
Martin Wiegand,1,2 Sarah L Cowan,3 Claire S Waddington,4 David J Halsall,3 Victoria L Keevil,5,6 Brian D M Tom,2 Vince Taylor,6 Effrossyni Gkrania-Klotsas,7 Jacobus Preller,3 Robert J B Goudie2

ABSTRACT

Objectives To develop a disease stratification model for COVID-19 that updates according to changes in a patient’s condition while in hospital to facilitate patient management and resource allocation.

Design In this retrospective cohort study, we adopted a landmarking approach to dynamic prediction of all-cause in-hospital mortality over the next 48 hours. We accounted for informative predictor missingness and selected predictors using penalised regression.

Setting All data used in this study were obtained from a single UK teaching hospital.

Participants We developed the model using 473 consecutive patients with COVID-19 presenting to a UK hospital between 1 March 2020 and 12 September 2020; and temporally validated using data on 1119 patients presenting between 13 September 2020 and 17 March 2021.

Primary and secondary outcome measures The primary outcome is all-cause in-hospital mortality within 48 hours of the prediction time. We accounted for the competing risks of discharge from hospital alive and transfer to a tertiary intensive care unit for extracorporeal membrane oxygenation.

Results Our final model includes age, Clinical Frailty Scale score, heart rate, respiratory rate, oxygen saturation/ fractional inspired oxygen ratio, white cell count, presence of acidosis (pH <7.35) and interleukin-6. Internal validation achieved an area under the receiver operating characteristic (AUROC) of 0.90 (95% CI 0.87 to 0.93) and temporal validation gave an AUROC of 0.86 (95% CI 0.83 to 0.88).

Conclusions Our model incorporates both static risk factors (eg, age) and evolving clinical and laboratory data, to provide a dynamic risk prediction model that adapts to both sudden and gradual changes in an individual patient’s clinical condition. On successful external validation, the model has the potential to be a powerful clinical risk assessment tool.

Trial registration The study is registered as ‘researchregistry5464’ on the Research Registry (www.researchregistry.com).

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Our dynamic prediction model can incorporate patient data as it accumulates throughout a hospital visit.
⇒ We use the established statistical landmarking approach to dynamic prediction; we account for competing risks for the primary outcome of in-hospital mortality, and the potentially informative availability of clinical and laboratory data.
⇒ The sample size of the first wave of patients admitted with severe COVID-19 was relatively low, due to the lower incidence in Cambridgeshire, but increased significantly during the winter months of 2020/2021, providing the opportunity to temporally validate the model.
⇒ As a single-centre study, the presented model will require external validation to assess its performance in other cohorts; and also if there are significant changes in the characteristics of new variants or the management thereof.
⇒ Our work also highlights the adaptability of the statistical landmarking framework to be used to model individual patient outcomes using frequently collected hospital data.

INTRODUCTION

SARS-CoV-2 virus infection, the cause of COVID-19, results in a spectrum of disease ranging from asymptomatic infection through to life-threatening disease requiring critical care and even death. For patients admitted to hospital, it is essential to identify who is at risk of deterioration and death to enable timely targeted interventions (such as immune modulation and mechanical ventilation), to facilitate appropriate resource allocation and patient flow, and to inform discussions with patients and families.
Most existing disease severity prediction models for COVID-19 use only data that are available at the time of admission to hospital. Such point-of-admission models have been proposed for both mortality and composite escalation/mortality outcomes, including new and repurposed severity and early warning scores and time-to-event models. While some markers of severity, such as sex and age can be assumed constant for the duration of the hospital visit, others, such as clinical observations and blood test results, can change markedly over the course of admission. COVID-19 is a dynamic disease in which patients may deteriorate over a short time period or suffer acute complications, for example, thromboembolism. This may have a significant effect on a patient’s prognosis that cannot be foreseen by a point-of-admission model.

A model with the ability to adjust predictions at arbitrary time points by including updated patient information could greatly aid in clinical decision-making. Dynamic models that assimilate clinical data as it accrues may provide more accurate and clinically useful prediction of a patient’s clinical course and prognosis over the subsequent days than point-of-admission models. Predictive models that incorporate post-admission information are limited in number and scope. Some models for predicting mortality or deterioration have used information after admission, but do not continue beyond the first few days of admission. More recent time-varying Cox models (for mortality and escalation) and machine learning models (for mortality) have used additional post-admission data. However, time-varying Cox models should not be used for prediction, because they require knowledge of clinical information from the future to calculate the hazard function, which is impossible in practice. Furthermore, while indicating promising discrimination, these models use clinically unjustifiable or unclear methods for handling missing data and censoring, and do not account for informative missingness or consider the effect of treatments. Informative missingness describes the fact that in routinely collected data the availability (or absence) of a result or observation may be related to the probability of the outcome. For example, a more extensive panel of investigations may be sent for patients thought more likely to benefit from escalation in care, such as transfer to an intensive care unit (ICU). While often ignored, such effects can be strong in electronic health record (EHR) data.

We propose a prognostic risk stratification score for hospital patients with COVID-19, based on prediction of mortality in the subsequent 48 hours, using routinely collected clinical data. Our model is based on a principled statistical approach called landmarking that allows inclusion of any time-varying clinical parameters recorded prior to the time of prediction, while appropriately accounting for censoring and changes in the set of patients at risk. The model accounts for informative missingness and competing risks, which arise when there are two or more mutually exclusive outcomes, for example, once a patient is discharged, the risk of inhospital mortality (during that admission) is removed. Therefore, discharge is a ‘competing risk’ when viewed from the perspective of in-hospital mortality. We account for competing risks within the landmarking framework using a recently proposed approach that has not previously been used to model individual patient outcomes using frequently collected EHR data.

**MATERIALS AND METHODS**

**Study design**

This is a retrospective cohort study of all patients presenting to Cambridge University Hospitals, a regional, tertiary care, university hospital in the East of England, between 1 March 2020 and 17 March 2021. This hospital is the sole admission hospital for patients in its immediate catchment population with COVID-19, and is a regional referral centre for a wide range of specialist services, which do not include extracorporeal membrane oxygenation (ECMO).

We report our findings according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis reporting guidelines.

**Study population**

All adults (≥18 years of age) presenting to hospital during the study period and diagnosed with COVID-19 were included. Diagnosis was based on either a positive diagnostic SARS-CoV-2 test during or up to 14 days prior to the hospital visit, or a clinical diagnosis of COVID-19 (online supplemental appendix 1). Patients with clinically diagnosed COVID-19 (based on symptoms, and the clinical opinion of the treating clinician) were included because diagnostic testing was limited during the early stages of the pandemic.

We include only the first hospital visit for each patient involving (or subsequent to) their first positive test; any readmissions were excluded. Nosocomial infection was defined as a first positive SARS-CoV-2 test or diagnosis more than 10 days after hospital admission. Since we first train our model at 6 hours (to allow time for laboratory investigations), patients who died, were discharged or were classified as end of life within 6 hours of presentation to hospital were excluded.

All patients were treated as per detailed local guidance in use in the hospital at the time. Patients were also eligible for inclusion in relevant clinical trials running at the hospital during the study period (online supplemental appendix 2).

**Outcomes**

Throughout each patient’s hospital visit, we aim to predict all-cause in-hospital mortality during the next 48 hours, a time period that we refer to as the ‘prediction horizon’. We also considered two competing risks: transfer to a tertiary ICU for ECMO; and discharge from the hospital...
due to clinical improvement. Patients were followed up until 19 March 2021.

**Patient and public involvement**

No patients were involved in the design of this study.

**Model development**

A list of 59 candidate clinical parameters (online supplemental etable 1) were chosen based on existing point-of-admission prediction models for COVID-19; components of scores predicting mortality in critical illness more generally, such as the Sequential Organ Failure Assessment and the Acute Physiology and Chronic Health Evaluation II; and clinical opinion as to other likely predictors. These are divided into five categories: demographics, comorbidities, observations, laboratory tests, and treatments, interventions and level of care.

Basic patient demographics were extracted from the hospital EHR: sex, ethnicity and age at hospital presentation; and deviation from standard ranges of body mass index, using the most recent measurement (up to 1 year old).

Twelve comorbidities that have previously been associated with COVID-19 severity were identified by the presence of the corresponding ICD-10 codes entered in the EHR prior to the time at which the prediction is made (either before or during the hospital visit). Online supplemental etable 2 provides the ICD-10 codes used to define each comorbidity. In addition to specific comorbidities, frailty among patients over 65 years old was assessed by the Clinical Frailty Scale (CFS) score (online supplemental appendix 3).

We included the following observations that are regularly recorded in the EHR: heart rate (HR), mean arterial pressure, temperature and respiratory rate (RR). Oxygen saturation (SpO2)/fractional inspired oxygen (FiO2) ratio was calculated (where SpO2 and FiO2 were available at the same time point) to indicate the severity of hypoxia. SpO2 itself was not included as a potential predictor as our exploratory work suggested that, without accounting for FiO2, this largely reflected a patient’s assigned oxygen saturation targets, and therefore, acted as a proxy for underlying respiratory disease (eg, patients with chronic obstructive pulmonary disease being assigned a lower SpO2 target). Where only oxygen flow rate was available, FiO2 was estimated according to the Sequential Organ Failure Assessment and the Acute Physiology and Chronic Health Evaluation II; and clinical opinion as to other likely predictors. These make it possible to include repeatedly measured clinical parameters into the prediction model, so that predictors reflect any changes in the trajectory of the patient, while appropriately accounting for censoring and changes in the at-risk population. If the primary outcome was recorded within the prediction horizon of a landmark time, we recorded the outcome at the relative time from landmark to event; events after the prediction horizon were censored. Patients who had had any event prior to the landmark time were excluded, since these patients were no longer at risk. The first landmark is 6 hours after presentation to allow time for clinical information to accrue, or at the point of COVID-19 diagnosis for nosocomial patients (diagnosed 10 days or more after presentation). We only used data at each landmark time from patients being actively treated for COVID-19 at that point in time: landmark times after transition to end of life care were omitted, meaning that no predictions were made at these timepoints, although events was made. Where more than one result was available, we used the most recent result. In addition, for seven of the most frequently measured blood tests (C reactive protein (CRP), white cell count (WCC), platelets, haemoglobin, creatinine, sodium, potassium), we included the median trend. The neutrophil/lymphocyte and interleukin-6/interleukin-10 (serum IL-6/IL-10) ratios have previously been identified as prognostic, therefore we also considered these as potential predictors.

We included five indicators of treatments, interventions and levels of care. The level of care of the patient was summarised by whether the patient had been in an ICU bed in the previous 24 hours. Mechanical ventilation was defined as patients receiving invasive ventilation during the previous 24 hours, either via endotracheal tube or tracheostomy. The use of renal replacement therapy during the last 24 hours was identified from the EHR. Cardiovascular support was defined as the administration of any vasopressors or inotropes in the last 24 hours. Steroid administration has been shown to reduce 28-day mortality in patients with COVID-19.

**Models**

We used a landmarking approach, which has been proposed for dynamic prediction. At intervals of 24 hours (referred to as the ‘landmark times’), we trained time-to-event models, using clinical parameters recorded before (or at) the landmark time as predictors. This makes it possible to include repeatedly measured clinical parameters into the prediction model, so that predictors reflect any changes in the trajectory of the patient, while appropriately accounting for censoring and changes in the at-risk population. If the primary outcome was recorded within the prediction horizon of a landmark time, we recorded the outcome at the relative time from landmark to event; events after the prediction horizon were censored. Patients who had had any event prior to the landmark time were excluded, since these patients were no longer at risk. The first landmark is 6 hours after presentation to allow time for clinical information to accrue, or at the point of COVID-19 diagnosis for nosocomial patients (diagnosed 10 days or more after presentation). We only used data at each landmark time from patients being actively treated for COVID-19 at that point in time: landmark times after transition to end of life care were omitted, meaning that no predictions were made at these timepoints, although events was made. Where more than one result was available, we used the most recent result. In addition, for seven of the most frequently measured blood tests (C reactive protein (CRP), white cell count (WCC), platelets, haemoglobin, creatinine, sodium, potassium), we included the median trend. The neutrophil/lymphocyte and interleukin-6/interleukin-10 (serum IL-6/IL-10) ratios have previously been identified as prognostic, therefore we also considered these as potential predictors.

We included five indicators of treatments, interventions and levels of care. The level of care of the patient was summarised by whether the patient had been in an ICU bed in the previous 24 hours. Mechanical ventilation was defined as patients receiving invasive ventilation during the previous 24 hours, either via endotracheal tube or tracheostomy. The use of renal replacement therapy during the last 24 hours was identified from the EHR. Cardiovascular support was defined as the administration of any vasopressors or inotropes in the last 24 hours. Steroid administration has been shown to reduce 28-day mortality in patients with COVID-19.

We therefore include an indicator of whether the patient had received treatment dose steroids (defined as 6 mg dexamethasone daily or an equivalent dose of prednisolone, hydrocortisone or methylprednisolone) during their hospital admission prior to the time the prediction is made.

**Open access**

BMJ Open: first published as 10.1136/bmjopen-2021-060026 on 5 September 2022. Downloaded from http://bmjopen.bmj.com/ on September 17, 2023 by guest. Protected by copyright.
occuring within the existing prediction horizon were still included. Patients receiving end of life care were identified from their recorded escalation status in the EHR and by manual review of patient notes containing relevant keywords (online supplemental appendix 4). We used a supermodel approach in which the time-to-event model is assumed constant across landmark times. 39

Specifically, we used a competing-risks landmarking approach, 27 41 which uses a Fine-Gray competing risk model to predict in-hospital death, and account for the competing risks of hospital discharge and transfer for ECMO. 42 43 A Fine-Gray model uses subdistribution hazards, which are directly related to the cumulative incidence function, by which the probability of an event of interest occurring can be estimated. While the cause-specific hazard function used by Cox models is preferable for inferring biological mechanisms, subdistribution hazard-based models are preferable for prediction. 44

**Missing values**
We handled missing data using the missingness indicator approach since the recording in the EHR of a clinical parameter, regardless of the value, is often indicative of the treating health professional’s contemporaneous view of the patient’s condition. 45 46 Conceptually our approach, as well as estimating the ‘effect’ of a unit increase of a particular clinical parameter (as is standard in all regression approaches), estimates the ‘effect’ of a variable being ‘missing’. A clinical parameter may be ‘missing’ (also referred to as ‘not recorded’ or ‘not available’) due to not being measured or due to not being documented in the EHR. For each potential predictor in the model we also include a missingness indicator, which indicates that no data were recorded during the corresponding time period. This approach allows clinical parameters with an incomplete record to be included in our model and avoids the need to make the missing at random (MAR) assumption that is unlikely to hold in these data. 47 For each parameter, ultimately one of the two expressions is used for prediction for each patient at each timepoint: either a coefficient describing the relationship with the clinical parameter when it is recorded, or a fixed value if the clinical parameter is missing.

Blood tests are considered missing if the most recent measurement was collected more than 48 hours prior to the landmark time. When a blood test is repeated during the previous 48 hours, the most recent result is used. The vital signs we considered did not have missing values at any included landmark.

**Predictor selection**
To select the most predictive parameters for the model we adopted an approach proposed for Fine-Gray models 48 that uses standard penalised variable selection, specifically smoothly clipped absolute deviations (SCAD), with the tuning parameter chosen to minimise the Bayesian information criterion. We paired parameters together with their corresponding missingness indicator to prevent inclusion of an incompletely recorded parameter without its missingness indicator, using the group SCAD 49; but also allowed for the missingness indicators to be included by themselves.

The development and validation of the model has been carried out in R V.3.6. 50

**Model assessment**
Quantitative assessment of discrimination was performed using the area under the receiver operating characteristic (AUROC) curve, in which 0.5 indicates no discrimination and 1.0 indicates perfect discrimination. For validation of the performance of the model on the training data, in addition to the unadjusted AUROC, we also performed repeated fivefold (split into 80% training, 20% validation data) cross-validation to account for uncertainty and over-optimism due to the complete model building process (including variable selection). 51 We also calculated the precision-recall (PR) curve and the area under the PR curve (AUPRC) since it provides a clearer performance summary than AUROC when the primary outcome has low incidence, as here. 42 We assessed clinical benefit visually via the number needed to evaluate (NNE), defined as 1/positive predictive value (1/PPV), against the sensitivity. We also calculated the net benefit curve. 53

We assessed calibration visually using a calibration plot of predicted risk against observed mortality rate. We also quantitatively assessed the calibration slope and calibration-in-the-large. 54

**Sensitivity analyses**
To assess whether the model is unduly influenced by patients with long hospitalisations we retrained the model using only each patient’s first 28 landmark times (spanning 28 days). We also assessed the sensitivity of our model assessment by stratifying by whether COVID-19 was confirmed by a positive SARS-CoV-2 diagnostic test, and according to whether patients had received at least a single COVID-19 vaccination dose prior to admission (Oxford-AstraZeneca Covishield, Pfizer-BionTech Comirnaty or Moderna Spikevax). We also assessed the discrimination of the model if we removed all laboratory tests. Finally, we explored the implications of replacing the 24-hour prediction horizon with a 72-hour prediction horizon.

**RESULTS**

**Development of prediction model**
We developed the model using data from wave 1 (1 March 2020 to 12 September 2020), with the end date chosen since only a single patient remained in hospital with COVID-19 on this date. A total of 519 patients presented to hospital with COVID-19 during wave 1, of whom 46 were excluded due to discharge (34), death (2) or transition to end-of-life care (10) prior to the first landmark time (ie, within 6 hours of presentation). The
Table 1  Cohort demographics and clinical features in wave 1 and wave 2

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Wave 1 (training data set)</th>
<th>Wave 2 (validation data set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of patients</td>
<td>473</td>
<td>1119</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>196 (41.4)</td>
<td>579 (48.3)</td>
</tr>
<tr>
<td>Age at admission, median (IQR), years</td>
<td>69 (55–81)</td>
<td>65 (49–79)</td>
</tr>
<tr>
<td>Admission BMI, median (IQR), kg/m² *</td>
<td>25.7 (22.1–29.9)</td>
<td>27.35 (23.5–32.1)</td>
</tr>
<tr>
<td>Clinical Frailty Score at admission for over 65 year olds, median (IQR)</td>
<td>5 (3–6)</td>
<td>5 (4–6)</td>
</tr>
<tr>
<td>Nosocomial infection, n (%)</td>
<td>36 (7.6)</td>
<td>116 (9.7)</td>
</tr>
<tr>
<td>Length of stay, median (IQR), days</td>
<td>10.8 (3.9–19.8)</td>
<td>8.0 (2.95–18.5)†</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>354 (74.8)</td>
<td>792 (66.1)</td>
</tr>
<tr>
<td>Asian</td>
<td>24 (5.1)</td>
<td>61 (5.1)</td>
</tr>
<tr>
<td>Black</td>
<td>10 (2.1)</td>
<td>13 (1.1)</td>
</tr>
<tr>
<td>Other</td>
<td>11 (2.3)</td>
<td>39 (3.3)</td>
</tr>
<tr>
<td>Prefer not to say/not recorded</td>
<td>74 (15.6)</td>
<td>294 (24.5)</td>
</tr>
<tr>
<td>Outcomes, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deceased in-hospital‡</td>
<td>99 (20.9)</td>
<td>145 (12.1)</td>
</tr>
<tr>
<td>Transferred for ECMO</td>
<td>5 (1.1)</td>
<td>1 (0.0)</td>
</tr>
<tr>
<td>Discharged alive</td>
<td>369 (78.0)</td>
<td>967 (80.7)</td>
</tr>
<tr>
<td>Remain in hospital on 19 March 2021</td>
<td>0 (0)</td>
<td>86 (7.2)</td>
</tr>
<tr>
<td>Support/treatments received during hospital stay, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICU admission</td>
<td>103 (21.8)</td>
<td>238 (19.8)</td>
</tr>
<tr>
<td>Invasive mechanical ventilation</td>
<td>82 (17.3)</td>
<td>172 (14.3)</td>
</tr>
<tr>
<td>Non-invasive ventilation (CPAP or Bi-PAP)</td>
<td>37 (7.8)</td>
<td>144 (12.9)</td>
</tr>
<tr>
<td>Cardiovascular support</td>
<td>86 (18.2)</td>
<td>170 (14.2)</td>
</tr>
<tr>
<td>Renal replacement therapy</td>
<td>32 (6.8)</td>
<td>39 (3.3)</td>
</tr>
<tr>
<td>Steroid treatment</td>
<td>120 (25.4)</td>
<td>619 (55.3)</td>
</tr>
<tr>
<td>Admission observations*, median (IQR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean arterial pressure, mm Hg</td>
<td>86 (76–96)</td>
<td>88 (78–98)</td>
</tr>
<tr>
<td>Heart rate, beats/min</td>
<td>83 (72–93)</td>
<td>81 (72–91)</td>
</tr>
<tr>
<td>Temperature, °C</td>
<td>37.2 (36.7, 37.8)</td>
<td>36.9 (36.5, 37.4)</td>
</tr>
<tr>
<td>Respiratory rate, breaths/min</td>
<td>19 (17, 22)</td>
<td>18 (17, 20)</td>
</tr>
<tr>
<td>SpO2, %</td>
<td>96 (94, 97)</td>
<td>95 (94, 97)</td>
</tr>
<tr>
<td>SpO2/FiO2 ratio</td>
<td>448 (337, 457)</td>
<td>452 (400, 462)</td>
</tr>
</tbody>
</table>

*First after positive SARS-CoV-2 test result for hospital-acquired COVID-19 cases.†For wave 2, includes stays completed by 19 March 2021 only.‡Note that because follow-up continued until 19 March 2021, the outcome during the 48-hour prediction horizon was known for all landmark times up to 17 March 2021, and so patients remaining in hospital at the study end-date could be included in the validation.‡Total in-hospital mortality including those patients who were classified as ‘end of life’ prior to death. Bi-PAP, Bi-level positive airway pressure; BMI, body mass index; CPAP, Continuous positive airway pressure; CRP, C reactive protein; ECMO, extracorporeal membrane oxygenation; FiO2, fractional inspired oxygen; ICU, intensive care unit; IL-6, interleukin-6; SpO2, oxygen saturation; WCC, white cell count.

In total, we included 6846 landmark times for training the model, with a median of 9 (IQR 5–17) landmark times per patient. In the 48-hour prediction horizon following these landmark times, there were 119 in-hospital death events (1.7% of landmarks), 658 hospital discharge events (9.6%) and 10 transfers for ECMO (0.1%). Note that, since landmarks occur every 24 hours and the prediction horizon is 48 hours, patient events will usually occur within the prediction horizon of two adjacent landmark times. Online supplemental etable 1 reports summary statistics, missingness and the number of measurements available per landmark time for each predictor. No patients were excluded due to missing data.

Model results

Our proposed model (table 2 and online supplemental etable 3) for 48-hour in-hospital mortality includes age, CFS score,55 HR, RR, SpO2/FiO2 ratio,34,35 WCC, acidosis (pH < 7.35) and interleukin-6 (IL-6). The mortality probability can be calculated using the calculator at http://shiny.mrc-bsu.cam.ac.uk/apps/covid19mortalityrisk/; see online supplemental appendix 5 for details.
Internal performance assessment

Figure 1 shows the internal performance metrics for the model in table 2 (using the training data). The unadjusted internal area under receiver operating characteristic curve (AUROC) was 0.90 (95% CI 0.87 to 0.93) and the median cross-validation AUROC was 0.87, both indicating good discrimination (figure 1A). The PR curve (figure 1B) also showed good discrimination, with an AUPRC of 0.31 in a population with 48 hour in-hospital mortality of 0.017 (1.7%), and we observe that the NNE <10 for sensitivity less than 0.75 (figure 1C). Figure 1D shows the calibration plot. The calibration intercept was −0.02 (95% CI −0.22 to 0.17) indicating that the mean predicted probabilities matched the mean observed mortality, while the calibration slope was 1.16 (95% CI 1.02 to 1.31) suggesting that the observed mortality in high predicted risk patients slightly exceeded the predicted mortality risk, although the CIs are wide. The net benefit curve for risk stratification by the proposed model is clearly higher than for the two non-model alternatives of classifying everyone or no-one as high risk patients, indicating the clinical utility of the dynamic model (online supplemental efigure 1).

Temporal validation of prediction model

We assessed the performance of the model by applying it to held-out data corresponding to admissions during wave 2 (13 September 2020 and 17 March 2021). A total of 1119 patients presented to the study hospital during this period. In total, we tested the model using 12 981 landmark times, with a median of 6 (2–14) landmark times per patient. In the 48-hour prediction horizon following these landmark times, there were 253 in-hospital death events (1.9% of landmarks), 1615 hospital discharge events (12.4%) and 2 transfers for ECMO (0.015%). Forty-seven landmark times were omitted due to missing vital sign data. Characteristics are summarised in table 1. Of note, compared with wave 1, patients presenting in wave 2 were slightly younger and more likely to be female, evidenced by a more balanced data set.

Figure 2 shows the temporal validation performance metrics, obtained by applying the trained model (table 2) to the wave 2 patients. The receiver operating characteristic (ROC) curve (figure 2A) shows the model continues to discriminate well, with AUROC 0.86 (95% CI 0.83 to 0.88). The PR curve (figure 2B) shows that PPV was consistently well above the 48-hour in-hospital mortality incidence of 0.019 (1.9%) in the wave 2 cohort across all sensitivities. The AUPRC 0.15 is reduced and the NNE is higher than in the training data, although NNE <10 for sensitivities between 0.02 and 0.65 (figure 2C). Figure 2D shows the calibration plot, which shows a tendency of the model to underpredict risk in the higher risk patients: calibration-in-the-large was 0.35 (95% CI 0.21 to 0.47), suggesting the mean of the predicted probabilities was lower than the mean observed mortality, and calibration slope was 0.90 (95% CI 0.82 to 0.99), indicating that the spread of predicted risk was significantly smaller than the spread of observed mortality. The calibration plot shows a good correspondence between observed mortality rate and predicted mortality risk for the lower-risk patients (<0.4), but due to the low incidence of mortality events among the landmarks corresponds less well for the higher-risk patients. This is evidenced by the considerable CIs. The net benefit curves for the proposed model surpasses both alternatives of classifying everyone and no-one as high-risk patients (online supplemental efigure 2).

Sensitivity analyses

The model trained on data from the first 28 landmarks for each patient closely resembled the model in table 2, although the Clinical Frailty Score and IL-6 were not selected (online supplemental etable 4).

We did not find evidence that the discrimination of the model was affected by the presence or absence of positive diagnostic SARS-CoV-2 results (rather than solely a clinical diagnosis). In patients with a positive SARS-CoV-2 test the AUROC was 0.90 (95% CI 0.87 to 0.93) in the training dataset and 0.85 (95% CI 0.83 to 0.88) in the validation dataset; whereas for patients with only a clinical diagnosis the AUROC was 0.94 (95% CI 0.88 to 1.00) in the training dataset and 0.88 (95% CI 0.79 to 0.99) in the validation dataset.

A small number of patients (65 patients, for whom 874 landmarks were available) in our validation dataset had received a COVID-19 vaccine at the time of analysis: the AUROC of 0.88 (95% CI 0.71 to 0.99) for these patients suggested good discrimination and is consistent with the
unvaccinated patients (AUROC 0.85 (95% CI 0.83 to 0.88)).

To assess whether a simpler model could perform similarly we considered removing the laboratory tests from our model (online supplemental etables 5 and 6). The resulting model provided slightly worse discrimination with an AUROC of 0.88 (95% CI 0.85 to 0.90) in the training data and an AUROC of 0.85 (95% CI 0.82 to 0.87) in the validation data.

To assess whether CRP could serve as a proxy for IL-6 in our model when it is not available, we refitted the model with CRP in place of IL-6 (online supplemental etables 7 and 8). The AUROC was slightly lower on both training (0.89, 95% CI 0.85 to 0.93) and validation (0.84, 95% CI 0.81 to 0.87) data.

To assess the feasibility of our approach with an extended time horizon and the stability of the predictors selected, we refitted the model with a 72-hour prediction horizon. The resulting model matched the previous AUROC in the training data of 0.90 and the AUROC of 0.85 in the validation data set. The full performance metrics are available in online supplemental figures 3 and 4, and the model coefficients in online supplemental etable 9.

**DISCUSSION**

SARS-CoV-2 causes a wide spectrum of disease that can evolve over time, and may necessitate critical care management and even result in death. In light of the threat of further waves of coronavirus infections there is still a pressing clinical need to be able to anticipate disease severity and the trajectory of illness to facilitate patient management and resource allocation. The model described herein incorporates both static admission risk factors (age and CFS) and evolving clinical and laboratory data, providing a dynamic 48-hour risk prediction model that can adapt to both sudden and gradual changes in an individual patient’s clinical condition. The data used in the model were routinely collected demographic and clinical data from during the patient’s hospitalisation, automatically extracted from patient EHRs. As such, this
model could be readily incorporated into routine clinical care.

An extensive literature of risk stratification models for COVID-19 has developed since the start of the pandemic for a variety of endpoints and time horizons, although many have been described as unreliable in a large systematic review and remain unvalidated. Most of the literature considers point-of-admission scores, including the 4C Mortality score, which is based on demographic information, comorbidities and blood tests taken on admission. This score performed well in both large development and validation cohorts (AUROC 0.77 and 0.79, respectively) and has now been externally validated in several countries. Similarly the point-of-admission ISARIC 4C Deterioration score has been externally validated. Promising machine-learning alternatives have also been recently proposed, although these have not yet been independently validated in contrast to the 4C scores. Another approach has been to repurpose existing early warning scores, particularly the National Early Warning Score (NEWS) 2 score. The discrimination of NEWS2 for predicting 14-day deterioration (a composite outcome of transfer to ICU or death) has been described as moderate (AUROC 0.70), whereas for 24-hour clinical deterioration (initiation of ventilatory support or death) it has been described as reasonable (AUROC 0.78). The addition of age and additional physiological parameters and blood tests led to an improvement in discrimination (AUROC 0.735 and 0.78). NEWS2 is a core component of escalation pathways in most UK hospitals, with a national recommendation that patients with NEWS2 score of 7 or above should be reviewed by the intensive care team. The fact that NEWS2 can, to a degree, predict intensive care admission is therefore unsurprising.

The literature for risk prediction using dynamic, post-admission in-hospital data for patients with COVID-19 is much smaller. Some models have used only data from the first few days of admission. Others have used additional postadmission data in time-varying Cox models (for mortality and escalation), which require knowledge of clinical information from the future and so should

---

**Figure 2** Performance metrics for in-hospital mortality in the validation dataset. (A) Receiver operating characteristic plot, with labels indicating the corresponding threshold and the dashed line indicating the line of no discrimination. (B) Precision-recall plot, with the 3.1% observed incidence indicated by the dashed line. (C) NNE against sensitivity. (D) Calibration plot (with 95% CI), by tenths of predicted risk and a LOESS interpolation (grey), with the dashed line indicating perfect calibration. AUROC, area under the receiver operating characteristic; FPR, false positive rate; LOESS, locally estimated scatterplot smoothing; NNE, number needed to evaluate; PPV, positive predictive value; TPR, true positive rate.
for mortality and respiratory support have also been reported to have good discrimination, although missing data are either discarded or imputed, which ignores the problem of informative missingness and may limit their applicability on an individual patient level. They also do not account for competing risks, for example, hospital discharge. While this paper has been under review, a promising random forest-based dynamic model predicting the need for respiratory support (encompassing high flow nasal oxygen, and both invasive and non-invasive ventilation) and death has been proposed. This model reports high internal discrimination (AUROC 0.89, 95% CI 0.88 to 0.90) although PR is not reported. The most important predictors identified in this model were the SpO2/FiO2 ratio and the RR, both of which also feature in our model.

Both these observations play an important role in the decision to escalate the level of respiratory support, and so it is not clear whether for the 1 day horizon this model offers much in prediction that is not already apparent clinically (predicting the need for additional respiratory support after that decision has already been made serves little purpose). In addition, it is again unclear how the model handles missing data. Markov models have also been proposed that seek to model the trajectory of patients rather than focus on prediction as we do in this paper.

Our approach has several methodological strengths which we believe sets it apart from previous models. First, we accounted for competing risks, whereby the outcome (risk) of interest (in this case in-hospital mortality) can only happen while the patient is in hospital, and therefore the outcome of interest is ‘competing’ against the risk of transfer to another hospital and/or discharge from hospital. Allowance for this is important in predictive modeling. Second, we allowed for the possibility that the availability of observations and investigations may in itself reflect disease severity, as the fact that specific tests have been requested at a given time provides an insight into the treating clinicians’ contemporaneous view of the patient’s condition. While multiple imputation is often used in clinical prediction models because it gives unbiased estimates under the MAR assumption, it is unlikely that this assumption holds in the routinely collected EHR data that we use. The missing indicator method that we adopted does not rely on the MAR assumption and can improve predictive performance in EHR data.

Furthermore, we validated our model using data from different waves. As each wave included COVID-19 variants of different infectiousness that are potentially associated with different morbidity and mortality risks, the fact that the model performs reasonably (NNE <10 for sensitivities between 0.02 and 0.63) across waves further attests to the fact that our selected parameters are useful for prognostication in different clinical scenarios. Finally, we did not seek to make prognostic predictions for patients after clinicians have identified them as entering the last few hours or days of life. Since observations and investigations are often discontinued at the end of life, including these time periods would distort the model due to extreme missingness (in our data, no vital sign observations were recorded on 43% of days during end-of-life care, compared with 0% of days during active treatment). In addition, predicting end-of-life after it is clinically apparent would have little clinical utility.

Several predictors of disease severity included in our model have also been identified by point-of-admission severity models, and in epidemiological studies of risk factors for severe disease. Increasing age is widely recognised as being the strongest predictor of poor outcome from COVID-19. Frailty has similarly been shown to be a strong independent predictor of mortality in hospitalised older adults, including those with COVID-19.

Respiratory compromise is a common reason for hospital admission and markers of respiratory function, including RR, SpO2, oxygen requirement and SpO2/FiO2 ratio have been included in previous point-of-admission models. The SpO2/FiO2 ratio, as selected by our model, allows a fully quantitative rather than dichotomous measure of the need for additional oxygen, as well as allowing for the confounding effect of variation in the target oxygen saturations in different patient groups.

Acidosis frequently complicates respiratory, renal and advanced circulatory failure and has previously been noted as a marker of disease severity in COVID-19. The separate inclusion of the severity of acidosis and alkalosis in our set of candidate predictors allowed for pH changes in either direction to be accounted for and avoided, for example, a minor negative effect of alkalosis from masking a more major effect of acidosis.

Our model selected two markers of infection and inflammation: WCC and IL-6. This is consistent with other findings. IL-6 was included in the routine COVID-19 panel of blood tests at the study hospital but we recognise that this may be less commonly requested in other hospitals. The results when refitting the model with CRP in place of IL-6 produced a slightly weaker but potentially more broadly applicable model. The preference of the model for IL-6 over CRP may reflect the fact that IL-6 is responsible for the production of CRP and, as such, is an earlier and more dynamic marker of the inflammatory response.

There are several limitations to our study. We chose to include only laboratory results up to 48 hours and vital signs up to 24 hours before the landmark time; exploiting older data might improve the predictive ability of our model, at the expense of complexity and real-world utility. Our data were gathered from a single centre, and therefore, the generalisability of our findings to other centres and populations are pending external validation. Further, our model was generated from a relatively modest sample size due to the relatively low prevalence of COVID-19 patients in the catchment population of the hospital, particularly during the early months of the pandemic. One advantage of using this single dataset


BMJ Open: first published as 10.1136/bmjopen-2021-060026 on 5 September 2022. Downloaded from http://bmjopen.bmj.com/ on September 17, 2023 by guest. Protected by copyright.
from a large, tertiary hospital was that the hospital never became overwhelmed with patients, and therefore, it is considered that patients received care according to what was considered clinically appropriate rather than what resources permitted. Finally, while it is encouraging that the model continued to perform well in the wave 2 validation data, changes in clinical care of patients (notably use of steroids and IL-6 inhibitors) and the dominant virus strains (including the Delta and Omicron variants that emerged in the UK while this manuscript was under review) may influence the clinical picture of the disease, its severity and the risk factors for disease. The model will therefore likely need to be updated as the pandemic evolves, but the utilisation of routinely available data in this model makes this straightforward.

Author affiliations
1Faculty of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, UK
2MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
3Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
4Department of Medicine for the Elderly, Addenbrooke's Hospital, Cambridge, UK
5Department of Medicine, University of Cambridge, Cambridge, UK
6Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
7Department of Infectious Diseases, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

Acknowledgements We thank Christopher Osafor, Catriona Davidson, Alistair Mackett, Marie Coujon and Leelane Van Der Poel for assisting with the frailty substudy.

Contributors Conception: JP, RJBG, SLC, EG-K, CSW and MW conceived the research project and designed the study. Data collection: VT, MW, SLC and RJBG extracted and curated the dataset. VLK and DJH collected data. Analysis tools & Data interpretation: MW, RJBG and BDMT analysed and contributed to the selection and creation of analysis tools. JP, CW, EG-K, DJH, VLK and BDMT interpreted the data and results. Implementation: MW and RJBG performed the implementation and analysis of the proposed method. Draft writing: MW, CSW, EG-K, SLC and RJBG wrote the initial draft. All authors contributed to substantially revising the article for important intellectual content. All authors approved of the final version of this script to be published and are accountable for the work presented. MW is acting as the guarantor for the project.

Funding MW was funded by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). VLK was funded by the MRC/NICHR Clinical Academic Research Partnership Grant (CARP) (grant code MR/T023902/1). VT was funded by the Cancer Research UK Cambridge Centre. EG-K was supported by the NIHR Clinical Research Network (CRN) Greenshoots Award. BT and RJBG were funded by the UKRI Medical Research Council (MRC) (programme code MC_UU_00022/2) and supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014).

Disclaimer The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. The views expressed are those of the authors and not necessarily those of the NHS, the MRC, or the Department of Health and Social Care.

Competing interests None Declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by a UK Health Research Authority ethics committee (20/WM/0125). Patient consent was waived because the de-identified data presented here were collected during routine clinical practice; there was no requirement for informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The deidentified data that support the findings of this study are available from Cambridge University Hospitals but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available from the authors on reasonable request with permission of Cambridge University Hospitals.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

Author note JP and RJBG are last authors.

ORCID IDs
Martin Wiegand http://orcid.org/0000-0003-0276-658X
Sarah L Cowan http://orcid.org/0000-0001-9149-3530
Claire S Waddington http://orcid.org/0000-0002-4130-6027
Victoria L Kevill http://orcid.org/0000-0001-6148-0640
Brian D M Tom http://orcid.org/0000-0002-3335-9322
Effensoyli Gkrania-Klotas http://orcid.org/0000-0002-0930-8330
Jacobus Preller http://orcid.org/0000-0001-5706-816X
Robert J B Goudie http://orcid.org/0000-0001-9554-1499

REFERENCES
18 Crooks CJ, West J, Fogarty A. Predicting the need for escalation of care or death from repeated daily clinical observations and laboratory results in patients with SARS-CoV-2 during 2020: a retrospective population-based cohort study from the United Kingdom. medRxiv 2020;2024811.