EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study

Rishi K. Gupta, Michael Marks, Thomas H. A. Samuels, Akish Luintel, Tommy Rampling, Humayra Chowdhury, Matteo Quartagno, Arjun Nair, Marc Lipman, Ibrahim Abubakar, Maarten van Smeden, Wai Keong Wong, Bryan Williams, Mahdad Noursadeghi, on behalf of The UCLH COVID-19 Reporting Group

Please cite this article as: Gupta RK, Marks M, Samuels THA, *et al*. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study. *Eur Respir J* 2020; in press (https://doi.org/10.1183/13993003.03498-2020).

Title page

Title

Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study

Authors

Rishi K. Gupta[1,2], Michael Marks[2,3], Thomas H. A. Samuels[2], Akish Luintel[2], Tommy Rampling[2], Humayra Chowdhury[2], Matteo Quartagno[4], Arjun Nair[2], Marc Lipman[5], Ibrahim Abubakar[1], Maarten van Smeden[6], Wai Keong Wong[2], Bryan Williams[7,8] and Mahdad Noursadeghi[2,9] on behalf of The UCLH COVID-19 Reporting Group*

*Members listed in Acknowledgements

Affiliations

1. Institute for Global Health, University College London, London, UK
2. University College London Hospitals NHS Trust
3. Clinical Research Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK
4. MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK
5. UCL Respiratory, Division of Medicine, University College London, London, UK
6. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands
7. NIHR University College London Hospitals Biomedical Research Centre
8. University College London, London, UK
9. Division of Infection & Immunity, University College London, UK

Correspondence

Prof Mahdad Noursadeghi, Division of Infection & Immunity, Cruciform Building, University College London, London WC1E 6BT, United Kingdom. Telephone: +442031082128. Email: m.noursadeghi@ucl.ac.uk

Take home message

Oxygen saturation on room air and patient age are strong predictors of deterioration and mortality among hospitalised adults with COVID-19, respectively. None of the 22 prognostic models evaluated in this study add incremental value to these univariable predictors.

Abstract

Background:

The number of proposed prognostic models for COVID-19 is growing rapidly, but it is unknown whether any are suitable for widespread clinical implementation.

Methods:

We independently externally validated the performance candidate prognostic models, identified through a living systematic review, among consecutive adults admitted to hospital with a final diagnosis of COVID-19. We reconstructed candidate models as per original descriptions and evaluated performance for their original intended outcomes using predictors measured at admission. We assessed discrimination, calibration and net benefit, compared to the default strategies of treating all and no patients, and against the most discriminating predictor in univariable analyses.

Results:

We tested 22 candidate prognostic models among 411 participants with COVID-19, of whom 180 (43.8%) and 115 (28.0%) met the endpoints of clinical deterioration and mortality, respectively. Highest areas under receiver operating characteristic (AUROC) curves were achieved by the NEWS2 score for prediction of deterioration over 24 hours (0.78; 95% CI 0.73-0.83), and a novel model for prediction of deterioration <14 days from admission (0.78; 0.74-0.82). The most discriminating univariable predictors were admission oxygen saturation on room air for in-hospital deterioration (AUROC 0.76; 0.71-0.81), and age for in-hospital mortality (AUROC 0.76; 0.71-0.81). No prognostic model demonstrated consistently higher net benefit than these univariable predictors, across a range of threshold probabilities.

Conclusions:

Admission oxygen saturation on room air and patient age are strong predictors of deterioration and mortality among hospitalised adults with COVID-19, respectively. None of the prognostic models evaluated here offered incremental value for patient stratification to these univariable predictors.

Introduction

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), causes a spectrum of disease ranging from asymptomatic infection to critical illness. Among people admitted to hospital, COVID-19 has reported mortality of 21-33%, with 14-17% requiring admission to high dependency or intensive care units (ICU)[1–4]. Exponential surges in transmission of SARS-CoV-2, coupled with the severity of disease among a subset of those affected, pose major challenges to health services by threatening to overwhelm resource capacity[5]. Rapid and effective triage at the point of presentation to hospital is therefore required to facilitate adequate allocation of resources and to ensure that patients at higher risk of deterioration are managed and monitored appropriately. Importantly, prognostic models may have additional value in patient stratification for emerging drug therapies[6, 7].

As a result, there has been global interest in development of prediction models for COVID-19[8]. These include models aiming to predict a diagnosis of COVID-19, and prognostic models, aiming to predict disease outcomes. At the time of writing, a living systematic review has already catalogued 145 diagnostic or prognostic models for COVID-19[8]. Critical appraisal of these models using quality assessment tools developed specifically for prediction modelling studies suggests that the candidate models are poorly reported, at high risk of bias and over-estimation of their reported performance[8, 9]. However, independent evaluation of candidate prognostic models in unselected datasets has been lacking. It therefore remains unclear how well these proposed models perform in practice, or whether any are suitable for widespread clinical implementation. We aimed to address this knowledge gap by systematically evaluating the performance of proposed prognostic models, among consecutive patients hospitalised with a final diagnosis of COVID-19 at a single centre, when using predictors measured at the point of hospital admission.

Methods

Identification of candidate prognostic models

We used a published living systematic review to identify all candidate prognostic models for COVID-19 indexed in PubMed, Embase, Arxiv, medRxiv, or bioRxiv until 5th May 2020, regardless of underlying study quality[8]. We included models that aim to predict clinical deterioration or mortality among patients with COVID-19. We also included prognostic scores commonly used in clinical practice[10–12], but not specifically developed for COVID-19 patients, since these models may also be considered for use by clinicians to aid risk-stratification for patients with COVID-19. For each candidate model identified, we extracted predictor variables, outcome definitions (including time horizons), modelling approaches, and final model parameters from original publications, and contacted authors for additional information where required. We excluded scores where the underlying model parameters were not publicly available, since we were unable to reconstruct them, along with models for which included predictors were not available in our dataset. The latter included models that require computed tomography imaging or arterial blood gas sampling, since these investigations were not routinely performed among unselected patients with COVID-19 at our centre.

Study population

Our study is reported in accordance with transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidance for external validation studies[13]. We included consecutive adults admitted to University College Hospital London with a final diagnosis of PCR-confirmed (including all sample types) or clinically diagnosed COVID-19, between 1st February and 30th April 2020. Since we sought to use data from the point of hospital admission to predict outcomes, we excluded patients transferred in from other hospitals, and those with hospital-acquired COVID-19 (defined as 1st PCR swab sent >5 days from date of hospital admission, as a proxy for the onset of clinical suspicion of SARS-CoV-2 infection). Clinical COVID-19 diagnoses were made on the basis of manual record review by an infectious disease specialist, using clinical features, laboratory results and radiological appearances, in the absence of an alternative diagnosis. During the study period, PCR testing was performed on the basis of clinical suspicion, and no SARS-CoV-2 serology investigations were routinely performed.

Data sources and variables of interest

Data were collected by direct extraction from electronic health records, complemented by manual curation. Variables of interest in the dataset included: demographics (age, gender, ethnicity), comorbidities (identified through manual record review), clinical observations, laboratory measurements, radiology reports, and clinical outcomes. Each chest radiograph was reported by a single radiologist, who was provided with a short summary of the indication for the investigation at the time of request, reflecting routine clinical conditions. Chest radiographs were classified using British Society of Thoracic Imaging criteria, and using a modified version of the Radiographic Assessment of Lung Edema (RALE) score[14, 15]. For each predictor, measurements were recorded as part of routine clinical care. Where serial measurements were available, we included the measurement taken closest to the time of presentation to hospital, with a maximum interval between presentation and measurement of 24 hours.

Outcomes

For models that used ICU admission or death, or progression to 'severe' COVID-19 or death, as composite endpoints, we used a composite 'clinical deterioration' endpoint as the primary outcome. We defined clinical deterioration as initiation of ventilatory support (continuous positive airway pressure, non-invasive ventilation, high flow nasal cannula oxygen, invasive mechanical ventilation or extra-corporeal membrane oxygenation) or death, equivalent to World Health Organization Clinical Progression Scale $\geq$ 6[16]. This definition does not include standard oxygen therapy. We did not apply any temporal limits on (a) the minimum duration of respiratory support; or (b) the interval between presentation to hospital and the outcome. The rationale for this composite outcome is to make the endpoint more generalisable between centres, since hospital respiratory management algorithms may vary substantially. Defining the outcome based on level of support, as opposed to ward setting, also ensures that it is appropriate in the context of a pandemic, when treatments that would usually only be considered in an ICU setting may be administered in other environments due to resource constraints. Where models specified their intended time horizon in their original description, we used this timepoint in the primary analysis, in order to ensure unbiased assessment of model calibration. Where the intended time horizon was not specified, we assessed the model to predict in-hospital deterioration or mortality, as appropriate. All deterioration and mortality events were included, regardless of their clinical aetiology.

Participants were followed-up clinically to the point of discharge from hospital. We extended follow-up beyond discharge by cross-checking NHS spine records to identify reported deaths post-discharge, thus ensuring >30 days' follow-up for all participants.

Statistical analyses

For each prognostic model included in the analyses, we reconstructed the model according to authors' original descriptions, and sought to evaluate the model discrimination and calibration performance against our approximation of their original intended endpoint. For models that provide online risk calculator tools, we validated our reconstructed models against original authors' models, by cross-checking our predictions against those generated by the web-based tools for a random subset of participants.

For all models, we assessed discrimination by quantifying the area under the receiver operating characteristic curve (AUROC)[17]. For models that provided outcome probability scores, we assessed calibration by visualising calibration of predicted *vs.* observed risk using loess-smoothed plots, and by quantifying calibration slopes and calibration-in-the-large (CITL). A perfect calibration slope should be 1; slopes <1 indicate that risk estimates are too extreme, while slopes >1 reflect risk estimates not being extreme enough. Ideal CITL is 0; CITL>0 indicates that predictions are systematically too low, while CITL<0 indicates that predictions are too high. For models with points-based scores, we assessed calibration visually by plotting model scores *vs.* actual outcome proportions. For models that provide probability estimates, but where the model intercept was not available, we calibrated the model to our dataset by calculating the intercept when using the model linear predictor as an offset term, leading to perfect CITL. This approach, by definition, overestimated calibration with respect to CITL, but allowed us to examine the calibration slope in our dataset.

We also assessed the discrimination of each candidate model for standardised outcomes of: (a) our composite endpoint of clinical deterioration; and (b) mortality, across a range of pre-specified time horizons from admission (7 days, 14 days, 30 days and any time during hospital admission), by calculating time-dependent AUROCs (with cumulative sensitivity and dynamic specificity)[18]. The rationale for this analysis was to harmonise endpoints, in order to facilitate more direct comparisons of discrimination between the candidate models.

In order to further benchmark the performance of candidate prognostic models, we then computed AUROCs for a limited number of univariable predictors considered to be of highest importance *a priori*, based on clinical knowledge and existing data, for prediction of our composite endpoints of clinical deterioration and mortality (7 days, 14 days, 30 days and any time during hospital admission). The *a priori* predictors of interest examined in this analysis were age, clinical frailty scale, oxygen saturation at presentation on room air, C-reactive protein and absolute lymphocyte count[8, 19].

Decision curve analysis allows assessment of the clinical utility of candidate models, and is dependent on both model discrimination and calibration[20]. We performed decision curve analyses to quantify the net benefit achieved by each model for predicting the intended endpoint, in order to inform clinical decision making across a range of risk:benefit ratios for an intervention or 'treatment'[20]. In this approach, the risk:benefit ratio is analogous to the cut point for a statistical model above which the intervention would be considered beneficial (deemed the 'threshold probability'). Net benefit was calculated as sensitivity × prevalence – (1 – specificity) × (1 – prevalence) × w where w is the odds at the threshold probability and the prevalence is the proportion of patients who experienced the outcome[20]. We calculated net benefit across a range of clinically relevant threshold probabilities, ranging from 0 to 0.5, since the risk:benefit ratio may vary for any given intervention (or 'treatment'). We compared the utility of each candidate model against strategies of treating all and no patients, and against the best performing univariable predictor for in-hospital clinical deterioration, or mortality, as appropriate. To ensure that fair, head-to-head net benefit comparisons were made between multivariable probability based models, points score models and univariable predictors, we calibrated each of these to the validation dataset for the purpose of decision curve analysis. Probability-based models were recalibrated to the validation data by refitting logistic regression models with the candidate model linear predictor as the sole predictor. We calculated 'delta' net benefit as net benefit when using the index model minus net benefit when: (a) treating all patients; and (b) using most discriminating univariable predictor. Decision curve analyses were done using the *rmda* package in R[21].

We handled missing data using multiple imputation by chained equations[22], using the *mice* package in R[23]. All variables and outcomes in the final prognostic models were included in the imputation

model to ensure compatibility[22] . A total of 10 imputed datasets were generated; discrimination, calibration and net benefit metrics were pooled using Rubin's rules[24].

All analyses were conducted in R (version 3.5.1).

Sensitivity analyses

We recalculated discrimination and calibration parameters for each candidate model using (a) a complete case analysis (in view of the large amount of missingness for some models); (b) excluding patients without PCR-confirmed SARS-CoV-2 infection; and (c) excluding patients who met the clinical deterioration outcome within 4 hours of arrival to hospital. We also examined for non-linearity in the *a priori* univariable predictors using restricted cubic splines, with 3 knots. Finally, we estimated optimism for discrimination and calibration parameters for the *a priori* univariable predictors using bootstrapping (1,000 iterations), using the *rms* package in R[25].

Ethical approval

The pre-specified study protocol was approved by East Midlands - Nottingham 2 Research Ethics Committee (REF: 20/EM/0114; IRAS: 282900).

Results

Summary of candidate prognostic models

We identified a total of 37 studies describing prognostic models, of which 19 studies (including 22 unique models) were eligible for inclusion (Supplementary Figure 1 and Table 1). Of these, 5 models were not specific to COVID-19, but were developed as prognostic scores for emergency department attendees[26], hospitalised patients[12, 27], people with suspected infection[10] or community-acquired pneumonia[11], respectively. Of the 17 models developed specifically for COVID-19, most (10/17) were developed using datasets originating in China. Overall, discovery populations included hospitalised patients and were similar to the current validation population with the exception of one study that discovered a model using community data[28], and another that used simulated data[29]. A total of 13/22 models use points-based scoring systems to derive final model scores, with the remainder using logistic regression modelling approaches to derive probability estimates. A total of 12/22 prognostic models primarily aimed to predict clinical deterioration, while the remaining 10 sought to predict mortality alone. When specified, time horizons for prognosis ranged from 1 to 30 days. Candidate prognostic models not included in the current validation study are summarised in Supplementary Table 1.

Overview of study cohort

During the study period, 521 adults were admitted with a final diagnosis of COVID-19, of whom 411 met the eligibility criteria for inclusion (flowchart shown in Supplementary Figure 2). Median age of the cohort was 66 years (interquartile range (IQR) 53-79), and the majority were male (252/411; 61.3%). Table 2 shows the baseline demographics, comorbidities, laboratory results and clinical measurements of the study cohort, of whom most (370/411; 90.0%) had PCR-confirmed SARS-CoV-2 infection (315/370 (85.1%) were positive on their first PCR test). A total of 180 (43.8%) and 115 (28.0%) of participants met the endpoints of clinical deterioration and mortality, respectively, above the minimum requirement of 100 events recommended for external validation studies [30]. The risks of clinical deterioration and death declined with time since admission (median days to deterioration 1.4 (IQR 0.3-4.2); median days to death 6.6 (IQR 3.6-13.1); Supplementary Figure 3). Most variables required for calculation of the 22 prognostic model scores were available among the vast majority of participants. However, admission lactate dehydrogenase was only available for 183/411 (44.5%) and

D-dimer measured for 153/411 (37.2%), resulting in significant missingness for models requiring these variables (Supplementary Figure 4).

Evaluation of prognostic models for original primary outcomes

Table 3 shows discrimination and calibration metrics, where appropriate, for the 22 evaluated prognostic models in the primary multiple imputation analysis. The highest AUROCs were achieved by the NEWS2 score for prediction of deterioration over 24 hours (0.78; 95% CI 0.73 - 0.83), and the Carr 'final' model for prediction of deterioration over 14 days (0.78; 95% CI 0.74 - 0.82). Of the other prognostic scores currently used in routine clinical practice, CURB65 had an AUROC 0.75 for 30-day mortality (95% CI 0.70 - 0.80), while qSOFA discriminated in-hospital mortality with an AUROC of 0.6 (95% CI 0.55 - 0.65).

For all models that provide probability scores for either deterioration or mortality, calibration appeared visually poor with evidence of overfitting and either systematic overestimation or underestimation of risk (Figure 1). Supplementary Figure 5 shows associations between prognostic models with points-based scores and actual risk. In addition to demonstrating reasonable discrimination, the NEWS2 and CURB65 models demonstrated approximately linear associations between scores and actual probability of deterioration at 24 hours and mortality at 30 days, respectively.

Time-dependent discrimination of candidate models and a priori univariable predictors for standardised outcomes

Next, we sought to compare the discrimination of these models for both clinical deterioration and mortality across the range of time horizons, benchmarked against preselected univariable predictors associated with adverse outcomes in COVID-19[8, 19]. We recalculated time-dependent AUROCs for each of these outcomes, stratified by time horizon to the outcome (Supplementary Figures 6 and 7). These analyses showed that AUROCs generally declined with increasing time horizons. Admission oxygen saturation on room air was the strongest predictor of in-hospital deterioration (AUROC 0.76; 95% CI 0.71-0.81), while age was the strongest predictor of in-hospital mortality (AUROC 0.76; 95% CI 0.71-0.81).

Decision curve analyses to assess clinical utility

We compared net benefit for each prognostic model (for its original intended endpoint) to the strategies of treating all patients, treating no patients, and using the most discriminating univariable

predictor for either deterioration (i.e. oxygen saturation on air) or mortality (i.e. patient age) to stratify treatment (Supplementary Figure 8). Although all prognostic models showed greater net benefit than treating all patients at the higher range of threshold probabilities, none of these models demonstrated consistently greater net benefit than the most discriminating univariable predictor, across the range of threshold probabilities (Figure 2).

Sensitivity analyses

Recalculation of model discrimination and calibration metrics for prediction of the original intended endpoint using (a) a complete case analysis; (b) excluding patients without PCR-confirmed SARS-CoV-2 infection; and (c) excluding patients who met the clinical deterioration outcome within 4 hours of arrival to hospital revealed similar results to the primary multiple imputation approach, though discrimination was noted to be lower overall when excluding early events (Supplementary Tables 2a-c). Visual examination of associations between the most discriminating univariable predictors and log odds of deterioration or death using restricted cubic splines showed no evidence of non-linear associations (Supplementary Figure 9). Finally, internal validation using bootstrapping showed near zero optimism for discrimination and calibration parameters for the univariable models (Supplementary Table).

Discussion

In this observational cohort study of consecutive adults hospitalised with COVID-19, we systematically evaluated the performance of 22 prognostic models for COVID-19. These included models developed specifically for COVID-19, along with existing scores in routine clinical use prior to the pandemic. For prediction of both clinical deterioration or mortality, AUROCs ranged from 0.56-0.78. NEWS2 performed reasonably well for prediction of deterioration over a 24-hour interval, achieving an AUROC of 0.78, while the Carr 'final' model[31] also had an AUROC of 0.78, but tended to systematically underestimate risk. All COVID-specific models that derived an outcome probability of either deterioration or mortality showed poor calibration. We found that oxygen saturation (AUROC 0.76) and patient age (AUROC 0.76) were the most discriminating single variables for prediction of in-hospital deterioration and mortality respectively. These predictors have the added advantage that they are immediately available at the point of presentation to hospital. In decision curve analysis, which is dependent upon both model discrimination and calibration, no prognostic model demonstrated clinical utility consistently greater than using these univariable predictors to inform decision-making.

While previous studies have largely focused on novel model discovery, or evaluation of a limited number of existing models, this is the first study to our knowledge to evaluate systematically-identified candidate prognostic models for COVID-19. We used a comprehensive living systematic review[8] to identify eligible models and sought to reconstruct each model as per the original authors' description. We then evaluated performance against its intended outcome and time horizon, wherever possible, using recommended methods of external validation incorporating assessments of discrimination, calibration and net benefit[17]. Moreover, we used a robust approach of electronic health record data capture, supported by manual curation, in order to ensure a high-quality dataset, and inclusion of unselected and consecutive COVID-19 cases that met our eligibility criteria. In addition, we used robust outcome measures of mortality and clinical deterioration, aligning with the WHO Clinical Progression Scale[16].

A weakness of the current study is that it is based on retrospective data from a single centre, and therefore cannot assess between-setting heterogeneity in model performance. Second, due to the limitations of routinely collected data, predictor variables were available for varying numbers of participants for each model, with a large proportion of missingness for models requiring lactate

dehydrogenase and D-dimer measurements. We therefore performed multiple imputation, in keeping with recommendations for development and validation of multivariable prediction models, in our primary analyses[32]. Findings were similar in the complete case sensitivity analysis, thus supporting the robustness of our results. Future studies would benefit from standardising data capture and laboratory measurements prospectively to minimise predictor missingness. Thirdly, a number of models could not be reconstructed in our data. For some models, this was due the absence of predictors in our dataset, such as those requiring computed tomography imaging, since this is not currently routinely recommended for patients with suspected or confirmed COVID-19[15]. We were also not able to include models for which the parameters were not publicly available. This underscores the need for strict adherence to reporting standards in multivariable prediction models[13]. Finally, we used admission data only as predictors in this study, since most prognostic scores are intended to predict outcomes at the point of hospital admission. We note, however, that some scores are designed for dynamic in-patient monitoring, with NEWS2 showing reasonable discrimination for deterioration over a 24-hour interval, as originally intended[27]. Future studies may integrate serial data to examine model performance when using such dynamic measurements.

Despite the vast global interest in the pursuit of prognostic models for COVID-19, our findings show that none of the COVID-19-specific models evaluated in this study can currently be recommended for routine clinical use. In addition, while some of the evaluated models that are not specific to COVID-19 are routinely used and may be of value among in-patients[12, 27], people with suspected infection[10] or community-acquired pneumonia[11], none showed greater clinical utility than the strongest univariable predictors among patients with COVID-19. Our data show that admission oxygen saturation on air is a strong predictor of clinical deterioration and may be evaluated in future studies to stratify in-patient management and for remote community monitoring. We note that all novel prognostic models for COVID-19 assessed in the current study were derived from single-centre data. Future studies may seek to pool data from multiple centres in order to robustly evaluate the performance of existing and newly emerging models across heterogeneous populations, and develop and validate novel prognostic models, through individual participant data meta-analysis[33]. Such an approach would allow assessments of between-study heterogeneity and the likely generalisability of candidate models. It is also imperative that discovery populations are representative of target populations for model implementation, with inclusion of unselected cohorts. Moreover, we strongly

advocate for transparent reporting in keeping with TRIPOD standards (including modelling approaches, all coefficients and standard errors) along with standardisation of outcomes and time horizons, in order to facilitate ongoing systematic evaluations of model performance and clinical utility[13].

We conclude that baseline oxygen saturation on room air and patient age are strong predictors of deterioration and mortality, respectively. None of the prognostic models evaluated in this study offer incremental value for patient stratification to these univariable predictors when using admission data. Therefore, none of the evaluated prognostic models for COVID-19 can be recommended for routine clinical implementation. Future studies seeking to develop prognostic models for COVID-19 should consider integrating multi-centre data in order to increase generalisability of findings, and should ensure benchmarking against existing models and simpler univariable predictors.

Footnotes

Acknowledgements

The UCLH COVID-19 Reporting Group was comprised of the following individuals, who were involved in data curation as non-author contributors: Asia Ahmed, Ronan Astin, Malcolm Avari, Elkie Benhur, Anisha Bhagwanani, Timothy Bonnici, Sean Carlson, Jessica Carter, Sonya Crowe, Mark Duncan, Ferran Espuny-Pujol, James Fullerton, Marc George, Georgina Harridge, Ali Hosin, Rachel Hubbard, Adnan Hubraq, Prem Jareonsettasin, Zella King, Avi Korman, Sophie Kristina, Lawrence Langley, Jacques-Henri Meurgey, Henrietta Mills, Alfio Missaglia, Ankita Mondal, Samuel Moulding, Christina Pagel, Liyang Pan, Shivani Patel, Valeria Pintar, Jordan Poulos, Ruth Prendecki, Alexander Procter, Magali Taylor, David Thompson, Lucy Tiffen, Hannah Wright, Luke Wynne, Jason Yeung, Claudia Zeicu, Leilei Zhu

Author contributions

RKG and MN conceived the study. RKG conducted the analysis and wrote the first draft of the manuscript. All other authors contributed towards data collection, study design and/or interpretation. All authors have critically appraised and approved the final manuscript prior to submission. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Members of The UCLH COVID-19 Reporting contributed towards data curation and are non-author contributors/collaborators for this study.

Funding

The study was funded by National Institute for Health Research (DRF-2018-11-ST2-004 to RKG; NF-SI-0616-10037 to IA), the Wellcome Trust (207511/Z/17/Z to MN) and has been supported by the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre, in particular by the NIHR UCLH/UCL BRC Clinical and Research Informatics Unit.

This paper presents independent research supported by the NIHR. 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 funder had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Declaration of interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: non-financial support from AIDENCE BV (Dr Nair), outside the submitted work; no support from any organisation outside those declared above for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Data sharing statement

The conditions of regulatory approvals for the present study preclude open access data sharing to minimise 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 GDPR regulations.

References

1. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, Barnaby DP, Becker LB, Chelico JD, Cohen SL, Cookingham J, Coppa K, Diefenbach MA, Dominello AJ, Duer-Hefele J, Falzon L, Gitlin J, Hajizadeh N, Harvin TG, Hirschwerk DA, Kim EJ, Kozel ZM, Marrast LM, Mogavero JN, Osorio GA, Qiu M, Zanos TP. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* 2020; .
2. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, Holden KA, Read JM, Dondelinger F, Carson G, Merson L, Lee J, Plotkin D, Sigfrid L, Halpin S, Jackson C, Gamble C, Horby PW, Nguyen-Van-Tam JS, Ho A, Russell CD, Dunning J, Openshaw PJ, Baillie JK, Semple MG, ISARIC4C investigators. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* British Medical Journal Publishing Group; 2020; 369: m1985.
3. Grasselli G, Pesenti A, Cecconi M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy. *JAMA* American Medical Association; 2020; 323: 1545.
4. Imperial College COVID-19 response team. Report 17 - Clinical characteristics and predictors of outcomes of hospitalised patients with COVID-19 in a London NHS Trust: a retrospective cohort study | Faculty of Medicine | Imperial College London [Internet]. 2020 [cited 2020 May 14].Available from: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-17-clinical/.
5. Li R, Rivers C, Tan Q, Murray MB, Toner E, Lipsitch M. The demand for inpatient and ICU beds for COVID-19 in the US: lessons from Chinese cities. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.09.20033241.
6. Beigel JH, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, Hohmann E, Chu HY, Luetkemeyer A, Kline S, Lopez de Castilla D, Finberg RW, Dierberg K, Tapson V, Hsieh L, Patterson TF, Paredes R, Sweeney DA, Short WR, Touloumi G, Lye DC, Ohmagari N, Oh M, Ruiz-Palacios GM, Benfield T, Fätkenheuer G, Kortepeter MG, Atmar RL, Creech CB, Lundgren J, et al. Remdesivir for the Treatment of Covid-19 — Preliminary Report. *N. Engl. J. Med.* Massachusetts Medical Society; 2020; : NEJMoa2007764.
7. Horby P, Lim WS, Emberson J, Mafham M, Bell J, Linsell L, Staplin N, Brightling C, Ustianowski A, Elmahi E, Prudon B, Green C, Felton T, Chadwick D, Rege K, Fegan C, Chappell LC, Faust SN, Jaki T, Jeffery K, Montgomery A, Rowan K, Juszczak E, Baillie JK, Haynes R, Landray MJ, Group RC. Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.06.22.20137273.
8. Wynants L, Calster B Van, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Damen JAA, Debray TPA, Vos M De, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Kreuzberger N, Lohmann A, Luijken K, Ma J, Navarro CLA, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Kuijk SMJ van, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* British Medical Journal Publishing Group; 2020; 369.
9. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* 2019; 170: 51.
10. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS, Escobar GJ, Angus DC. Assessment of Clinical Criteria for Sepsis. *JAMA* American Medical Association; 2016; 315: 762.
11. Lim WS, Eerden MM van der, Laing R, Boersma WG, Karalus N, Town GI, Lewis SA, Macfarlane JT. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* BMJ Publishing Group Ltd; 2003; 58: 377–382.
12. Royal College of Physicians. National Early Warning Score (NEWS) 2 | RCP London [Internet]. [cited 2020 Jul 1].Available from: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2.
13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* British Medical Journal Publishing Group; 2015; 350.
14. Wong HYF, Lam HYS, Fong AH-T, Leung ST, Chin TW-Y, Lo CSY, Lui MM-S, Lee JCY, Chiu KW-H, Chung T, Lee EYP, Wan EYF, Hung FNI, Lam TPW, Kuo M, Ng M-Y. Frequency and

Distribution of Chest Radiographic Findings in COVID-19 Positive Patients. *Radiology* Radiological Society of North America ; 2019; : 201160.
15. COVID-19 Resources | The British Society of Thoracic Imaging [Internet]. [cited 2020 Jul 1].Available from: https://www.bsti.org.uk/covid-19-resources/.
16. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection JC, Murthy S, Diaz J, Adhikari N, Angus DC, Arabi YM, Baillie K, Bauer M, Berry S, Blackwood B, Bonten M, Bozza F, Brunkhorst F, Cheng A, Clarke M, Dat VQ, Jong M de, Denholm J, Derde L, Dunning J, Feng X, Fletcher T, Foster N, Fowler R, Gobat N, Gomersall C, Gordon A, Glueck T, Harhay M, Hodgson C, et al. A minimal common outcome measure set for COVID-19 clinical research. *Lancet. Infect. Dis.* Elsevier; 2020; 0.
17. Riley RD, Windt D van der, Croft P, Moons KGM. Prognosis research in healthcare : concepts, methods, and impact. .
18. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* BMC Med Res Methodol; 2017; 17: 53.
19. Hewitt J, Carter B, Vilches-Moraga A, Quinn TJ, Braude P, Verduri A, Pearce L, Stechman M, Short R, Price A, Collins JT, Bruce E, Einarsson A, Rickard F, Mitchell E, Holloway M, Hesford J, Barlow-Pay F, Clini E, Myint PK, Moug SJ, McCarthy K, COPE Study Collaborators C, Jones S, Lunstone K, Cavenagh A, Silver C, Telford T, Simmons R, Mutasem TEJ, et al. The effect of frailty on survival in patients with COVID-19 (COPE): a multicentre, European, observational cohort study. *Lancet. Public Heal.* Elsevier; 2020; 0.
20. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic Progn. Res.* BioMed Central; 2019; 3: 18.
21. Brown M. rmda: Risk Model Decision Analysis. 2018.
22. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* John Wiley & Sons, Ltd; 2011; 30: 377–399.
23. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 2011; 45(3): 1–67.
24. Rubin DB. Multiple imputation for nonresponse in surveys. Wiley-Interscience; 2004.
25. Harrell Jr FE. rms: Regression Modeling Strategies. 2019.
26. T O, A T, L L. Rapid Emergency Medicine Score: A New Prognostic Tool for In-Hospital Mortality in Nonsurgical Emergency Department Patients. *J. Intern. Med.* J Intern Med; 2004; 255.
27. 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* Elsevier; 2013; 84: 465–470.
28. Bello-Chavolla OY, Bahena-López JP, Antonio-Villa NE, Vargas-Vázquez A, González-Díaz A, Márquez-Salinas A, Fermín-Martínez CA, Naveja JJ, Aguilar-Salinas CA. Predicting mortality due to SARS-CoV-2: A mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico. *J. Clin. Endocrinol. Metab.* 2020; .
29. Caramelo F, Ferreira N, Oliveiros B. Estimation of risk factors for COVID-19 mortality - preliminary results. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.02.24.20027268.
30. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat. Med.* Stat Med; 2016; 35: 214–226.
31. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, Searle T, Kraljevic Z, Shek A, Phan HTT, Muruet W, Shinton AJ, Shi T, Zhang X, Pickles A, Stahl D, Zakeri R, O'Gallagher K, Folarin A, Roguski L, Borca F, Batchelor J, Wu X, Sun J, Pinto A, Guthrie B, Breen C, Douiri A, Wu H, Curcin V, et al. Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *medRxiv* 2020; .
32. KG M, DG A, JB R, JP I, P M, EW S, AJ V, DF R, GS C. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann. Intern. Med.* Ann Intern Med; 2015; 162.
33. Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPD Meta-analysis Methods group. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLOS Med.* 2015; 12: e1001886.
34. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* Oxford Academic; 2001; 94: 521–526.

35. Colombi D, Bodini FC, Petrini M, Maffi G, Morelli N, Milanese G, Silva M, Sverzellati N, Michieletti E. Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology* Radiological Society of North America ; 2020; : 201433.
36. Galloway JB, Norton S, Barker RD, Brookes A, Carey I, Clarke BD, Jina R, Reid C, Russell MD, Sneep R, Sugarman L, Williams S, Yates M, Teo J, Shah AM, Cantle F. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: An observational cohort study. *J. Infect.* W.B. Saunders; 2020; .
37. Guo Y, Liu Y, Lu J, Fan R, Zhang F, Yin X, Liu Z, Zeng Q, Yuan J, Hu S, Wang Q, Liao B, Huang M, Yin S, Zhang X, Xin R, Lin Z, Hu C, Zhao B, He R, Liang M, Zhang Z, Liu L, Sun J, Tang L, Deng L, Xia J, Tang X, Liu L, Hou J. Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.17.20064691.
38. Cambridge Clinical Trials Unit. TACTIC trial [Internet]. [cited 2020 Jul 1].Available from: https://cctu.org.uk/portfolio/COVID-19/TACTIC.
39. Chen X, Liu Z. Early prediction of mortality risk among severe COVID-19 patients using machine learning. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.13.20064329.
40. Huang H, Cai S, Li Y, Li Y, Fan Y, Li L, Lei C, Tang X, Hu F, Li F, Deng X. Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.28.20045989.
41. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, Chen G, Cheng G, Wang Y, Bi J, Tan L, Lau G, Qin E. Prediction for Progression Risk in Patients with COVID-19 Pneumonia: the CALL Score. *Clin. Infect. Dis.* 2020; .
42. Lu J, Hu S, Fan R, Liu Z, Yin X, Wang Q, Lv Q, Cai Z, Li H, Hu Y, Han Y, Hu H, Gao W, Feng S, Liu Q, Li H, Sun J, Peng J, Yi X, Zhou Z, Guo Y, Hou J. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.02.20.20025510.
43. 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* 24: 108.
44. Xie J, Hungerford D, Chen H, Abrams ST, Li S, Wang G, Wang Y, Kang H, Bonnett L, Zheng R, Li X, Tong Z, Du B, Qiu H, Toh C-H. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.28.20045997.
45. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M, Huang X, Xiao Y, Cao H, Chen Y, Ren T, Wang F, Xiao Y, Huang S, Tan X, Huang N, Jiao B, Cheng C, Zhang Y, Luo A, Mombaerts L, Jin J, Cao Z, Li S, Xu H, Yuan Y. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* Nature Publishing Group; 2020; 2: 283–288.
46. Zhang H, Shi T, Wu X, Zhang X, Wang K, Bean D, Dobson R, Teo JT, Sun J, Zhao P, Li C, Dhaliwal K, Wu H, Li Q, Guthrie B. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.28.20082222.
47. Hu H, Yao N, Qiu Y. Comparing Rapid Scoring Systems in Mortality Prediction of Critically Ill Patients With Novel Coronavirus Disease. Burton JH, editor. *Acad. Emerg. Med.* John Wiley & Sons, Ltd; 2020; 27: 461–468.

Table 1: Characteristics of studies describing prognostic models included in systematic evaluation.

MEWS = modified early warning score; qSOFA = quick sequential (sepsis-related) organ failure assessment; REMS = rapid emergency medicine score; NEWS = national early warning score; TACTIC = therapeutic study in pre-ICU patients admitted with COVID-19; AVPU = Alert / responds to voice / responsive to pain / unresponsive; CRP = C-reactive protein; LDH = lactate dehydrogenase; RALE = radiographic assessment of lung edema; ARDS = acute respiratory distress syndrome; ICU = intensive care unit; ECMO = extra-corporeal membrane oxygenation.

Units, unless otherwise specified, are: age in years; respiratory rate in breaths per minute; heart rate in beats per minute; blood pressure in mmHg; temperature in °C; oxygen saturation in %; CRP in mg/L; LDH in U/L; neutrophils, lymphocytes, total white cell count and platelets x 10^9/L; D-dimer in ng/mL; creatinine in µmol/L; estimated glomerular filtration rate in mL/min/1.73 m2, albumin in g/L.

Authors	Score name	Country of derivation	Development population	Pre-existing or COVID-specific?	Model outcome	Predictors	Original modelling approach	How are predictors combined?
Subbe et al[34]	MEWS#	UK	Hospital inpatients	Pre-existing (hospital patients)	Mortality, ICU admission or cardiac arrest (no specified timepoint)	Systolic blood pressure, pulse rate, respiratory rate, temperature, AVPU score	Clinical consensus	Points-based score
Olsson et al[26]	REMS#	Sweden	Patients presenting to emergency department	Pre-existing (emergency department patients)	Mortality (in-hospital)	Blood pressure, respiratory rate, pulse rate, Glasgow coma scale, oxygen saturation, age	Logistic regression	Points-based score
Seymour et al[10]	qSOFA	USA	Electronic health record encounters	Pre-existing (suspected infection)	Mortality (in-hospital)	Systolic hypotension [≤100 mm Hg], tachypnoea [≥22/min], altered mentation	Logistic regression	Points-based score
Lim et al[11]	CURB65	UK, New Zealand, Netherlands	Patients with community acquired pneumonia	Pre-existing (community-acquired pneumonia)	Mortality (30 days)	Confusion, urea >7 mmol/L, respiratory rate >30/min, low systolic (<90 mm Hg) or diastolic (<60 mm Hg) blood pressure), age > 65 years	Logistic regression	Points-based score
Royal College of Physicians[12]	NEWS2^	UK	Hospital admissions	Pre-existing (hospital patients)	Mortality, ICU admission or cardiac arrest (24h)	Respiratory rate, oxygen saturation, systolic blood pressure, pulse rate, level of consciousness or new confusion, temperature	Clinical consensus	Points-based score
Bello-Chavolla et al[28]	Bello-Chavolla	Mexico	Confirmed COVID-19 patients	COVID-specific	Mortality (30 day)	Age ≥65 years, diabetes, early-onset diabetes, obesity, age <40	Cox regression	Points-based score

			presenting in primary care			years, chronic kidney disease, hypertension, immunosuppression (rheumatoid arthritis, lupus, HIV or immunosuppressive drugs)		
Caramelo et al[29]	Caramelo[$]	Simulated data	Simulated data	COVID-specific	Mortality (period unspecified)	Age, hypertension, diabetes, cardiovascular disease, chronic respiratory disease, cancer	Logistic regression	Logistic regression
Carr et al[31]	'Carr final', 'Carr threshold'	UK	Inpatients with confirmed COVID-19	COVID-specific	ICU admission or death (14 days from symptom onset)	NEWS2, CRP, neutrophils, estimated glomerular filtration rate, albumin, age	Regularized logistic regression with LASSO estimator	Regularized logistic regression
Colombi et al[35]	Colombi_clinical[$] (clinical model only)	Italy	Inpatients with confirmed COVID-19	COVID-specific	ICU admission or in-hospital mortality (period unspecified)	Age > 68 years, cardiovascular disease, CRP > 76 mg/L, LDH > 347 U/L, platelets > 180 x 10^9/L	Logistic regression	Logistic regression
Galloway et al[36]	Galloway	UK	Inpatients with confirmed COVID-19	COVID-specific	ICU admission or death during admission	Modified RALE score >3, oxygen saturation < 93%, creatinine > 100 µmol/L, neutrophils > 8 x 10^9/L, age > 40 years, chronic lung disease, CRP > 40 mg/L, albumin < 34g/L, male gender, non-white ethnicity, hypertension, diabetes.	Logistic regression (LASSO)	Points-based score
Guo et al[37]	Guo	China	Inpatients with confirmed COVID-19	COVID-specific	Deterioration within 14 days of admission	Age >50, underlying chronic disease (not defined), neutrophil/lymphocyte ratio > 5, CRP > 25 mg/L, d-dimer > 800 ng/mL	Cox regression	Points-based score
Hall et al[38]	TACTIC	UK	Inpatients with confirmed COVID-19	COVID-specific	Admission to ICU or death during admission	Modified RALE score >3, age >40 years, male sex, non-white ethnicity, diabetes, hypertension, neutrophils > 8 x 10^9/L, CRP > 40 mg/L	Logistic regression (LASSO)	Points-based score
Hu et al[39]	Hu	China	Inpatients with confirmed COVID-19	COVID-specific	Mortality (in-hospital)	Age, CRP, lymphocytes, d-dimer (µg/mL)	Logistic regression	Logistic regression
Huang et al[40]	Huang	China	Inpatients with confirmed COVID-19	COVID-specific	Progression to severe COVID (defined as respiratory rate ≥ 30, oxygen saturation ≤ 93% in the resting state or arterial blood oxygen partial pressure / oxygen concentration (FiO2) ≤ 300mmHg), 3-7 days from admission	CRP > 10 mg/L, LDH > 250 U/L, respiratory rate > 24/min, comorbidity (hypertension, coronary artery disease, diabetes, obesity, chronic obstructive pulmonary disease, chronic kidney disease, obstructive sleep apnoea)	Logistic regression	Logistic regression
Ji et al[41]	Ji	China	Inpatients with	COVID-specific	Progression to severe COVID-	Age (> 60 years), lymphocytes (≤1	Cox regression	Points-based

			confirmed COVID-19		19 at 10 days (defined as respiratory rate ≥ 30, resting oxygen saturation ≤ 93%, PaO2/FiO2 ≤ 300 mmHg, requirement of mechanical ventilation or worsening of lung CT findings)	x 10^9/L) LDH (<250, 250-500, >500 U/L), comorbidity (hypertension, diabetes, cardiovascular disease, chronic lung disease, or HIV)		score
Lu et al[42]	Lu	China	Inpatients with suspected or confirmed COVID-19	COVID-specific	Mortality (12 days)	Age ≥ 60 years, CRP ≥ 34 mg/L	Cox regression	Points-based score
Shi et al[43]	Shi	China	Inpatients with confirmed COVID-19	COVID-specific	Death or 'severe' COVID-19 (not defined) over unspecified period	Age>50 years, male sex, hypertension	Not specified	Points-based score
Xie et al[44]	Xie	China	Inpatients with confirmed COVID-19	COVID-specific	Mortality (in-hospital)	Age, lymphocytes, LDH, oxygen saturation	Logistic regression	Logistic regression
Yan et al[45]	Yan	China	Inpatients suspected of COVID-19	COVID-specific	Mortality (period unspecified)	LDH > 365 U/L, CRP > 41.2 mg/L, lymphocyte percentage > 14.7%	Decision-tree model with XG boost	Points-based score
Zhang et al[46]	'Zhang poor', 'Zhang death'	China	Inpatients with confirmed COVID-19	COVID-specific	Mortality and poor outcome (ARDS, intubation or ECMO, ICU admission) as separate models; no timepoint specified	Age, sex, neutrophils, lymphocytes, platelets, CRP, creatinine	Logistic regression (LASSO)	Logistic regression

[#]MEWS and REMS were evaluated among people with COVID-19 by Hu et al[47], and thus were included in the present study.

[$]No model intercept was available; the intercepts for these models were therefore calibrated to the validation dataset, using the model linear predictors as offset terms.

^Using oxygen scale 1 for all participants, except for those with target oxygen saturation ranges of 88–92%, e.g. in hypercapnic respiratory failure, when scale 2 is used, as recommended[12].

Table 2: Baseline characteristics of hospitalised adults with COVID-19 included in systematic evaluation cohort.

Laboratory and physiological measurements reflect parameters at the time of hospital admission. N column shows number of participants with available data for each variable. Data are shown as N (%) for categorical data or median (interquartile range (IQR)) for continuous variables.

Variable	n	Level	Overall
			411
Demographics			
Age (years)	411 (100)		66.0 [53.0, 79.0]
Gender	411 (100)	Female	159 (38.7)
		Male	252 (61.3)
Ethnicity	390 (94.9)	Asian	52 (13.3)
		Black	56 (14.4)
		White	234 (60.0)
		Mixed	7 (1.8)
		Other	41 (10.5)
Clinical frailty scale	411 (100)		2.0 [1.0, 6.0]
Comorbidities			
Hypertension	411 (100)		172 (41.8)
Chronic cardiovascular disease	410 (99.8)		108 (26.3)
Chronic respiratory disease	411 (100)		99 (24.1)
Diabetes	411 (100)		105 (25.5)
*Obesity^*	411 (100)		83 (20.2)
Chronic kidney disease	410 (99.8)		40 (9.8)
Laboratory measurements			
C-reactive protein (mg/L)	403 (98.1)		96.7 [45.2, 178.7]
Lymphocytes (x 10^9)	410 (99.8)		0.9 [0.6, 1.4]
Lactate dehydrogenase (U/L)	183 (44.5)		395.0 [309.0, 511.0]
D-dimer (ng/mL)	153 (37.2)		1070.0 [640.0, 2120.0]
SARS CoV-2 PCR	411 (100)		370 (90.0)
Physiological measurements			
Respiratory rate (per min)	410 (99.8)		24.0 [20.0, 28.0]
Heart rate (per min)	410 (99.8)		94.0 [81.2, 107.0]
Systolic blood pressure (mmHg)	411 (100)		131.0 [115.0, 143.0]
Oxygen saturation (%; on air)	403 (98.1)		91.0 [86.0, 95.0]
Outcome			
Deteriorated	411 (100)		180 (43.8)
Died	411 (100)		115 (28.0)

^Clinician-defined obesity.

Table 3: Validation metrics of prognostic scores for COVID-19, using primary multiple imputation analysis (n=411).

For each model, performance is evaluated for its original intended outcome, shown in 'Primary outcome' column. AUROC = area under the receiver operating characteristic curve; CI = confidence interval.

Score	Primary outcome	AUROC (95% CI)	Calibration slope (95% CI)	Calibration in the large (95% CI)
NEWS2	Deterioration (1 day)	0.78 (0.73 - 0.83)		
Ji	Deterioration (10 days)	0.56 (0.5 - 0.62)		
Carr_final	Deterioration (14 days)	0.78 (0.74 - 0.82)	1.04 (0.8 - 1.28)	0.33 (0.11 - 0.55)
Carr_threshold	Deterioration (14 days)	0.76 (0.71 - 0.81)	0.85 (0.65 - 1.05)	-0.34 (-0.57 - -0.12)
Guo	Deterioration (14 days)	0.67 (0.61 - 0.73)		
Zhang_poor	Deterioration (in-hospital)	0.74 (0.69 - 0.79)	0.33 (0.22 - 0.43)	0.56 (0.3 - 0.81)
Galloway	Deterioration (in-hospital)	0.72 (0.68 - 0.77)		
TACTIC	Deterioration (in-hospital)	0.7 (0.65 - 0.75)		
Colombi_clinical	Deterioration (in-hospital)	0.69 (0.63 - 0.74)	0.53 (0.35 - 0.71)	0 (-0.23 - 0.23)
Huang	Deterioration (in-hospital)	0.67 (0.61 - 0.73)	0.18 (0.1 - 0.26)	-4.26 (-4.61 - -3.91)
Shi	Deterioration (in-hospital)	0.61 (0.56 - 0.66)		
MEWS	Deterioration (in-hospital)	0.6 (0.54 - 0.65)		
Lu	Mortality (12 days)	0.72 (0.67 - 0.76)		
CURB65	Mortality (30 days)	0.75 (0.7 - 0.8)		
BelloChavolla	Mortality (30 days)	0.66 (0.6 - 0.72)		
REMS	Mortality (in-hospital)	0.76 (0.71 - 0.81)		
Xie	Mortality (in-hospital)	0.76 (0.69 - 0.82)	0.83 (0.51 - 1.15)	0.41 (0.16 - 0.66)
Hu	Mortality (in-hospital)	0.74 (0.68 - 0.79)	0.33 (0.2 - 0.45)	-1.07 (-1.37 - -0.77)
Caramelo	Mortality (in-hospital)	0.71 (0.66 - 0.76)	0.53 (0.36 - 0.69)	0 (-0.25 - 0.25)
Zhang_death	Mortality (in-hospital)	0.7 (0.65 - 0.76)	0.29 (0.19 - 0.4)	0.89 (0.6 - 1.19)
qSOFA	Mortality (in-hospital)	0.6 (0.55 - 0.65)		
Yan	Mortality (in-hospital)	0.58 (0.49 - 0.67)		

Figure 1: Calibration plots for prognostic models estimating outcome probabilities.

For each plot, the blue line represents a Loess-smoothed calibration curve from the stacked multiply imputed datasets and rug plots indicate the distribution of data points. No model intercept was available for the Caramelo or Colombi 'clinical' models; the intercepts for these models were calibrated to the validation dataset, by using the model linear predictors as offset terms. The primary outcome of interest for each model is shown in the plot sub-heading.

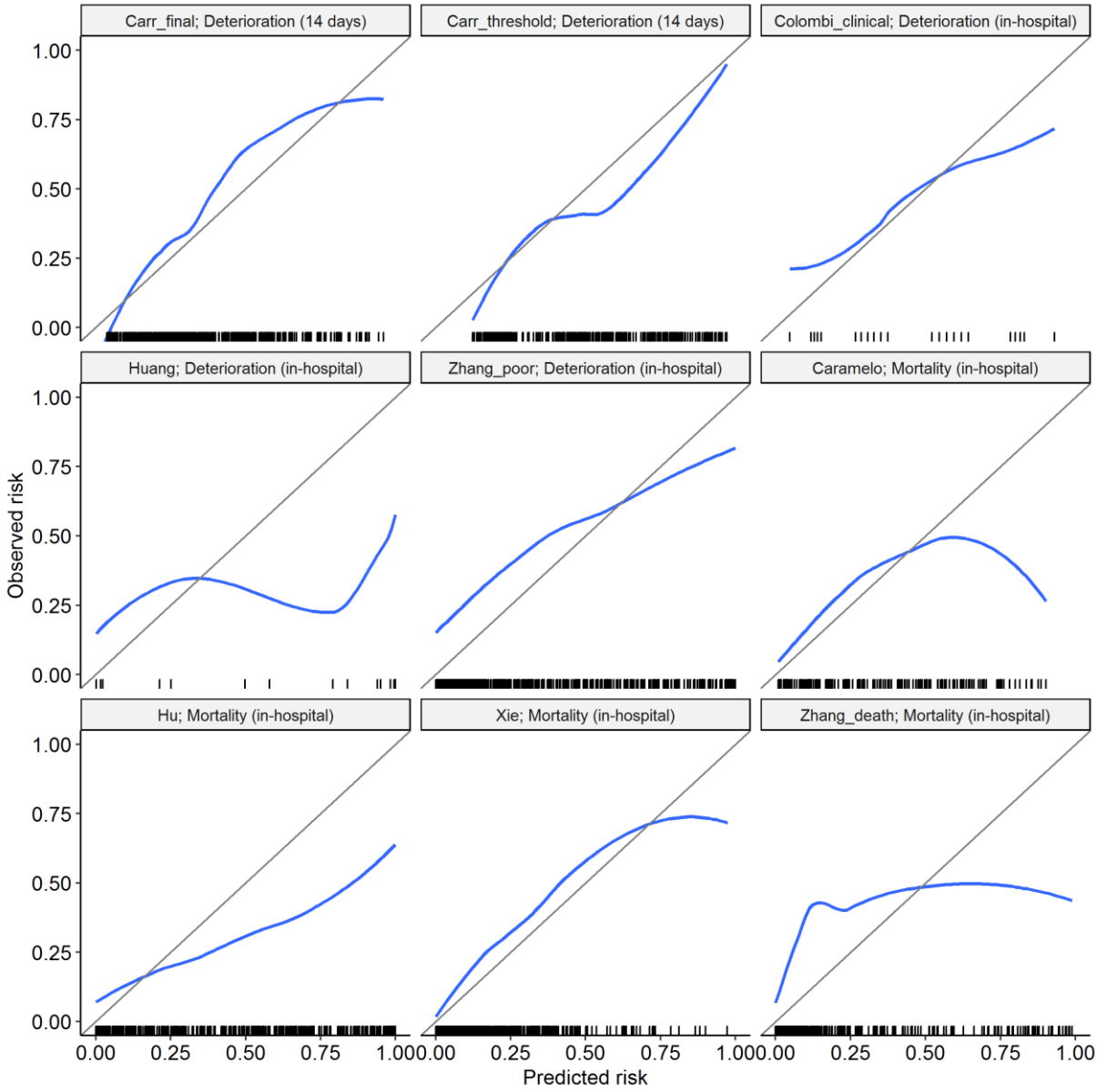
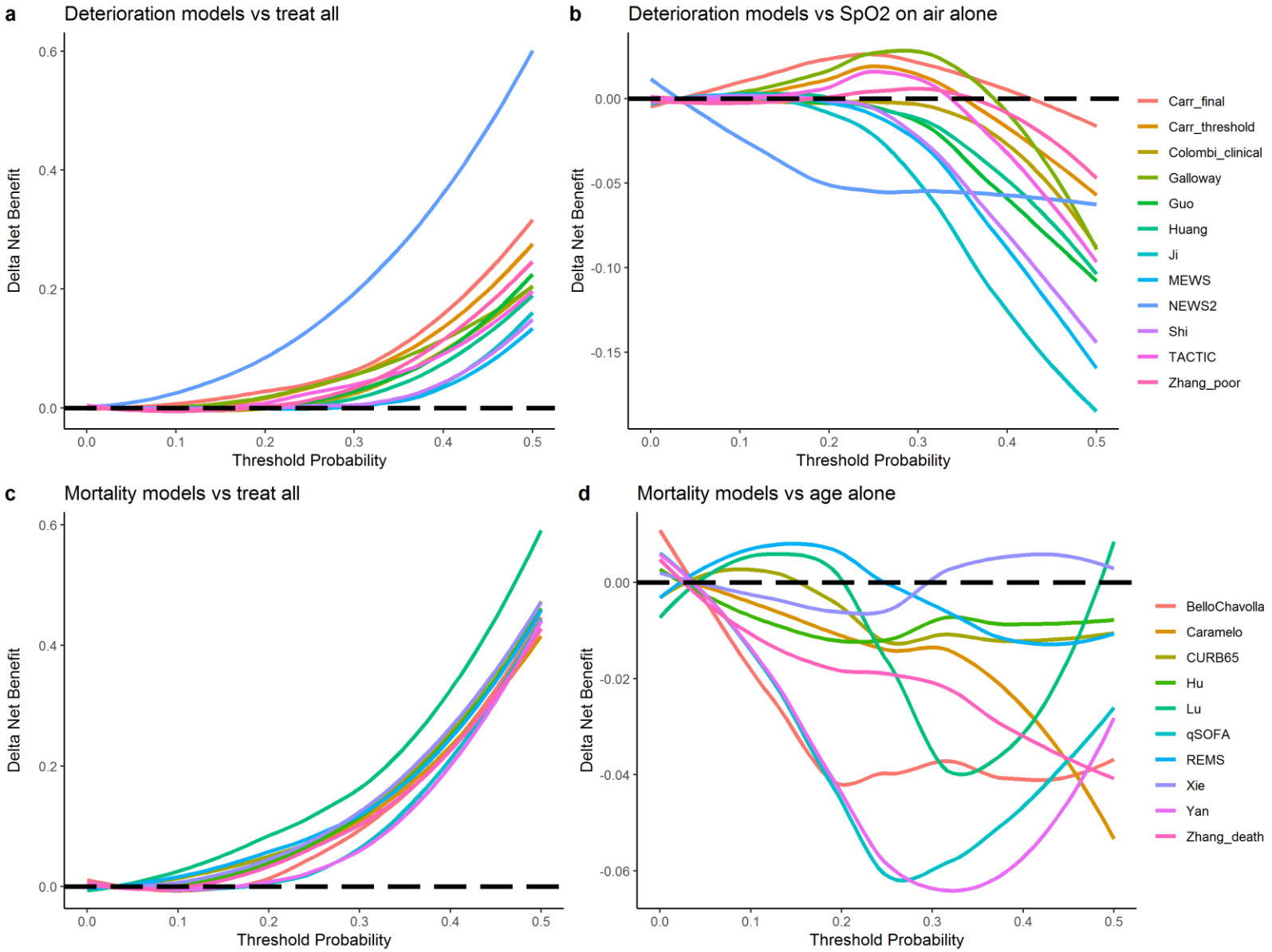


Figure 2: Decision curve analysis showing delta net benefit of each candidate model, compared to treating all patients and best univariable predictors.

For each analysis, the endpoint is the original intended outcome and time horizon for the index model. Each candidate model and univariable predictor was calibrated to the validation data during analysis to enable fair, head-to-head comparisons. Delta net benefit is calculated as net benefit when using the index model minus net benefit when: (1) treating all patients; and (2) using the most discriminating univariable predictor. The most discriminating univariable predictor is admission oxygen saturation (SpO2) on room air for deterioration models and patient age for mortality models. Delta net benefit is shown with Loess-smoothing. Black dashed line indicates threshold above which index model has greater net benefit than the comparator. Individual decision curves for each candidate model are shown in Supplementary Figure 8.


a
Deterioration models vs treat all
Delta Net Benefit
Threshold Probability
b
Deterioration models vs SpO2 on air alone
Delta Net Benefit
Threshold Probability
Carr_final
Carr_threshold
Colombi_clinical
Galloway
Guo
Huang
Ji
MEWS
NEWS2
Shi
TACTIC
Zhang_poor
c
Mortality models vs treat all
Delta Net Benefit
Threshold Probability
d
Mortality models vs age alone
Delta Net Benefit
Threshold Probability
BelloChavolla
Caramelo
CURB65
Hu
Lu
qSOFA
REMS
Xie
Yan
Zhang_death

Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study

Supplementary Material

Supplementary Table 1: Characteristics of candidate prognostic models for COVID-19 not included in current external validation study.

All candidate models included in a living systematic review were considered at high risk of bias[1]. ARDS = acute respiratory distress syndrome; ICU = intensive care unit; CT = computed tomography.

Authors	Pre-existing or COVID-specific?	Model outcome	Reason for exclusion
Bai et al[2]	COVID-specific	Deterioration	CT imaging required
Barda et al[3]	COVID-specific	Mortality	Multiple predictors not available in validation data
Chassagnon et al[4]	COVID-specific	Death or invasive mechanical ventilation	CT imaging required
Das et al[5]	COVID-specific	Mortality	Province required - not generalisable, and not reproducible
Gong et al[6]	COVID-specific	Risk of severe disease	Model parameters not publicly available
Jiang et al[7]	COVID-specific	Development of ARDS by Berlin criteria	Model parameters not publicly available
Levy et al[8]	COVID-specific	Mortality	Emergency severity index not available in validation data
Liang et al[9]	COVID-specific	Deterioration	Symptom data not available in validation data
Liu et al[10]	COVID-specific	Severity	T lymphocyte subsets not available in validation data
McRae et al[11]	COVID-specific	Mortality	Pro-calcitonin and myoglobin not available; exact model coefficients also not provided
Pourhomayoun et al[12]	COVID-specific	Mortality	Model parameters not publicly available
Qi et al[13]	COVID-specific	Length of hospital stay	CT imaging required
Sarkar et al[14]	COVID-specific	Mortality	Not generalisable outside Wuhan and model not reproducible
Singh et al[15]	Pre-existing (EPIC deterioration index)	ICU admission, mechanical ventilation or death	Model parameters not publicly available
Vaid et al[16]	COVID-specific	Mortality and 'critical events'	Model parameters not publicly available
Vazquez et al[17]	COVID-specific	Mortality (in-hospital)	Intended for ICU admissions only
Yuan et al[18]	COVID-specific	Mortality	CT imaging required
Zeng et al[19]	COVID-specific	Progression to severe disease	CT imaging required and no reproducible model

Supplementary Tables 2a-c: Validation metrics of prognostic scores for COVID-19, using (a) complete case sensitivity analysis (n=411); (b) excluding patients without PCR-confirmed SARS-CoV-2 infection; and (c) excluding patients who met the clinical deterioration outcome within 4 hours of arrival to hospital..

For each model, performance is evaluated for an approximation of its original intended outcome, shown in 'Primary outcome' column. AUROC = area under the receiver operating characteristic curve; CI = confidence interval.

(a) Complete case analysis

Score	Primary outcome	n	AUROC (95% CI)	Calibration slope (95% CI)	Calibration in the large (95% CI)
NEWS2	Deterioration (1 day)	404	0.78 (0.73 - 0.83)		
Ji	Deterioration (10 days)	183	0.61 (0.53 - 0.69)		
Carr_final	Deterioration (14 days)	381	0.75 (0.71 - 0.8)	0.93 (0.7 - 1.18)	0.46 (0.23 - 0.69)
Carr_threshold	Deterioration (14 days)	381	0.74 (0.69 - 0.79)	0.78 (0.59 - 0.99)	-0.29 (-0.52 - -0.05)
Guo	Deterioration (14 days)	153	0.7 (0.62 - 0.78)		
Zhang_poor	Deterioration (in-hospital)	400	0.74 (0.69 - 0.78)	0.31 (0.21 - 0.42)	0.59 (0.32 - 0.83)
Colombi_clinical	Deterioration (in-hospital)	182	0.72 (0.64 - 0.79)	0.64 (0.38 - 0.93)	0 (-0.34 - 0.33)
Galloway	Deterioration (in-hospital)	351	0.7 (0.65 - 0.75)		
Huang	Deterioration (in-hospital)	182	0.7 (0.63 - 0.78)	0.27 (0.16 - 0.38)	-5.09 (-5.54 - -4.65)
TACTIC	Deterioration (in-hospital)	366	0.68 (0.63 - 0.73)		
Shi	Deterioration (in-hospital)	411	0.61 (0.56 - 0.66)		
MEWS	Deterioration (in-hospital)	405	0.59 (0.54 - 0.65)		
Lu	Mortality (12 days)	403	0.71 (0.67 - 0.76)		
CURB65	Mortality (30 days)	374	0.75 (0.7 - 0.8)		
BelloChavolla	Mortality (30 days)	385	0.66 (0.6 - 0.72)		
Xie	Mortality (in-hospital)	183	0.78 (0.7 - 0.86)	0.91 (0.55 - 1.31)	0.03 (-0.37 - 0.41)
REMS	Mortality (in-hospital)	404	0.76 (0.71 - 0.81)		
Hu	Mortality (in-hospital)	153	0.75 (0.66 - 0.84)	0.38 (0.21 - 0.58)	-1.61 (-2.11 - -1.14)
Caramelo	Mortality (in-hospital)	408	0.71 (0.66 - 0.77)	0.53 (0.36 - 0.7)	0 (-0.26 - 0.25)
Zhang_death	Mortality (in-hospital)	400	0.7 (0.64 - 0.76)	0.28 (0.18 - 0.39)	0.92 (0.63 - 1.21)
qSOFA	Mortality (in-hospital)	405	0.6 (0.54 - 0.65)		
Yan	Mortality (in-hospital)	182	0.6 (0.52 - 0.68)		
Age	Mortality (in-hospital)	411	0.76 (0.71 - 0.81)		
SpO2 on air	Deterioration (in-hospital)	403	0.76 (0.71 - 0.81)		

(b) Restriction to PCR-confirmed cases (n=370)

Score	Primary outcome	AUROC (95% CI)	Calibration slope (95% CI)	Calibration in the large (95% CI)
NEWS2	Deterioration (1 day)	0.79 (0.74 - 0.84)		
Ji	Deterioration (10 days)	0.58 (0.52 - 0.64)		
Carr_final	Deterioration (14 days)	0.78 (0.73 - 0.83)	1.02 (0.77 - 1.27)	0.38 (0.14 - 0.61)
Carr_threshold	Deterioration (14 days)	0.76 (0.71 - 0.81)	0.83 (0.62 - 1.04)	-0.3 (-0.54 - -0.06)
Guo	Deterioration (14 days)	0.68 (0.62 - 0.74)		
Zhang_poor	Deterioration (in-hospital)	0.74 (0.69 - 0.79)	0.3 (0.19 - 0.41)	0.64 (0.38 - 0.91)
Galloway	Deterioration (in-hospital)	0.73 (0.68 - 0.78)		
TACTIC	Deterioration (in-hospital)	0.71 (0.65 - 0.76)		
Colombi_clinical	Deterioration (in-hospital)	0.68 (0.63 - 0.74)	0.52 (0.33 - 0.71)	0.02 (-0.23 - 0.26)
Huang	Deterioration (in-hospital)	0.67 (0.61 - 0.73)	0.18 (0.1 - 0.26)	-4.2 (-4.58 - -3.81)
Shi	Deterioration (in-hospital)	0.61 (0.56 - 0.67)		
MEWS	Deterioration (in-hospital)	0.6 (0.55 - 0.66)		
Lu	Mortality (12 days)	0.73 (0.69 - 0.78)		
CURB65	Mortality (30 days)	0.74 (0.69 - 0.79)		
BelloChavolla	Mortality (30 days)	0.66 (0.6 - 0.72)		
REMS	Mortality (in-hospital)	0.76 (0.71 - 0.81)		
Xie	Mortality (in-hospital)	0.75 (0.68 - 0.82)	0.78 (0.46 - 1.1)	0.46 (0.2 - 0.71)
Hu	Mortality (in-hospital)	0.73 (0.67 - 0.79)	0.33 (0.2 - 0.46)	-0.99 (-1.3 - -0.67)
Zhang_death	Mortality (in-hospital)	0.71 (0.65 - 0.76)	0.3 (0.19 - 0.41)	0.99 (0.69 - 1.29)
Caramelo	Mortality (in-hospital)	0.7 (0.65 - 0.76)	0.5 (0.32 - 0.67)	0.02 (-0.24 - 0.28)
qSOFA	Mortality (in-hospital)	0.6 (0.55 - 0.66)		
Yan	Mortality (in-hospital)	0.58 (0.49 - 0.67)		
SpO2 on air	Deterioration (in-hospital)	0.76 (0.71 - 0.81)		
Age	Mortality (in-hospital)	0.75 (0.7 - 0.81)		

(c) Excluding deterioration events <4 hours from admission (n=371)

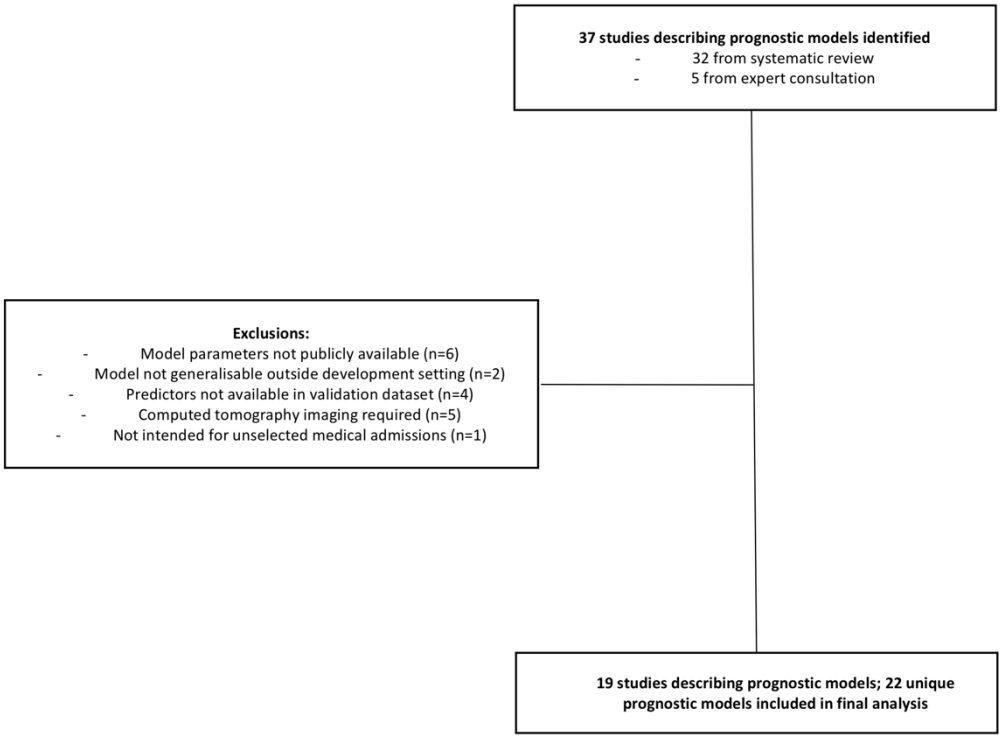
Score	Primary outcome	AUROC (95% CI)	Calibration slope (95% CI)	Calibration in the large (95% CI)
NEWS2	Deterioration (1 day)	0.74 (0.67 - 0.81)		
Ji	Deterioration (10 days)	0.55 (0.49 - 0.62)		
Carr_final	Deterioration (14 days)	0.74 (0.69 - 0.79)	0.87 (0.62 - 1.12)	0.14 (-0.1 - 0.37)
Carr_threshold	Deterioration (14 days)	0.73 (0.67 - 0.78)	0.73 (0.52 - 0.94)	-0.54 (-0.78 - -0.29)
Guo	Deterioration (14 days)	0.65 (0.58 - 0.71)		
Zhang_poor	Deterioration (in-hospital)	0.72 (0.67 - 0.78)	0.39 (0.27 - 0.51)	0.3 (0.03 - 0.58)
Galloway	Deterioration (in-hospital)	0.69 (0.64 - 0.75)		
Colombi_clinical	Deterioration (in-hospital)	0.68 (0.62 - 0.74)	0.48 (0.29 - 0.66)	-0.25 (-0.5 - -0.01)
TACTIC	Deterioration (in-hospital)	0.66 (0.6 - 0.71)		
Huang	Deterioration (in-hospital)	0.63 (0.56 - 0.71)	0.14 (0.06 - 0.23)	-4.49 (-4.84 - -4.15)
Shi	Deterioration (in-hospital)	0.59 (0.54 - 0.65)		
MEWS	Deterioration (in-hospital)	0.56 (0.5 - 0.62)		
Lu	Mortality (12 days)	0.73 (0.69 - 0.78)		
CURB65	Mortality (30 days)	0.74 (0.69 - 0.8)		
BelloChavolla	Mortality (30 days)	0.65 (0.59 - 0.72)		
REMS	Mortality (in-hospital)	0.76 (0.71 - 0.81)		
Xie	Mortality (in-hospital)	0.75 (0.68 - 0.82)	0.83 (0.52 - 1.13)	0.32 (0.05 - 0.59)
Hu	Mortality (in-hospital)	0.73 (0.67 - 0.79)	0.34 (0.21 - 0.48)	-1.07 (-1.38 - -0.75)
Caramelo	Mortality (in-hospital)	0.71 (0.65 - 0.77)	0.51 (0.33 - 0.69)	-0.18 (-0.46 - 0.09)
Zhang_death	Mortality (in-hospital)	0.69 (0.63 - 0.75)	0.29 (0.17 - 0.4)	0.88 (0.57 - 1.19)
qSOFA	Mortality (in-hospital)	0.59 (0.53 - 0.66)		
Yan	Mortality (in-hospital)	0.57 (0.48 - 0.66)		
Age	Mortality (in-hospital)	0.77 (0.71 - 0.82)		
SpO2 on air	Deterioration (in-hospital)	0.7 (0.64 - 0.76)		

Supplementary Table 3: Optimism estimates for most discriminating univariable predictors of clinical deterioration and mortality

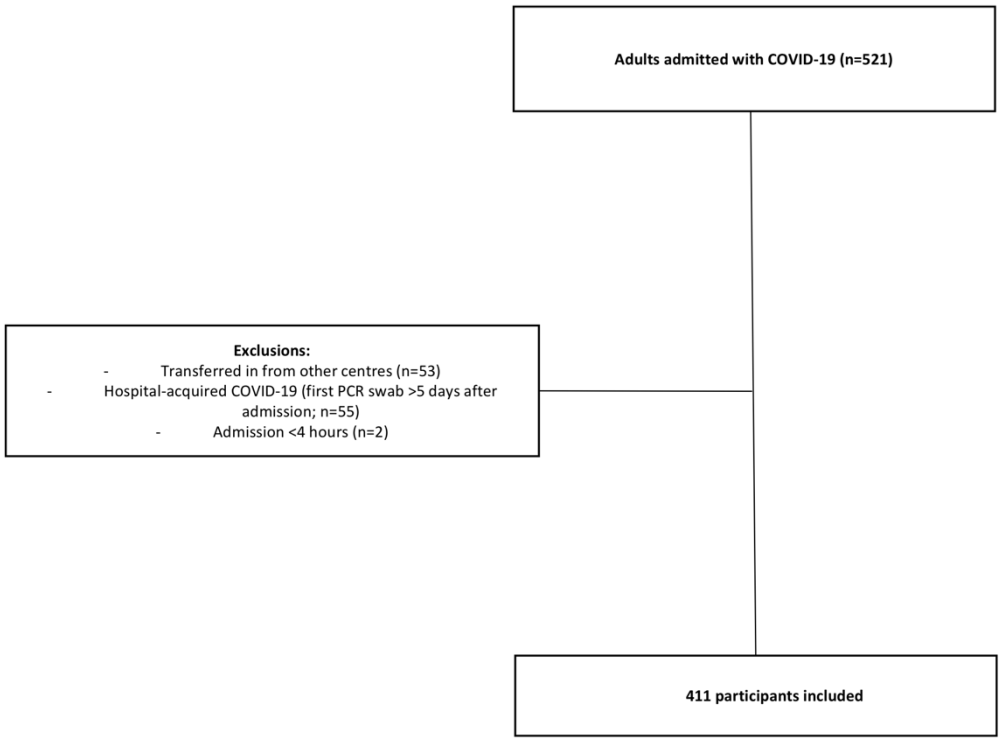
Optimism is calculated using bootstrapping with 1,000 iterations. AUROC = area under the receiver operating characteristic curve; CI = confidence interval. Dxy = Somers' Delta, which is a measure of agreement between pairs of ordinal variables, ranging from -1 (no agreement) to +1 (complete agreement).

Predictor	**Outcome**	**AUROC (95% CI)**	**Optimism**	
			Dxy	*Slope*
Age	Mortality (in-hospital)	0.76 (0.71 - 0.81)	0.000	-0.009
Oxygen saturation (room air)	Deterioration (in-hospital)	0.76 (0.71 - 0.81)	-0.001	-0.011

Supplementary Figure 1: Flowchart showing prognostic models included in systematic evaluation.

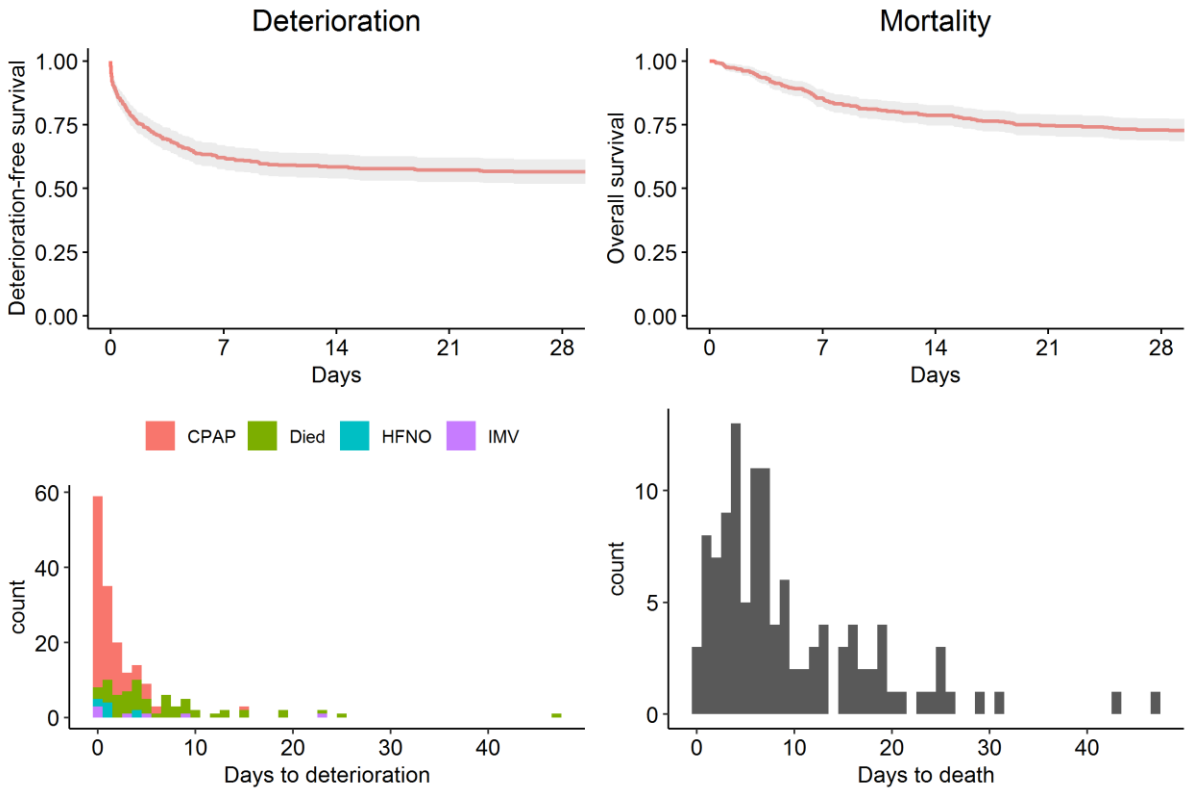


Supplementary Figure 2: Flowchart showing study participants included in analysis.



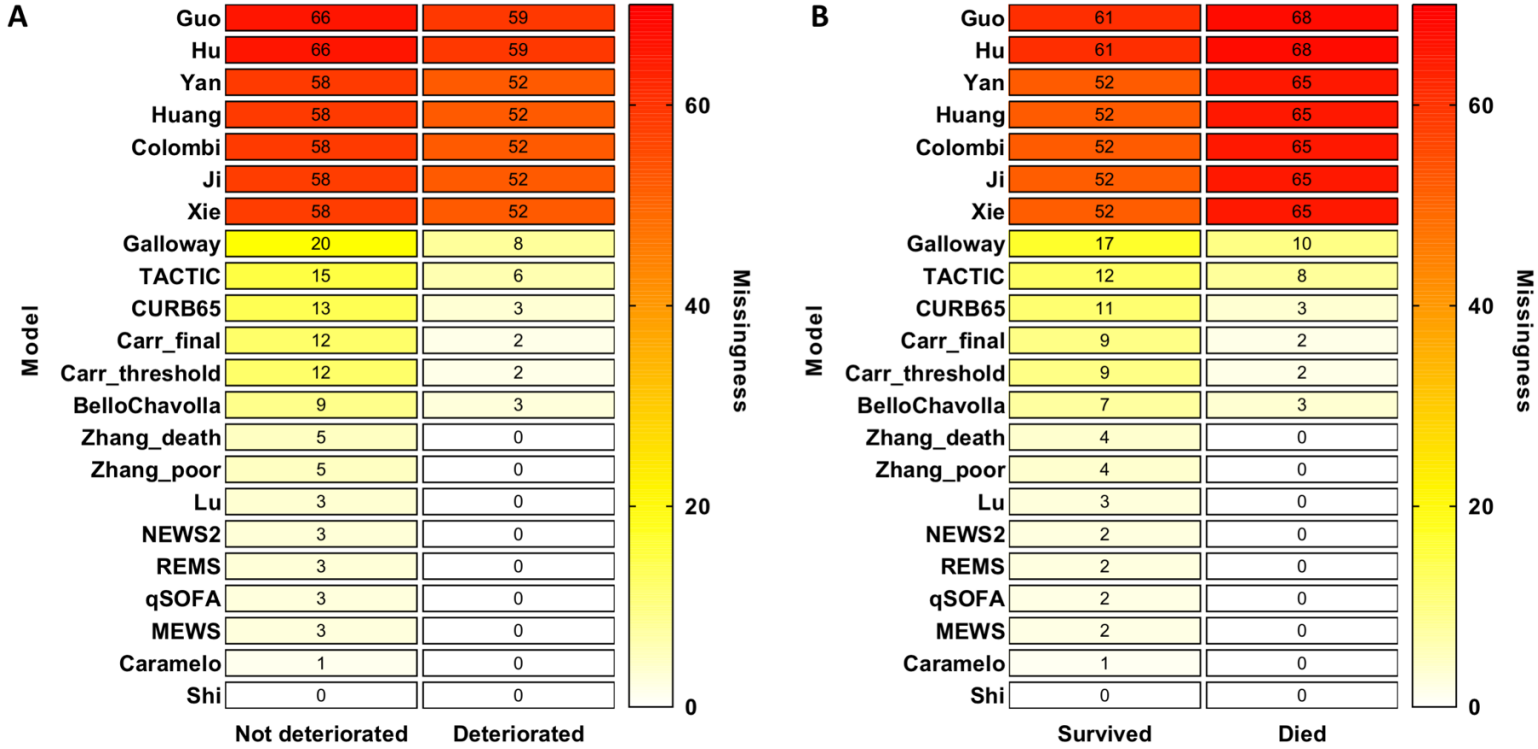
Supplementary Figure 3: Timing of clinical deterioration and death following hospital admission among patients with COVID-19

Shown as Kaplan-Meier plots and histograms. Days to deterioration histogram reflects the first time that the endpoint was met, with the criteria for meeting the endpoint indicated by colour. CPAP = continuous positive airway pressure; HFNO = high flow nasal cannula oxygen; IMV = invasive mechanical ventilation.



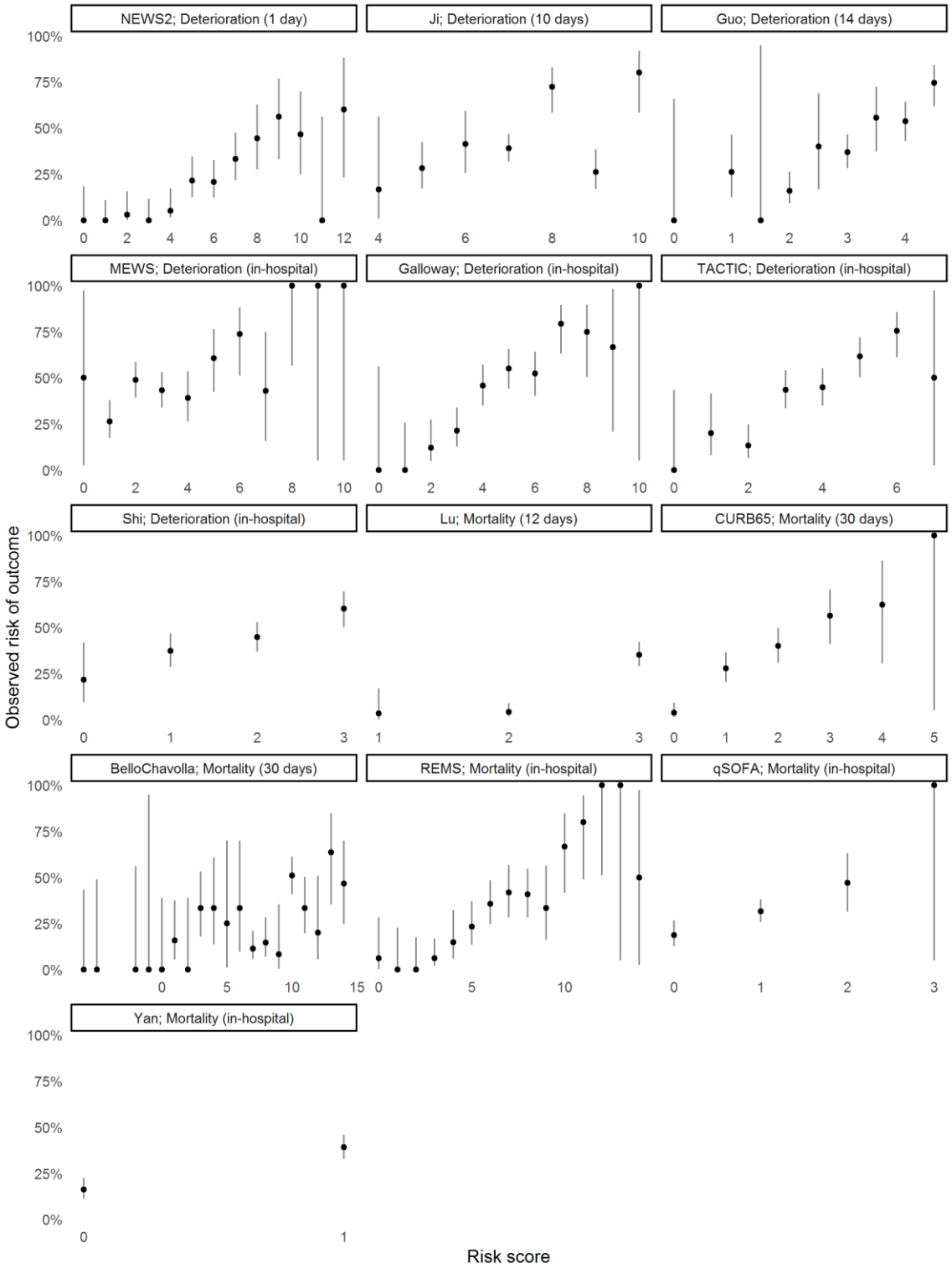
Supplementary Figure 4: Missingness of candidate prognostic models.

Shown as column-wise percentage missing, stratified by (a) composite outcome of clinical deterioration; and (b) mortality during hospital admission.



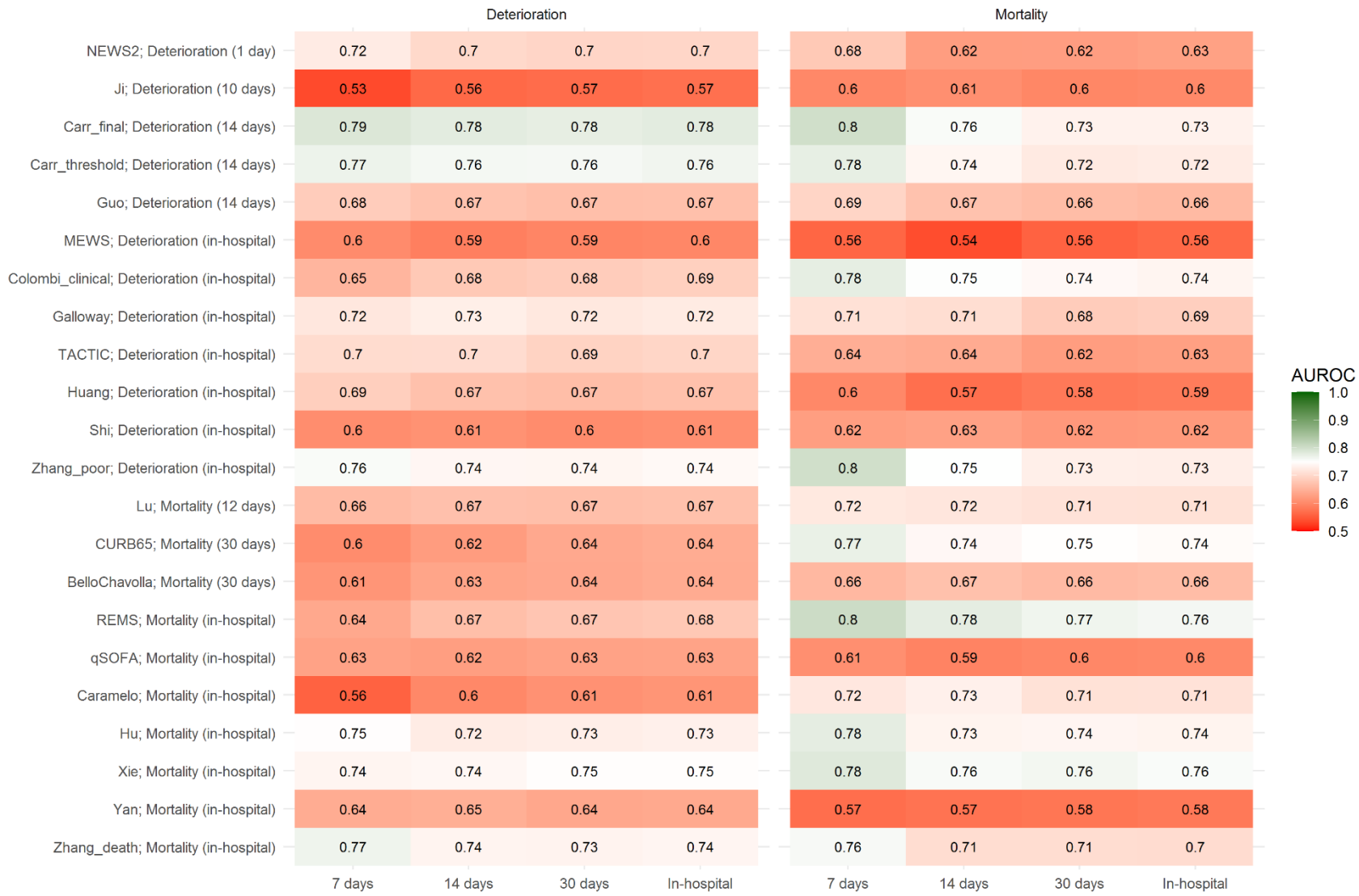
Supplementary Figure 5: Plots showing risk scores vs observed prevalence of outcomes for points-based prognostic scores.

The primary outcome of interest for each model is shown in the plot sub-heading. Individual predictions for each prognostic model were averaged across imputations for each participant in the dataset in order to generate these pooled plots.


NEWS2; Deterioration (1 day)
Ji; Deterioration (10 days)
Guo; Deterioration (14 days)
MEWS; Deterioration (in-hospital)
Galloway; Deterioration (in-hospital)
TACTIC; Deterioration (in-hospital)
Shi; Deterioration (in-hospital)
Lu; Mortality (12 days)
CURB65; Mortality (30 days)
BelloChavolla; Mortality (30 days)
REMS; Mortality (in-hospital)
qSOFA; Mortality (in-hospital)
Yan; Mortality (in-hospital)
Observed risk of outcome
Risk score

Supplementary Figure 6: Heat map showing time-dependent receiver operating characteristic areas under the curve for each prognostic score to predict (a) deterioration or (b) mortality,

For each model, discrimination is stratified by interval from hospital admission to the outcome event. The original intended primary outcome for the model is shown in brackets in the y-axis labels. AUROC = area under the receiver operating characteristic curve.


Deterioration
Mortality
NEWS2; Deterioration (1 day) 0.72 0.7 0.7 0.7 0.68 0.62 0.62 0.63
Ji; Deterioration (10 days) 0.53 0.56 0.57 0.57 0.6 0.61 0.6 0.6
Carr_final; Deterioration (14 days) 0.79 0.78 0.78 0.78 0.8 0.76 0.73 0.73
Carr_threshold; Deterioration (14 days) 0.77 0.76 0.76 0.76 0.78 0.74 0.72 0.72
Guo; Deterioration (14 days) 0.68 0.67 0.67 0.67 0.69 0.67 0.66 0.66
MEWS; Deterioration (in-hospital) 0.6 0.59 0.59 0.6 0.56 0.54 0.56 0.56
Colombi_clinical; Deterioration (in-hospital) 0.65 0.68 0.68 0.69 0.78 0.75 0.74 0.74
Galloway; Deterioration (in-hospital) 0.72 0.73 0.72 0.72 0.71 0.71 0.68 0.69
TACTIC; Deterioration (in-hospital) 0.7 0.7 0.69 0.7 0.64 0.64 0.62 0.63
Huang; Deterioration (in-hospital) 0.69 0.67 0.67 0.67 0.6 0.57 0.58 0.59
Shi; Deterioration (in-hospital) 0.6 0.61 0.6 0.61 0.62 0.63 0.62 0.62
Zhang_poor; Deterioration (in-hospital) 0.76 0.74 0.74 0.74 0.8 0.75 0.73 0.73
Lu; Mortality (12 days) 0.66 0.67 0.67 0.67 0.72 0.72 0.71 0.71
CURB65; Mortality (30 days) 0.6 0.62 0.64 0.64 0.77 0.74 0.75 0.74
BelloChavolla; Mortality (30 days) 0.61 0.63 0.64 0.64 0.66 0.67 0.66 0.66
REMS; Mortality (in-hospital) 0.64 0.67 0.67 0.68 0.8 0.78 0.77 0.76
qSOFA; Mortality (in-hospital) 0.63 0.62 0.63 0.63 0.61 0.59 0.6 0.6
Caramelo; Mortality (in-hospital) 0.56 0.6 0.61 0.61 0.72 0.73 0.71 0.71
Hu; Mortality (in-hospital) 0.75 0.72 0.73 0.73 0.78 0.73 0.74 0.74
Xie; Mortality (in-hospital) 0.74 0.74 0.75 0.75 0.78 0.76 0.76 0.76
Yan; Mortality (in-hospital) 0.64 0.65 0.64 0.64 0.57 0.57 0.58 0.58
Zhang_death; Mortality (in-hospital) 0.77 0.74 0.73 0.74 0.76 0.71 0.71 0.7
7 days
14 days
30 days
In-hospital
7 days
14 days
30 days
In-hospital
AUROC
1.0
0.9
0.8
0.7
0.6
0.5

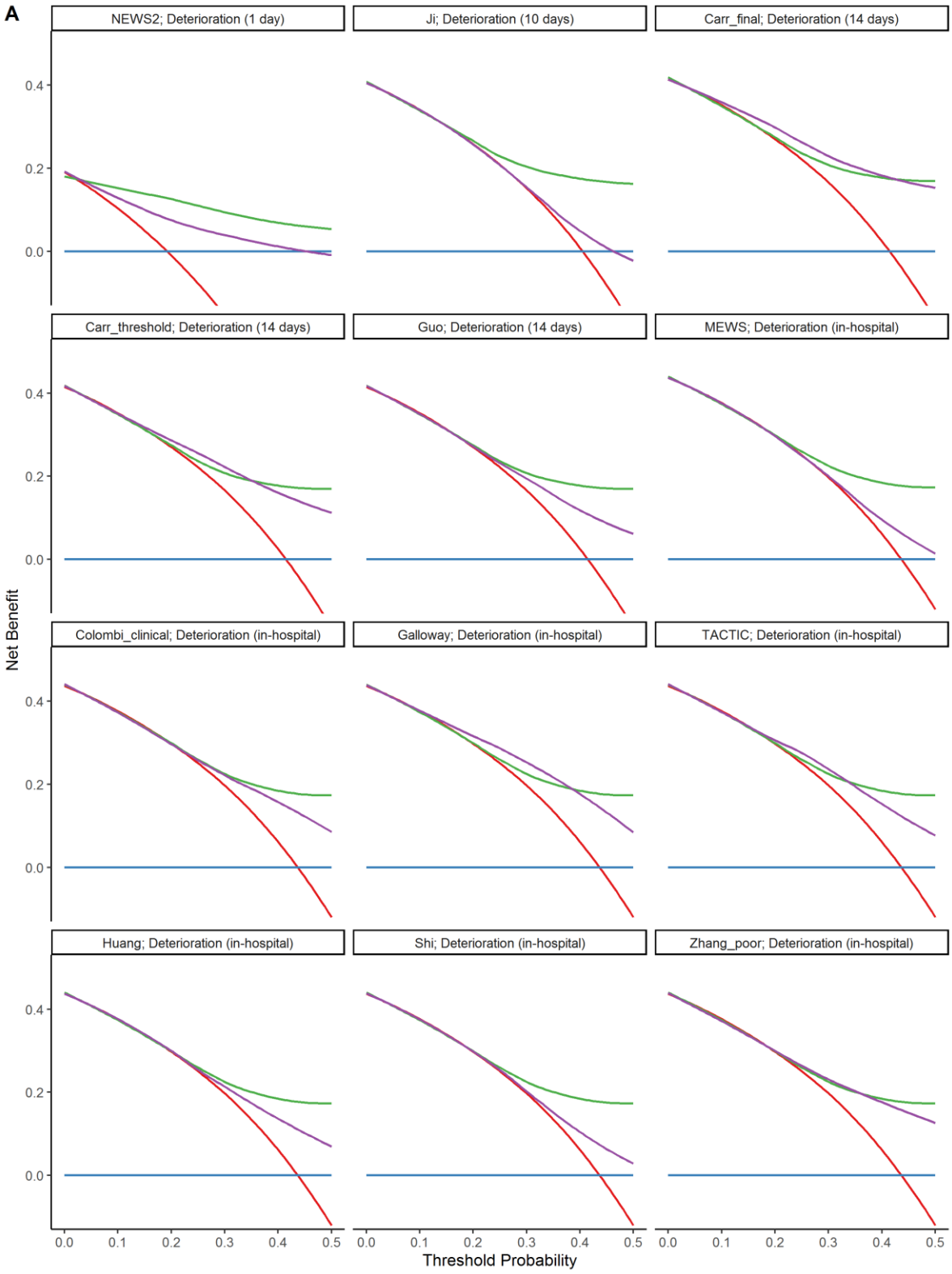
Supplementary Figure 7: Heat map showing time-dependent receiver operating characteristic areas under the curve for a priori clinical predictors of interest in univariable analyses to predict (a) clinical deterioration or (b) mortality

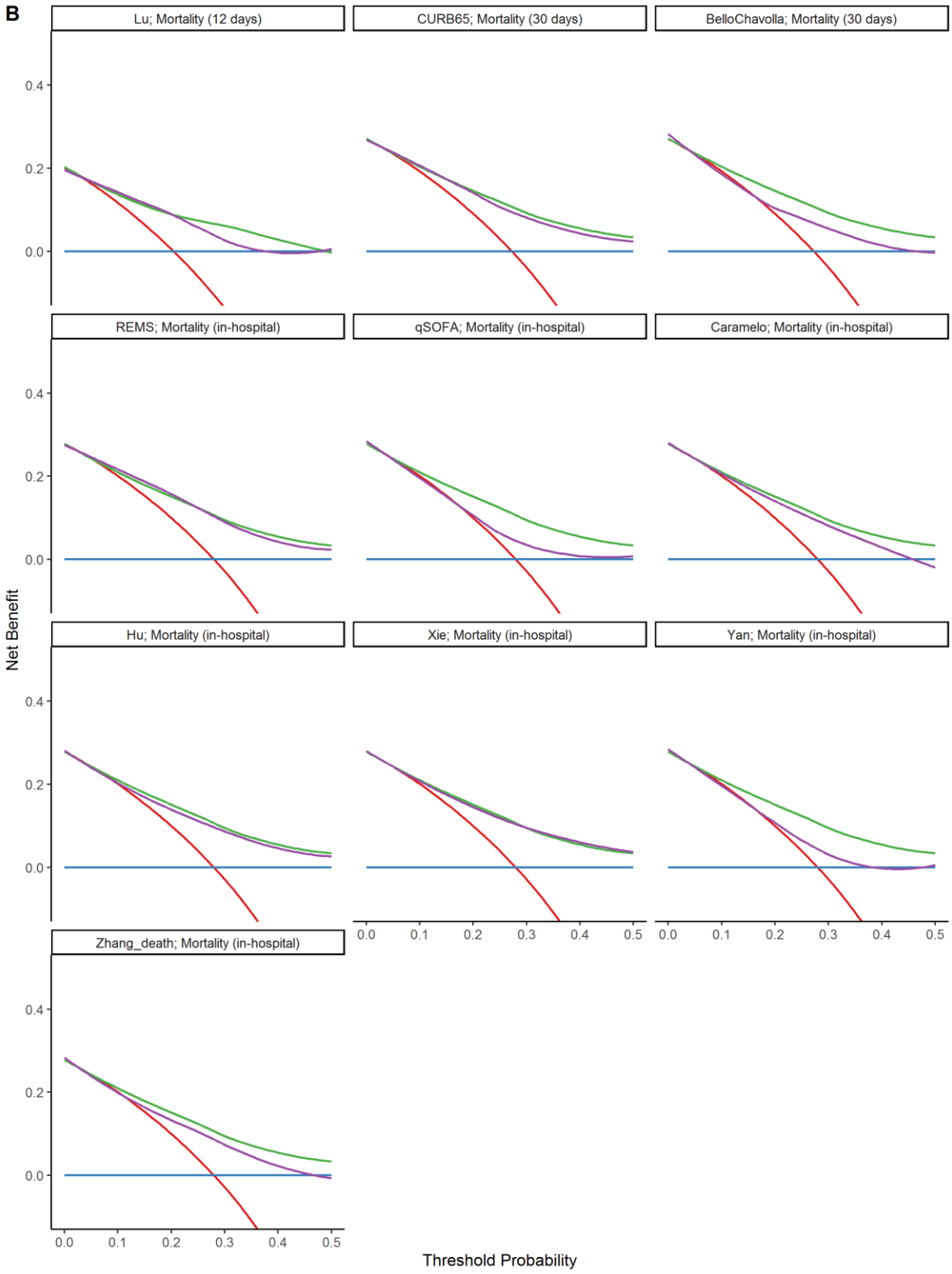
For each predictor, discrimination is stratified by interval from hospital admission to the outcome event. AUROC = area under the receiver operating characteristic curve.


Deterioration
Mortality
Age
Clinical frailty scale
SpO2 on air
CRP
Lymphocytes
0.57
0.6
0.63
0.63
0.79
0.78
0.77
0.76
0.51
0.55
0.57
0.57
0.73
0.73
0.71
0.71
0.79
0.77
0.76
0.76
0.68
0.64
0.65
0.65
0.76
0.73
0.72
0.73
0.67
0.64
0.64
0.65
0.65
0.62
0.62
0.62
0.64
0.61
0.62
0.61
7 days
14 days
30 days
In-hospital
7 days
14 days
30 days
In-hospital
AUROC
1.0
0.9
0.8
0.7
0.6
0.5

Supplementary Figure 8: Decision curve analysis comparing net benefit of each candidate model for (A) clinical deterioration; and (B) mortality.

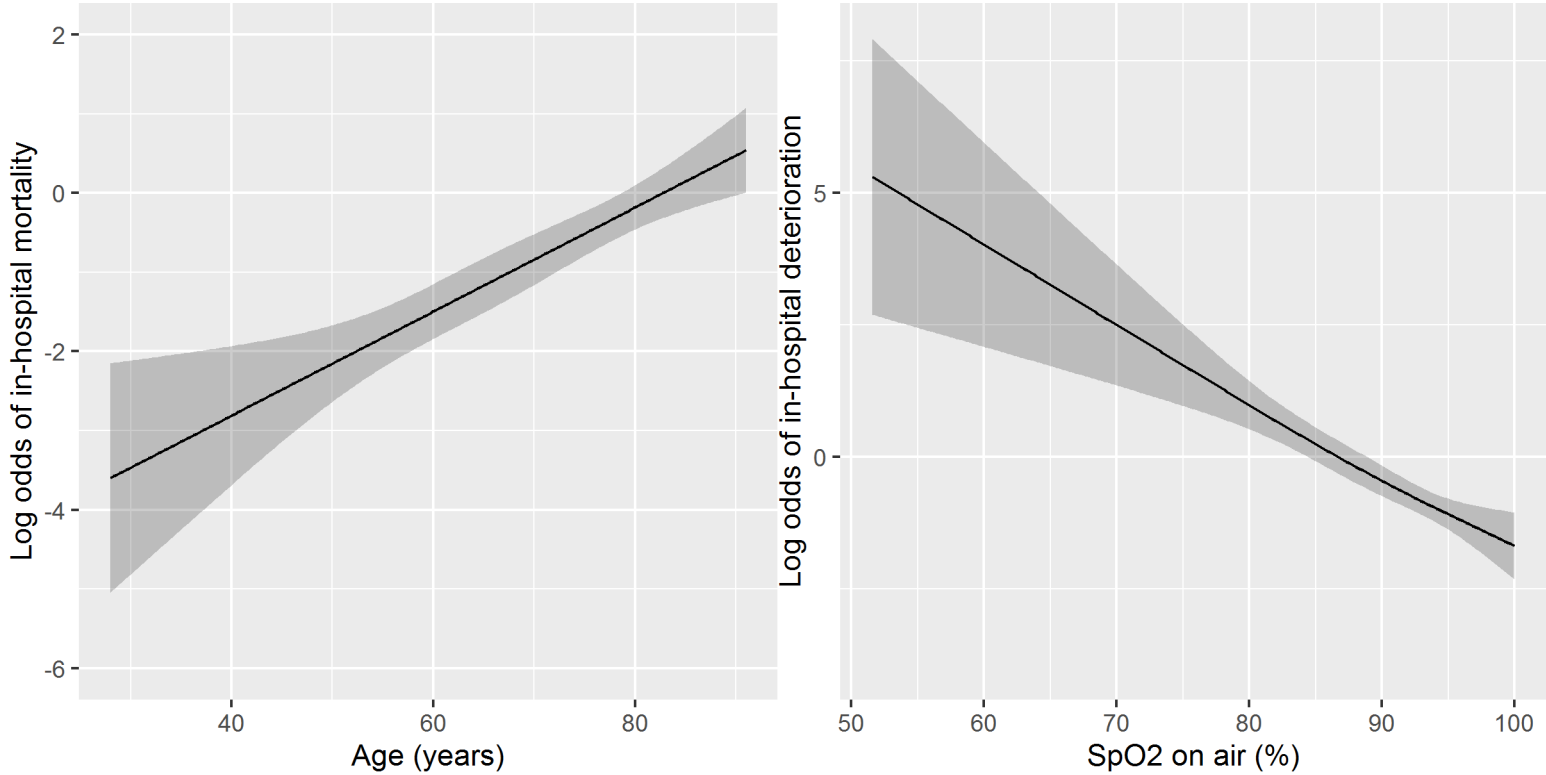
For each analysis, the endpoint is the original intended endpoint for the index model (purple line; endpoint shown in plot subheading). Comparisons are made to the strategies of treating all patients (red); treating no patients (blue); and offering treatment based on the most discriminating univariable predictor (green; admission oxygen saturation on room air for deterioration models (A); patient age for mortality models (B)). Each candidate model and univariable predictor was recalibrated to the validation data during analysis to enable fair, head-to-head comparisons. All curves are shown with Loess-smoothing.


A
NEWS2; Deterioration (1 day)
Ji; Deterioration (10 days)
Carr_final; Deterioration (14 days)
Carr_threshold; Deterioration (14 days)
Guo; Deterioration (14 days)
MEWS; Deterioration (in-hospital)
Colombi_clinical; Deterioration (in-hospital)
Galloway; Deterioration (in-hospital)
TACTIC; Deterioration (in-hospital)
Huang; Deterioration (in-hospital)
Shi; Deterioration (in-hospital)
Zhang_poor; Deterioration (in-hospital)
Net Benefit
0.0
0.2
0.4
0.0
0.1
0.2
0.3
0.4
0.5
Threshold Probability


B
Lu; Mortality (12 days)
CURB65; Mortality (30 days)
BelloChavolla; Mortality (30 days)
REMS; Mortality (in-hospital)
qSOFA; Mortality (in-hospital)
Caramelo; Mortality (in-hospital)
Hu; Mortality (in-hospital)
Xie; Mortality (in-hospital)
Yan; Mortality (in-hospital)
Zhang_death; Mortality (in-hospital)
Net Benefit
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
Threshold Probability

Supplementary Figure 9: Restricted cubic splines plots showing associations between most discriminating univariable predictors of clinical deterioration and mortality and log odds of outcome, respectively.

Age was the most discriminating univariable predictor of in-hospital mortality, while oxygen saturation (SpO2) on air was the most discriminating predictor of in-hospital clinical deterioration. For both univariable predictors, associations appear approximately linear on the log odds scale.



References

1. Wynants L, Calster B Van, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Damen JAA, Debray TPA, Vos M De, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Kreuzberger N, Lohmann A, Luijken K, Ma J, Navarro CLA, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Kuijk SMJ van, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* British Medical Journal Publishing Group; 2020; 369.
2. Bai X, Fang C, Zhou Y, Bai S, Liu Z, Chen Q, Xu Y, Xia T, Gong S, Xie X, Song D, Du R, Zhou C, Chen C, Nie D, Tu D, Zhang C, Liu X, Qin L, Chen W. Predicting COVID-19 malignant progression with AI techniques. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.20.20037325.
3. Barda N, Riesel D, Akriv A, Levi J, Finkel U, Yona G, Greenfeld D, Sheiba S, Somer J, Bachmat E, Rothblum GN, Shalit U, Netzer D, Balicer R, Dagan N. Performing risk stratification for COVID-19 when individual level data is not available, the experience of a large healthcare organization. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.23.20076976.
4. Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi T-N, Dangeard S, Deutsch E, Andre F, Guillo E, Halm N, Hajj S El, Bompard F, Neveu S, Hani C, Saab I, Campredon A, Koulakian H, Bennani S, Freche G, Barat M, Lombard A, Fournier L, Monnier H, Grand T, Gregory J, Nguyen Y, Khalil A, Mahdjoub E, Brillet P-Y, Ba ST, et al. Holistic AI-Driven Quantification, Staging and Prognosis of COVID-19 Pneumonia. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.17.20069187.
5. DAS A, Mishra S, Gopalan SS. Predicting community mortality risk due to CoVID-19 using machine learning and development of a prediction tool. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.27.20081794.
6. Gong J, Ou J, Qiu X, Jie Y, Chen Y, Yuan L, Cao J, Tan M, Xu W, Zheng F, Shi Y, Hu B. A Tool for Early Prediction of Severe Coronavirus Disease 2019 (COVID-19): A Multicenter Study Using the Risk Nomogram in Wuhan and Guangdong, China. *Clin. Infect. Dis.* Oxford University Press; 2020; 71: 833–840.
7. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y. Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. *Comput. Mater. Contin.* 2020; 62: 537–551.
8. Levy TJ, Richardson S, Coppa K, Barnaby DP, McGinn T, Becker LB, Davidson KW, Cohen SL, Hirsch JS, Zanos T, Consortium N & MC-19 R. Development and Validation of a Survival Calculator for Hospitalized Patients with COVID-19. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.22.20075416.
9. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, Li Y, Guan W, Sang L, Lu J, Xu Y, Chen G, Guo H, Guo J, Chen Z, Zhao Y, Li S, Zhang N, Zhong N, He J, COVID-19 for the CMTEG for. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA Intern. Med.* American Medical Association; 2020; 180: 1081.
10. Liu Q, Fang X, Tokuno S, Chung U, Chen X, Dai X, Liu X, Xu F, Wang B, Peng P. Prediction of the clinical outcome of COVID-19 patients using T lymphocyte subsets with 340 cases from Wuhan, China: a retrospective cohort study and a web visualization tool. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.06.20056127.
11. McRae MP, Simmons GW, Christodoulides NJ, Lu Z, Kang SK, Fenyo D, Alcorn T, Dapkins IP, Sharif I, Vurmaz D, Modak SS, Srinivasan K, Warhadpande S, Shrivastav R, McDevitt JT. Clinical Decision Support Tool and Rapid Point-of-Care Platform for Determining Disease Severity in Patients with COVID-19. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.16.20068411.
12. Pourhomayoun M, Shakibi M. Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.30.20047308.
13. Qi X, Jiang Z, YU Q, Shao C, Zhang H, Yue H, Ma B, Wang Y, Liu C, Meng X, Huang S, Wang J, Xu D, Lei J, Xie G, Huang H, Yang J, Ji J, Pan H, Zou S, Ju S. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.02.29.20029603.
14. Sarkar J, Chakrabarti P. A Machine Learning Model Reveals Older Age and Delayed Hospitalization as Predictors of Mortality in Patients with COVID-19. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.25.20043331.
15. Singh K, Valley TS, Tang S, Li BY, Kamran F, Sjoding MW, Wiens J, Otles E, Donnelly JP, Wei MY, McBride JP, Cao J, Penoza C, Ayanian JZ, Nallamothu BK. Evaluating a Widely Implemented Proprietary

Deterioration Index Model Among Hospitalized COVID-19 Patients. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.24.20079012.

16. Vaid A, Somani S, Russak AJ, Freitas JK De, Chaudhry FF, Paranjpe I, Johnson KW, Lee SJ, Miotto R, Zhao S, Beckmann N, Naik N, Arfer K, Kia A, Timsina P, Lala A, Paranjpe M, Glowe P, Golden E, Danieletto M, Singh M, Meyer D, O'Reilly PF, Huckins LH, Kovatch P, Finkelstein J, Freeman RM, Argulian E, Kasarskis A, Percha B, et al. Machine Learning to Predict Mortality and Critical Events in COVID-19 Positive New York City Patients. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.26.20073411.
17. Guillamet CV, Guillamet RV, Kramer AA, Maurer PM, Menke GA, Hill CL, Knaus WA. TOWARD A COVID-19 SCORE-RISK ASSESSMENTS AND REGISTRY. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.04.15.20066860.
18. Yuan M, Yin W, Tao Z, Tan W, Hu Y. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. Schildgen O, editor. *PLoS One* Public Library of Science; 2020; 15: e0230548.
19. Zeng L, Li J, Liao M, Hua R, Huang P, Zhang M, Zhang Y, Shi Q, Xia Z, Ning X, Liu D, Mo J, Zhou Z, Li Z, Fu Y, Liao Y, Yuan J, Wang L, He Q, Liu L, Qiao K. Risk assessment of progression to severe conditions for patients with COVID-19 pneumonia: a single-center retrospective study. *medRxiv* Cold Spring Harbor Laboratory Press; 2020; : 2020.03.25.20043166.