PLOS ONE

RESEARCH ARTICLE

Regional performance variation in external validation of four prediction models for severity of COVID-19 at hospital admission: An observational multi-centre cohort study

Kristin E. Wickstrøm[1,2], Valeria Vitelli[3], Ewan Carr[4], Aleksander R. Holten[2,5], Rebecca Bendayan[4,6], Andrew H. Reiner[7], Daniel Bean[4,8], Tom Searle[4,6], Anthony Shek[9], Zeljko Kraljevic[4], James Teo[9,10], Richard Dobson[4,6,8,11,12], Kristian Tonby[2,13], Alvaro Köhn-Luque[2], Erik K. Amundsen[1,14]*

1 Department of Medical Biochemistry, Blood Cell Research Group, Oslo University Hospital, Oslo, Norway, **2** Institute of Clinical Medicine, University of Oslo, Oslo, Norway, **3** Oslo Centre for Biostatistics and Epidemiology, Faculty of Medicine, University of Oslo, Oslo, Norway, **4** Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, **5** Department of Acute Medicine, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway, **6** NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, United Kingdom, **7** Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway, **8** Health Data Research UK London, University College London, London, United Kingdom, **9** Department of Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, **10** King's College Hospital NHS Foundation Trust, London, United Kingdom, **11** Institute of Health Informatics, University College London, London, United Kingdom, **12** NIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust, London, United Kingdom, **13** Department of Infectious Diseases, Oslo University Hospital, Oslo, Norway, **14** Department of Life Sciences and Health, Oslo Metropolitan University, Oslo, Norway

* uxamue@ous-hf.no

Citation: Wickstrøm KE, Vitelli V, Carr E, Holten AR, Bendayan R, Reiner AH, et al. (2021) Regional performance variation in external validation of four prediction models for severity of COVID-19 at hospital admission: An observational multi-centre cohort study. PLoS ONE 16(8): e0255748. https://doi.org/10.1371/journal.pone.0255748

Editor: Chiara Lazzeri, Azienda Ospedaliero Universitaria Careggi, ITALY

Received: March 20, 2021

Accepted: July 22, 2021

Published: August 25, 2021

Data Availability Statement: Code and pre-trained models are available at https://github.com/ocbe-uio/validation-covid19-OUH-KCL for model validation and openly shared for testing in other COVID-19 datasets. KCH: Source text from patient records used at all sites in the study will not be available due to inability to safely fully anonymise up to the Information Commissioner Office (ICO) standards and would be likely to contain strong identifiers (e.g. names, postcodes) and highly sensitive data (e.g. diagnoses). A subset of the

Abstract

Background

Prediction models should be externally validated to assess their performance before implementation. Several prediction models for coronavirus disease-19 (COVID-19) have been published. This observational cohort study aimed to validate published models of severity for hospitalized patients with COVID-19 using clinical and laboratory predictors.

Methods

Prediction models fitting relevant inclusion criteria were chosen for validation. The outcome was either mortality or a composite outcome of mortality and ICU admission (severe disease). 1295 patients admitted with symptoms of COVID-19 at Kings Cross Hospital (KCH) in London, United Kingdom, and 307 patients at Oslo University Hospital (OUH) in Oslo, Norway were included. The performance of the models was assessed in terms of discrimination and calibration.

KCH dataset limited to anonymisable information (e.g. only SNOMED codes and aggregated demographics) is available on request to researchers with suitable training in information governance and human confidentiality protocols subject to approval by the King's College Hospital Information Governance committee; applications for research access should be sent to kch-tr.cogstackrequests@nhs.net. This dataset cannot be released publicly due to the risk of re-identification of such granular individual-level data, as determined by the King's College Hospital Caldicott Guardian. OUH: The OUH dataset cannot be released publicly due to the risk of reidentification of such granular individual-level data. Researchers can contact the data protection officer at personvern@oslo-universitetssykehus.no for any enquiries.

Funding: One of the authors (JTHT) have previously received research support and funding from InnovateUL, Bristol-Myers-Squibb, iRhytm Technologies, and hold shares under £5000 in Glaxo SmithKline and Biogen. No external commercial financial support was received for this work. Institutional support and grants are stated under funding information. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: One of the authors (JTHT) have received research support and funding from InnovateUL, Bristol-Myers- Squibb, iRhytm Technologies, and hold shares <£5000 in Glaxo SmithKline and Biogen. No external commercial financial support was received for this work. Institutional support and grants are stated under funding information. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The funders did not alter our adherence to PLOS ONE policies on sharing data and materials.

Results

We identified two models for prediction of mortality (referred to as Xie and Zhang1) and two models for prediction of severe disease (Allenbach and Zhang2). The performance of the models was variable. For prediction of mortality Xie had good discrimination at OUH with an area under the receiver-operating characteristic (AUROC) 0.87 [95% confidence interval (CI) 0.79–0.95] and acceptable discrimination at KCH, AUROC 0.79 [0.76–0.82]. In prediction of severe disease, Allenbach had acceptable discrimination (OUH AUROC 0.81 [0.74–0.88] and KCH AUROC 0.72 [0.68–0.75]). The Zhang models had moderate to poor discrimination. Initial calibration was poor for all models but improved with recalibration.

Conclusions

The performance of the four prediction models was variable. The Xie model had the best discrimination for mortality, while the Allenbach model had acceptable results for prediction of severe disease.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was discovered in Wuhan, China in December 2019. The virus was shown to cause viral pneumonia, later designated as coronavirus disease 2019 (COVID-19) [1]. The disease has evolved as a pandemic with an extensive amount of severe cases with high mortality [2]. Several biomarkers, clinical and epidemiological parameters have been associated with disease severity [3, 4]. Practical tools for prediction of prognosis in COVID-19 patients are still lacking in clinical practice [5, 6]. We observed that many laboratory tests are ordered for patients with COVID-19 due to their predictive value. Very likely, there is redundancy in the information from the different tests and it could be possible to improve the prediction by using a multivariable model and reducing the number of redundantly ordered tests. Prediction models can be crucial to prioritize patients needing hospitalization, intensive care treatment, or future individualized therapy.

Since the onset of the pandemic, the number of prediction models for COVID-19 patients has been continuously growing [7]. Prediction models should be validated in different populations with a sufficient number of patients reaching the outcome before implementation [8–10]. A validation study of 22 prediction models at one site was recently published [6]. Interestingly, this study found that none of the models performed better than oxygen saturation alone, even though the performance at the original study sites in most cases was much better.

This study aimed to validate published prediction models of severity and mortality for hospitalized patients based on laboratory and clinical values in COVID-19 cohorts from London (United Kingdom) and Oslo (Norway).

The study is reported according to the guidelines in "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD) [11] and has also followed recommendations from "Prediction Model Risk of Bias Assessment Tool" (PROBAST) [12].

Methods

Study design and participants

The study was performed as a retrospective validation study with adult patients hospitalized with COVID-19. Two cohorts were included: (1) Oslo University Hospital (OUH) in Norway,

(2) Kings Cross Hospital (KCH) in London, United Kingdom. The patients included were all adult inpatients testing positive for SARS-CoV-2 by real-time polymerase chain reaction (RT-PCR) with symptoms consistent with COVID-19 at admission. SARS-CoV-2 -positive patients admitted for conditions not related to COVID-19 were excluded, e.g. pregnancy-related conditions or trauma. Patients referred from other hospitals were also excluded, as we did not have access to measurements from the first hospital admission.

OUH cohort. OUH is a large urban university hospital. Patients admitted between 6th March and 31th December 2020 were included. The OUH project protocol was approved by the Regional Ethical Committee of South East Norway (Reference 137045). All patients with confirmed COVID-19 were included in the quality registry "COVID19 OUS", approved by the data protection officer (Reference 20/08822). Informed consent was waived because of the strictly observational nature of the project. Demographics, clinical variables and hospital stay information were manually recorded in the registry and merged with laboratory results exported from the laboratory information system in Microsoft Excel.

KCH cohort. In the KCH cohort patients were admitted between 23rd February to 1st May 2020 at two hospitals (King's College Hospital and Princess Royal University Hospital) in South East London (UK) of Kings College Hospital NHS Foundation Trust.

Data (demographics, emergency department letters, discharge summaries, lab results, vital signs) were retrieved from components of the electronic health record (EHR) using a variety of natural language processing (NLP) informatics tools belonging to the CogStack ecosystem [13]. The study adhered to the principles of the UK Data Protection Act 2018, UK National Health Service (NHS) information governance requirements, and the Declaration of Helsinki. De-identified data from patients admitted to KCHFT were analysed under London SE Research Ethics Committee approval (reference 18/LO/2048) granted to the King's Electronic Records Research Interface (KERRI) which waives the consent requirements. Specific work on COVID-19 on the de-identified data was approved by the KERRI committee which included patients and the Caldicott Guardian in March 2020 and reaffirmed in May 2020. Data from this cohort has been published in prior studies [14, 15].

Selection of prediction models

A literature search was performed to select prediction models for validation. Published articles or preprint manuscripts were included until 29.05.2020. A structured search was performed in PubMed with the words "COVID-19" and "prediction model" or "machine learning" or "prognosis model". Prediction models included in the review by Wynants et al. [7] published April 7th 2020 were also investigated, as well as search for articles/preprints citing Wynants et. al. using Google Scholar 18.05.2020.

The inclusion criteria for selection of multivariable prediction models were: (1) Symptomatic hospitalized patients over 18 years with PCR confirmed COVID-19; (2) outcomes including respiratory failure or intensive care unit (ICU) admission or death or composite outcomes of these. (3) The predictive models had to include at least one laboratory test as we wanted to explore models that combined clinical and laboratory variables (4). All variables had to be available in the datasets and the model had to be described in adequate detail.

Missing values

Predictive variables were collected from the admission to the emergency department (ED). If not available in the ED, the first available values within 24 hours from hospital admission were used. Missing values (i.e. no recorded values within 24 hours) were imputed using both simple (k-nearest neighbors (KNN) and random forest) and multiple imputation (Bayesian ridge and

Gaussian process) [16, 17], using the multivariate imputation by chained equations method implemented in the Python function *IterativeImputer* available in the scikit-learn package, version 0.24.2 [18].

Statistical analyses and performance measurements for the prediction models

Univariate comparisons between patients with 'mild' versus 'severe' disease were carried out for continuous (Wilcoxon rank-sum test) and binary (X^2 test) measures. Severe disease was defined as transfer to ICU or in-hospital mortality.

Validation of the selected prediction models was assessed with discrimination and calibration as recommended in TRIPOD [11]. Discrimination is the ability of the model to differentiate between those who do or do not experience the outcome. It is commonly estimated by concordance index (c-index) which is identical to the area under the receiver-operating characteristic curve (AUROC) for models with binary endpoints. The discrimination for the models at OUH and KCH was also compared to the discrimination in the original development cohort and to the external validation by Gupta et al. [6]. Calibration is the agreement between the observed outcomes and the outcome predictions from the model. It is preferably reported by a calibration plot, intercept and slope.

The mortality rate and the rate of 'poor outcome' varied between the cohorts. The models were therefore recalibrated by adjusting the intercept of the logistic regression models according to the frequency of outcomes at each study site [19]. Validation of the recalibration was not performed. All statistical analyses were conducted in Python 3.7 and R 3.4 [20].

Results

Selection of prediction models

Four publications comprising five prediction models met our inclusion criteria [14, 21–23]. The inclusion process is illustrated in Fig 1. However, since one of the models was developed at KCH and validated at OUH in a previous publication [14], only four models are presented here. The four models are referred to as 'Xie'[21], 'Zhang1', 'Zhang2'[22] and 'Allenbach'[23].

Information on the predictor variables and outcomes of the four models are summarized in Table 1.

All predictors were measured at hospital admission. Treatment of missing values in the development cohorts was not well described and imputation methods were not mentioned. The Xie model had hospital mortality as the only outcome. Zhang presented two models with different outcomes: (1) Mortality and (2) Composite outcome of mortality or 'poor outcome'. Poor outcome was defined as acute respiratory distress syndrome (ARDS), intubation or extracorporeal membrane oxygenation (ECMO) treatment, ICU admission or death. The Allenbach model used a composite outcome of transfer to ICU or mortality within 14 days of hospital admission. There were no details of the censoring date in the original studies. Mortality during the hospital stay was used for the OUH cohort and for the KCH cohort hospital mortality at data collection time.

All prediction models were based on multiple logistic regression and presented coefficients and intercepts for the different variables that enabled the calculation of risk prediction for our cohorts. Allenbach additionally provided an 8-point scoring system derived from the logistic regression model. However, we chose to use the regression model for calculation as this retains as much information as possible.

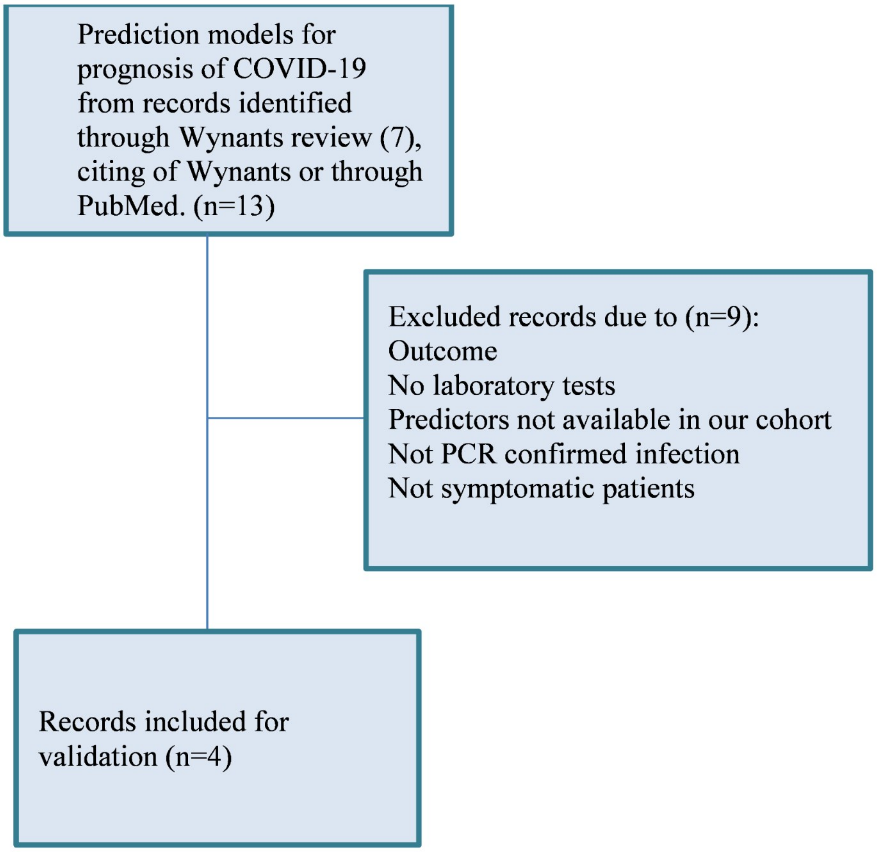
Prediction models for prognosis of COVID-19 from records identified through Wynants review (7), citing of Wynants or through PubMed. (n=13)
Excluded records due to (n=9):
Outcome
No laboratory tests
Predictors not available in our cohort
Not PCR confirmed infection
Not symptomatic patients
Records included for validation (n=4)

Fig 1. Selection of prediction models for validation.

https://doi.org/10.1371/journal.pone.0255748.g001

Description of the cohorts

Patient characteristics for the three development cohorts and the KCH and OUH cohorts are shown in S1 Table.

Since the three models use different outcomes and timeframes, the number of patients included in each validation is not the same. An overview of missing values is presented in Table 2. Missing values were imputed via simple imputation and multiple imputations [17]. Preliminary analyses showed no differences between AUROCs calculated with different

Table 1. Predictors and outcomes in the four prediction models.

	Zhang models	Xie model	Allenbach model
Country of development cohort	China	China	France
Predictors	Age, sex, neutrophil count, lymphocyte count, platelets count, CRP and creatinine at admission.	Age, LDH, SpO2, lymphocyte count (log, due to extreme value).	CRP (per 100mg/L), age, lymphocyte count, WHO scale (22) by admission.
Outcome	1. Mortality	1. Hospital mortality	1. ICU transfer or death by 14 days after admission.
	2. Poor outcome, defined as developing ARDS, receiving intubation or ECMO treatment, ICU admission and death.		

CRP; C-reactive protein, LDH; lactate dehydrogenase, SpO2; Peripheral oxygen saturation, WHO; World Health Organization, ICU; Intensive care unit, NEWS2; National Early Warning score 2, eGFR; estimated glomerular filtration rate, ECMO; extracorporeal membrane oxygenation, ARDS; acute respiratory distress syndrome.

https://doi.org/10.1371/journal.pone.0255748.t001

Table 2. Validation of the four prediction models.

	Zhang1			Zhang2			Xie			Allenbach		
	Validation		Dev.	Validation		Dev.	Validation		Dev.	Validation		Dev.
	Oslo	London	Wuhan	Oslo	London	Wuhan	Oslo	London	Wuhan	Oslo	London	Paris
Participants, n	307	1244	775	307	1244	775	307	1286	299	307	1248	152
Outcome, n (%)	32 (10)	333 (27)	33 (4.3)	66 (22)	419 (34)	75 (9.7)	32 (10)	333 (26)	155 (52)	62 (20)	389 (31)	47 (32)
Missing values Predictors (%)												
• ALC	3.9	4.6	*	3.9	4.6	*	3.9	7.7	*	3.9	4.9	*
• ANC	3.9	4.7	NA	3.9	4.7	NA	NA	NA	NA	NA	NA	NA
• Platelets	NA	4.5	NA	NA	4.5	NA	NA	NA	NA	NA	NA	NA
• CRP	NA	3.3	NA	NA	3.3	NA	NA	NA	NA	NA	3.6	NA
• LDH	NA	NA	NA	NA	NA	NA	12.4	87.8	NA	NA	NA	NA
• Crea.	NA	3.5	NA	NA	3.5	NA	NA	NA	NA	NA	NA	NA
• SaO2	NA	NA	NA	NA	NA	NA	NA	33.3	NA	NA	NA	NA
• WHO	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	33.8	NA
Outcome (%)	None	0.07	NA	None	4.1	0.065	None	0.07	NA	None	3.8	0.03
AUROC (C-index)	0.72 [0.62–0.82]	0.64 [0.60–0.68]	0.91	0.77 [0.70–0.84]	0.67 [0.64–0.70]	0.88	0.87 [0.79–0.95]	0.79 [0.76–0.82]	0.89 [0.86–0.93]	0.81 [0.74–0.88]	0.72 [0.68–0.75]	0.79
Calibration Slope	0.57 [0.27–0.87]	0.37 [0.18–0.56]	0.98 [0.24–1.72]	1.18 [0.59–1.77]	0.75 [0.58–0.92]	1.04 [0.79–1.29]	0.86 [0.71–1.00]	1.03 [0.89–1.17]	1.00 [0.77–0.26]	1.03 [0.79–1.28]	0.76 [0.62–0.89]	0.89
Calibration intercept	0.04 [-0.01–0.09]	0.17 [0.10–0.23]	0.19 [-0.10–0.48]	-0.02 [-0.17–0.12]	0.10 [0.04–0.16]	0.01 [-0.13–0.15]	-0.02 [-0.05–0.01]	-0.04 [-0.09–0.01]	0.00 [-0.33–0.33]	0.00 [-0,06–0.07]	0.09 [0.04–0.14]	-0.06
Calibration before recal.;slope	0.47 [0.24–0.71]	0.38 [0.18–0.58]	-	1.56 [0.68–2.43]	0.92 [0.66–1.17]	-	0.53 [0.29–0.77]	0.87 [0.73–1.00]	-	1.19 [0.94–1.45]	0.87 [0.70–1.00]]	-
Calibration before recal.; intercept	0.03 [-0.02–0.08]	0.17 [0.11–0.24]	-	0.01 [-0.13–0.15]	0.16 [0.10–0.22]	-	-0.06 [-0.16–0.03]	-0.12 [-0.19--0.06]	-	0.02 [-0.04–0.07]	0.12 [0.07–0.17]	-

* Information missing.

Dev. = Development, ALC = Absolute lymphocyte count, ANC = Absolute neutrophil count, Crea. = Creatinine, recal. = recalibration.

https://doi.org/10.1371/journal.pone.0255748.t002

imputation methods (see S2 Table). Thus, the simple imputation method k-nearest neighbor was used for the rest of this paper. At KCH the number of missing values was very high for LDH (87.8%) and relatively high for SpO2 (33.3%) and WHO scale (33.8%).

The OUH cohort consisted of 307 patients while the KCH cohort consisted of 1295 patients (S1 Fig). For the OUH cohort median age was 60 years with 57% males, while in the KCH cohort the median age was 69 with 59% males. In the OUH cohort, 32 patients died in the hospital (10.4%), while 333 (26.8%) had died at the hospital by data collection time in the KCH cohort. For the composite outcome death or ICU transfer, the number of patients with the outcome was 66 (21.5%) at OUH and 419 (33.7%) at KCH.

The percentage of patients with hypertension and diabetes was higher in the KCH cohort (54% and 35%, respectively) than in the OUH cohort (34% and 21%, respectively). The patients at KCH also had higher levels of CRP, creatinine, LDH, and possibly a lower number of lymphocytes than the OUH patients; all of which are known predictors for severe COVID-19.

In Table 3, univariate associations are presented for mild/moderate and severe groups for the KCH and OUH cohorts. In general, the same variables were predictive for severe disease at

Table 3. Univariate analysis of predictors in mild/moderate and severe disease.

	OUH cohort				KCH cohort			
	N	Mild/Moderate disease	Severe disease	P-value	N	Mild/moderate disease	Severe disease	P-value
Age	307	55 [46–70]	68 [58–78]	<0.01	1295	67 [53–82]	75 [62–86]	<0.01
Male sex (%)	307	129 (54)	46 (70)	0.02	1295	463 (56)	271(65)	0.01
Hypertension (%)	307	75 (31)	29 (44)	0.05	1295	428 (52)	244 (58)	0.04
Diabetes (%)	307	46 (19)	18 (27)	0.15	1295	282 (34)	154 (37)	0.40
Ischemic heart disease (%)	307	20 (8)	13 (20)	0.01	1295	105 (13)	66 (16)	0.17
Chronic lung disease (%)	307	61 (25)	22 (33)	0.19	1295	82 (10)	52 (12)	0.22
Days at hospital	307	5.0 [2.0–9.0]	16.5 [8.0–24.0]	<0.01	854	7.0 [3.0–12.0]	16.0 [10.5–31.1]	<0.01
Temperature (celcius)	306	37.1 [36.5–37.8]	37.8 [36.8–38.8]	<0.01	864	36.9 [36.6–37.4]	37.0 [36.6–37.5]	0.22
Resp/min (highest)	303	22 [18–28]	28 [22–32]	<0.01	860	19 [18–20]	20 [19–24]	<0.01
NEWS2 score	299	4 [2–6]	7 [5–10]	<0.01	815	2 [1–4]	4 [2–6]	<0.01
SpO2 [1]	307	96.0 [93.0–98.0]	92.0 [87.3–95.0]	<0.01	858	96.0 [95.0, 98.0]	96.0 [94.0, 97.0]	<0.01
CRP (mg/L)	307	34 [10–74]	93 [45–154]	<0.01	1203	73 [33–128]	118 [59–196]	<0.01
Creatinine (μmol/L)	307	77 [64–94]	97 [71–128]	<0.01	1200	87 [69–118]	108 [83–166]	<0.01
LDH (U/L)	269	237 [188–305]	329 [242–499]	<0.01	157	349 [277–431]	532 [393–706]	<0.01
Neutrophils (10^9/L)	295	4.0 [2.0–5.9]	5.7 [2.8–7.9]	<0.01	1186	5.1 [3.6–7.2]	6.4 [4.5–8.8]	<0.01
Lymphocytes (10^9/L)	295	1.1 [0.8–1.6]	0.8 [0.6–1.1]	<0.01	1187	1.0 [0.7–1.3]	0.9 [0.6–1.3]	<0.01
Platelets (10^9/L)	307	215 [173–279]	179 [135–244]	<0.01	1188	214 [165–269]	205 [153–272]	0.15

https://doi.org/10.1371/journal.pone.0255748.t003

KCH and OUH; except for ischemic heart disease, temperature and platelets which were associated with severe disease at OUH, but not KCH.

Performance of the prediction models

The validation of the four prediction models with both the OUH and KCH cohorts is presented in terms of discrimination (AUROC) and calibration (slope and intercept) in Table 2 and Figs 2 and 3, respectively. For the models predicting mortality, the Xie model had the highest AUROC both in the KCH cohort (0.79; 95% CI 0.76–0.82) and the OUH cohort (0.87; 95% CI 0.79–0.95). The Zhang1 model had a lower AUROC at both KHC (0.64; 95% CI 0.60–0.68) and OUH (0.72; 95% CI 0.62–0.82).

For 'severe disease', discrimination was highest in the Allenbach model with AUROCs 0.72 (95% CI 0.68–0.75) for KCH and 0.81 (95% CI 0.74–0.88) for OUH. For the Zhang2 model, the AUROC was 0.67 (95% CI 0.64–0.70) for KCH and 0.77 (95% CI 0.70–0.84) for OUH. For the Xie and Allenbach models, discrimination at OUH was similar to the development cohorts (Fig 2). We compared the AUROCs between the KCH and OUH cohorts using the bootstrap method implemented in the pROC R package [24]. The results indicated that there was a statistically significant difference in the AUROCs between KCH and OUH for the Xie model (p = 0.01), Allenbach model (p = 0.009), and the Zhang2 model (p = 0.007), but not for the Zhang1 model (p = 0.140).

The calibration plots are shown in Fig 3 (after recalibration). S3 Fig shows the calibration results before and after recalibration for the Xie and Allenbach models. Recalibration will not render models with poor discrimination more useful. Thus, we focused on the recalibration of the Xie and Allenbach models as these had the best discrimination. Recalibration improved the predictions for both the Xie and Allenbach models at OUH and the Xie model at KCH, and the slope and intercept were acceptable for both models at both hospitals after recalibration.

Continuous variables in median [IQR] and categorical variables in number (percent). P-values are calculated with the Pearson X^2 test for categorical variables, and with the Wilcoxon

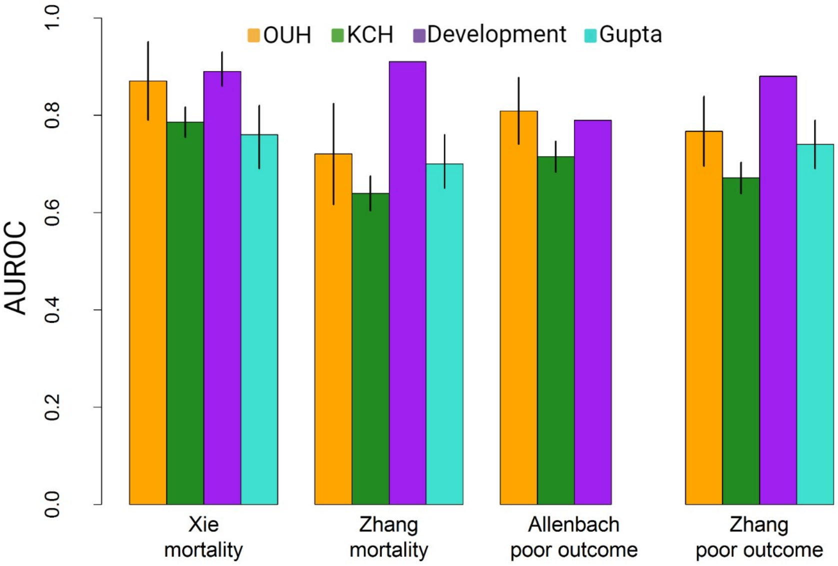

Fig 2. Discrimination of the models at the different sites. AUROCs from validation of the four models at the KCH and OUH cohorts, and the original AUROC from development cohorts [21–23]. Also shown are the results from the external validation of the Xie and Zhang models by Gupta et al. [6]. Lines represent the 95% CIs of the AUROCs. For the development cohorts only Xie reported confidence intervals.

https://doi.org/10.1371/journal.pone.0255748.g002

rank-sum test for continuous variables. SaO_2 value under oxygen treatment was registered if oxygen was applied, there are also values for patients without oxygen in this registration. ICU; intensive care unit, ACE; angiotensin converting enzyme, NEWS; National early warning score, CRP; C-reactive protein. LDH; Lactate dehydrogenase, eGFR; estimated glomerular filtration rate.

Discussion

In this study, we validated four prediction models for prognosis in hospitalized COVID-19 patients from London, UK and Oslo, Norway. We found varying performance of the models in the two cohorts. The models performed better in the OUH cohort with similar discrimination to the original studies. The Xie and Allenbach models had the best performance for prediction of death and severe disease, respectively.

Initial calibration was poor for all models, but improved after recalibration of the intercept according to the frequency of the outcome in our cohorts. This improves the accuracy of the prediction for each patient without affecting the discrimination and is recommended in several publications [5, 11, 19]. Local or possibly regional/national recalibration is likely to be important for COVID-19 prediction models since there is a large variation in the frequency of severe disease and death in different studies. However, ideally the local recalibration should also be tested for optimism using methods for internal validation such as bootstrapping or cross validation.

In some cases, we found poorer discrimination in the validation cohorts compared to the development cohorts. This is consistent with past evidence showing discrimination in development cohorts to be better than at external validation due to overfitting and differences in characteristics of the cohorts [25]. The cohorts in the original studies and at KCH and OUH had many differences such as mortality, age and frequencies of severe disease and comorbidities. UK and Norway differ in the structures of their healthcare systems, and the incidence of

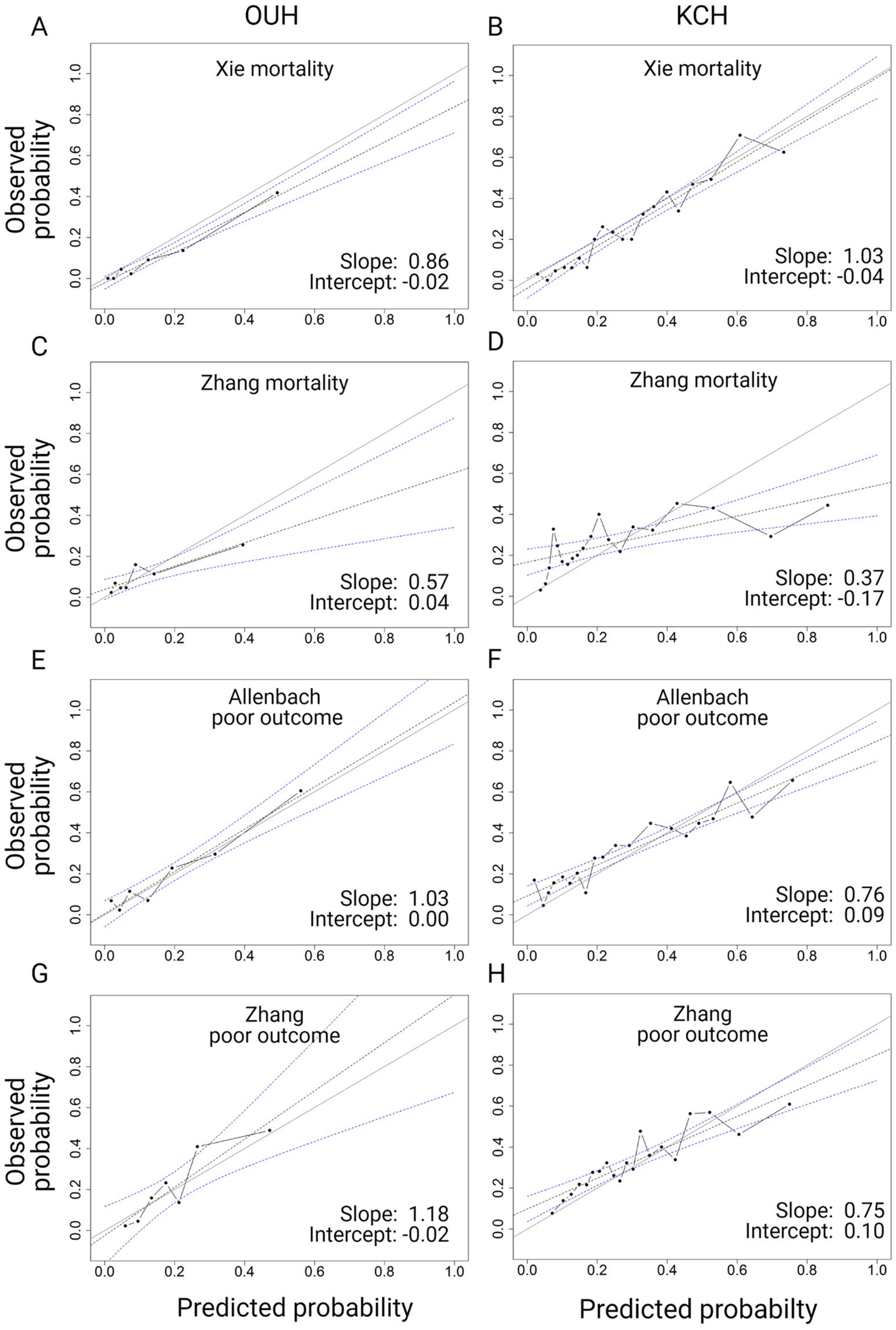

Fig 3. Calibration plots for OUH and KCH after recalibration.

https://doi.org/10.1371/journal.pone.0255748.g003

COVID-19 has been far higher in the UK. These factors may have affected the selection of patients for hospital and ICU admission, which might have resulted in a more homogenous patient population in regards to severity at KCH. It is to be expected that discrimination will be less good when the population is more homogenous.

The findings underline the importance of validation at several external sites. This is particularly true for a new disease like COVID-19, with rapidly developing treatment guidelines, and with an overwhelming effect on healthcare resources in some locations, but not at others.

The Xie model had the best results compared to the other models. The differences in the performance of the prediction models might have several reasons. Firstly, the predictors used in one model might have better predictive value than predictors used in others. SaO_2, which is included in the Xie model, is a strong clinical indicator of the severity of disease, and often indicates a need for ICU transfer. Secondly, there might be weaknesses in the models, as bias is common in prediction models [12]. To date, only the Allenbach study is published in a peer-reviewed journal, while Xie and Zhang are preprints. Thirdly, criteria for ICU admittance might vary across sites. The fact that we and other studies generally find better discrimination for mortality than for severe disease (often defined by ICU admittance) supports this hypothesis. For instance, patients with short life expectancy will often not be admitted to the ICU, but given oxygen therapy in a hospital ward and transferred to nursing homes for palliative care. These patients, not fulfilling the criteria for severe disease, often have predictors that indicate severe disease at admission.

Many prediction models have been published, but few have been systematically validated [26]. To our knowledge, only one study to date has validated COVID-19 prediction models; Gupta et al recently validated 22 prognostic models [6], including the Xie and Xhang models. For the OUH cohort, we found substantially better discrimination for the Xie and Allenbach models for the prediction of mortality and severe disease, respectively. The performance of the models at KCH was more similar to the results in the Gupta study, also performed at a London hospital. The rate of severe disease, mortality and the characteristics of the London cohorts are quite similar which might explain the similar performance at these two sites.

Several other prediction models have been recently published, such as models based on NEWS2 or the ISARIC model [14, 27]. The AUROCs of the models are in the range of 0.75 to 0.80, which is not a substantial improvement over single univariate predictors of severity. Thus, the finding that the Xie and Allenbach models perform well at both the original study site and at our validation cohort at OUH might indicate that it is possible to achieve higher AUROCs with relatively simple prediction models.

Our study has several strengths. Validation was performed at two sites in different countries with consistent inclusion and exclusion criteria. We included all eligible patients admitted to the hospital during the study period therefore the cohorts should be representative of the study sites. Moreover, the study was conducted and reported according to the TRIPOD guidelines. However, there are also some weaknesses. Firstly, the OUH cohort is not very large with relatively few patients meeting the outcomes. Some publications recommend including at least 100 patients with the relevant outcome [10]. However, studies with lower numbers may still contain useful information. Furthermore, the KCH cohort is probably one of the largest cohorts analyzed in prediction models for severe COVID-19 in hospital. Secondly, Gupta et al. included 22 models in their validation study, while we ended our inclusion of models in May 2020, and included only four models in this study. Whereas it could be interesting to include more models we think that the results for the Xie and Allenbach models at OUH indicate that further studies of these models could be interesting. Thirdly, there was a relatively high number of missing values for LDH and SpO_2 at KCH. It is uncertain how much this affected the results. Both are included in the Xie model and SpO_2 is a strong predictor for mortality, while

LDH is probably a weaker predictor (6). The number of missing values at OUH was low and probably did not affect the validation.

In conclusion, following the TRIPOD guidelines, our study validated developed models for prediction of prognosis in COVID-19, and showed that these models have a variable performance in different cohorts. The Xie model and Allenbach model clearly had the best performance, and we suggest that these models should be included in future studies of COVID-19 prediction models. However, the performance of these models at our two validation sites was not similar, which underlines the importance of external validation of prediction models at several study sites before their implementation in the clinical practice.

Supporting information

S1 Table. Patient characteristics of the development/validation cohorts.
(DOCX)

S2 Table. Different methods of imputation.
(DOCX)

S1 Fig. Flowchart of the included patients at OUH for validation of the four prediction models.
(JPEG)

S2 Fig. Flowchart of included patientes in the KCH cohort for validation of the four predcition models.
(JPEG)

S3 Fig. Results for calibration of the Xie and Allenbach models at KCH and OUH before and after recalibration.
(JPEG)

S1 Checklist. TRIPOD checklist.
(DOCX)

Acknowledgments

We would like to thank Prof. Anne Ma Dyrhol Riise and Dr. Ane M. Andersson at the Department of Infectious Diseases, OUH, for their support with the quality registry "COVID19 OUS".

Author Contributions

Conceptualization: Kristin E. Wickstrøm, Valeria Vitelli, Aleksander R. Holten, Kristian Tonby, Alvaro Köhn-Luque, Erik K. Amundsen.

Data curation: Kristin E. Wickstrøm, Valeria Vitelli, Ewan Carr, Rebecca Bendayan, Andrew H. Reiner, Daniel Bean, Anthony Shek, Zeljko Kraljevic, Kristian Tonby, Alvaro Köhn-Luque, Erik K. Amundsen.

Formal analysis: Kristin E. Wickstrøm, Valeria Vitelli, Ewan Carr, Andrew H. Reiner, Alvaro Köhn-Luque.

Investigation: Kristin E. Wickstrøm.

Methodology: Kristin E. Wickstrøm, Valeria Vitelli, Kristian Tonby, Alvaro Köhn-Luque, Erik K. Amundsen.

Project administration: Kristian Tonby, Erik K. Amundsen.

Resources: James Teo, Richard Dobson, Erik K. Amundsen.

Software: Valeria Vitelli, Ewan Carr, Andrew H. Reiner, Daniel Bean, Anthony Shek, Zeljko Kraljevic, Alvaro Köhn-Luque.

Supervision: James Teo, Richard Dobson, Erik K. Amundsen.

Validation: Kristin E. Wickstrøm, Valeria Vitelli, Ewan Carr, Andrew H. Reiner, Alvaro Köhn-Luque.

Visualization: Kristin E. Wickstrøm, Valeria Vitelli, Andrew H. Reiner, Alvaro Köhn-Luque, Erik K. Amundsen.

Writing – original draft: Kristin E. Wickstrøm, Erik K. Amundsen.

Writing – review & editing: Valeria Vitelli, Ewan Carr, Aleksander R. Holten, Rebecca Bendayan, Andrew H. Reiner, Daniel Bean, Tom Searle, Anthony Shek, Zeljko Kraljevic, James Teo, Richard Dobson, Kristian Tonby, Alvaro Köhn-Luque.

References

1. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020; 395(10229):1054–62. https://doi.org/10.1016/S0140-6736(20)30566-3 PMID: 32171076.
2. Weiss P, Murdoch DR. Clinical course and mortality risk of severe COVID-19. Lancet. 2020; 395 (10229):1014–5. https://doi.org/10.1016/S0140-6736(20)30633-4 PMID: 32197108.
3. Henry BM, de Oliveira MHS, Benoit S, Plebani M, Lippi G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. Clinical chemistry and laboratory medicine: CCLM / FESCC. 2020. https://doi.org/10.1515/cclm-2020-0369 PMID: 32286245.
4. Zeng F, Li L, Zeng J, Deng Y, Huang H, Chen B, et al. Can we predict the severity of COVID-19 with a routine blood test? Polish archives of internal medicine. 2020. Epub 2020/05/02. https://doi.org/10.20452/pamw.15331 PMID: 32356642.
5. Martin GP, Sperrin M, Sotgiu G. Performance of prediction models for COVID-19: the Caudine Forks of the external validation. The European respiratory journal. 2020; 56(6). https://doi.org/10.1183/13993003.03728-2020 PMID: 33060155
6. Gupta RK, Marks M, Samuels THA, Luintel A, Rampling T, Chowdhury H, et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. The European respiratory journal. 2020; 56(6). https://doi.org/10.1183/13993003.03498-2020 PMID: 32978307
7. 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; 369:m1328. https://doi.org/10.1136/bmj.m1328 PMID: 32265220
8. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. Journal of clinical epidemiology. 2003; 56(5):441–7. https://doi.org/10.1016/s0895-4356(03)00047-7 PMID: 12812818.
9. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ (Clinical research ed). 2009; 338:b605. Epub 2009/05/30. https://doi.org/10.1136/bmj.b605 PMID: 19477892.
10. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine. 2016; 35(2):214–26. Epub 2015/11/11. https://doi.org/10.1002/sim.6787 PMID: 26553135.
11. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015; 162(1):W1–W73. https://doi.org/10.7326/M14-0698 PMID: 25560730

12. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine. 2019; 170(1):51–8. https://doi.org/10.7326/M18-1376 PMID: 30596875

13. Jackson R, Kartoglu I, Stringer C, Gorrell G, Roberts A, Song X, et al. CogStack—Experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Medical Informatics and Decision Making. 2018; 18. https://doi.org/10.1186/s12911-018-0623-9 PMID: 29941004

14. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, et al. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. BMC Medicine. 2021; 19(1):23. https://doi.org/10.1186/s12916-020-01893-3 PMID: 33472631

15. Zakeri R, Pickles A, Carr E, Bean DM, O'Gallagher K, Kraljewic Z, et al. Biological responses to COVID-19: Insights from physiological and blood biomarker profiles. Curr Res Transl Med. 2021; 69(2):103276. https://doi.org/10.1016/j.retram.2021.103276 PMID: 33588321.

16. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006; 59(10):1087–91. Epub 2006/09/19. https://doi.org/10.1016/j.jclinepi.2006.01.014 PMID: 16980149.

17. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. 2011; 30(4):377–99. https://doi.org/10.1002/sim.4067 PMID: 21225900

18. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011; 45(3).

19. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008; 61(1):76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018 PMID: 18083464.

20. team Rc. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. https://www.R-project.org/.

21. Xie J, Hungerford D, Chen H, Abrams ST, Li S, Wang G, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. medRxiv. 2020:2020.03.28.20045997. https://doi.org/10.1101/2020.03.28.20045997

22. Zhang H, Shi T, Wu X, Zhang X, Wang K, Bean D, et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. medRxiv. 2020:2020.04.28.20082222. https://doi.org/10.1101/2020.04.28.20082222

23. Allenbach Y, Saadoun D, Maalouf G, Vieira M, Hellio A, Boddaert J, et al. Multivariable prediction model of intensive care unit transfer and death: a French prospective cohort study of COVID-19 patients. medRxiv. 2020:2020.05.04.20090118. https://doi.org/10.1371/journal.pone.0240711 PMID: 33075088

24. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. https://doi.org/10.1186/1471-2105-12-77 PMID: 21414208.

25. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of clinical epidemiology. 2015; 68 (1):25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007 PMID: 25441703.

26. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021; 14(1):49–58. https://doi.org/10.1093/ckj/sfaa188 PMID: 33564405.

27. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, 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. https://doi.org/10.1136/bmj.m3339 PMID: 32907855.