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350**: g7594.
2. Organisation WH. WHO R&D Blueprint novel Coronavirus COVID-19 Therapeutic Trial Synopsis. Geveva, Switzerland: WHO, 2020.
3. Force ADT, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012; **307**(23): 2526-33.
4. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**(3): 261-72.
5. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol* 2002; **20**(2): 96-107.
6. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 2011*; **12**(Oct), 2825–2830.