

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. Samuels², Akish Luintel², Tommy Rampling², Humayra Chowdhury², Matteo Quartagno⁴, Arjun Nair², Marc Lipman⁵, Ibrahim Abubakar¹, Maarten van Smeden⁶, Wai Keong Wong², 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

Word count

3,167

Abstract

Background

The number of proposed prognostic models for COVID-19, which aim to predict disease outcomes, is growing rapidly. It is not known whether any are suitable for widespread clinical implementation. We addressed this question by independent and systematic evaluation of their performance among hospitalised COVID-19 cases.

Methods

We conducted an observational cohort study to assess candidate prognostic models, identified through a living systematic review. We included consecutive adults admitted to a secondary care hospital with PCR-confirmed or clinically diagnosed community-acquired COVID-19 (1st February to 30th April 2020). We reconstructed candidate models as per their original descriptions and evaluated performance for their original intended outcomes (clinical deterioration or mortality) and time horizons. We assessed discrimination using the area under the receiver operating characteristic curve (AUROC), and calibration using calibration plots, slopes and calibration-in-the-large. We calculated net benefit compared to the default strategies of treating all and no patients, and against the most discriminating predictor in univariable analyses, based on a limited subset of *a priori* candidates.

Results

We tested 22 candidate prognostic models among a cohort of 411 participants, of whom 180 (43.8%) and 115 (28.0%) met the endpoints of clinical deterioration and mortality, respectively. 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 a novel model for prediction of deterioration <14 days from admission (0.78; 0.74-0.82). Calibration appeared generally poor for models that used probability outcomes. In univariable analyses, admission oxygen saturation on room air was the strongest predictor of in-hospital deterioration (AUROC 0.76; 0.71-0.81), while age was the strongest predictor of in-hospital mortality (AUROC 0.76; 0.71-0.81). No prognostic model demonstrated consistently higher net benefit than using the most discriminating univariable predictors to stratify treatment, across a range of threshold probabilities.

Conclusions

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 offer 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 capacity⁵. 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-19⁸. 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-19⁸. 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 for COVID-19 at a single centre.

Methods

Identification of candidate prognostic models

We used a published living systematic review to identify candidate prognostic models for COVID-19 indexed in PubMed, Embase, Arxiv, medRxiv, or bioRxiv until 5th May 2020⁸. 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. 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 studies¹³. We included consecutive adults admitted to University College Hospital London with a final diagnosis of PCR-confirmed 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).

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, clinical observations, laboratory measurements, radiology reports, and clinical outcomes. We defined 'clinical deterioration' as a composite outcome including 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 ≥ 6 ¹⁴. The rationale for this composite outcome is to make the endpoint more generalisable between centres, since hospital respiratory management algorithms may vary substantially. Each chest radiograph was reported by a single radiologist, reflecting routine clinical conditions, using British Society of Thoracic Imaging criteria, and using a modified version of the Radiographic Assessment of Lung Edema (RALE) score^{15,16}. 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 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 models that used ICU admission or death, or 'severe' COVID-19 or death, as composite endpoints, we used our 'clinical deterioration' endpoint as the primary outcome, as defined above. 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.

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 and quartile plots, and by quantifying calibration slopes and calibration-in-the-large (CITL). A perfect calibration slope should be 1; slopes <1 indicate overfitting, while slopes >1 reflect underfitting. 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}.

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 $\times$ prevalence $- (1 - \text{specificity}) \times (1 - \text{prevalence}) \times w$ where w is the odds at the threshold probability and the prevalence is the proportion of patients who experienced the outcome²⁰. 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. 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 R²¹.

We handled missing data using multiple imputation by chained equations²², using the *mice* package in R²³. All variables in the final prognostic models were included in the imputation model to ensure compatibility²². A total of 10 imputed datasets were generated; discrimination and calibration metrics were pooled using Rubin's rules²⁴. Individual predictions for each prognostic model were averaged across imputations for each participant in order to generate pooled calibration plots, ROC curves and decision curves.

All analyses were conducted in R (version 3.5.1).

Sensitivity analyses

We recalculated discrimination and calibration parameters for each candidate model using a complete case analysis. 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 R²⁵.

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 attendees²⁶, hospitalised patients^{12,27}, people with suspected infection¹⁰ or community-acquired pneumonia¹¹, respectively. Of the 17 models developed specifically for COVID-19, most (10/17) were developed using datasets originating in China. 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.

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 (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. 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 studies²⁸. 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%). Supplementary Figure 4 shows missingness of each prognostic model in the complete case dataset, stratified by the outcomes of interest, due to unavailability of predictor variables.

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 was noted to have reasonable discrimination for 30-day mortality (AUROC 0.75; 95% CI 0.70 - 0.80), and qSOFA discriminated in-hospital mortality poorly (AUROC 0.6; 95% CI 0.55 - 0.65). ROC curves are shown for each candidate model in Supplementary Figure 5.

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 6 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 different outcomes 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 7 and 8). 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 9). 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 complete case analysis revealed similar results to the primary multiple imputation approach (Supplementary Table 1). 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 10). Finally, internal validation using bootstrapping showed near zero optimism for discrimination and calibration parameters for the univariable models (Supplementary Table 2).

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, discrimination appeared modest or poor for most models. NEWS2 performed reasonably well for prediction of deterioration over a 24-hour interval, achieving an AUROC of 0.78, while the Carr ‘final’ model²⁹ also had reasonable discrimination (AUROC 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, no prognostic model demonstrated clinical utility consistently greater than using oxygen saturation on room air to predict deterioration, or patient age to predict mortality.

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 review⁸ 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 benefit¹⁷. 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 Scale¹⁴

A weakness of the current study is that it is based on 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. We therefore performed multiple imputation, in keeping with recommendations for

development and validation of multivariable prediction models, in our primary analyses³⁰. Findings were similar in the complete case sensitivity analysis, thus supporting the robustness of our results. Thirdly, a number of models could not be reconstructed in our data. For some models, this was due to 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-19¹⁶. 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 models¹³. 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 (such as NEWS2) are designed for dynamic in-patient monitoring. 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 no COVID-19-specific models can currently be recommended for routine clinical use. 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 models across heterogeneous populations, and develop and validate novel prognostic models, through individual participant data meta-analysis³¹. 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 utility¹³.

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. 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 Predecki, 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.

Ethical approval

This study was approved by East Midlands - Nottingham 2 Research Ethics Committee (REF: 20/EM/0114).

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. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* (2020). doi:10.1001/jama.2020.6775
2. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
3. Grasselli, G., Pesenti, A. & Cecconi, M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy. *JAMA* **323**, 1545 (2020).
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. (2020). Available at: <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-17-clinical/>. (Accessed: 14th May 2020)
5. Li, R. *et al.* The demand for inpatient and ICU beds for COVID-19 in the US: lessons from Chinese cities. *medRxiv* 2020.03.09.20033241 (2020). doi:10.1101/2020.03.09.20033241
6. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Preliminary Report. *N. Engl. J. Med.* NEJMoa2007764 (2020). doi:10.1056/NEJMoa2007764
7. Horby, P. *et al.* Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report. *medRxiv* 2020.06.22.20137273 (2020). doi:10.1101/2020.06.22.20137273
8. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, (2020).
9. Wolff, R. F. *et al.* PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **170**, 51 (2019).
10. Seymour, C. W. *et al.* Assessment of Clinical Criteria for Sepsis. *JAMA* **315**, 762 (2016).
11. Lim, W. S. *et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* **58**, 377–382 (2003).
12. Royal College of Physicians. National Early Warning Score (NEWS) 2 | RCP London. Available at: <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>. (Accessed: 1st July 2020)
13. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, (2015).
14. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection, J. C. *et al.* A minimal common outcome measure set for COVID-19 clinical research. *Lancet. Infect. Dis.* **0**, (2020).
15. Wong, H. Y. F. *et al.* Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients. *Radiology* 201160 (2019). doi:10.1148/radiol.2020201160
16. COVID-19 Resources | The British Society of Thoracic Imaging. Available at: <https://www.bsti.org.uk/covid-19-resources/>. (Accessed: 1st July 2020)
17. Riley, R. D., Windt, D. van der, Croft, P. & Moons, K. G. M. *Prognosis research in healthcare*: concepts, methods, and impact.
18. Kamarudin, A. N., Cox, T. & Kolamunnage-Dona, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* **17**, 53 (2017).
19. Hewitt, J. *et al.* The effect of frailty on survival in patients with COVID-19 (COPE): a multicentre, European, observational cohort study. *Lancet. Public Heal.* **0**, (2020).
20. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic Progn. Res.* **3**, 18 (2019).
21. Brown, M. rmda: Risk Model Decision Analysis. (2018).
22. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399 (2011).
23. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45(3)**, 1–67 (2011).
24. Rubin, D. B. *Multiple imputation for nonresponse in surveys*. (Wiley-Interscience, 2004).
25. Harrell Jr, F. E. 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.* **255**, (2004).
27. Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. & Featherstone, P. I. 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* **84**, 465–470 (2013).
28. Collins, G. S., Ogundimu, E. O. & Altman, D. G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat. Med.* **35**, 214–26 (2016).
29. Carr, E. *et al.* Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *medRxiv* (2020).
30. KG, M. *et al.* Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann. Intern. Med.* **162**, (2015).
31. Debray, T. P. A. *et al.* Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLOS Med.* **12**, e1001886 (2015).
32. Subbe, C. P., Kruger, M., Rutherford, P. & Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM* **94**, 521–526 (2001).
33. Bello-Chavolla, O. Y. *et al.* 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). doi:10.1210/clinem/dgaa346
34. Caramelo, F., Ferreira, N. & Oliveiros, B. Estimation of risk factors for COVID-19 mortality - preliminary results. *medRxiv* 2020.02.24.20027268 (2020). doi:10.1101/2020.02.24.20027268
35. Colombi, D. *et al.* Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology* 201433 (2020). doi:10.1148/radiol.2020201433
36. Galloway, J. B. *et al.* 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.* (2020). doi:10.1016/J.JINF.2020.05.064
37. Guo, Y. *et al.* Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. *medRxiv* 2020.04.17.20064691 (2020). doi:10.1101/2020.04.17.20064691
38. Cambridge Clinical Trials Unit. TACTIC trial. Available at: <https://cctu.org.uk/portfolio/COVID-19/TACTIC>. (Accessed: 1st July 2020)
39. Chen, X. & Liu, Z. Early prediction of mortality risk among severe COVID-19 patients using machine learning. *medRxiv* 2020.04.13.20064329 (2020). doi:10.1101/2020.04.13.20064329
40. Huang, H. *et al.* Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis. *medRxiv* 2020.03.28.20045989 (2020). doi:10.1101/2020.03.28.20045989
41. Ji, D. *et al.* Prediction for Progression Risk in Patients with COVID-19 Pneumonia: the CALL Score. *Clin. Infect. Dis.* (2020). doi:10.1093/cid/ciaa414
42. Lu, J. *et al.* 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* 2020.02.20.20025510 (2020). doi:10.1101/2020.02.20.20025510
43. Shi, Y. *et al.* 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. *et al.* Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. *medRxiv* 2020.03.28.20045997 (2020). doi:10.1101/2020.03.28.20045997
45. Yan, L. *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288 (2020).
46. Zhang, H. *et al.* Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. *medRxiv* 2020.04.28.20082222 (2020). doi:10.1101/2020.04.28.20082222
47. Hu, H., Yao, N. & Qiu, Y. Comparing Rapid Scoring Systems in Mortality Prediction of Critically Ill Patients With Novel Coronavirus Disease. *Acad. Emerg. Med.* **27**, 461–468 (2020).

Tables

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 10⁹/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 ³²	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 ³³	Bello-Chavolla	Mexico	Confirmed COVID-19 patients presenting in	COVID-specific	Mortality (30 day)	Age ≥65 years, diabetes, early-onset diabetes, obesity, age <40 years, chronic kidney disease,	Cox regression	Points-based score

primary care						hypertension, immunosuppression (rheumatoid arthritis, lupus, HIV or immunosuppressive drugs)		
Caramelo et al ³⁴	Caramelo ⁵	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 ²⁹	'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 ⁵ (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 confirmed COVID-19	COVID-specific	Progression to severe COVID-19 at 10 days (defined as respiratory rate ≥ 30, resting	Age (> 60 years), lymphocytes (≤1 x 10 ⁹ /L) LDH (<250, 250-500, >500 U/L), comorbidity (hypertension,	Cox regression	Points-based score

					oxygen saturation $\leq$ 93%, PaO ₂ /FIO ₂ $\leq$ 300 mmHg, requirement of mechanical ventilation or worsening of lung CT findings)	diabetes, cardiovascular disease, chronic lung disease, or HIV)		
Lu et al ⁴²	Lu	China	Inpatients with suspected or confirmed COVID- 19	COVID-specific	Mortality (12 days)	Age $\geq$ 60 years, CRP $\geq$ 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

9 [#]MEWS and REMS were evaluated among people with COVID-19 by Hu et al⁴⁷, and thus were included in the present study.

10 ^{\$}No model intercept was available; the intercepts for these models were therefore calibrated to the validation dataset, using the model linear predictors as
11 offset terms.

12

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

14 Laboratory and physiological measurements reflect parameters at the time of hospital admission. N column shows number of participants with available data
15 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 ($\times 10^9$)	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)

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

17 For each model, performance is evaluated for its original intended outcome, shown in 'Primary outcome' column. AUROC = area under the receiver operating
18 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)	N/A
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)	N/A
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)		

Figures

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

For each plot, the blue line represents a Loess-smoothed calibration curve, and scatter points show quartiles of predicted risk. 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. Calibration-in-the-large is therefore not shown for these models, since it is zero by definition. 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 calibration plots.

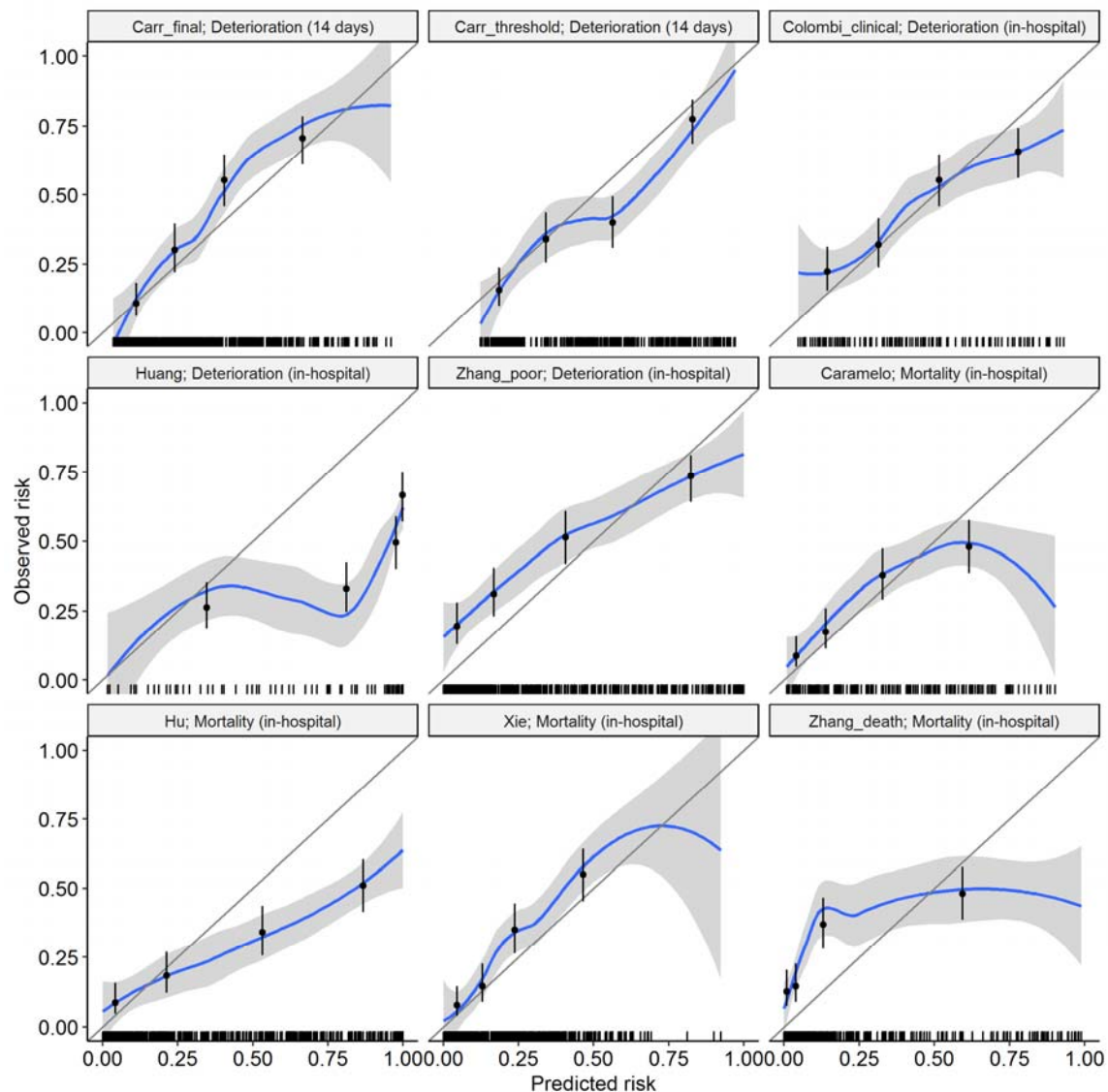


Figure 2: Decision curve analysis showing delta net benefit of each candidate model, compared to treating all patients and best univariable predictors.

For each analysis, the endpoint is the original intended outcome and time horizon for the index model. 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. Individual predictions for each prognostic model were averaged across imputations for each participant in the dataset in order to generate pooled decision curve plots. Delta net benefit is shown with Loess-smoothing. Black dashed line indicates threshold above which index model has greater net benefit than the comparator. Full decision curves for each candidate model are shown in Supplementary Figure 9.

