Carr et al. BMC Medicine (2021) 19:23
https://doi.org/10.1186/s12916-020-01893-3

BMC Medicine

RESEARCH ARTICLE Open Access

Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study

Ewan Carr[1*†], Rebecca Bendayan[1,2†], Daniel Bean[1,3], Matt Stammers[4,5,6], Wenjuan Wang[7], Huayu Zhang[8], Thomas Searle[1,2], Zeljko Kraljevic[1], Anthony Shek[9], Hang T. T. Phan[4,5], Walter Muruet[7], Rishi K. Gupta[10], Anthony J. Shinton[6], Mike Wyatt[11], Ting Shi[8], Xin Zhang[12], Andrew Pickles[1,2], Daniel Stahl[1], Rosita Zakeri[13,14], Mahdad Noursadeghi[15], Kevin O'Gallagher[13,14], Matt Rogers[11], Amos Folarin[1,3,16,17], Andreas Karwath[18,19,20], Kristin E. Wickstrøm[21], Alvaro Köhn-Luque[22], Luke Slater[18,19,20], Victor Roth Cardoso[18,19,20], Christopher Bourdeaux[11], Aleksander Rygh Holten[23], Simon Ball[20,24], Chris McWilliams[25], Lukasz Roguski[3,16,19], Florina Borca[4,5,6], James Batchelor[4], Erik Koldberg Amundsen[21], Xiaodong Wu[26,27], Georgios V. Gkoutos[18,19,20,24], Jiaxing Sun[26], Ashwin Pinto[6], Bruce Guthrie[8], Cormac Breen[7], Abdel Douiri[7], Honghan Wu[3,16], Vasa Curcin[7], James T. Teo[9,13†], Ajay M. Shah[13,14†] and Richard J. B. Dobson[1,2,3,16,17†]

Abstract

Background: The National Early Warning Score (NEWS2) is currently recommended in the UK for the risk stratification of COVID-19 patients, but little is known about its ability to detect severe cases. We aimed to evaluate NEWS2 for the prediction of severe COVID-19 outcome and identify and validate a set of blood and physiological parameters routinely collected at hospital admission to improve upon the use of NEWS2 alone for medium-term risk stratification.

Methods: Training cohorts comprised 1276 patients admitted to King's College Hospital National Health Service (NHS) Foundation Trust with COVID-19 disease from 1 March to 30 April 2020. External validation cohorts included 6237 patients from five UK NHS Trusts (Guy's and St Thomas' Hospitals, University Hospitals Southampton, University Hospitals Bristol and Weston NHS Foundation Trust, University College London Hospitals, University Hospitals Birmingham), one hospital in Norway (Oslo University Hospital), and two hospitals in Wuhan, China (Wuhan Sixth Hospital and Taikang Tongji Hospital). The outcome was severe COVID-19 disease (transfer to intensive care unit (ICU) or death) at 14 days after hospital admission. Age, physiological measures, blood biomarkers, sex, ethnicity, and comorbidities (hypertension, diabetes, cardiovascular, respiratory and kidney diseases) measured at hospital admission were considered in the models.

(Continued on next page)

* Correspondence: ewan.carr@kcl.ac.uk
†James T Teo, Ajay M Shah, and Richard J B Dobson are joint last authors.
†Ewan Carr and Rebecca Bendayan are joint first authors.
[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, 16 De Crespigny Park, London SE5 8AF, UK
Full list of author information is available at the end of the article

(Continued from previous page)

Results: A baseline model of 'NEWS2 + age' had poor-to-moderate discrimination for severe COVID-19 infection at 14 days (area under receiver operating characteristic curve (AUC) in training cohort = 0.700, 95% confidence interval (CI) 0.680, 0.722; Brier score = 0.192, 95% CI 0.186, 0.197). A supplemented model adding eight routinely collected blood and physiological parameters (supplemental oxygen flow rate, urea, age, oxygen saturation, C-reactive protein, estimated glomerular filtration rate, neutrophil count, neutrophil/lymphocyte ratio) improved discrimination (AUC = 0.735; 95% CI 0.715, 0.757), and these improvements were replicated across seven UK and non-UK sites. However, there was evidence of miscalibration with the model tending to underestimate risks in most sites.

Conclusions: NEWS2 score had poor-to-moderate discrimination for medium-term COVID-19 outcome which raises questions about its use as a screening tool at hospital admission. Risk stratification was improved by including readily available blood and physiological parameters measured at hospital admission, but there was evidence of miscalibration in external sites. This highlights the need for a better understanding of the use of early warning scores for COVID.

Keywords: NEWS2 score, Blood parameters, COVID-19, Prediction model

Key messages

- The National Early Warning Score (NEWS2), currently recommended for stratification of severe COVID-19 disease in the UK, showed poor-to-moderate discrimination for medium-term outcomes (14-day transfer to intensive care unit (ICU) or death) amongst COVID-19 patients.
- Risk stratification was improved by the addition of routinely measured blood and physiological parameters routinely at hospital admission (supplemental oxygen, urea, oxygen saturation, C-reactive protein, estimated glomerular filtration rate, neutrophil count, neutrophil/lymphocyte ratio) which provided moderate improvements in a risk stratification model for 14-day ICU/death.
- This improvement over NEWS2 alone was maintained across multiple hospital trusts, but the model tended to be miscalibrated with risks of severe outcomes underestimated in most sites.
- We benefited from existing pipelines for informatics at King's College Hospital such as CogStack that allowed rapid extraction and processing of electronic health records. This methodological approach provided rapid insights and allowed us to overcome the complications associated with slow data centralisation approaches.

Background

As of 9 December 2020, there have been > 67 million confirmed cases of COVID-19 disease worldwide [1]. While approximately 80% of infected individuals have mild or no symptoms [2], some develop severe COVID-19 disease requiring hospital admission. Within the subset of those requiring hospitalisation, early identification of those who deteriorate and require transfer to an intensive care unit (ICU) for organ support or may die is vital.

Currently, available risk scores for deterioration of acutely ill patients include (i) widely used generic ward-based risk indices such as the National Early Warning Score (NEWS2, [3]), (ii) the Modified Sequential Organ Failure Assessment (mSOFA) [4] and Quick Sequential Organ Failure Assessment [5] scoring systems, and (iii) the pneumonia-specific risk index, CURB-65 [6] which combines physiological observations with limited blood markers and comorbidities. NEWS2 is a summary score of six physiological parameters or 'vital signs' (respiratory rate, oxygen saturation, systolic blood pressure, heart rate, level of consciousness, temperature and supplemental oxygen dependency) used to identify patients at risk of early clinical deterioration in the United Kingdom (UK) National Health Service (NHS) hospitals [7, 8] and primary care. Some components (in particular, patient temperature, oxygen saturation, and supplemental oxygen dependency) have been associated with COVID-19 outcomes [2], but little is known about their predictive value for COVID-19 disease severity in hospitalised patients [9]. Additionally, a number of COVID-19-specific risk indices are being developed [10, 11] as well as unvalidated online calculators [12], but generalisability is unknown [13]. A Chinese study has suggested a modified version of NEWS2 with the addition of age only [14] but without any data on performance. With near-universal usage of NEWS2 in UK NHS Trusts since March 2019 [15], a minor adaptation to NEWS2 would be relatively easy to implement.

As the SARS-Cov2 pandemic has progressed, a number of risk prediction models to support clinical decisions, triage, and care in hospitalised patients have been proposed [13] incorporating potentially useful blood biomarkers [2, 16–19]. These include neutrophilia and lymphopenia, particularly in older adults [11, 18, 20, 21];

neutrophil-to-lymphocyte ratio [22]; C-reactive protein (CRP) [13]; lymphocyte-to-CRP ratio [22]; markers of liver and cardiac injury such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), and cardiac troponin [23]; and elevated D-dimers, ferritin and fibrinogen [2, 6, 8].

Our aim is to evaluate the NEWS2 score and identify which clinical and blood biomarkers routinely measured at hospital admission can improve medium-term risk stratification of severe COVID-19 outcome at 14 days from hospital admission. Our specific objectives were as follows:

1. To explore independent associations of routinely measured physiological and blood parameters (including NEWS2 parameters) at hospital admission with disease severity (ICU admission or death at 14 days from hospital admission), adjusting for demographics and comorbidities
2. To develop a prediction model for severe COVID-19 outcomes at 14 days combining multiple blood and physiological parameters
3. To compare the discrimination, calibration, and clinical utility of the resulting model with NEWS2 score and age alone using (i) internal validation and (ii) external validation at seven UK and international sites

A recent systematic review found that most existing prediction models for COVID-19 had a high risk of bias due to non-representative samples, model overfitting, or poor reporting [13]. The analyses presented here build upon our earlier work [24] which suggested that adding age and common blood biomarkers to the NEWS2 score could improve risk stratification in patients hospitalised with COVID-19. While incorporating external validation, this preliminary work was limited in that the training sample comprised 439 patients (the cohort available at the time of model development). In the present study, we (i) expand the cohort used for model development to all 1276 patients at King's College Hospital (KCH), (ii) use hospital admission (rather than symptom onset) as the index date, (iii) consider shorter-term outcomes (3-day ICU/death), (iv) improve the reporting of model calibration and clinical utility, and (v) increase the number of external sites from three to seven.

Methods

Study cohorts

The KCH training cohort ($n = 1276$) was defined as all adult inpatients testing positive for severe acute respiratory syndrome coronavirus 2 (SARS-Cov2) by reverse transcription polymerase chain reaction (RT-PCR) between 1 March and 31 April 2020 at two acute hospitals (King's College Hospital and Princess Royal University Hospital) in South East London (UK) of Kings College Hospital NHS Foundation Trust (KCH). All patients included in the study had symptoms consistent with COVID-19 (e.g. cough, fever, dyspnoea, myalgia, delirium, diarrhoea). For external validation purposes, we used seven cohorts:

1. Guy's and St Thomas' Hospital NHS Foundation Trust (GSTT) of 988 cases (3 March 2020 to 26 August 2020)
2. University Hospitals Southampton NHS Foundation Trust (UHS) of 633 cases (7 March to 6 June 2020)
3. University Hospitals Bristol and Weston NHS Foundation Trust (UHBW) of 190 cases (12 March to 11 June 2020)
4. University College Hospital London (UCH) of 411 cases (1 February to 30 April 2020)
5. University Hospitals Birmingham (UHB) of 1037 cases (1 March to 31 June 2020)
6. Oslo University Hospital (OUH) of 163 cases (6 March to 13 June 2020)
7. Wuhan Sixth Hospital and Taikang Tongji Hospital of 2815 cases (4 February 2020 to 30 March 2020)

Data were extracted from structured and/or unstructured components of electronic health records (EHR) in each site as detailed below.

Measures

Outcome

For all sites, the outcome was severe COVID-19 disease at 14 days following hospital admission, categorised as transfer to the ICU/death (WHO-COVID-19 Outcomes Scales 6–8) vs. not transferred to the ICU/death (scales 3–5) [25]. For nosocomial patients (patients with symptom onset after hospital admission), the endpoint was defined as 14 days after symptom onset. Dates of hospital admission, symptom onset, ICU transfer, and death were extracted from electronic health records or ascertained manually by a clinician.

Blood and physiological parameters

We included blood and physiological parameters that were routinely obtained at hospital admission and which are routinely available in a wide range of national and international hospital and community settings. Measures available for fewer than 30% of patients were not considered (including Troponin-T, Ferritin, D-dimers and glycated haemoglobin (HbA1c), Glasgow Coma Scale score). We excluded creatinine since this parameter correlates highly ($r > 0.8$) with, and is used in the derivation of, estimated glomerular filtration rate. We excluded

white blood cell count (WBCs) which is highly correlated with neutrophil and lymphocyte counts.

The candidate blood parameters therefore comprised albumin (g/L), C-reactive protein (CRP; mg/L), estimated glomerular filtration rate (GFR; mL/min), haemoglobin (g/L), lymphocyte count ($\times 10^9$/L), neutrophil count ($\times 10^9$/L), platelet count (PLT; $\times 10^9$/L), neutrophil-to-lymphocyte ratio (NLR), lymphocyte-to-CRP ratio [22], and urea (mmol/L). The candidate physiological parameters included the NEWS2 total score, as well as the following parameters: respiratory rate (breaths per minute), oxygen saturation (%), supplemental oxygen flow rate (L/min), diastolic blood pressure (mmHg), systolic blood pressure (mmHg), heart rate (beats/min), and temperature (°C). For all parameters, we used the first available measure up to 48 h following hospital admission.

Demographics and comorbidities

Age, sex, ethnicity and comorbidities were considered. Self-defined ethnicity was categorised as White vs. non-White (Black, Asian, or other minority ethnic) and patients with ethnicity recorded as 'unknown/mixed/other' were excluded ($n = 316$; 25%). Binary variables were derived for comorbidities: hypertension, diabetes, heart disease (heart failure and ischemic heart disease), respiratory disease (asthma and chronic obstructive pulmonary disease (COPD)), and chronic kidney disease.

Data processing

King's College Hospital

Data were extracted from the structured and unstructured components of the electronic health record (EHR) using natural language processing (NLP) tools belonging to the CogStack ecosystem [26], namely MedCAT [27] and MedCATTrainer [28]. The CogStack NLP pipeline captures negation, synonyms, and acronyms for medical Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) concepts as well as surrounding linguistic context using deep learning and long short-term memory networks. MedCAT produces unsupervised annotations for all SNOMED-CT concepts (Additional file 1: Table S1) under parent terms Clinical Finding, Disorder, Organism, and Event with disambiguation, pre-trained on MIMIC-III [29]. Starting from our previous model [30], further supervised training improved detection of annotations and meta-annotations such as experiencer (is the annotated concept experienced by the patient or other), negation (is the concept annotated negated or not), and temporality (is the concept annotated in the past or present) with MedCATTrainer. Meta-annotations for hypothetical, historical, and experiencer were merged into "Irrelevant" allowing us to exclude any mentions of a concept that did not directly relate to the patient currently. Performance of the NLP pipeline for comorbidities mentioned in the text was evaluated on 4343 annotations in 146 clinical documents by a clinician (JT). F1 scores, precision, and recall are presented in Additional file 2: Table S2.

Guy's and St Thomas' NHS Foundation Trust

Electronic health records from all patients admitted to Guy's and St Thomas' NHS Foundation Trust who had a positive COVID-19 test result between 3 March and 21 May 2020, inclusive, were identified. Data were extracted using structured queries from six complementary platforms and linked using unique patient identifiers. Data processing was performed using Python 3.7 [31]. The process and outputs were reviewed by a study clinician.

University Hospitals Southampton

Data were extracted from the structured components of the UHS CHARTS EHR system and data warehouse. Data were transformed into the required format for validation purposes using Python 3.7 [31]. Diagnosis and comorbidity data of interest were gathered from the International Statistical Classification of Diseases (ICD-10) coded data. No unstructured data extraction was required for validation purposes. The process and outputs were reviewed by an experienced clinician prior to analysis.

University Hospitals Bristol and Weston NHS Foundation Trust

Data were extracted from UHBW electronic health records system (Medway). ICD-10 codes were used for diagnosis and comorbidity data. Data were transformed in line with project specifications and exported for analysis in Python 3.7 [31].

University College Hospital London

Dates of hospital admission, symptom onset, ICU transfer, and death were extracted from electronic health records. The outcome (14-day ICU/death) was defined in UCLH as 'initiation of ventilatory support (continuous positive airway pressure, non-invasive ventilation, high-flow nasal cannula oxygen, invasive mechanical ventilation, or extracorporeal membrane oxygenation) or death' which is consistent WHO-COVID-19 Outcomes Scales 6–8.

Wuhan cohort

Demographic, premorbid conditions, clinical symptoms or signs at presentation, laboratory data, and 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.

University Hospitals Birmingham

Dates of hospital admission, symptom onset, ICU transfer, and death were extracted from electronic health records using the Prescribing Information and Communications System (PICS) system. The extracted data was transformed into the required format for validation purposes using Python 3.8 [31]. Diagnosis and comorbidity data of interest were gathered from ICD-10 coded data. The outcomes (3- and 14-day ICU/death) were defined consistent with WHO-COVID-19 Outcomes Scales 6–8.

Oslo University Hospital

All admitted patients with confirmed COVID-19 by positive SARS-CoV2 PCR were included in a quality registry. Data input into the register was manual. Register data was supplemented with test results from the laboratory information system (LIS) by matching exported Excel files from the register with exported Excel files from LIS. The fidelity of the match was checked against the original data source manually for a small number of patients. Only patients with symptoms consistent with COVID-19 were included in the study.

Statistical analyses

All continuous parameters were winsorized (at 1% and 99%) and scaled (mean = 0; standard deviation = 1) to facilitate interpretability and comparability [32]. Logarithmic or square root transformations were applied to skewed parameters. To explore independent associations of blood and physiological parameters with 14-day ICU/death (objective 1), we used logistic regression with Firth's bias reduction method [33]. Each parameter was tested independently, adjusted for age and sex (model 1), and then additionally adjusted for comorbidities (model 2). *P* values were adjusted using the Benjamini-Hochberg procedure to keep the false discovery rate (FDR) at 5% [34].

To evaluate NEWS2 and identify parameters that could improve prediction of severe COVID-19 outcomes (objectives 2 and 3), we used regularised logistic regression with a least absolute shrinkage and selection operator (LASSO) estimator that shrinks parameters according to their variance, reduces overfitting, and enables automatic variable selection [35]. The optimal degree of regularisation was determined by identifying a tuning parameter λ using cross-validation. To avoid overfitting and to reduce the number of false-positive predictors, λ was selected to give a model with an area under the receiver operating characteristic curve (AUC) one standard error below the 'best' model. To evaluate the predictive performance of our model on new cases of the same underlying population (internal validation), we performed nested cross-validation (10-folds the for inner loop; 10-folds/1000 repeats for the outer loop). Discrimination was assessed using AUC and Brier score. Missing feature information was imputed using *k*-nearest neighbour (kNN) imputation ($k = 5$). All steps (feature selection, winsorizing, scaling, and kNN imputation) were incorporated within the model development and selection process to avoid data leakage that would otherwise result in optimistic performance measures [36]. All analyses were conducted with Python 3.8 [31] using the statsmodels [37] and Scikit-Learn [38] packages.

We evaluated the transportability of the derived regularised logistic regression model in external validation samples from GSTT ($n = 988$), UHS ($n = 633$), UHBW ($n = 190$), UCH ($n = 411$), UHB ($n = 1037$), OUH ($n = 163$), and Wuhan ($n = 2815$). Validation used LASSO logistic regression models trained on the KCH training sample, with code and pre-trained models shared via GitHub.[1] Models were assessed in terms of discrimination (AUC, sensitivity, specificity, Brier score), calibration, and clinical utility (decision curve analysis, number needed to evaluate) [32, 39]. Moderate calibration was assessed by plotting model-predicted probabilities (*x*-axis) against observed proportions (*y*-axis) with locally estimated scatterplot smoothing (LOESS) and logistic curves [40]. Clinical utility was assessed using decision curve analysis where 'net benefit' was plotted against a range of threshold probabilities. Unlike diagnostic performance measures, decision curves incorporate preferences of the clinician and patient. The threshold probability (p_t) is where the expected benefit of treatment is equal to the expected benefit of avoiding treatment [41]. Net benefit was calculated by counting the number of true positives (predicted risk $> p_t$ and experienced severe COVID-19 outcome) and false positives (predicted risk $> p_t$ but did not experience severe COVID-19 outcome) and using the below formula:

$$\text{Net benefit} = \frac{\text{True positives}}{N} - \frac{\text{False positives}}{N} \times \frac{p_t}{1 - p_t}$$

Our model was developed as a screening tool, to identify at hospital admission patients at risk of more severe outcomes. The intended treatment for patients with a positive result from this model would be further examination by a clinician, who would make recommendations regarding appropriate treatment (e.g. earlier transfer to the ICU, intensive monitoring, treatment). We compared the decision curve from our model to two extreme cases of 'treat none' and 'treat all'. The 'treat none' (i.e. routine management) strategy implies that no patients would be

[1] https://github.com/ewancarr/NEWS2-COVID-19

selected for further examination by a clinician; the 'treat all' strategy (i.e. intensive management) implies that all patients would undergo further assessment. A model is clinically beneficial if the model-implied net benefit is greater than either the 'treat none' or 'treat all' strategies.

Since the intended strategy involves a further examination by a clinician, and is therefore low risk, our emphasis throughout is on avoiding false negatives (i.e. failing to detect a severe case) at the expense of false positives. We therefore used thresholds of 30% and 20% (for 14-day and 3-day outcomes, respectively) to calculate sensitivity and specificity. This gave a better balance of sensitivity vs. specificity and reflected the clinical preference to avoid false negatives for the proposed screening tool.

Sensitivity analyses

We conducted five sensitivity analyses. First, to explore the ability of NEWS2 to predict shorter-term severe COVID-19 outcome, we developed models for ICU transfer/death at 3 days following hospital admission. All steps described above were repeated, including training (feature selection) and external validation. Second, following recent studies suggesting sex differences in COVID-19 outcome [18], we tested interactions between each physiological and blood parameters and sex using likelihood-ratio tests. Third, we repeated all models with adjustment for ethnicity in the subset of individuals with available data for ethnicity ($n = 960$ in the KCH training sample). Fourth, to explore the differences between community-acquired vs. nosocomial infection, we repeated all models after excluding 153 nosocomial patients ($n = 1123$). Finally, we considered an alternative baseline model of 'NEWS2 only'. Our primary analyses used a baseline model of 'NEWS2 + age' because NEWS2 is rarely used in isolation for prognostication and treatment decisions will incorporate other patient characteristics such as age.

Results

Descriptive analyses

The KCH training cohort comprised 1276 patients admitted with a confirmed diagnosis of COVID-19 (from 1 March to 31 April 2020) of whom 389 (31%) were transferred to the ICU or died within 14 days of hospital admission, respectively. The validation cohorts comprised 6237 patients across seven sites. At UK NHS trusts, 30 to 42% of patients were transferred to the ICU or died within 14 days of admission. Disease severity was lower in the Wuhan sample, where 4% were transferred to the ICU or died. Table 1 presents the demographic and clinical characteristics of the training and validation cohorts. The UK sites were similar in terms of age and sex, with patients tending to be older (median age 59–74) and male (58 to 63%) but varied in the proportion of patients of non-White ethnicity (from 10% at UHS to 40% at KCH and UCH). Blood and physiological parameters were broadly consistent across UK sites.

Logistic regression models were used to assess independent associations between each variable and severe COVID-19 outcome (ICU transfer/death) in the KCH cohort. Additional file 3: Table S3 presents odds ratios adjusted for age and sex (model 1) and comorbidities (model 2), sorted by effect size. Increased odds of transfer to the ICU or death by 14 days were associated with NEWS2 score, oxygen flow rate, respiratory rate, CRP, neutrophil count, urea, neutrophil/lymphocyte ratio, heart rate, and temperature. Reduced odds of severe outcomes were associated with lymphocyte/CRP ratio, oxygen saturation, estimated GFR, and albumin.

Evaluating NEWS2 score for prediction of severe COVID-19 outcome

Logistic regression models were used to evaluate a baseline model containing hospital admission NEWS2 score and age for the prediction of severe COVID-19 outcomes at 14 days. Internally validated discrimination for the KCH training sample was moderate (AUC = 0.700; 95% confidence interval (CI) 0.680, 0.722; Brier score = 0.192; 0.186, 0.197; Table 2). Discrimination remained poor-to-moderate in UK validation sites (AUC = 0.623 to 0.729) but was moderate-to-good in Norway (AUC = 0.786) and Wuhan hospitals (AUC = 0.815) (Figs. 1 and 2). Calibration was inconsistent with risks underestimated in some sites (UHS, GSTT) and overestimated in others (UHBW, UHB; Fig. 2).

Supplementing NEWS2 with routinely collected blood and physiological parameters

We considered whether routine blood and physiological parameters could improve risk stratification for medium-term COVID-19 outcome (ICU transfer/death at 14 days). When adding demographic, blood, and physiological parameters to NEWS2, nine features were retained following LASSO regularisation, in order of effect size: NEWS2 score, supplemental oxygen flow rate, urea, age, oxygen saturation, CRP, estimated GFR, neutrophil count, and neutrophil/lymphocyte ratio. Notably, comorbid conditions were not retained when added in subsequent models, suggesting most of the variance explained was already captured by the included parameters. Internally validated discrimination in the KCH training sample was moderate (AUC = 0.735; 95% CI 0.715, 0.757) but improved compared to 'NEWS2 + age' (Table 2). This improvement over NEWS2 alone was replicated in validation samples (Fig. 1). The supplemented model continued to show evidence of substantial miscalibration.

Table 1 Patient characteristics of the training/validation cohorts

	Training cohort		Validation cohorts ($n = 6237$)													
	KCH ($n = 1276$)		UHS ($n = 633$)		UCH ($n = 411$)		GSTT ($n = 988$)		UHBW ($n = 190$)		Wuhan ($n = 2815$)		UHB ($n = 1037$)		Oslo ($n = 163$)	
COVID-19 WHO Score 6–8 (ICU/death)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)
3 days	1276	163 (12.8%)	633	109 (17.2%)	411	120 (29.0%)	988	289 (29.3%)	190	32 (16.8%)	2815	58 (2.1%)	1037	169 (16.3%)	163	27 (16.3%)
14 days	1276	389 (30.5%)	633	223 (35.2%)	411	171 (42.0%)	988	391 (39.6%)	190	56 (29.5%)	2815	118 (4.2%)	1037	310 (29.9%)	163	39 (23.9%)
Demographics	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)
Age (median [IQR])	1276	71.5 [57.1, 82.6]	633	73.0 [56.0, 84.0]	411	66.0 [53.0, 79.0]	988	59.0 [46.0,75.0]	190	73.5 [59.3, 82.0]	2815	60.0 [50.0, 68.0]	1037	70.0 [57.0, 82.0]	163	60.0 [48.0–74.0]
Sex (male)	1276	742 (58.2%)	633	364 (57.5%)	411	252 (61.0%)	988	581 (58.8%)	190	120 (63.1%)	2815	1437 (51.0%)	1037	573 (55.3%)	163	95 (58.3%)
Non-White ethnicity	960	379 (39.5%)	546	55 (10.0%)	390	156 (40.0%)	817	607 (74.3%)	190	46 (24.2%)	2815	2815 (100.0%)	892	306 (34.3%)	–	–
Comorbidities	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)	*N* avail.	*N* (%)
Hypertension	1276	695 (54.5%)	633	321 (50.7%)	411	172 (42.0%)	988	309 (31.3%)	190	117 (61.6%)	2815	821 (29.2%)	1037	637 [61.4%)	163	55 (33.7%)
Diabetes mellitus	1276	439 (34.4%)	633	163 (25.8%)	411	105 (26.0%)	988	286 (28.9%)	190	71 (37.4%)	2815	371 (13.2%)	1037	358 (34.5%)	163	27 (16.6%)
Heart failure	1276	117 (9.2%)	633	137 (21.6%)	410	–	988	52 (5.3%)	190	33 (17.4%)	2815	236 (8.4%)[2]	1037	178 (17.2%)	163	15 (9.2%)
Ischaemic heart diseases	1276	185 (14.5%)	633	152 (24.0%)	409	108 (26.0%)[1]	–	–	190	52 (27.4%)	–	–	1037	245 (23.6%)	163	21 (12.9%)
COPD	1276	141 (11.1%)	633	115 (18.2%)	409	27 (6.6%)	988	64 (6.5%)	190	41 (21.6%)	2815	17 (0.6%)	1037	152 (14.7%)	–	–
Asthma	1276	174 (13.6%)	633	112 (17.7%)	409	41 (10.0%)	988	85 (8.6%)	190	27 (14.2%)	–	–	1037	169 (16.3%)	–	–
Chronic kidney disease	1276	234 (18.3%)	633	111 (17.5%)	410	40 (9.8%)	988	110 (11.1%)	190	59 (31.1%)	2815	56 (2.0%)	1037	274 (26.4%)	163	9 (5.5%)
Blood biomarker	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]
Albumin (g/L)	1153	37.0 [33.0, 40.0]	501	32.0 [29.0, 36.0]	390	38.0 [35.0, 42.0]	863	36.0 [31.0, 40.0]	190	30.0 [27.0, 33.0]	2404	38.1 [35.1, 40.5]	993	32.0 [28.0, 35.0]	120	39.0 [36.0–43.0]
C-reactive protein (CRP, mg/L)	1240	80.0 [36.0, 141.6]	545	75.0 [25.0, 150.0]	403	97.0 [45.0, 179.0]	974	76.5 [25.0, 153.8]	190	77.0 [36.3, 138.3]	2393	2.3 [0.8, 9.0]	970	89.0 [33.2, 157.0]	163	48.0 [16.0–113.0]
Urea (mmol/L)	1221	7.1 [4.6, 11.7]	563	6.95 [4.8, 10.6]	375	6.0 [4.0, 9.4]	489	7.4 [4.6, 12.5]	–	–	–	–	1018	6.8 [4.6, 11.8]	154	5.5 [4.3–7.5]
Estimated GFR	1254	65.0 [41.0, 86.0]	377	62.0 [40.0, 81.0]	407	77.0 [54.0, 96.0]	965	74.0 [49.0, 100.0]	190	68.0 [43.3. 88.0]	2433	103.1 [88.2, 117.5]	757	54.0 [29.0, 72.0]	163	84.0 [57.0–99.0]
Haemoglobin (g/L)	1223	127.0 [112.0, 141.0]	561	128.0 [111.0, 143.0]	410	130.0 [112.0, 143.0]	987	125.0 [108.0, 139.0]	190	129.0 [110.0, 141.0]	2584	124.0 [113.0, 135.0]	1009	130.0 [113.0, 144.0]	163	139.0 [129.0–148.0]
Lymphocyte count ($\times 10^9$/L)	1221	0.9 [0.7, 1.3]	561	1.0 [0.7, 1.4]	410	0.9 [0.6, 1.4]	987	0.9 [0.6, 1.3]	190	0.9 [0.6, 1.2]	2584	1.5 [1.1, 1.9]	1011	0.9 [0.7, 1.4]	153	1.1 [0.8–1.5]
Neutrophil count ($\times 10^9$/L)	1220	5.4 [3.8, 7.7]	560	5.8 [4.2, 8.8]	410	5.9 [3.9, 8.2]	986	5.0 [3.5, 8.1]	190	5.2 [3.5, 7.4]	2584	3.5 [2.7, 4.7]	1011	5.5 [3.8, 8.2]	153	4.4 [3.0–7.1]
Neutrophil/lymphocyte ratio	1218	5.6 [3.4, 9.5]	559	5.8 [3.4, 10]	410	6.0 [4.0, 10.0]	986	5.6 [3.2, 10.1]	190	5.7 [3.6, 9.8]	2584	2.3 [1.7, 3.5]	1011	5.6 [3.2, 10.2]	153	4.3 [2.4–7.7]
Lymphocyte/CRP ratio	1196	1.2 [0.6, 3.2]	559	1.3 [0.5, 4.6]	402	1.0 [0.4, 2.4]		0.0 [0.0, 0.0]	190	1.1 [0.5, 2.7]	2362	0.7 [0.1, 2.0]	962	1.1 [0.5, 3.3]	–	–
Platelet count ($\times 10^9$/L)	1224	213.0 [161.8, 274.0]	560	231 [176.8, 303.5]	409	221.0 [169.0, 280.0]	986	209.0 [161.0, 275.8]	190	207.5 [150.3, 268.5]	2584	223.0 [179.8, 273.0]	1008	218.0 [165.0, 287.2]	163	205.0 [160.0–279.0]
Physiological parameters	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]	*N* avail.	Median [IQR]
NEWS2 Total Score	1262	2.0 [1.0, 4.0]	529	3.0 [2.0, 5.0]	404	5.0 [3.0, 7.0]	744	3.0 [1.0, 5.0]	190	3.0 [2.0, 5.0]	2804	1.0 [0.0, 3.0]	1019	4.0 [2.0, 7.0]	163	5.0 [3.0–7.0]

Table 1 Patient characteristics of the training/validation cohorts *(Continued)*

	Training cohort		Validation cohorts (*n* = 6237)											
	KCH (*n* = 1276)		UHS (*n* = 633)		UCH (*n* = 411)		GSTT (*n* = 988)		UHBW (*n* = 190)		Wuhan (*n* = 2815)		UHB (*n* = 1037)		Oslo (*n* = 163)	
Heart rate	1273	85.0 [75.0, 94.0]	560	90.5 [82.0, 102.0]	410	94.0 [81.0, 107.0]	752	85.0 [75.0, 95.0]	190	82.0 [71.0, 95.0]	2812	81.0 [76.9, 85.8]	1028	90.0 [79.0, 104.0]	160	89.5 [75.3–100.8]
Oxygen saturation	1273	96.0 [95.0, 98.0]	561	97.0 [96.0, 99.0]	410	96.0 [94.0, 98.0]	712	96.0 [95.0, 97.0]	190	95.0 [94.0, 96.0]	2797	97.8 [97.0, 98.2]	1029	96.0 [94.0, 98.0]	163	95.0 [92.0–97.0]
Oxygen flow rate (L/min)	1271	0.0 [0.0, 4.0]	260	3.0 [2.0, 8.0]	403	2.0 [0.0, 10.0]	978	0.0 [0.0, 0.0]	190	2.0 [0.0, 3.0]	–	–	1017	0.0 [0.0, 4.0]	125	1.0 [0.0–2.0]
Respiration rate	1273	19.0 [18.0, 21.0]	561	20.0 [19.0, 24.0]	410	24.0 [20.0, 28.0]	755	19.0 [18.0, 22.0]	190	20.0 [18.0, 21.0]	2811	20.0 [19.0, 21.0]	1020	20.0 [18.0, 25.0]	160	24.0 [20.0–28.0]
Systolic blood pressure	1273	125.0 [112.0, 139.0]	555	137.0 [123.0, 152.0]	411	131.0 [115.0, 143.0]	751	125.0 [115.0, 140.0]	190	123.0 [111.0, 140.8]	1431	120.0 [110.0, 128.0]	1022	128.0 [113.0, 144.0]	160	129 [116.0–142.8]
Diastolic blood pressure	1273	71.0 [62.0, 80.0]	555	78.0 [70.0, 85.0]	411	73.0 [64.0, 81.0]	751	74.0 [66.0, 81.0]	190	72.0 [64.3, 82.0]	1433	71.0 [65.0, 78.0]	1022	75.0 [67.0, 84.0]	160	77.0 [69.0–87.0]
Temperature	1273	36.9 [36.6, 37.4]	558	36.9 [36.7, 37.5]	410	37.3 [36.8, 38.1]	750	36.9 [36.4, 37.5]	190	37.2 [36.7, 37.9]	2815	36.5 [36.3, 36.7]	1029	36.8 [36.2, 37.5]	162	37.2 [36.5–38.3]

[1]Measured as 'cardiovascular disease' at UCH because separate measures of 'heart failure' and 'ischaemic heart diseases' were unavailable
[2]Measured as overall 'heart disease' in the Wuhan cohort

Table 2 KCH internally validated predictive performance ($n = 1276$) based on nested repeated cross-validation

		NEWS2 + age, mean (95% CI)	All features, mean (95% CI)
14-day ICU/death	AUC	0.700 [0.680, 0.722]	0.735 [0.715, 0.757]
	Brier score	0.192 [0.186, 0.197]	0.183 [0.177, 0.189]
	Sensitivity[1]	0.778 [0.747, 0.815]	0.735 [0.702, 0.772]
	Specificity[1]	0.478 [0.445, 0.509]	0.592 [0.562, 0.621]

[1]Calculated at 30% probability threshold. AUC based on repeated, nested cross-validation (inner loop, 10-folds; outer loop = 10-folds/1000 repeats). Missing values imputed at each outer loop with *k*-nearest neighbour (kNN) imputation

Sensitivity analyses

For the 3-day endpoint, 13% of patients at KCH ($n = 163$) and between 16 and 29% of patients in the UK and Norway were transferred to the ICU or died (Table 1). The 3-day model retained just two parameters following regularisation: NEWS2 score and supplemental oxygen flow rate. For the baseline model ('NEWS2 + age'), discrimination was moderate at internal validation (AUC = 0.764; 95% CI 0.737, 0.794; Additional file 4: Table S4) and external validation (AUC = 0.673 to 0.755), but calibration remained poor (Additional file 5: Figure S1). Moreover, the supplemented model ('NEWS2 + oxygen flow rate') showed smaller improvements in discrimination compared to those seen at 14 days. For the KCH training cohort, internally validated AUC increased by 0.025: from 0.764 (95% CI 0.737, 0.794) for 'NEWS2 + age' to 0.789 (0.763, 0.819) for the supplemented model ('NEWS2 + oxygen flow rate'). At external validation, improvements were modest (UHBW, OUH) or negative (GSTT) in some sites, but more substantial in others (UHS, UCH). Moreover, model calibration was considerably worse for the supplemented 3-day model (Additional file 5: Figure S1).

We found no evidence of difference by sex (results not shown) and the findings were consistent when additionally adjusting for ethnicity in the subset of individuals with ethnicity data and when excluding nosocomial patients (Additional file 6: Table S5). Discrimination for the alternative baseline model of 'NEWS2 only' (Additional file 7: Table S6) showed a similar pattern of results as those for 'NEWS2 + age', except that improvements in discrimination for the supplemented model ('All features') were larger in most sites.

Decision curve analysis

Decision curve analysis for the 14-day endpoint is presented in Fig. 3. At KCH, the baseline model ('NEWS2 + age') offered small increments in net benefit compared to the 'treat all' and 'treat none' strategies for risk thresholds in the range 25 to 60%. This was replicated in all validation cohorts except for UHBW and OUH where the net benefit for 'NEWS2 + age' was lower than the 'treat none' strategy beyond the 40% risk threshold. The supplemented model ('All features') improved upon 'NEWS2 + age' and the two default strategies in most sites across the range 20 to 80%, except for (i) UHBW, where 'treat none' was superior beyond thresholds of

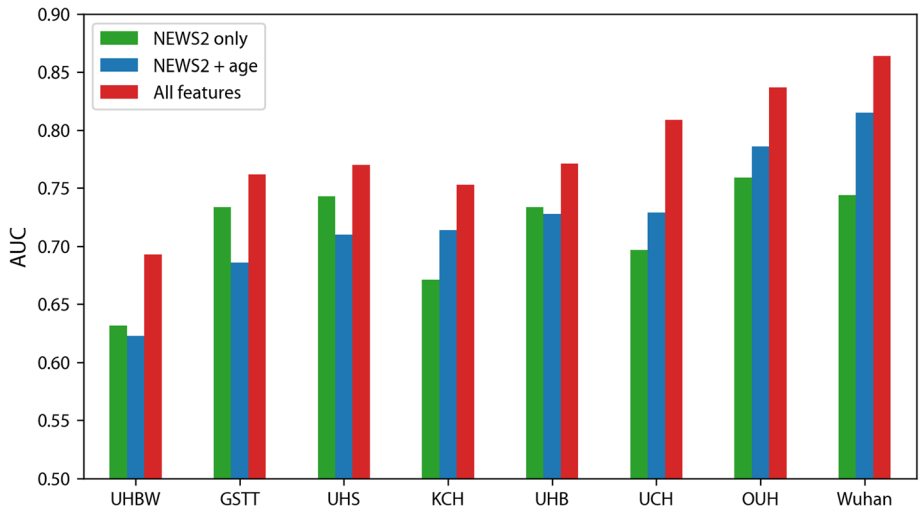

Fig. 1 Improvement in the area under the curve (AUC) for supplemented NEWS2 model for 14-day ICU/death at training and validation sites

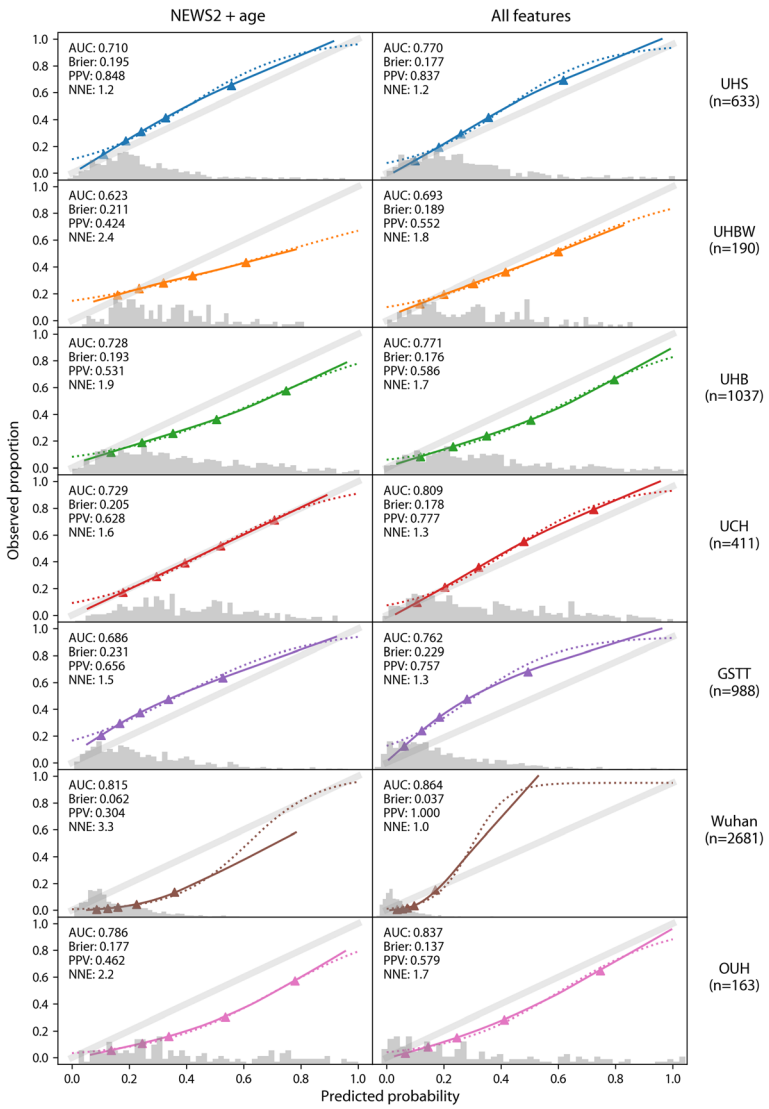

Fig. 2 Calibration (logistic and LOESS curves) of supplemented NEWS2 model for 14-day ICU/death model at validation sites

55%, and (ii) GSTT, where 'treat all' was superior up to a threshold of 30% and no improvement was seen for the supplemented model.

For the 3-day endpoint, the improvement in net benefit for the supplemented model over the two default strategies was smaller, compared to the improvements seen at 14 days (Additional file 8: Figure S2). At three sites (UHBW, GSST, and Wuhan), neither the baseline ('NEWS2 + age') nor the supplemented ('All features') models offered any improvement over the 'treat all' or 'treat none' strategies. At KCH and UHS, net benefit for 'NEWS2 + age' was higher than the default strategies for a range of risk thresholds but was not increased further by the supplemented ('NEWS2 + oxygen flow rate') model.

Discussion

Principal findings

This study is amongst the first to systematically evaluate NEWS2 for severe COVID-19 outcome and carry out external validation at multiple international sites (five UK NHS Trusts, one hospital in Norway, and two hospitals in Wuhan, China). We found that while 'NEWS2 + age' had moderate discrimination for short-term COVID-19 outcome (3-day ICU transfer/death), it showed poor-to-moderate discrimination for the medium-term outcome (14-day ICU transfer/death). Thus, while NEWS2 may be effective for short-term (e.g. 24 h) prognostication, our results question its suitability as a screening tool for medium-term COVID-19 outcome. Risk stratification was improved by adding routinely collected blood and physiological parameters, and

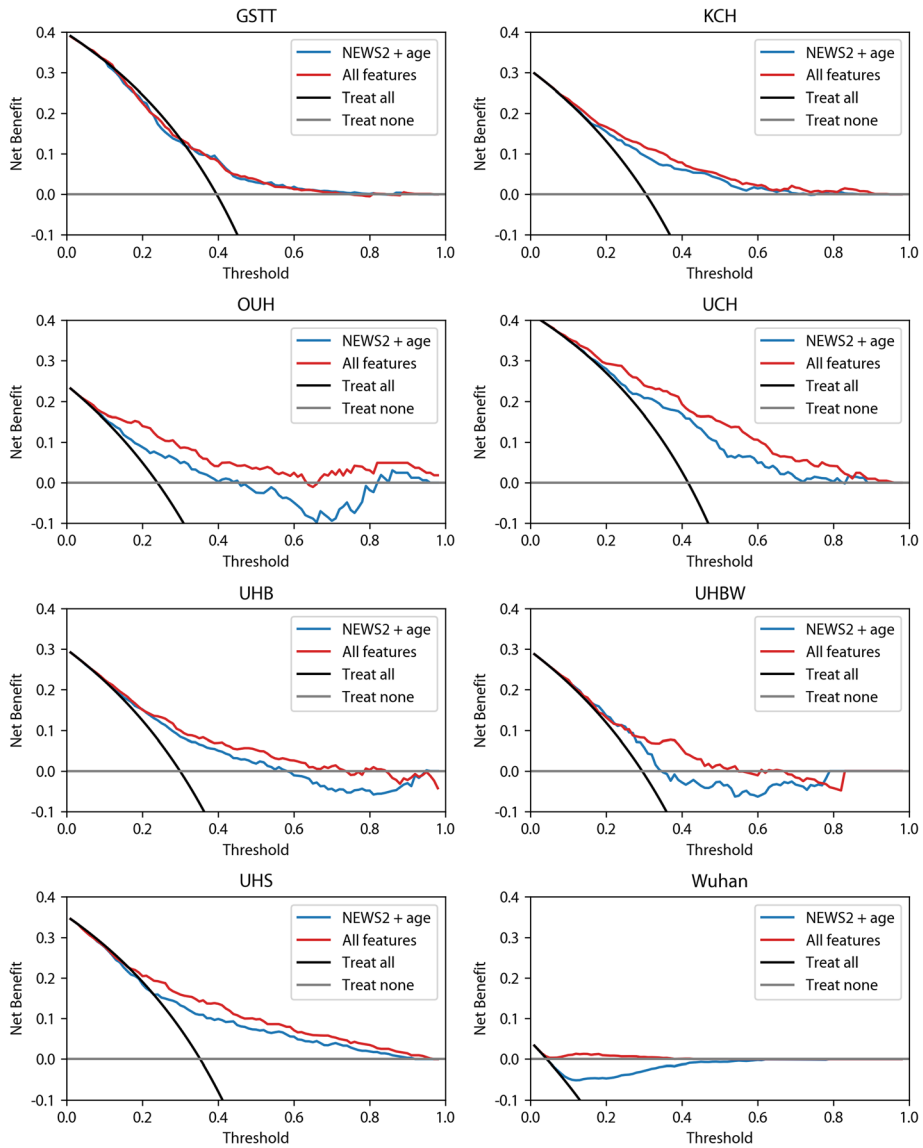

Fig. 3 Net benefit of supplemented NEWS2 model for 14-day ICU/death compared to default strategies ('treat all' and 'treat none') at training and validation sites

discrimination in supplemented models was moderate-to-good. However, the model showed evidence of miscalibration, with a tendency to underestimate risks in external sites. The derived model for 14-day ICU transfer/death included nine parameters: NEWS2 score, supplemental oxygen flow rate, urea, age, oxygen saturation, CRP, estimated GFR, neutrophil count, and neutrophil/lymphocyte ratio. Notably, pre-existing comorbidities did not improve risk prediction and were not retained in the final model. This was unexpected but may indicate that the effect of pre-existing health conditions could be manifest through some of the included blood or physiological markers.

Overall, this study overcomes many of the factors associated with a high risk of bias in the development of prognostic models for COVID-19 [13] and provides some evidence to support the supplementation of NEWS2 for clinical decisions with these patients.

Comparison with other studies

A systematic review of 10 prediction models for mortality in COVID-19 infection [10] found broad similarities with the features retained in our models, particularly regarding CRP and neutrophil levels. However, existing prediction models suffer several methodological

weaknesses including overfitting, selection bias, and reliance on cross-sectional data without accounting for censoring. Additionally, many existing studies have relied on single-centre or ethnically homogenous Chinese cohorts, whereas the present study shows validation across multiple and diverse populations. A key strength of our study is the robust and repeated external validation across national and international sites; however, evidence of miscalibration suggests we should be cautious when attempting to generalise these findings. Future research should include larger collaborations and aim to develop 'from onset' population predictions.

NEWS2 is a summary score derived from six physiological parameters, including oxygen supplementation. Lack of evidence for NEWS2 use in COVID-19 especially in primary care has been highlighted [9]. The oxygen saturation component of physiological measurements added value beyond NEWS2 total score and was retained following regularisation for 14-day endpoints. This suggests some residual association over and above what is captured by the NEWS2 score and reinforces Royal College of Physicians guidance that the NEWS2 score ceilings with respect to respiratory function [42].

Cardiac disease and myocardial injury have been described in severe COVID-19 cases in China [2, 23]. In our model, blood Troponin-T, a marker of myocardial injury, had additional salient signal but was only measured in a subset of our cohort at admission, so it was excluded from our final model. This could be explored further in larger datasets.

Strengths and limitations

Our study provides a risk stratification model for which we obtained generalisable and robust results across seven national and international sites with differing geographical catchment and population characteristics. It is amongst the first to evaluate NEWS2 at hospital admission for severe COVID-19 outcome and amongst a handful to externally validate a supplemented model across multiple sites.

However, some limitations must be acknowledged. First, there are likely to be other parameters not measured in this study that could substantially improve the risk stratification model (e.g. radiological features, obesity, or comorbidity load). These parameters could be explored in future work but were not considered in the present study to avoid limiting the real-world implementation of the risk stratification model. Second, our models showed better performance in UK secondary care settings amongst populations with higher rates of severe COVID-19 disease. Therefore, further research is needed to investigate the suitability of our model for primary care settings which have a high prevalence of mild disease severities and in community settings. This would allow us to capture variability at earlier stages of the disease and trends in patients not requiring hospital admission. Third, while external validation across multiple national and international sites represents a key strength, we did not have access to individual participant data and model development was limited to a single site (KCH). Although we benefited from existing infrastructure to support rapid data analysis, we urgently need infrastructure to support data sharing between sites to address some of the limitations of the present study (e.g. miscalibration) and improve the transferability of these models. Not only would this facilitate external validation, but more importantly, it would allow multi-site prediction models to be developed using pooled, individual participant data [43]. Fourth, our analyses would have excluded patients who experienced severe COVID-19 outcome at home or at another hospital, after being discharged from a participating hospital. Fifth, our model was restricted to blood and physiological parameters measured at hospital admission. This was by design and reflected the aim of developing a screening tool for risk stratification at hospital admission. However, future studies should explore the extent to which risk stratification could be improved by incorporating repeated measures of NEWS2 and relevant biomarkers.

Conclusions

The NEWS2 early warning score is in near-universal use in UK NHS Trusts since March 2019 [15], but little is known about its use for COVID-19 patients. Here, we showed that NEWS2 and age at hospital admission had poor-to-moderate discrimination for medium-term (14-day) severe COVID-19 outcome, questioning its use as a tool to guide hospital admission. Moreover, we showed that NEWS2 discrimination could be improved by adding eight blood and physiological parameters (supplemental oxygen flow rate, urea, age, oxygen saturation, CRP, estimated GFR, neutrophil count, neutrophil/lymphocyte ratio) that are routinely collected and readily available in healthcare services. Thus, this type of model could be easily implemented in clinical practice, and predicted risk score probabilities of individual patients are easy to communicate. At the same time, although we provided some evidence of improved discrimination vs. NEWS2 and age alone, given miscalibration in external sites, our proposed model should be used as a complement and not as a replacement for clinical judgement.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-020-01893-3.

Additional file 1: Table S1. SNOMED terms.

Additional file 2: Table S2. F1, precision and recall for NLP comorbidity detection.

Additional file 3: Table S3. Logistic regression models for each blood and physiological measure tested separately in the KCH training cohort, for 14- and 3-day ICU/death.

Additional file 4: Table S4. Internally validated discrimination for KCH training sample based on nested repeated cross-validation.

Additional file 5: Figure S1. Calibration (logistic and LOESS curves) of supplemented NEWS2 model for 3-day ICU/death model at validation sites.

Additional file 6: Table S5. Univariate logistic regression models for sensitivity analyses showing odds ratios of ICU/death at 3- and 14-days for subsets of the training cohort.

Additional file 7: Table S6. Discrimination for all models in training and validation cohorts, including alternative baseline model of 'NEWS2 only'.

Additional file 8: Figure S2. Net benefit of supplemented NEWS2 model for 3-day ICU/death compared to default strategies ('treat all' and 'treat none') at training and validation sites.

Acknowledgements

This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centres at South London and Maudsley NHS Foundation Trust, London AI Medical Imaging Centre for Value-Based Healthcare, and Guy's and St Thomas' NHS Foundation Trust, both with King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would also like to thank all the clinicians managing the patients, the patient experts of the KERRI committee, Professor Irene Higginson, Professor Alastair Baker, Professor Jules Wendon, Dan Persson, and Damian Lewsley for their support.

The authors acknowledge the use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's and St Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's and St Thomas' Charity (TR130505). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, King's College London, or the Department of Health and Social Care.

GVG also acknowledges the support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032), and the NIHR Birmingham Biomedical Research Centre. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council, or the Department of Health. The funding organisations had no role in the design of this study, data collection, analysis or interpretation, or preparation of the manuscript and did not approve or disapprove of or delay publication of the work. Furthermore, the UHB data collection was supported by the PIONEER Acute Care Hub and HDR-UK Better Care Programme. This work uses data provided by patients and collected by the NHS as part of their care and support. We would like to acknowledge the contribution of all staff, key workers, patients, and the community who have supported our hospitals and the wider NHS at this time.

Authors' contributions

The corresponding author, Dr. Ewan Carr, is the guarantor of the manuscript. JT, AMS, RD, EC, and RB conceived the study design and developed the study objectives. JT, RD, AF, LR, DB, ZK, TS, and AS were the leads to develop the CogStack platform. DB, ZK, TS, and AS were responsible for the data extraction and preparation. EC, RB, AP, and DS contributed to the statistical analyses. All authors contributed to the interpretation of the data. AMS, JT, KO, and RZ provided clinical input. All authors contributed to interpret the data and draft the article and provided final approval of the manuscript. DMB, ZK, AS, TS, JTHT, LR, and KN performed the data processing and software development. KOG, RZ, and JTHT performed the data validation. At GSTT, WW and WM were responsible for the data extraction and preparation. WW performed the model validation. AD and VC contributed to the interpretation of the data. At UHS, MS and FB were responsible for the data extraction and preparation. MS, HP, and AS contributed to the statistical analysis. All authors contributed to the interpretation of the data. MS and AP provided clinical input. MS and HP performed the data/model validation. At UCH, RKG and MN were responsible for the data extraction, preparation, and model validation. At UHBW, MR and MW were responsible for the data extraction and preparation. CM and CB conducted the data and model validation. For the Wuhan cohort, XZ, XW, and JS extracted the data from the EHR system. HW and HZ preprocessed the raw data and conducted the prediction model validations. BG, HW, HZ, TS, and JS interpreted the data and results. At UHB/UoB, UHB IT and AK were responsible for the data extraction and preparation. AK, LS, VRC, and GVG performed the model validation. AK, GVG, and SB contributed to the interpretation of the data. At OUH, KW, EKA, and ARH were responsible for the data extraction and preparation. AKL contributed to the statistical analysis and performed the model validation. All authors contributed to the interpretation of the data. The views expressed are those of the authors and not necessarily those of the MRC, NHS, the NIHR, or the Department of Health and Social Care. The funders of the study had no role in the study design, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the article for publication.

Funding

DMB is funded by a UKRI Innovation Fellowship as part of the Health Data Research UK MR/S00310X/1 (https://www.hdruk.ac.uk).

RB is funded in part by grant MR/R016372/1 for the King's College London MRC Skills Development Fellowship programme funded by the UK Medical Research Council (MRC, https://mrc.ukri.org) and by grant IS-BRC-1215-20018 for the National Institute for Health Research (NIHR, https://www.nihr.ac.uk) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London.

RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust; (3) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; (4) the National Institute for Health Research University College London Hospitals Biomedical Research Centre; (5) the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; (7) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.

KOG is supported by an MRC Clinical Training Fellowship (MR/R017751/1).

WW is supported by the Health Foundation grant.

AD and VC acknowledge the support from the National Institute for Health Research (NIHR) Applied Research Collaboration (ARC) South London at King's College Hospital NHS Foundation Trust and the Royal College of Physicians, as well as the support from the NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. VC is additionally supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the

Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust.
RZ is supported by a King's Prize Fellowship.
AS is supported by a King's Medical Research Trust studentship.
JTHT is supported by London AI Medical Imaging Centre for Value-Based Healthcare (AI4VBH) and the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.
FB and PTTH are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre, Data Sciences at University Hospital Southampton NHS Foundation Trust, and the Clinical Informatics Research Unit, University of Southampton.
JB is funded by the Clinical Informatics Research Unit, University of Southampton, and part-funded by the Global Alliance for Chronic Disease (GDAC).
A Pinto is part-funded by UHS Digital, University Hospital Southampton, Tremona Road, Southampton.
AJS is supported by a Digital Health Fellowship through Health Education England (Wessex).
HW and HZ are supported by the Medical Research Council and Health Data Research UK Grant (MR/S004149/1), Industrial Strategy Challenge Grant (MC_PC_18029), and Wellcome Institutional Translation Partnership Award (PIII054). XW is supported by the National Natural Science Foundation of China (grant number 81700006).
AMS is supported by the British Heart Foundation (CH/1999001/11735), the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London (IS-BRC-1215-20006), and the Fondation Leducq. AP is partially supported by NIHR NF-SI-0617-10120. This work was supported by the National Institute for Health Research (NIHR) University College London Hospitals (UCH) Biomedical Research Centre (BRC) Clinical and Research Informatics Unit (CRIU), NIHR Health Informatics Collaborative (HIC), and by awards establishing the Institute of Health Informatics at University College London (UCL). This work was also supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and the Wellcome Trust.
RKG is funded by the NIHR (DRF-2018-11-ST2-004). MN is funded by the Wellcome Trust (207511/Z/17/Z).
The work was supported by MRC Health Data Research UK (HDRUK/CFC/01), an initiative funded by the UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. AK is funded by a MRC Rutherford Fellowship MR/S003991/1 (as part of Health Data Research UK https://www.hdruk.ac.uk).

Availability of data and materials

Code and pre-trained models are available at https://github.com/ewancarr/NEWS2-COVID-19 and openly shared for testing in other COVID-19 datasets.
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 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.
The GSTT dataset cannot be released publicly due to the risk of re-identification of such granular individual-level data, as determined by the Guy's and St Thomas's Trust Caldicott Guardian.
The UHS dataset cannot be released publicly due to the risk of re-identification of such granular individual-level data, as determined by the University Hospital Southampton Caldicott Guardian.
The UCH data cannot be released publicly due to conditions of regulatory approvals that preclude open access data sharing to minimise the risk of patient identification through granular individual health record data. The authors will consider specific requests for data sharing as part of academic collaborations subject to ethical approval and data transfer agreements in accordance with the GDPR regulations.
The Wuhan dataset used in the study will not be available due to the inability to fully anonymise in line with ethical requirements. Applications for research access should be sent to TS and details will be made available via https://covid.datahelps.life/prediction/.
The OUH dataset cannot be released publicly due to the risk of re-identification of such granular individual-level data.

Ethics approval and consent to participate

The KCH component of the project 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. The UHS validation was performed as part of an urgent service evaluation agreed with approval from trust research leads and the Caldicott Guardian. For UCH, ethical approval was given by East Midlands - Nottingham 2 Research Ethics Committee (REF: 20/EM/0114; IRAS: 282900). The UHB component was operated under the PIONEER Health Data Research Hub in Acute Care ethical approval provided by the East Midlands Derby REC (reference: 20/EM/0158). For UHBW, the project was considered as service evaluation by the organisational review board. Informed consent was deemed unnecessary due to the retrospective observational nature of the data. Ethical approval for GSTT was granted by the London Bromley Research Ethics Committee (reference 20/HRA/1871) to the King's Health Partners Data Analytics and Modelling COVID-19 Group to collect clinically relevant data points from patient's electronic health records. The Wuhan validation was approved by the Research Ethics Committee of Shanghai Dongfang Hospital and Taikang Tongji Hospital. For the OUH validation, a project protocol was approved by the Regional Ethical Committee of South-East Norway (Reference number 137045) and the OUH data protection officer (Reference number 20/08822). Informed consent in the OUH cohort was waived because of the strictly observational nature of the project.

Consent for publication

Not applicable.

Competing interests

JTHT received research support and funding from InnovateUK, Bristol-Myers-Squibb, iRhythm Technologies, and holds shares < £5000 in Glaxo Smithkline and Biogen.

Author details

[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, 16 De Crespigny Park, London SE5 8AF, UK. [2]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK. [3]Health Data Research UK London, University College London, London, UK. [4]Clinical Informatics Research Unit, University of Southampton, Coxford Rd., Southampton SO16 5AF, UK. [5]NIHR Biomedical Research Centre at University Hospital Southampton NHS Trust, Coxford Road, Southampton, UK. [6]UHS Digital, University Hospital Southampton, Tremona Road, Southampton SO16 6YD, UK. [7]School of Population Health and Environmental Sciences, King's College London, London, UK. [8]Usher Institute, University of Edinburgh, Edinburgh, UK. [9]Department of Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [10]UCL Institute for Global Health, University College London Hospitals NHS Trust, London, UK. [11]University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK. [12]Department of Pulmonary and Critical Care Medicine, People's Liberation Army Joint Logistic Support Force 920th Hospital, Kunming, Yunnan, China. [13]King's College Hospital NHS Foundation Trust, London, UK. [14]School of Cardiovascular Medicine & Sciences, King's College London British Heart Foundation Centre of Excellence, London SE5 9NU, UK. [15]UCL Division of Infection and Immunity, University College London Hospitals NHS Trust, London, UK. [16]Institute of Health Informatics, University College London, London, UK. [17]NIHR Biomedical Research Centre at University College London Hospitals

NHS Foundation Trust, London, UK. [18]College of Medical and Dental Sciences, Institute of Cancer and Genomics, University of Birmingham, Birmingham, UK. [19]Institute of Translational Medicine, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [20]Health Data Research UK Midlands, Birmingham, UK. [21]Department of Medical Biochemistry, Blood Cell Research Group, Oslo University Hospital, Oslo, Norway. [22]Oslo Centre for Biostatistics and Epidemiology, Faculty of Medicine, University of Oslo, Oslo, Norway. [23]Department of Acute Medicine, Oslo University Hospital and Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [24]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [25]Department of Engineering Mathematics, University of Bristol, Bristol, UK. [26]Department of Pulmonary and Critical Care Medicine, Shanghai East Hospital, Tongji University, Shanghai, China. [27]Department of Pulmonary and Critical Care Medicine, Taikang Tongji Hospital, Wuhan, China.

Received: 29 September 2020 Accepted: 16 December 2020
Published online: 21 January 2021

References

1. WHO. WHO COVID-19 dashboard. 2020. https://who.sprinklr.com/.
2. 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;395:1054–62.
3. Scott LJ, Redmond NM, Tavaré A, Little H, Srivastava S, Pullyblank A. Association between National Early Warning Scores in primary care and clinical outcomes: an observational study in UK primary and secondary care. Br J Gen Pract. 2020. https://doi.org/10.3399/bjgp20X709337.
4. Lambden S, Laterre PF, Levy MM, Francois B. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. Crit Care. 2019;23:374.
5. Liu S, Yao N, Qiu Y, He C. Predictive performance of SOFA and qSOFA for in-hospital mortality in severe novel coronavirus disease. Am J Emerg Med. 2020. https://doi.org/10.1016/j.ajem.2020.07.019.
6. Lim WS, van der Eerden MM, Laing R, Boersma WG, Karalus N, Town GI, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. Thorax. 2003;58:377–82.
7. Royal College of Physicians. National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS. Updated report of a working party. London: RCP; 2017.
8. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation. 2013;84:465–70.
9. Greenhalgh T, Treadwell J, Burrow R. NEWS (or NEWS2) score when assessing possible COVID-19 patients in primary care. Cent Evid-Based Med Nuffield Dep Prim Care Health Sci Univ Oxf. 2020;20. https://www.cebm.net/covid-19/should-we-use-the-news-or-news2-score-when-assessing-patients-with-possible-covid-19-inprimary-care/.
10. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL score. Clin Infect Dis. https://doi.org/10.1093/cid/ciaa414.
11. Shi Y, Yu X, Zhao H, Wang H, Zhao R, Sheng J. Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. Crit Care. 2020;24:108.
12. COVIDAnalytics. https://www.covidanalytics.io/calculator. Access date 21 April 2020.
13. Wynants L, Calster BV, Bonten MMJ, Collins GS, Debray TPA, Vos MD, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369. https://doi.org/10.1136/bmj.m1328.
14. Liao X, Wang B, Kang Y. Novel coronavirus infection during the 2019–2020 epidemic: preparing intensive care units—the experience in Sichuan Province, China. Intensive Care Med. 2020;46:357–60.
15. NHS England » National Early Warning Score (NEWS). https://www.england.nhs.uk/ourwork/clinical-policy/sepsis/nationalearlywarningscore/. Access date 23 April 2020.
16. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395:497–506.
17. Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Invest Radiol. 2020;Publish Ahead of Print. doi:https://doi.org/10.1097/RLI.0000000000000672.
18. Xie J, Tong Z, Guan X, Du B, Qiu H. Clinical characteristics of patients who died of coronavirus disease 2019 in China. JAMA Netw Open. 2020;3:e205619.
19. Zhang J-J, Dong X, Cao Y-Y, Yuan Y-D, Yang Y-B, Yan Y-Q, et al. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. Allergy. 2020;75:1730-41.
20. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive Care Med. 2020;46:846–8.
21. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020. https://doi.org/10.1056/NEJMoa2002032.
22. Lagunas-Rangel FA. Neutrophil-to-lymphocyte ratio and lymphocyte-to-C-reactive protein ratio in patients with severe coronavirus disease 2019 (COVID-19): a meta-analysis. J Med Virol. 2020;92:1733-4.
23. Guo T, Fan Y, Chen M, Wu X, Zhang L, He T, et al. Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19). JAMA Cardiol. 2020. https://doi.org/10.1001/jamacardio.2020.1017.
24. 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. medRxiv. 2020. https://www.medrxiv.org/content/10.1101/2020.04.24.20078006v4.article-info.
25. Marshall JC, Murthy S, Diaz J, Adhikari NK, Angus DC, Arabi YM, et al. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020;20:e192–7.
26. 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 Med Inform Decis Mak. 2018;18. https://doi.org/10.1186/s12911-018-0623-9.
27. Kraljevic Z, Bean D, Mascio A, Roguski L, Folarin A, Roberts A, et al. MedCAT -- Medical Concept Annotation Tool. ArXiv191210166 Cs Stat. 2019. https://arxiv.org/abs/1912.10166.
28. Searle T, Kraljevic Z, Bendayan R, Bean D, Dobson R. MedCATTrainer: a biomedical free text annotation interface with active learning and research use case specific customisation. ArXiv190707322 Cs. 2019. https://arxiv.org/abs/1907.07322.
29. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:1–9.
30. Bean D, Kraljevic Z, Searle T, Bendayan R, Pickles A, Folarin A, et al. Treatment with ACE-inhibitors is associated with less severe disease with SARS-Covid-19 infection in a multi-site UK acute Hospital Trust. medRxiv. 2020;:2020.04.07.20056788. https://www.medrxiv.org/content/10.1101/2020.04.07.20056788v1.full-text.
31. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley: CreateSpace; 2009.
32. Steyerberg E. Clinical prediction models. 2nd ed. Cham: Springer; 2019.
33. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80:27–38.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach. J R Stat Soc Ser B-Methodol. 1995;57:289–300.
35. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58:267–88.
36. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
37. Seabold S, Perktold J. statsmodels: econometric and statistical modeling with Python. In: 9th Python in Science conference. Austin: SciPy; 2010. http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf.
38. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
39. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care Lond Engl. 2015;19:285.
40. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak. 2006;26:565–74.

42. NEWS2 and deterioration in COVID-19. RCP London. 2020. https://www.rcplondon.ac.uk/news/news2-and-deterioration-covid-19. Access date 24 April 2020.
43. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.

Publisher's Note