Supplementing the National Early Warning Score (NEWS2) for anticipating early deterioration among patients with COVID-19 infection

Authors

Ewan Carr*+[1], Rebecca Bendayan*[1,4], Daniel Bean[1,2], Kevin O'Gallagher[5,6], Andrew Pickles[1,4], Daniel Stahl[1], Rosita Zakeri[5,6], Thomas Searle[1,4], Anthony Shek[8], Zeljko Kraljevic[1], James T. Teo[5,8#], Ajay M. Shah[5,6#], Richard JB Dobson[1,2,3,4,7#]

*#joint authors
+corresponding authors: Dr Ewan Carr (ewan.carr@kcl.ac.uk); Dr James Teo (jamesteo@nhs.net); Prof Ajay M. Shah (ajay.shah@kcl.ac.uk); Prof Richard Dobson (richard.j.dobson@kcl.ac.uk)

Affiliations

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, U.K.
[2] Health Data Research UK London, University College London, London, U.K.
[3] Institute of Health Informatics, University College London, London, U.K.
[4] NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, U.K.
[5] King's College Hospital NHS Foundation Trust, London, U.K.
[6] School of Cardiovascular Medicine & Sciences, King's College London British Heart Foundation Centre of Excellence, London, SE5 9NU, U.K.
[7] NIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust, London, U.K.
[8] Dept of Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, U.K.

Abstract

Importance An early minimally symptomatic phase is often followed by deterioration in patients with COVID-19 infection. This study shows that the addition of age and a minimal set of common blood tests taken in patients on admission to hospital significantly improves the National Early Warning Score (NEWS2) for risk-stratification of severe COVID disease.

Objective To supplement the NEWS2 score with a small number of easily obtained additional demographic, physiological and blood variables indicative of severity of COVID-19 infection.

Design Retrospective observational cohort with internal and temporal held-out external validation.

Setting Acute secondary care.

Participants 708 patients admitted to an acute multi-site UK NHS hospital with confirmed COVID-19 disease from 1st March to 5th April 2020.

Intervention Not applicable.

Main outcome and measures The primary outcome was patient status at 14 days after symptom onset categorised as severe disease (WHO-COVID-19 Outcomes Scales 6-8: i.e. transferred to intensive care unit or death). 218 of the 708 patients reached the primary end point. A range of physiological and blood biomarkers as well age, gender, ethnicity and comorbidities (hypertension, diabetes, heart, respiratory and kidney diseases) were assessed for their association with the primary outcome.

Results NEWS2 total score on admission was a weak predictor for severity of COVID-19 infection at 14 days (internally validated AUC = 0.628). The addition of age and common blood tests (CRP, neutrophil count, estimated GFR and albumin) provided substantial improvements to a risk stratification model but performance was still only moderate (AUC = 0.75). Common comorbidities hypertension, diabetes, heart, respiratory and kidney diseases have minor additional predictive value.

Conclusions and relevance Adding age and a minimal set of common blood parameters to NEWS2 improves the risk stratification of patients likely to develop severe COVID-19 outcomes. The addition of a few common parameters is likely to be much easier to implement in a short time-scale than a novel risk-scoring system.

Introduction

While approximately 80% of individuals with COVID-19 infection have mild or no symptoms[1], some develop severe COVID-19 disease requiring hospital admission. As of 23rd April 2020, there have been >2.5 million confirmed cases worldwide[2]. 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 invaluable[12].

Currently available risk scores for deterioration of acutely ill patients include (1) widely-used generic ward-based risk indices such as the National Early Warning Score (NEWS2)[3] or modified sequential organ failure assessment (mSOFA)[4]; and (2) the pneumonia-specific risk index, CURB-65[5] which usefully capture a combination of physiological observations with limited blood markers and comorbidities. The 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 UK NHS hospitals[6,7]. The physiological parameters assessed in the NEWS2 score - particularly patient temperature, oxygen saturations and the supplemental oxygen dependency - have been associated with COVID-19 outcomes[1]; however, little is known about their predictive value for the severity of COVID-19 disease. Additionally, a number of COVID-19-specific risk indices are being developed[8–10] as well as unvalidated online calculators[11] but generalisability is not yet known[10]. A Chinese study has suggested a modified version of NEWS2 with addition of age only[12] but without any data on performance. With near universal usage of NEWS2 in UK NHS Trusts since March 2019[13], minor adaptation to NEWS2 would be relatively easy to implement.

As the SARS-Cov2 pandemic has progressed, evidence has emerged regarding potentially useful blood biomarkers[1,14–17]. Although most of these early reports contain data from small numbers of patients, a number of markers have been found to be associated with severity. These include neutrophilia and lymphopenia, particularly in older adults[9,16,18,19], neutrophil-to-lymphocyte ratio[20], raised C-Reactive Protein (CRP) and lymphocyte-to-CRP ratio[20], markers of liver and cardiac injury such as alanine aminotransferase (ALT), aspartate aminotransferase (AST) and cardiac troponin[21] and elevated D-dimers, ferritin and fibrinogen[2,5,7]. Furthermore, plasma levels of cytokines such as IL-6 have been found to be higher in COVID-19 patients compared to controls[1].

Our aim is to understand the performance of NEWS2 and identify a supplemental combination of simple clinical and blood biomarkers routinely measured in hospitals to supplement the NEWS2 score to improve prediction of a severe disease outcome at 14 days from symptom onset. To reach this aim, our specific objectives were:

1. To explore independent associations of routinely measured physiological and blood parameters (including NEWS2 parameters) at or near hospital admission with disease

severity (i.e., ICU admission or death), adjusting for socio-demographics and comorbidities.

2. To examine which minimal combination of these potential determinants of disease severity (physiological and blood parameters, sociodemographics and comorbidities) are the best predictors of disease severity at 14 days since symptom onset; and
3. To compare the predictive value of the resulting model with a model based on the NEWS2 total score alone.

Methods

Patients

The study cohort was defined as all adult inpatients testing positive for SARS-Cov2 by reverse transcription polymerase chain reaction (RT-PCR) between 1st March to 5st April 2020 at a multi-site acute NHS hospital in South East London (UK). The catchment area of King's College Hospital NHS Foundation Trust includes the most severely affected part of the UK during the current pandemic. All patients included in the study had symptoms consistent with COVID-19 disease (e.g. cough, fever, dyspnoea, myalgia, delirium). We excluded subjects who were seen in the emergency department but not admitted. For purposes of temporal external validation, detailed below, patients were split into training and temporal external validation samples, with those tested positive before 31st March 2020 assigned to training, and those tested positive on/after 31st March 2020 assigned to validation.

This 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.

Data Processing

The data (demographics, emergency department letters, discharge summaries, clinical notes, lab results, vital signs) were retrieved and analyzed in near real-time from the structured and unstructured components of the electronic health record (EHR) using a variety of natural language processing (NLP) informatics tools belonging to the CogStack ecosystem[22], namely MedCAT[23] and MedCATTrainer[24]. The CogStack NLP pipeline captures negation, synonyms, and acronyms for medical 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 under parent terms Clinical Finding, Disorder, Organism, and Event with disambiguation, pre-trained on MIMIC-III[25]. The annotated SNOMED-CT terms are summarised in Supplementary Table 1.

Starting from our previous model[26], further supervised training improved detection of annotations and meta-annotations such as experiencer (is the concept annotated 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 do not directly relate to the patient currently. Performance of the MedCAT NLP pipeline for disorders 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 Supplementary Table 2.

Measures

Outcome. The primary outcome was patient status at 14 days after symptom onset, or admission to hospital where symptom onset was missing, categorised as transfer to ICU/death (WHO-COVID-19 Outcomes Scales 6-8) vs. not ICU/death (Scales 3-5). The WHO-COVID-19 Outcome Scales 6-7 incorporate admission to an ICU while Outcome Scale 8 indicates death. Date of symptom onset, date of ICU transfer and date of death were ascertained and verified manually by a clinician.

Blood parameters. We focused on biomarkers that were routinely obtained at or shortly after admission and were therefore available for the vast majority of patients. These comprised: albumin (g/L), alanine aminotransferase (ALT; IU/L), creatinine (µmol/L), C-reactive protein (CRP; mg/L), estimated Glomerular Filtration Rate (eGFR; mL/min), Haemoglobin (g/L), lymphocyte count (x 10^9/L), neutrophil count (x 10^9/L), and platelet count (PLT; x 10^9/L). We also derived the neutrophil-to-lymphocyte ratio (NLR) and the lymphocyte-to-CRP ratio[13]. Troponin-T (ng/L) and Ferritin (ug/L) were included, although these measures were only available for a subset of participants. D-dimers and HbA1c were excluded since they were measured in very few patients at admission and insufficient samples were available for analysis.

Physiological parameters. We included the six physiological parameters that form the basis of the NEWS2 score, namely, respiratory rate (breaths per minute), oxygen saturation (%), systolic blood pressure (mmHg), heart rate (beats/min), temperature (°C), and consciousness (measured by Glasgow Coma Scale (GCS) total score). All were measured at or shortly after admission. We assessed these parameters individually as well as a NEWS2 total score. Diastolic blood pressure, which is not part of the NEWS2 score, was also included in the analyses.

Demographics and comorbidities. Age, sex, ethnicity and comorbidities were considered. Where ethnicity data was available this was categorised as caucasian vs. BAME (Black, Asian and minority ethnic). For supplementary models adjusting for ethnicity, patients with ethnicity reported as ‘unknown/mixed/other’ were excluded. We included binary measures (present vs. not present) of relevant comorbid chronic health conditions derived from the NLP pipeline described above: hypertension, diabetes, heart disease (heart failure and ischemic heart disease), respiratory disease (asthma and chronic obstructive pulmonary disease, COPD) and chronic kidney disease.

Statistical analyses

Preliminary descriptive and exploratory analyses were performed. To address our first objective – exploring independent associations of physiological and blood parameters with 14-day death/ICU – we used penalised maximum likelihood logistic regression which reduces bias due to small sample size[27]. Each parameter was tested independently, adjusted for age and sex (Model 1) and then additionally adjusted for comorbidities (Model 2). Parameters exhibiting skewed distributions were transformed before modelling with logarithmic or square-root transformations. All parameters were scaled (mean = 0, standard deviation = 1) to improve interpretability. Outlying high values for some blood parameters were retained after individual examination by clinicians who ascertained their plausibility. We used the maximal available sample when testing each parameter. Given the number of tests conducted, *P*-values were adjusted using the Benjamini-Hochberg procedure to keep the False Discovery Rate (FDR) at 5%[28]. These models were conducted with R 3.6[23] using the logistf[24] package.

To address our second and third objectives – which combination of parameters performed best in predicting the 14-day outcome over and above NEWS2 – we estimated models combining all parameters using regularized logistic regression with a LASSO (Least Absolute Shrinkage and Selection Operator) estimator which shrinks parameters according to their variance, reduces overfitting and enables automatic variable selection[29]. The optimal degree of regularization was determined by identifying a tuning parameter λ using cross-validation[30]. LASSO regression provides a sparse, interpretable model, which allows us to predict individual risk scores (i.e. probability of severe outcome). Starting from an initial model with NEWS2 total score only, sets of features were added in order of (i) age and sex, (ii) blood and physiological parameters; (iii) comorbid conditions. A final model was estimated using NEWS2 total score alongside the top five most influential features from previous models. To estimate the predictive performance of our model on new unseen cases of the same underlying population, we performed internal nested cross-validation (10 folds and 20 repeats for the inner loop; 10 folds and 100 repeats for the outer loop). Overall discrimination was assessed based on the area under the curve (AUC). All continuous features were scaled (mean = 0, standard deviation = 1). Missing feature information was imputed (after scaling) using k-Nearest Neighbours imputation (k=5). Scaling and kNN imputation were incorporated within the model development and selection process to avoid data leakage which would otherwise result in optimistic performance measures[31].

To assess whether a more complex machine learning estimator would improve predictive performance, we repeated this set of models using gradient boosted trees implemented in the XGBoost library[32]. Procedures for internally validating these models were equivalent to those described above for regularized logistic regression except the imputation step was omitted due to the ability of XGBoost to handle missing data.

The predictive performance of the derived regularized logistic regression model was then evaluated by temporal external validation[33] with a hold-out sample of 256 patients who were admitted to hospital after the training sample (see Supplementary Figure 1). This involved

estimating the original model exactly as presented, including scaling and imputation models derived in the training data set. Discrimination performance was assessed using AUC, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Model calibration was assessed using a calibration plot (model predicted probability vs. true probability). These models were estimated in Python 3.6[34] using NumPy[35], and Scikit-Learn[36].

Sensitivity analyses were performed to account for potential demographic variability. Recent evidence suggest sex differences with men more likely to experience worse outcomes[16]. Therefore, in separate models, we tested interactions between each physiological and blood parameter and sex using likelihood-ratio tests (comparing a null model with the main effects only vs. a model additionally including the interaction term). In addition, we replicated all models with adjustment for ethnicity in the subset of individuals with available data for ethnicity (n=285 in training sample).

Results

The initial inpatient cohort comprised 452 inpatients testing positive for COVID-19 of whom 159 (35%) were transferred to ICU or died (COVID-19 WHO Score 6-8) within 14 days of symptom onset. Table 1 describes the clinical characteristics of the cohort: the mean age was 67 years (standard deviation = 18.5); 54% (n=248) were male; 42% (n=120) were categorised as BAME. Patients associated with a more severe outcome were significantly older (71 vs. 65 years; $p = 0.004$) but there was no evidence of differences by sex or ethnicity. There were some differences between groups in the prevalence of comorbidities but these did not reach statistical significance after multiple testing correction. For example, compared to patients with less severe outcomes, those who transferred to ICU or died had higher rates of hypertension (60% vs. 50%; $p = 0.11$), diabetes (38% vs. 32%; $p = 0.33$), heart failure (16% vs. 11%; $p = 0.33$) and chronic kidney disease (24% vs. 16%; $p = 0.11$). Rates of other comorbidities were similar between the two groups. There were differences between outcome groups for most blood and physiological parameters. Patients who had transferred to ICU or died within 14 days had, at admission, lower levels of Albumin, ALT, and estimated GFR; and elevated levels of CRP, creatinine, Ferritin, and Neutrophils. Mean NEWS2 total scores were significantly different (3.4 vs 2.1; $p < 0.001$; corresponding to Cohen's d of -0.57) in patients who transferred to ICU or died, compared to inpatients experiencing less severe outcomes.

Logistic regression models were used to assess independent associations between each physiological and blood parameter and disease severity measured as transfer to ICU or death (Table 2). Individuals were more likely to have transferred to ICU/died within 14 days of symptom onset if: they had higher CRP, NEWS2 score, heart rate, neutrophils, neutrophil-lymphocyte ratio, respiration rate; or if they had lower lymphocyte/CRP ratios, eGFR, creatinine, and oxygen saturation. These associations remained after adjustment for age, sex and comorbidities. There was no evidence of differences by sex (results not presented) and findings were consistent when additionally adjusting for ethnicity in secondary analyses using the subset of individuals with ethnicity data (Supplementary Table 3).

Combining physiological and blood parameters to assess ability to improve on NEWS2 in predicting 14-day outcome

To identify which minimal set of parameters were best able to improve on NEWS2 in predicting the 14-day outcome (ICU/death vs. not ICU/death), we combined all predictors in a single logistic regression model using LASSO regularisation. Internally validated predictive performance based on the area under the ROC curve (AUC) is presented in Table 3 for different feature sets. NEWS2 shows poor discrimination with an AUC of 0.628. Adding age and sex to a baseline model of NEWS2 total score only increased the AUC by 0.025 to 0.653 (+/- 2SD range: 0.639, 0.667). Further adding in all other blood and physiological parameters (except NEWS2) increased the AUC further by 0.089, to 0.742 (+/- 2SD: 0.726, 0.758). Additionally including comorbidities in this model did not improve performance. A final model was estimated including NEWS2 and the top five most important features taken from Model 4. This simpler model resulted in a slightly larger AUC of 0.751 (+/- 2SD range: 0.737, 0.764) which may indicate some overfitting due to the pre-selection of variables from previous analyses. Results were consistent when repeating these models in the subset of patients with information available on ethnicity (Supplementary Table 5).

Figure 1 summarises feature importances from the LASSO logistic regression models. When adding demographic (age and sex), blood and physiological parameters to NEWS2 ('NEWS2 + DBP'), 8 features were retained, in order of effect sizes: NEWS2 total score, CRP, neutrophils, estimated GFR, albumin, age, Troponin T, and oxygen saturation. Notably, when additionally considering comorbid conditions ('NEWS2 + DBPC'), the retained features were similar, and no comorbid conditions were retained. This suggests that most of the variance is already captured by the top 5 parameters.

When these models were repeated using a more complex estimator (gradient boosted trees, using XGBoost[32]) the pattern of results was consistent with those from regularized logistic regression (Supplementary Table 5). Namely, the internally validated AUC improved from 0.646 for a model with NEWS2 alone, to 0.722 for a model that additionally included the five parameters: CRP, neutrophils, estimated GFR, albumin, and age. Importantly, while the pattern of results was consistent, a more complex machine learning estimator produced no improvements to predictive performance.

Temporal external validation was conducted on a hold-out sample of 256 patients. This sample was similar to the training sample on all parameters (Supplementary Table 6) except the proportion who transferred to ICU or died was lower. Overall, results from the hold-out sample were consistent with those from internal validation. The AUC for NEWS2 alone was 0.700, and this improved to 0.730 when adding all blood and physiological parameters (sensitivity = 0.441; specificity = 0.873). The AUC for the simplified final model including NEWS2 and the top five features (CRP, neutrophils, estimated GFR, albumin and age) was similar (AUC = 0.730; sensitivity = 0.458; specificity = 0.873) (Supplementary Table 7). Calibration for these models

(Supplementary Figure 2) was acceptable but showed some consistent overestimation of risk probabilities.

Discussion

To our knowledge this study is the first to evaluate and then systematically attempt to improve performance of NEWS2 specifically for COVID-19. We found that the NEWS2 score shows overall poor discrimination with high specificity but poor sensitivity for severe outcomes in COVID-19 infection (transfer to ICU or death). However, its value for risk stratification (especially sensitivity) can be significantly improved by adding age and a small number of additional blood parameters (CRP, neutrophils, estimated GFR and albumin). A number of blood measures previously linked with more severe outcomes – such as lymphocyte and ALT[14], or transformations of inflammatory markers such as CRP/lymphocyte or neutrophil/lymphocyte ratio – did not provide additional value to the model over and above the existing features. Moreover, cardiac disease and myocardial injury has been described in severe COVID-19 cases in China[1,21]. 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 not included in our final model. This would have to be explored further in larger datasets. 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 over-fitting, selection bias, and reliance on cross-sectional data without accounting for censoring. Additionally, almost all existing studies have relied on ethnically homogenous Chinese cohorts and thus may be unrepresentative of other global populations.

Pre-existing disease comorbidities (hypertension, diabetes mellitus, heart failure, ischaemic heart disease, COPD, asthma and chronic kidney disease) were more common in patients with severe outcomes but had minimal contribution to the risk prediction and were not retained in the final model. This was unexpected and suggests potential shared variance between pre-existing health conditions and some of the included blood or physiological markers. Future research could explore the potential underlying shared mechanisms that predict deterioration.

NEWS2 is a summary score derived from six physiological parameters, including oxygen supplementation (yes/no). While NEWS2 total score was one of the most influential parameters in our models, the oxygen saturation component of our physiological measurements added value beyond NEWS total score and was retained following regularisation (i.e. model 'NEWS + DBP'). 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[37].

Strengths and limitations

Our study included data from a large sample of patients admitted to hospital with high rates of the primary outcome (transfer to ICU or death) and considered a large number of potential

predictors including demographics, physiological and blood parameters and comorbidities. However, some limitations should be acknowledged. First, there are likely to be other parameters not measured in this study that could improve the risk stratification model substantially (e.g. radiological features, other comorbidities or comorbidity load). This could be addressed by future work, but these were not considered in the present study to avoid limiting the real-world implementation of the risk stratification model; a complex model with many parameters will be harder to implement in clinical practice. Second, we used a 14-day time window from the symptom onset date as this provides a balance between medium-term prognostication and actionable risk stratification at the usual period of deterioration. Longer timeframes may be useful for prognostication but are harder to generalise due to the greater number of factors affecting outcomes, including institutional, regional or national policies. Since NEWS2 score is optimised for very near-term deterioration at 24 hours[7], a 14-day window was used as a compromise. Third, while the hold-out sample used for temporal external validation was similar in terms of demographics, blood and physiological parameters, the rate of more severe outcomes differed significantly. Perhaps due to changes in hospital procedures over time, this again suggests the need to validate these models in other hospitals or regions. Finally, while the model was derived from two hospital sites providing a mixed population, this study highlights that initial prediction models still have poor sensitivity and recalibration would be required before implementation as a risk model in clinical practice. Validation across datasets from a wider geographical region will be necessary to ensure generalisability.

Conclusion

In conclusion, this study suggests that the simple addition of a limited number of blood parameters to the existing and widely implemented NEWS2 system can contribute to improved risk stratification among COVID-19 patients. Our model can be easily implemented in clinical practice and predicted risk score probabilities of individual patients are easy to communicate. The additional parameters are widely collected on patients at hospital admission, and with near universal usage of NEWS2 in NHS Trusts since March 2019[13], a minor adaptation to NEWS2 is substantially easier to implement in a variety of health settings than a bespoke risk score.

Acknowledgments

DMB is funded by a UKRI Innovation Fellowship as part of 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: 1. 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. 2. 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. 3. The National Institute for Health Research University College London Hospitals Biomedical Research Centre. 4. National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. KO'G is supported by an MRC Clinical Training Fellowship. RZ is supported by a King's Prize Fellowship.

AS is supported by a King's Medical Research Trust studentship.
KO is supported by grant MR/R017751/1

AMS is supported by the British Heart Foundation (CH/1999001/11735), the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's & 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 (UCLH) 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.

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, and Guy's & St Thomas' NHS Foundation Trust, both with King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in 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.

References

1. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. 2020;395(10229):1054-1062. doi:10.1016/S0140-6736(20)30566-3
2. WHO. WHO COVID-19 Dashboard. https://who.sprinklr.com/. Published 2020. Accessed April 20, 2020.
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*. April 2020. doi: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(1):374. doi:10.1186/s13054-019-2663-7
5. Lim WS, Eerden MM van der, Laing R, et al. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*. 2003;58(5):377-382. doi:10.1136/thorax.58.5.377
6. 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.
7. 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(4):465-470. doi:10.1016/j.resuscitation.2012.12.016
8. Ji D, Zhang D, Xu J, et al. Prediction for Progression Risk in Patients with COVID-19 Pneumonia: the CALL Score. *Clin Infect Dis*. doi:10.1093/cid/ciaa414
9. 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(1):108. doi:10.1186/s13054-020-2833-7
10. Wynants L, Calster BV, Bonten MMJ, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*. 2020;369. doi:10.1136/bmj.m1328
11. COVIDAnalytics. https://www.covidanalytics.io/calculator. Accessed April 21, 2020.
12. 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(2):357-360. doi:10.1007/s00134-020-05954-2
13. NHS England » National Early Warning Score (NEWS). https://www.england.nhs.uk/ourwork/clinical-policy/sepsis/nationalearlywarningscore/. Accessed April 23, 2020.
14. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5
15. Li K, Wu J, Wu F, et al. The Clinical and Chest CT Features Associated with Severe and Critical COVID-19 Pneumonia. *Invest Radiol*. 2020;Publish Ahead of Print. doi:10.1097/RLI.0000000000000672
16. 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(4):e205619-e205619.

doi:10.1001/jamanetworkopen.2020.5619
17. Zhang J-J, Dong X, Cao Y-Y, et al. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy*. February 2020. doi:10.1111/all.14238
18. 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*. March 2020. doi:10.1007/s00134-020-05991-x
19. Guan W, Ni Z, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med*. February 2020. doi:10.1056/NEJMoa2002032
20. 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*. April 2020. doi:10.1002/jmv.25819
21. Guo T, Fan Y, Chen M, et al. Cardiovascular Implications of Fatal Outcomes of Patients With Coronavirus Disease 2019 (COVID-19). *JAMA Cardiol*. March 2020. doi:10.1001/jamacardio.2020.1017
22. Jackson R, Kartoglu I, Stringer C, 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. doi:10.1186/s12911-018-0623-9
23. Kraljevic Z, Bean D, Mascio A, et al. MedCAT -- Medical Concept Annotation Tool. *ArXiv191210166 Cs Stat*. December 2019. http://arxiv.org/abs/1912.10166. Accessed April 17, 2020.
24. 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*. July 2019. http://arxiv.org/abs/1907.07322. Accessed April 17, 2020.
25. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):1-9. doi:10.1038/sdata.2016.35
26. Bean D, Kraljevic Z, Searle T, 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*. April 2020:2020.04.07.20056788. doi:10.1101/2020.04.07.20056788
27. Firth D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika*. 1993;80(1):27-38. doi:10.2307/2336755
28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach. *J R Stat Soc Ser B-Methodol*. 1995;57(1):289-300.
29. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x
30. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York: CRC Press; 2015.
31. Kuhn M, Johnson K. *Applied Predictive Modeling*. Vol 26. Springer; 2013.
32. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: ACM; 2016:785-794. doi:10.1145/2939672.2939785
33. Steyerberg E. *Clinical Prediction Models*. Second Edition. Cham, Switzerland: Springer; 2019.
34. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace; 2009.
35. Oliphant T. NumPy: A guide to NumPy. 2006. http://www.numpy.org/.
36. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

37. NEWS2 and deterioration in COVID-19. RCP London. https://www.rcplondon.ac.uk/news/news2-and-deterioration-covid-19. Published April 14, 2020. Accessed April 24, 2020.

Tables

Table 1: Patient characteristics at hospital admission

	N avail.	All patients	Patients by status at 14-day endpoint		FDR-adjusted *P*-value for test between outcome groups[1]
			WHO-COVID-19 Outcomes Scales 3-5 (no ICU/death) n=393	WHO-COVID-19 Outcomes Scales 6-8 (ICU/death) n=159	
Age	452	67.00 [28.00]	64.87 [30.00]	70.92 [27.00]	0.004
Sex (male) N (%)	452	248 (54.9%)	157 (53.6%)	91 (57.2%)	0.682
BAME N (%)	285	120 (42.1%)	73 (41.0%)	47 (43.9%)	0.770
Comorbidities	N avail.	N (%)			*P*-value
Hypertension	452	243 (53.8%)	147 (50.2%)	96 (60.4%)	0.106
Diabetes mellitus	452	154 (34.1%)	93 (31.7%)	61 (38.4%)	0.325
Heart Failure	452	57 (12.6%)	32 (10.9%)	25 (15.7%)	0.325
Ischaemic Heart Diseases	452	85 (18.8%)	55 (18.8%)	30 (18.9%)	1.000
COPD	452	48 (10.6%)	27 (9.2%)	21 (13.2%)	0.366
Asthma	452	65 (14.4%)	44 (15.0%)	21 (13.2%)	0.770
Chronic Kidney Disease	452	84 (18.6%)	46 (15.7%)	38 (23.9%)	0.105
Blood biomarkers	N avail.	Mean [IQR]			*P*-value
Albumin	322	37.11 [7.00]	38.05 [7.00]	35.48 [7.00]	<0.001
Alanine aminotransferase (ALT)	184	54.83 [33.00]	60.34 [30.50]	46.45 [34.00]	0.386
C-reactive protein (CRP)	419	93.55 [106.70]	72.99 [84.90]	130.41 [135.62]	<0.001
Creatinine	420	121.67 [49.00]	105.86 [40.50]	150.42 [72.00]	0.001
Estimated GFR	334	63.75 [40.00]	68.01 [36.00]	56.05 [44.50]	<0.001
Ferritin	122	1356.01 [1165.25]	1272.35 [1149.75]	1442.45 [902.50]	0.016

Haemoglobin	419	125.05 [30.00]	125.52 [30.00]	124.21 [28.75]	0.770
Lymphocyte count	419	1.45 [0.67]	1.10 [0.69]	2.09 [0.67]	0.695
Neutrophil count	418	5.72 [3.53]	5.06 [3.01]	6.91 [5.31]	<0.001
Neutrophil/lymphocyte ratio	418	6.80 [5.01]	5.81 [4.22]	8.58 [6.26]	<0.001
Lymphocyte/CRP ratio	416	0.07 [0.04]	0.08 [0.05]	0.05 [0.02]	<0.001
Platelet count	421	226.68 [103.00]	228.34 [102.50]	223.69 [104.25]	0.958
Troponin T	141	33.92 [29.00]	30.40 [26.00]	37.92 [38.50]	0.351
Physiological parameters	N avail.	Mean [IQR]			*P*-value
NEWS2 Total Score	401	2.51 [3.00]	2.10 [3.00]	3.40 [4.00]	<0.001
Heart rate	405	85.35 [20.00]	84.49 [19.00]	87.15 [23.50]	0.359
Oxygen saturation	404	96.22 [3.00]	96.54 [2.00]	95.56 [3.00]	0.008
Respiration rate	405	19.84 [2.00]	19.42 [2.00]	20.72 [3.00]	0.008
GCS score	172	14.12 [1.00]	14.20 [1.00]	13.95 [1.00]	0.117
Systolic blood pressure	405	127.39 [29.00]	127.09 [26.50]	128.00 [32.00]	0.770
Diastolic blood pressure	405	72.69 [18.00]	73.20 [18.00]	71.63 [19.00]	0.325
Temperature	405	37.12 [0.90]	37.12 [0.90]	37.11 [1.00]	0.682

Notes.
[1] *Wilcoxon test for continuous variables; X^2 test for binary variables. FDR-corrected P-values based on the Benjamini–Hochberg correction.*

Table 2: Logistic regression models for each blood and physiological measure tested separately, sorted by effect size

	N avail.	**Model 1: Age, sex only**		**Model 2: + all comorbidities**	
		Odds Ratio [95% C.I.]	FDR-adjusted *P*-value[1]	Odds Ratio [95% C.I.]	FDR-adjusted *P*-value[1]
CRP	419	2.04 [1.64, 2.57]	<0.001	2.06 [1.65, 2.60]	<0.001
NEWS2 Total Score	401	1.82 [1.46, 2.30]	<0.001	1.83 [1.46, 2.31]	<0.001
Lymphocyte/CRP ratio	416	0.56 [0.44, 0.71]	<0.001	0.56 [0.44, 0.71]	<0.001
Troponin T	141	1.51 [1.02, 2.30]	0.119	1.69 [1.08, 2.78]	0.073
Neutrophil count	418	1.66 [1.33, 2.09]	<0.001	1.68 [1.35, 2.12]	<0.001
Ferritin	122	1.55 [1.05, 2.40]	0.098	1.60 [1.07, 2.54]	0.073
Estimated GFR	334	0.65 [0.51, 0.83]	0.004	0.66 [0.49, 0.87]	0.023
Respiration rate	405	1.47 [1.19, 1.83]	0.002	1.46 [1.19, 1.82]	0.003
Albumin	322	0.68 [0.53, 0.87]	0.010	0.69 [0.53, 0.89]	0.024
Oxygen saturation	404	0.72 [0.57, 0.89]	0.010	0.71 [0.56, 0.88]	0.013
Neutrophil/lymphocyte ratio	418	1.35 [1.09, 1.70]	0.026	1.36 [1.09, 1.72]	0.028
Creatinine	420	1.35 [1.09, 1.69]	0.024	1.35 [1.04, 1.76]	0.073
Heart rate	405	1.30 [1.05, 1.62]	0.068	1.32 [1.06, 1.65]	0.050
ALT	184	1.17 [0.86, 1.60]	0.923	1.22 [0.88, 1.68]	0.682
Temperature	405	1.09 [0.88, 1.36]	1.000	1.10 [0.88, 1.36]	0.999
Diastolic blood pressure	405	0.90 [0.73, 1.11]	0.952	0.92 [0.74, 1.13]	0.999
Platelet count	421	0.95 [0.77, 1.16]	1.000	0.94 [0.76, 1.15]	0.999
Lymphocyte count	419	1.05 [0.86, 1.29]	1.000	1.05 [0.86, 1.29]	0.999

GCS score	172	0.95 [0.70, 1.31]	1.000	0.96 [0.70, 1.32]	0.999
Hemoglobin	419	0.98 [0.79, 1.20]	1.000	1.03 [0.83, 1.27]	0.999
Systolic blood pressure	405	0.97 [0.78, 1.20]	1.000	0.98 [0.78, 1.21]	0.999

Notes.
[1]FDR-corrected P-values based on the Benjamini–Hochberg correction.
Odds ratios represent a one standard deviation change in the respective blood and clinical measure at admission (tested in separate models). Model 1 adjusted for age and sex. Model 2 additionally adjusted for comorbidities (hypertension, diabetes, heart diseases, respiratory diseases and chronic kidney disease).

Table 3: Internally validated predictive performance (n=452)

Notes. AUC based on repeated, nested cross-validation (inner loop: 10-fold, 20 repeats; outer loop = 10-fold, 100 repeats). Missing values imputed at each outer loop with k-Nearest Neighbours (KNN) imputation.

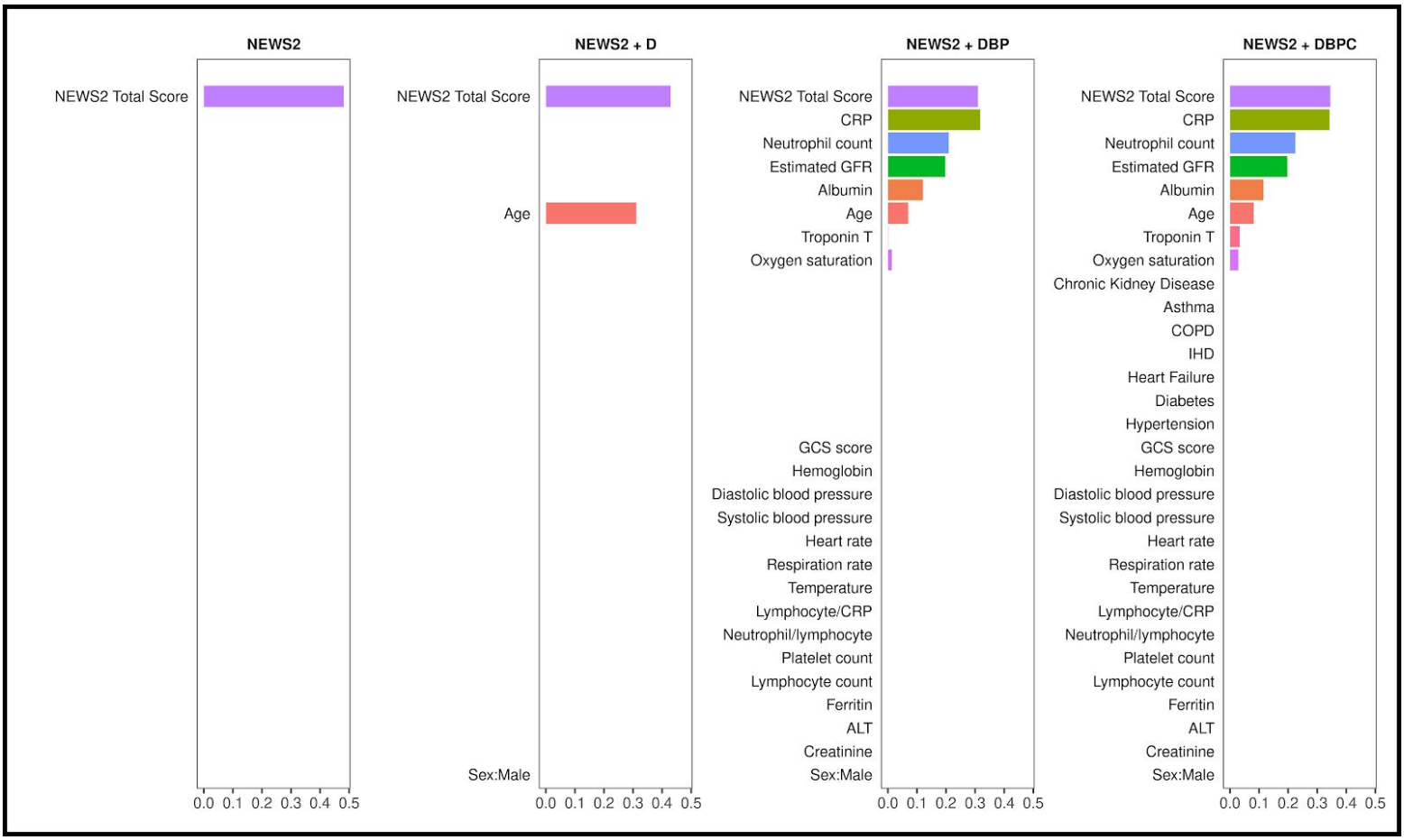
	Included features	**Internally validated AUC**			**Sensitivity**	**Specificity**	**PPV**	**NPV**
		Mean	-2SD	+2SD				
1	NEWS2	0.628	0.619	0.637	0.180	0.950	0.664	0.681
2	NEWS2 + **D**	0.653	0.639	0.667	0.189	0.929	0.597	0.678
3	NEWS2 + D**BP**	0.742	0.726	0.758	0.400	0.857	0.585	0.723
4	NEWS2 + DBP**C**	0.737	0.721	0.753	0.385	0.854	0.588	0.719
5	NEWS2 + CRP + Neutrophil + eGFR + Albumin + Age	0.751	0.737	0.764	0.415	0.842	0.589	0.727

D = Age, sex
C = comorbidities (8 features)
B = bloods (10 features)
P = physiological parameters (7 features)

Figures

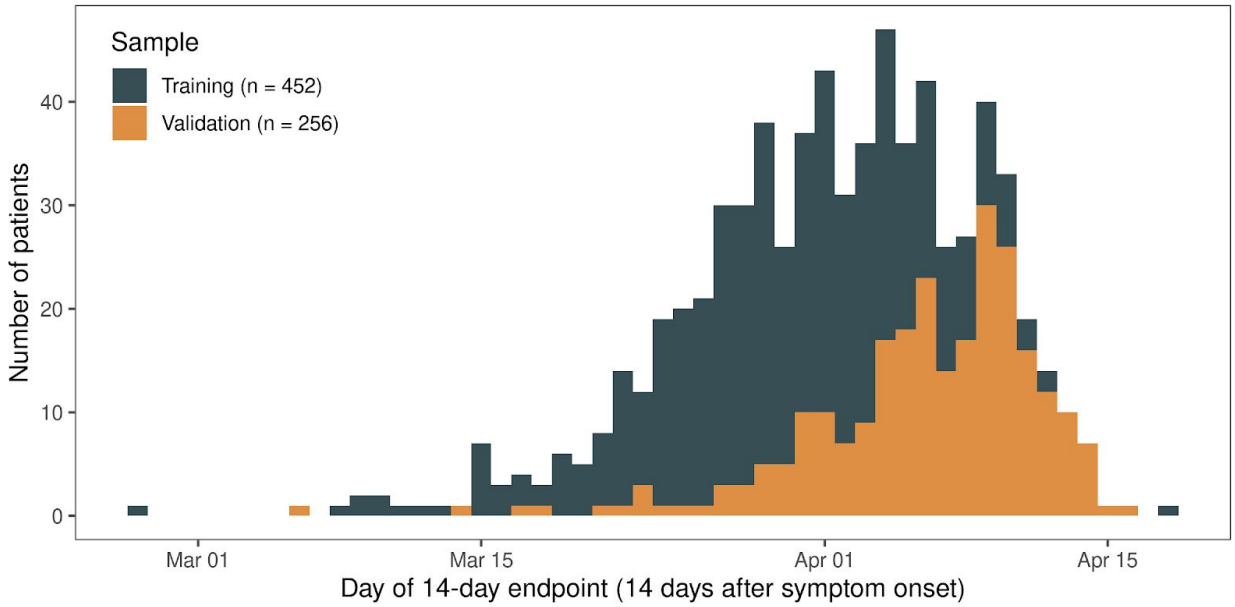
Figure 1: Feature importances from LASSO logistic regression in training sample (n=452)

Notes. Feature importances refer to absolute values of standardised coefficients from logistic regression, sorted by effect size in model 'NEWS2 + DBPC'. Where a feature is labelled on the y-axis, it was entered into the model. Features retained following LASSO regularisation are represented by a coloured bar; the absence of a bar indicates that this feature was omitted during regularisation.

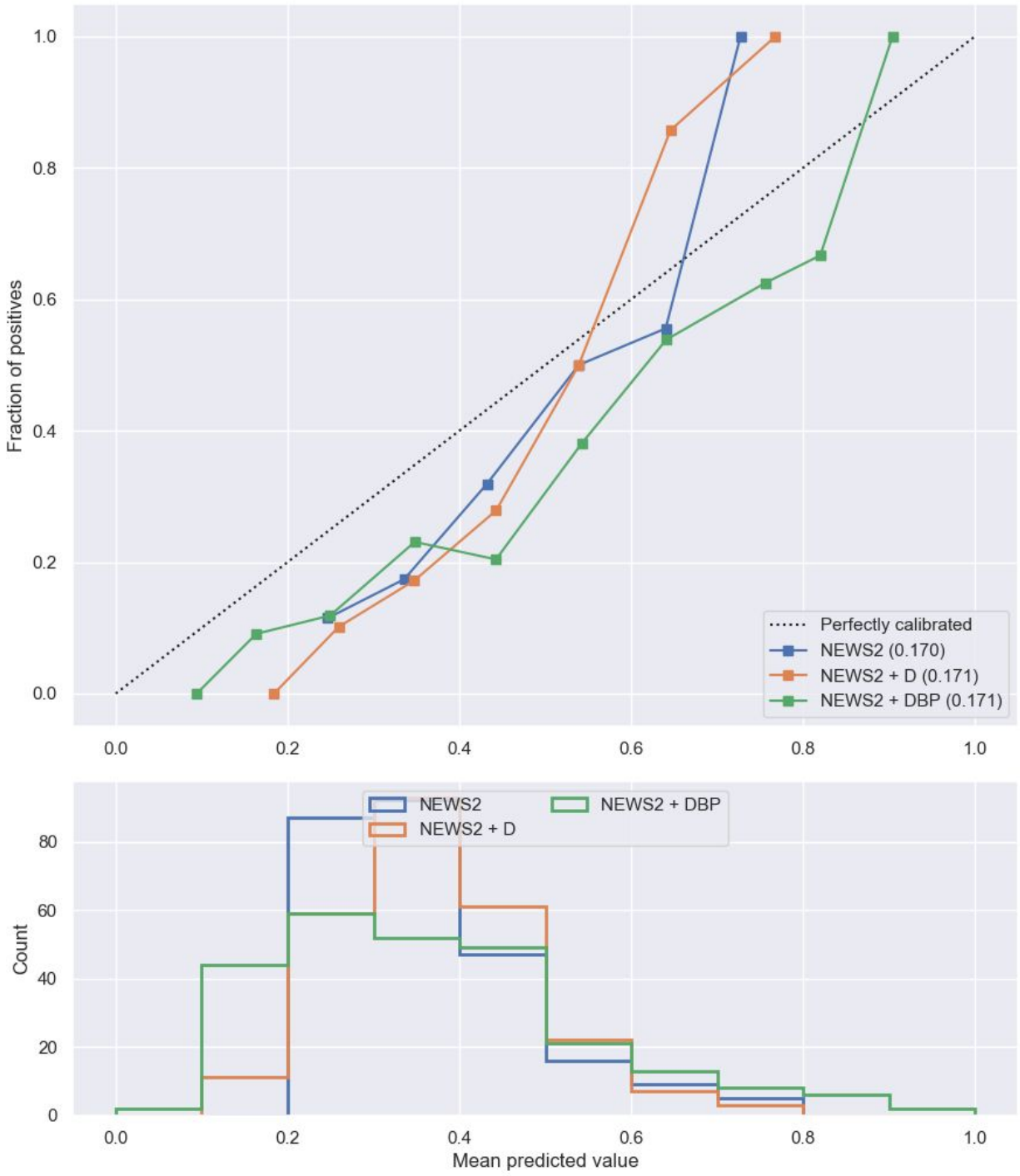


Supplementary Materials

Supplementary Figure 1: Timing of 14-day endpoints for training (n=452) and validation (n=256) samples



Supplementary Figure 2: Calibration plot from temporal external validation



Supplementary Table 1: SNOMED terms

SNOMED concept name	SNOMED concept IDs
Diabetes	S-230572002, S-44054006, S-237599002, S-49455004
Heart Failure	S-42343007, S-426263006, S-48447003, S-418304008, S-10633002
IHD	S-401314000, S-194828000, S-233839009, S-414545008 S-394659003, S-1755008, S-413838009
Hypertension	S-59621000
COPD	S-13645005, S-313297008
Asthma	S-195967001
CKD	S-433144002, S-90688005, S-709044004

Supplementary Table 2: F1, precision and recall for NLP co-morbidity detection

MedCATTrainer[24] was used to collect manual annotations for 146 clinical documents totalling 4343 annotations. Each co-morbidity is defined using one or more SNOMED terms. Predicted true positive labels (TP), precision (P), recall (R), F1-score (F1) are shown for these aggregated concepts. These results only consider entity detection and not meta annotation.

	TP	F1	P	R	SNOMED terms
Diabetes mellitus	73	0.936	0.924	0.948	S-230572002, S-44054006, S-237599002, S-49455004
Heart Failure	11	0.893	0.786	1.000	S-42343007, S-426263006 S-48447003, S-418304008 S-10633002
IHD	23	0.979	0.958	1.000	S-401314000, S-194828000 S-233839009, S-414545008 S-394659003, S-1755008 S-413838009
Hypertension	84	0.883	0.988	0.778	S-59621000
COPD	14	0.967	0.933	1.000	S-13645005, S-313297008

Asthma	11	1.000	1.000	1.000	S-195967001
CKD	15	0.938	0.938	0.938	S-433144002, S-90688005 S-709044004

Supplementary Table 3: Logistic regression models for each blood measure tested separately, adjusted for ethnicity for patients with information on ethnicity

		Model 1: Age, sex, ethnicity		**Model 2: + all comorbidities**	
Measure	N avail.	OR [95% C.I.]	FDR-adjusted P-value	OR [95% C.I.]	FDR-adjusted *P*-value
CRP	263	2.15 [1.63, 2.91]	<0.001	2.24 [1.68, 3.07]	<0.001
NEWS2 Total Score	250	2.06 [1.56, 2.79]	<0.001	2.04 [1.54, 2.77]	<0.001
Troponin T	84	1.62 [0.94, 2.99]	0.394	1.86 [1.01, 3.60]	0.210
Lymphocyte/CRP ratio	260	0.57 [0.41, 0.76]	0.001	0.56 [0.41, 0.76]	0.001
Neutrophil count	262	1.57 [1.20, 2.12]	0.007	1.56 [1.19, 2.10]	0.009
Oxygen saturation	252	0.63 [0.47, 0.83]	0.009	0.66 [0.49, 0.87]	0.022
Heart rate	253	1.46 [1.12, 1.93]	0.029	1.45 [1.11, 1.92]	0.037
Respiration rate	253	1.46 [1.15, 1.90]	0.012	1.44 [1.14, 1.87]	0.021
GCS score	109	0.70 [0.43, 1.11]	0.440	0.70 [0.42, 1.14]	0.527
Albumin	191	0.71 [0.51, 0.97]	0.162	0.71 [0.51, 0.99]	0.210
Creatinine	264	1.24 [0.95, 1.65]	0.440	1.33 [0.97, 1.87]	0.341
Estimated GFR	199	0.81 [0.58, 1.11]	0.594	0.77 [0.53, 1.12]	0.553
ALT	130	1.14 [0.73, 1.80]	1.000	1.26 [0.79, 2.04]	0.950
Neutrophil/lymphocyte ratio	262	1.24 [0.95, 1.65]	0.440	1.22 [0.94, 1.62]	0.527
Temperature	253	1.18 [0.92, 1.52]	0.594	1.18 [0.92, 1.53]	0.573

Ferritin	81	1.08 [0.64, 1.81]	1.000	1.17 [0.69, 2.00]	1.000
Platelet count	265	0.89 [0.67, 1.15]	1.000	0.89 [0.67, 1.15]	1.000
Diastolic blood pressure	253	0.89 [0.67, 1.17]	1.000	0.91 [0.68, 1.20]	1.000
Lymphocyte count	263	1.08 [0.85, 1.37]	1.000	1.08 [0.85, 1.38]	1.000
Hemoglobin	265	1.07 [0.83, 1.38]	1.000	1.07 [0.83, 1.40]	1.000
Systolic blood pressure	253	0.90 [0.69, 1.17]	1.000	0.93 [0.71, 1.21]	1.000

Notes.
Odds ratios for 1 SD change in each blood measure at admission (tested in separate models)
Model 1 adjusted for age and sex and ethnicity. Model 2 additionally adjusted for comorbidities (hypertension, diabetes, heart diseases, respiratory diseases and chronic kidney disease)

Supplementary Table 4: Internally validated predictive performance, adjusted for ethnicity for patients with information on ethnicity (n=285)

Included features		**Internally validated AUC**			**Sensitivity**	**Specificity**	**PPV**	**NPV**
		Mean	-2SD	+2SD				
1	NEWS2	0.663	0.641	0.648	0.256	0.889	0.582	0.665
2	NEWS2 + **D**	0.654	0.628	0.680	0.283	0.878	0.585	0.671
3	NEWS2 + D**BP**	0.722	0.693	0.750	0.432	0.805	0.571	0.702
4	NEWS2 + DBP**C**	0.710	0.681	0.740	0.434	0.794	0.559	0.700
5	NEWS2 + CRP + Neutrophil + eGFR + Albumin + Age	0.734	0.713	0.756	0.414	0.797	0.549	0.693

D = Age, sex
C = comorbidities (8 features)
B = bloods (10 features)
P = physiological parameters (7 features)

Supplementary Table 5: Internally validated predictive performance using XGBoost (Gradient Boosting Trees) (n=452)

AUC based on repeated, nested cross-validation (inner loop: 10-fold, 20 repeats; outer loop = 10-fold, 100 repeats).

	Included features	**Internally validated AUC**			Sensitivity	Specificity	PPV	NPV
		Mean	-2SD	+2SD				
1	NEWS2	0.646	0.626	0.666	0.364	0.880	0.624	0.718
2	NEWS2 + **D**	0.667	0.652	0.682	0.344	0.910	0.680	0.719
3	NEWS2 + D**BP**	0.728	0.700	0.755	0.452	0.837	0.601	0.739
4	NEWS2 + DBP**C**	0.719	0.693	0.745	0.428	0.839	0.591	0.731
5	NEWS2 + CRP + Neutrophil + eGFR+ + Albumin + Age	0.722	0.660	0.785	0.480	0.836	0.615	0.748

D = Age, sex
C = comorbidities (8 features)
B = bloods (10 features)
P = physiological parameters (7 features)

Supplementary Table 6: Comparison of training and held-out validation samples

	Training sample (n=452)		**Validation sample (n=256)**		*P*-value for test of difference between samples[1]
14-day outcome	N avail.	N (%)	N avail.	N (%)	
COVID-19 WHO Score 6-8 (ICU/death)	452	159 (35.2%)	256	59 (23.0%)	0.001
Demographics					
Age	452	67.0 [28.0]	256	67.9 [25.5]	0.822
Sex (male) N (%)	452	248 (54.9%)	256	137 (53.5%)	0.788
BAME N (%)	285	120 (42.1%)	206	86 (41.7%)	0.999
Comorbidities	N avail.	N (%)			*P*-value
Hypertension	452	243 (53.8%)	256	146 (57.0%)	0.446
Diabetes	452	154 (34.1%)	256	85 (33.2%)	0.879
Heart Failure	452	57 (12.6%)	256	24 (9.4%)	0.239
Ischaemic Heart Diseases	452	85 (18.8%)	256	43 (16.8%)	0.572
COPD	452	48 (10.6%)	256	30 (11.7%)	0.746
Asthma	452	65 (14.4%)	256	37 (14.5%)	0.999
Chronic Kidney Disease	452	84 (18.6%)	256	39 (15.2%)	0.304
Blood biomarkers	N avail.	Mean [IQR]			*P*-value
Albumin	322	37.1 [7.0]	219	36.4 [6.0]	0.079
ALT	184	54.8 [33.0]	105	42.8 [31.0]	0.889
CRP	419	93.5 [106.7]	224	97.7 [94.2]	0.341
Creatinine	420	121.7 [49.0]	226	147.1 [62.8]	0.190
Estimated GFR	334	63.7 [40.0]	225	59.7 [44.0]	0.076
Ferritin	122	1356.0 [1165.2]	78	1668.8 [1258.2]	0.702
Haemoglobin	419	125.1 [30.0]	226	125.3 [31.0]	0.919

Lymphocyte count	419	1.5 [0.7]	226	1.3 [0.6]	0.247
Neutrophil count	418	5.7 [3.5]	226	5.7 [3.9]	0.952
Neutrophil/lymphocyte ratio	418	6.8 [5.0]	226	6.8 [4.7]	0.387
Lymphocyte/CRP ratio	416	0.1 [0.0]	224	0.0 [0.0]	0.191
Platelet count	421	226.7 [103.0]	226	223.7 [124.2]	0.652
Troponin T	141	33.9 [29.0]	94	87.8 [45.2]	0.414
Physiological parameters	N avail.	Mean [IQR]			*P*-value
NEWS2 Total Score	401	2.5 [3.0]	253	2.7 [3.0]	0.283
Heart rate	405	85.4 [20.0]	254	85.3 [19.0]	0.894
Oxygen saturation	404	96.2 [3.0]	254	96.1 [3.0]	0.562
Respiration rate	405	19.8 [2.0]	254	20.4 [2.0]	0.161
GCS score	172	14.1 [1.0]	103	14.3 [1.0]	0.432
Systolic blood pressure	405	127.4 [29.0]	254	127.4 [25.0]	0.834
Diastolic blood pressure	405	72.7 [18.0]	254	72.7 [17.0]	0.721
Temperature	405	37.1 [0.9]	254	37.0 [0.7]	0.101

Notes.
[1] *Wilcoxon test for continuous variables; X^2 test for binary variables.*

Supplementary Table 7: Temporal external validation, using hold-out sample (n=256)

Included features	**AUC**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
NEWS2	0.700	0.305	0.939	0.600	0.819
NEWS2 + D**BP**	0.730	0.441	0.873	0.510	0.839
NEWS2 + CRP + Neutrophil + eGFR + Albumin + Age	0.730	0.458	0.873	0.519	0.843

D = Age, sex
C = comorbidities (8 features)
B = bloods (10 features)
P = physiological parameters (7 features)