**Appendix Table of Contents:**

**Appendix A: eMethods**

Study Workflow and Model Deployment

Component Models

Variable Definitions

Missing Data, Repeated Data, and Outliers

Time-Integrated Mortality Prediction

Statistical Analysis and Software

**Appendix B: eResults**

Confusion Matrix Assessments

Models Trained on COVID-19 Specific Data

**Appendix C: Additional Ethical Considerations in Score Development**

**Appendix D: References**

**eTable 1: Retrospective versus Prospective Cohort Demographic Characteristics and Hospital Course**

**eTable 2: Mortality Model Input Summary Measures, Retrospective versus Prospective Cohorts**

**eTable 3: Mortality Model Design and Inputs**

**eTable 4:** **Actual Mortality by Predicted Mortality Decile in the Prospective Cohort**

**eTable 5: Stacked Model Coefficients in the Retrospective, Prospective, and COVID-19 Cohorts**

**eTable 6: Mortality Prediction 3 and 7 Days After Admission**

**eTable 7: Model Performance Metrics**

**eTable 8: Calibration metrics for constituent and ensemble models**

**Appendix A: Additional Methods**

*Study Workflow and Model Deployment*

Rapid development and implementation of a new score in a real-time EHR requires a full clinical and informatics pipeline including skilled data warehousing, data wrangling, machine learning, health system information technology (IT), and clinical and ethics personnel working in sync. In this study, we deployed a data science (DS) team, a clinical data warehouse (CDW) team, and a clinical informatics operations (OPS) team (Figure 1A). Clinical expertise was present on all three teams and some individuals were present on multiple teams. The CDW and OPS teams worked to identify the clinical variables necessary. The CDW team constructed the retrospective and prospective data sets. The creation of the data sets and models required both the operational data and the CDW data to identify the correct data elements to ensure accuracy and robustness. The DS team constructed and validated the models. The DS and OPS teams worked to integrate the model within the operational EHR. All three teams were required for model validation within the operational system.

*Component Models*

The following variables were included in each component model:

1. SOFA: PaO2 to FiO2 ratio (PF), mechanical ventilation, platelets, Glasgow coma scale (GCS), bilirubin, mean arterial pressure (MAP), vasoactive medications, creatinine
2. qSOFA: GCS, systolic blood pressure (SBP), respiratory rate (RR)
3. CURB-65: age, systolic and diastolic blood pressure (SBP and DBP), blood urea nitrogen (BUN), Glasgow coma scale (GCS), and respiratory rate
4. ARDS Mortality Score: This was limited to a subset of variables available in structured form in the EHR. This includes arterial pH, bilirubin, PF ratio, SpO2, mechanical ventilation, transfusion of fresh frozen plasma (FFP), transfusion of red blood cells (PRBC), GCS, shock (defined as lactate >2.0) and acute kidney injury (defined as a creatinine >= 2.0).
5. Charlson Comorbidity Index (CCI): For retrospective data, this was calculated using ICD-9 and ICD-10 data available from previous encounters in the EHR. The open source R package ICD was used to transform these codes into a final Charlson Index.[1] In the live EHR, ICD-9 and ICD-10 codes are not available in real-time during an active encounter. Consequently, the health system developed and implemented a system to capture CCI data on all new admissions. Mining historical records for data relevant to the CCI calculation was possible, but bias against patients with a long history at our institution compared to patients who have never been treated at UCHealth.
6. Novel Variables: age (continuous), absolute lymphocyte count (ALC), alanine aminotransferase (ALT), brain natriuretic peptide (BNP), BUN, creatinine kinase (CK), D-dimer, lactate, lactate dehydrogenase (LDH), and troponin. Ferritin and CRP were not included as novel variables, despite data suggesting that they may predict outcomes in COVID-19 because too many patients were lacking data.

*Variable Definitions*

1. All vital sign, medication, and laboratory variables were directly extracted from the EHR. If not directly recorded in the EHR, MAP was calculated from SBP and DBP.
2. Mechanical Ventilation: Defined as the presence of an endotracheal tube or tracheostomy in the Line, Drain, Airway (LDA) table from the EHR
   1. Placement date-time set to admission date-time if missing or before admission
   2. Removal date-time set to discharge date-time if missing
   3. Overlapping ventilation episodes were collapsed into one episode, e.g., the temporal order of tube 1 placed > tube 2 placed > tube 1 removed > tube 2 removed was formatted into a single episode starting with the placement of tube 1 and concluding with the removal of tube 2.
   4. Distinct events with less than two hours lapsing between removal of initial tube and placement of second tube were collapsed into a single mechanical ventilation episode.
3. PaO2 to FiO2 Ratios (PF ratio): PF ratio was calculated in a stepwise fashion. First, if the PaO2 and FiO2 were both known, it was calculated directly. This occurred if a patient had a recent ABG and was on invasive mechanical ventilation, non-invasive positive pressure ventilation, or heated high flow nasal cannula – all modes where FiO2 is set directly. Second, if the patient was on nasal cannula or face mask, the FiO2 needed to be estimated based on flow rate (see below). Third, if the PaO2 was not directly measured, the PF ratio was estimated from the SpO2 to FiO2 ratio as described elsewhere.[2–5] For SpO2 and FiO2 outside provided ranges, the following logic was used: If FiO2 < 30 and SpO2 between 80 and 96, set PF ratio to '399' which results in a 1 for SOFA respiratory score. If FiO2 < 30 and SpO2 > 96, set PF ratio to '401' which results in a 0 for SOFA respiratory score. Finally, if the imputed SpO2 to FiO2 could not be imputed, the PaO2 was estimated from the FiO2 using the alveolar gas equation (assuming atmospheric pressure of 620, pCO2=40, R=0.8, and A-a gradient of 10.

*Missing Data, Repeated Data, and Outliers*

Vital signs, labs, and other covariates are collected throughout the encounter. The date and time of known values that a given covariate was measured was used to construct a time series of data. Because the collection of information is not synchronized in time, at any given time, many of the covariates are missing data. For example, a PaO2 taken at 13:01:00 and respiratory rate taken at 13:04:00 would result in the PaO2 value missing in the row for time 13:04:00 and respiratory rate missing in the row for time 13:01:00.

Four solutions are commonly utilized. The first is to impute the linear interpolation between known values. The second is to use the last known observation within a specific window of time (last observation carried forward, LOCF). The third is to use the next observation within a specified window of time (next observation carried backward, NOCB). The fourth is to use a 'normal’ default if no values were present during the entire hospitalization and it could not otherwise be estimated. This explicitly assumes that if a value was not checked, it was because the clinical team did not think it was going to be sufficiently different from normal to change management. We did the following:

*Indicator variables:* LOCF as appropriate for temporal associations, e.g., actively intubated, or set for the entire encounter record if no temporal considerations were needed.  
  
*Vital Signs* (SBP, DPB, MAP, RR, Spo2): NOCB was utilized until the time of the first measured, linear interpolation was utilized between serial measurements, and LOCF from the last measurement to the end of the encounter.

*Medications*: LOCF was used for all medication doses.

*Bilirubin, BUN, Platelets, and Lactate*: Linear interpolation was utilized between subsequent values.

*D*-*dimer, ALC, troponin, BNP, CK, arterial pH:* NOCB was utilized until the time of the first measured, linear interpolation was utilized between serial measurements, and LOCF from the last measurement to the end of the encounter.

*GCS*: GCS required special attention. It is commonly obtained in the emergency department, but in our health system is less likely to be used once the patient is intubated. Consequently, we used LOCF within a 24h window if it was recorded. If the patient was intubated, GCS was assumed to be 11. Otherwise, GCS was assumed to be 15.

*Normal values:*

1. PaO2 = 540 (mmHg)
2. FiO2 = 0.21
3. SpO2 = 95
4. Platelet = 250 (109 /L)
5. Bilirubin = 0.2 (mg / dL)
6. BUN = 14 mg/dL
7. Creatinine = 0.75
8. Dobutamine = 0 (mcg / kg / min)
9. Epinephrine = 0 (mcg / kg / min)
10. Norepinephrine = 0 (mcg / kg / min)
11. D-dimer = 250 (ng / mL)
12. LDH = 210 (U / L)
13. Absolute Lymphocyte Count = 1.4 (109 /L)
14. Troponin = 0.02 (ng / mL)
15. BNP = 100 (pg / mL)
16. Lactate = 0.75 (mmol / L)
17. CK = 110 (U / L)
18. ALT = 20 (U / L)
19. Arterial pH = 7.40

Sometimes multiple values for the same variable were available in the EHR at the same instance in time. In this case, different logic was applied to the variable in question to reduce values to one observation.

1. *PaO2, FiO2, SpO2:* For a given instant of time, keep largest value. If an FiO2 was present at the same time as a L/min flowrate, the FiO2 value was used.
2. *Absolute Lymphocyte Count, ALT, Arterial pH, Bilirubin, BNP, BUN, CK, Creatinine, D-dimer, Lactate, LDH, Platelet count, Troponin:* For a given instant of time, keep lab with first result time (most labs used the time the lab was obtained)
3. *Blood Pressure:* For timepoints with multiple SBP, DBP, and MAP values, keep value with highest MAP.
4. *Respiratory Rate:* Keep “Actual RR” over “Observed RR”; otherwise, keep highest value.
5. *GCS*: Prefer MD stated over other values; keep largest value if still duplicates
6. *COVID Status:* If multiple values present, prefer 'Positive' first, if 'Positive' not present, then prefer 'Negative'.
7. *Medications:* Dose (mainly drip rates) were tracked in 15-minute increments to avoid rapid fluctuations when changing bags.

Outliers were observed primarily in the vital sign data. For instance: a respiratory rate of 120 is likely 12 or 20; a MAP of 350 is likely a measurement being recorded while an arterial catheter is being flushed. To account for this, we set reasonable ranges on the values for vital signs. Measurements outside this range were clipped to the minimum or maximum value.

1. Respiratory Rate: Values below 0 or above 85 were discarded. Remaining values were clipped to the range of 4 to 60
2. Systolic Blood Pressure: Values below 11 or above 350 were discarded. Remaining values were clipped to the range of 30 to 300
3. Diastolic Blood Pressure: Values below 10 were discarded. Palpation values were set to 20. Remaining values were clipped to the range of 20 to 250
4. Mean Arterial Blood Pressure: MAP is normally machine reported. If missing or if MAP above SBP and below DBP, MAP was calculated according to the formula (SBP + 2 \* DBP) / 3. Remaining values were clipped to the range 20 to 250
5. SpO2: 0 to 100
6. FIO2: Acceptable rage 0.21 to 1.0. Additionally, for patients on nasal cannula or simple face mask: if O2 flow rate above 15 L/min, cap at 15 L/min. Where FiO2 was missing but L/min rate present, used the formula to convert. Invalid flow rates (negative values, above 100) were first discarded. If Nasal cannula FiO2 (either reported or calculated) above 60%, cap to 60%.
7. GCS: integer values between 3 and 15
8. Vasopressor medications: mcg/kg/min above 100 were dropped.

*Time-integrated mortality prediction*:

From a crisis triage standpoint, all critical care interventions are considered a time limited trial. Therefore, in addition to an initially estimate of mortality, a score to facilitate reassessment and ongoing utility of any intervention is necessary. Consequently, we also predicted mortality after three and seven days of hospitalization. This allowed us to incorporate change in the patient condition over time—the acuity trajectory—as a predictor of mortality, rather than simply utilizing physiologic data from a single point in time. To incorporate the acuity trajectory, we estimated the point-wise mortality score every time a covariate was updated during the patient encounter, creating a time series of predicted mortalities. We then calculated seven parameters and used them as covariates in a stacked ridge logistic regression predicting mortality. These included the initial and maximum SOFA score over the timespan of interest. Second, we estimated the integral of the stack-model-estimated point-wise mortality over time. Third, we estimated the linear trend of point-wise predicted mortality over time with linear regression and included both the slope and the standard error of the slope. Finally, we included both the first point-wise predicted mortality and the maximum point-wise predicted mortality over the time span of interest.

*Statistical Analysis and Software*

We performed univariable comparisons of the demographic and baseline characteristics of the retrospective and prospective cohorts and between patients with and without COVID-19 in the prospective cohort using chi-square tests for categorical data and unpaired t-tests for continuous variables. All analyses were performed using R version 3.6.[6]

**Appendix B: Additional Results**

*Confusion Matrix Assessments*

The intended use of the model is to rank patients for triage. As such, the most important metric is the relative ranking of cases (encounters ending in expiration of the patient) and non-cases (encounters concluding with the patient discharged alive). If we were to randomly select an encounter ending in death and an encounter ending in survival, then our model should provide a higher probability of mortality estimate for the former than the latter. The value of the estimated probability is less relevant for this intended use. If the encounter ending in mortality had an estimated probability of mortality of 0.08 and the encounter ending in survival had an estimated probability of mortality of 0.05, then our model is correctly ranking patients. This would also be true if the respective probability estimates were 0.99 and 0.90. The AUROC is a measure of the probability of correctly ranking a case and non-case.

A health system might choose to activate a clinical decision support alert at an estimated mortality probability. Assessing the model in correctly discriminating between encounters ending in mortality/survival is dependent on arbitrary thresholds to turn the continuous probability of mortality estimate to a binary survived/expired. Figure 3 and eTable 6 show common confusion matrix statistics as a function of the threshold.

*Models Trained on COVID-19 Specific Data*

To understand how patients with COVD-19 may differ from the population as whole, in sensitivity analyses we retrained the stacked model to predict point-wise mortality using patients in the prospective cohort and on patients with COVID-19 only. We divided the COVID-19 positive patient data 40%-40%-20% for training the logistic regression models, training the stacked model, and evaluating the stacked model, respectively. We estimated the stacked models with regularized (ridge) logistic regression and used 3-fold cross-validation to select a regularization parameter. The final stacked model was evaluated using bootstrap-estimated confidence intervals (CIs) and a primary metric of area under the receiver operator curve (AUROC). These cohorts are substantially smaller than the retrospective cohort, and care must be used when interpreting AUROC results.

However, we noted some qualitative differences in the weights given to different models between training groups. In general, SOFA and the ARDS mortality model received similar weights across all three cohorts, suggesting it remains a robust predictor of mortality. QSOFA was weighted substantially more, while CURB-65 and the Charlson Comorbidity Index were all weighted substantially less when trained with COVID-19 only cohorts. Interestingly, the novel variables were also weighted less in the COVID-19 only cohort (eTable 6). We hypothesize that these weights change between training cohorts because the novel variables and SOFA becomes more closely correlated in patients with COVID-19 than in patients without COVID-19.

In patients with COVID-19, models trained on COVID-19 patients had an AUROC for SOFA, CURB-65, the Charlson Comorbidity Index, and novel variables of 0.86, 0.90, 0.75, and 0.89 respectively. In this subset of patients, the stacked model predicted mortality with a AUROC of 0.95. The stacked model continued to predict mortality with narrowest 95% confidence intervals at the extremes of predicted mortality but given the smaller sample size at moderate predicted mortalities, 95% confidence intervals had a width of less of thirty percentage points (data not shown).

**Appendix C: Additional Ethical Considerations in Score Development**

First, a triage scoring system needs to identify and rank patients most likely to benefit from scarce resources regardless of whether they have one condition versus another. In other words, the method must rank both those patients who are going to die regardless of the resources give to them and those patients who will live regardless of the resources give to them. The idea is to accurately identify those patients who will survive a hospitalization if and only if the patient receives those scarce resources. An EWS system is not designed to accomplish this goal, but rather to identify all patients who may worsen and need advanced care. We felt this one-sided bias would be difficult to overcome and did not include scores from an EWS framework. Historically, the SOFA score has been utilized in a triage-type role, but its accuracy in viral pandemics such as influenza was moderate at best. Because our goal was to predict mortality risk, we chose to add additional scores that had been previously validated to predict mortality for conditions we expected to encounter in a primarily (although not exclusively) respiratory pandemic – namely pneumonia and ARDS. As an aside, the overall methodology was constructed such that any score could be included if it was desired or found to be ethically defensible.

Secondly, raw scores are biased by gender, race, socioeconomic status, and other factors, and rarely fairly select patients. This is because the scores were originally validated without addressing bias from such factors. Thus, these scores could not be used “out of the box”, but needed to be revalidated without such factors included.

Third, to ensure the ethical application of a triage process and to cope with a potentially rapidly changing situation, the score needed to be calculated frequently for all patients in the hospital, even though the majority of patients would be unlikely to utilize limited resources such as mechanical ventilation. This meant that scoring needed to use only factors routinely and accurately collected in an EHR (asking providers to provide additional inputs in a crisis scenario was felt likely to lead to inaccuracy or omissions). Because the score was to be used on all patients, the score was validated against all hospital admissions.

Finally, most U.S. States issued Crisis Standards of Care for the Triage of Scarce Resources. These protocols were frequently based on the SOFA score. Thus, we felt from a legal standpoint that the SOFA score should lie at the base of our score, only adding or changing the score when it could be shown, in real time, to produce better results. Hence, we added additional component scores to the SOFA in a stacked model, knowing that at a minimum, our score would perform no worse than the SOFA score, the standard proposed by many states.

**Ethical Considerations around 3-Day and 7-Day Sensitivity Analyses**

The 3-day and 7-day results (eTable 5) are from sensitivity analyses predicting 3-day and 7-day mortality instead of overall mortality. These sensitivity analyses were motivated by the following. First, medical triage needs to accurately identify those patients who will survive if and only if they receive scarce resources. It is well recognized that this is difficult to predict and that patients given a limited resource would need to be reevaluated after some time to evaluate this prediction. This reevaluation will depend on the situation in the hospital and the patient. In other words, because a patient is given a limited resource does not mean they will have it indefinitely. However, the ethics and practicality of medicine demand that the patient receive a fair trial of a given intervention before the intervention is continued or abandoned.

Given that our primary goal was the triage of ventilators, historical data from a national database (Vizient, not shown) suggested that overall average ventilator duration was 1-3 days. Our retrospective data suggested that median ventilator duration was 1 day. We chose to evaluate mortality at 3 and 7 days. As a multidisciplinary group, we decided that a 3-day trial of life support before revaluation was an adequate trial and sought to inform the triage team with an updated and hopefully more thorough prediction of mortality after this time as passed. When we developed the score in March 2020, we did not know if COVID-19 would have substantially longer ventilator times and thus also evaluated a 7-day window for that eventuality.

We agree that 30-day predicted mortality would have utility, even for triage purposes. One can argue that triage should attempt to identify the patient who will live if and only if given a scarce resource and also the patient who will have the most life-days saved. 30-day predicted mortality may address some of this concern. However, the data with which we can estimate the model, like all EHR and non-trial clinical data, lacks long-term post-discharge mortality, precluding this evaluation. Simultaneously, at the state level, the Crisis Standards of Care working groups realized the technical limitations of including life-days saved and dropped this from state requirements.

**Appendix D: References**

1 Wasey JO, R Core Team. *ICD: Comorbidity Calculations and Tools for ICD-9 and ICD-10 Codes*. 2019. https://jackwasey.github.io/icd/

2 Brown SM, Grissom CK, Moss M, *et al.* Nonlinear Imputation of PaO2/FiO2 From SpO2/FiO2 Among Patients With Acute Respiratory Distress Syndrome. *Chest* 2016;**150**:307–13. doi:10.1016/j.chest.2016.01.003

3 Chen W, Janz DR, Shaver CM, *et al.* Clinical Characteristics and Outcomes Are Similar in ARDS Diagnosed by Oxygen Saturation/FiO2 Ratio Compared With PaO2/FiO2 Ratio. *Chest* 2015;**148**:1477–83. doi:10.1378/chest.15-0169

4 Tripathi RS, Blum JM, Rosenberg AL, *et al.* Pulse oximetry saturation to fraction inspired oxygen ratio as a measure of hypoxia under general anesthesia and the influence of positive end-expiratory pressure. *Journal of Critical Care* 2010;**25**:542.e9-542.e13. doi:10.1016/j.jcrc.2010.04.009

5 Rice TW, Wheeler AP, Bernard GR, *et al.* Comparison of the SpO2/FIO2 ratio and the PaO2/FIO2 ratio in patients with acute lung injury or ARDS. *Chest* 2007;**132**:410–7. doi:10.1378/chest.07-0617

6 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2015. http://www.R-project.org/

7 Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958;**45**:562–5.

**eTable 1: Retrospective versus Prospective Cohort Demographic Characteristics and Hospital Course**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Retrospective Encounters  (N = 82,087) | Prospective Encounters  (N = 27,296) | P-value |
| Age (SD) | 58.1 (17.8) | 54.3 (20.4) | *P* < 0.001 |
| Female | 35,826 (43.6%) | 15,660 (57.4%) | *P* < 0.001 |
| Race |  |  | *P* < 0.001 |
| White or Caucasian | 58,915 (71.8%) | 20,430 (74.8%) |  |
| Black or African American | 8,651 (10.5%) | 1,964 (7.2%) |  |
| Other | 12,763 (15.5%) | 4,481 (16.4%) |  |
| Unknown | 1,758 (2.1%) | 421 (1.5%) |  |
| Ethnicity |  |  | *P* < 0.001 |
| Non-Hispanic | 69,772 (85.0%) | 22,496 (82.4%) |  |
| Hispanic | 10,489 (12.8%) | 4,398 (16.1%) |  |
| Unknown | 1,826 (2.2%) | 402 (1.5%) |  |
|  |  |  |  |
| Supplemental Oxygen | 66,013 (80.4%) | 16,052 (58.8%) | *P* < 0.001 |
| High Flow Nasal Cannula | 5,729 (7.0%) | 1,398 (5.1%) | *P* < 0.001 |
| Non-Invasive Ventilation | 5,925 (7.2%) | 1,482 (5.4%) | *P* < 0.001 |
|  |  |  |  |
| Median Hospital Days  (IQR) | 4.04 [2.2, 7.0] | 3.0 (2.0, 5.2) | *P* < 0.001 |
| Overall Mortality | 5,726 (7%) | 717 (2.6%) | *P* < 0.001 |
|  |  |  |  |
| All Mechanical Ventilation | 14,847 (18.1%) | 1,480 (5.4%) | *P* < 0.001 |
| Median Hospital Days  (IQR) | 6.3 (3.3, 9.3) | 8.4 (4.6, 15.1) | *P* < 0.001 |
| Median ICU Days (IQR) | 2.3 (1.1, 4.3) | 3.6 (1.6, 7.8) | *P* < 0.001 |
| Median Ventilator Days  (IQR) | 1.0 (0.5, 2.3) | 1.8 (0.7, 5.7) | *P* < 0.001 |
| Mortality | 4,055 (27.3%) | 408 (27.6%) | *P* = 0.86 |

**eTable 2: Mortality Model Input Summary Measures, Retrospective versus Prospective Cohorts**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Retrospective Encounters  (N = 82,087) | Prospective Encounter  (N = 27,296) | P-value |
| Scores |  |  |  |
| Median qSOFA (IQR) | 0.9 [0.0. 1.0] | 0.0 [0.0, 1.0] | *P* < 0.001 |
| Median SOFA (IQR) | 3.00 [2.0, 6.0] | 2.0 [2.0, 4.0] | *P* < 0.001 |
| Median CURB-65  (IQR) | 1.6 [1.0, 2.2] | 1.0 [0.1, 2.0] | *P* < 0.001 |
| Charlson  Comorbidity  Index (IQR) | 2.0 [1.0, 4.0] | 1.0 [0.0, 3.0] | *P* < 0.001 |
| ARDS Mortality Model |  |  |  |
| Transfusion FFP | 264 (0.3%) | 59 (0.2%) | *P* = 0.007 |
| Transfusion PRBC | 853 (1.0%) | 396 (1.5%) | *P* < 0.001 |
| GCS ≤ 8 | 2,676 (3.3%) | 264 (1.0%) | *P* < 0.001 |
| Lactate > 2 | 12,428 (15.1%) | 2,676 (9.8%) | *P* < 0.001 |
| Creatinine ≥ 2 | 11,860 (14.4%) | 2,486 (9.1%) | *P* < 0.001 |
| Mean Bilirubin (SD) | 0.8 ± 2.4 | 0.7 ± 2.0 | *P* < 0.001 |
| Mean Arterial pH  (SD) | 7.4 ± 0.1 | 7.4 ± 0.0 | *P* < 0.001 |
| Mean PF (SD) | 426.6 ± 575.9 | 335.7 ± 212.7 | *P* < 0.001 |
| Mean SpO2 (SD) | 94.7 ± 3.2 | 94.7 ± 2.4 | *P* = 0.74 |
| Novel Predictors |  |  |  |
| Mean D-Dimer (SD) | 391.9 ± 3,075.1 | 405.0 ± 3,699.8 | *P* = 0.60 |
| Mean LDH (SD) | 222.7 ± 221.6 | 229.1 ± 214.9 | *P* < 0.001 |
| Mean ALC (SD) | 1.4 ± 3.3 | 1.4 ± 2.0 | *P* = 0.81 |
| Mean BUN (SD) | 22.3 ± 18.1 | 19.4 ± 15.1 | *P* < 0.001 |
| Mean Troponin (SD) | 1.8 ± 15.6 | 0.5 ± 9.0 | *P* < 0.001 |
| Mean CK (SD) | 195.4 ± 2,240.9 | 173.7 ± 1,612.7 | *P* = 0.08 |
| Mean ALT (SD) | 23.2 ± 73.2 | 21.1 ± 20.6 | *P* < 0.001 |
| Mean Lactate (SD) | 1.3 ± 1.9 | 1.0 ± 1.1 | *P* < 0.001 |

The covariates included in the stacked model at the time of the maximum SOFA score for each encounter. FFP: fresh frozen plasm, PRBC: packed red blood cells, GCS: Glasgow comas score, PF: PaO2 to FiO2 ratio, LDH: lactate dehydrogenase, ALC: absolute lymphocyte count, BUN: blood urea nitrogen, CK: creatinine kinase, ALT: alanine aminotransferase

**eTable 3: Mortality Model Design and Inputs**

|  |  |  |
| --- | --- | --- |
| **Model** | **Formula** | **Notes** |
| qSOFA total | mortality ~ qsofa\_total | Uses the sum of the three elements |
| qSOFA system | mortality ~ qsofa\_GCS + qsofa\_RR + qsofa\_SBP | Uses each element as an explicit predictor |
| qSOFA total + age | mortality ~ qsofa\_total + age |  |
| qSOFA system + age | mortality ~ qsofa\_GCS + qsofa\_RR + qsofa\_SBP + age |  |
| CURB-65 total | mortality ~ curb65\_total |  |
| CURB-65 system | mortality ~ curb65\_GCS + curb65\_BUN + curb65\_RR + curb65\_BP + curb65\_age |  |
| SOFA total | mortality ~ sofa\_total |  |
| SOFA system | mortality ~ sofa\_resp + sofa\_hema + sofa\_hep + sofa\_cardio + sofa\_cns + sofa\_renal |  |
| Charlson Comorbidity Index | mortality ~ charlson\_score |  |
| ARDS Mortality | mortality ~ transfusion\_FFP + transfusion\_red\_blood\_cells + wash\_GCS + wash\_shock + wash\_renal + bilirubin + arterial\_ph + PFratio + SpO2 | Vasopressors omitted |
| Novel | mortality ~ ddimer +  LDH +  ALC +  BUN + CK + ALT + lactate + age + Charlson\_score |  |

This table shows the covariates included and the design for each component model in the stacked model. See Figure GCS: Glasgow Coma Scale, RR: Respiratory Rate, SBP: Systolic Blood Pressure, BUN: blood urea nitrogen, FFP: fresh frozen plasm, PRBC: packed red blood cells, PF: PaO2 to FiO2 ratio, LDH: lactate dehydrogenase, ALC: absolute lymphocyte count, CK: creatinine kinase, ALT: alanine aminotransferase.

**eTable 4: Actual Mortality by Predicted Mortality Decile in the Prospective Cohort**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Retrospective | | Prospective | | COVID-19 Positive | |
| Predicted Mortality | **Encounters** | **Expired** | **Encounters** | **Expired** | **Encounters** | Expired |
| <10% | 14,598 (88.9%) | 329 (2.3%) | 26,118 (95.7%) | 269 (1.0%) | 1,160 (85.4%) | 69 (5.9%) |
| 11 - 20% | 638 (3.9%) | 139 (21.8%) | 442 (1.6%) | 72 (16.3%) | 64 (4.7%) | 10 (15.6%) |
| 21 - 30% | 260 (1.6%) | 62 (23.8%) | 172 (0.6%) | 46 (26.7%) | 22 (1.6%) | 10 (45.5%) |
| 31 - 40% | 137 (0.8%) | 55 (40.1%) | 116 (0.4%) | 38 (32.8%) | 15 (1.1%) | 5 (33.3%) |
| 41 - 50% | 104 (0.6%) | 56 (53.8%) | 72 (0.3%) | 30 (41.7%) | 18 (1.3%) | 7 (38.9%) |
| 51 - 60% | 88 (0.5%) | 50 (56.8%) | 63 (0.2%) | 24 (38.1%) | 14 (1.0%) | 9 (64.3%) |
| 61 - 70% | 94 (0.6%) | 56 (59.6%) | 46 (0.2%) | 26 (56.5%) | 7 (0.5%) | 4 (57.1%) |
| 71 - 80% | 99 (0.6%) | 62 (62.6%) | 50 (0.2%) | 36 (72.0%) | 13 (1.0%) | 10 (76.9%) |
| 81 - 90% | 114 (0.7%) | 81 (71.1%) | 67 (0.2%) | 42 (62.7%) | 13 (1.0%) | 10 (76.9%) |
| 91 - 100% | 286 (1.7%) | 264 (92.3%) | 150 (0.5%) | 134 (89.3%) | 32 (2.4%) | 32 (100.0%) |

The percentages reported for encounters are column percentages, i.e., percentage of all encounters. The percentages for expired are row percentages, i.e., the percentage of the encounters noted in the row which concluded in the patient expiring.

**eTable 5: Mortality Prediction 3 and 7 Days After Admission**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Retrospective | | Prospective | | COVID-19 Positive | |
|  | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| First 3 days | 0.84 | 0.42 | 0.83 | 0.31 | 0.77 | 0.45 |
| First 7 days | 0.82 | 0.42 | 0.84 | 0.43 | 0.80 | 0.56 |

Table cells show AUROC and AUPRC in each cohort at two time points. The point-wise mortality scores were calculated at all time points with complete data. Using this time series, a logistic regression model using several features from the mortality time-series was trained and validated with the retrospective cohort (in a 40%-40%-20% split). The prospective and COVID-19 positive cohorts were used to validate the retrospectively trained logistic regression model.

**eTable 6: Stacked Model Coefficients in the Retrospective, Prospective, and COVID-19 Cohorts**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Retrospective Trained | Prospective Trained | COVID-19 Trained |
| SOFA | 1.22 | 1.76 | 0.86 |
| qSOFA | -0.10 | -2.45 | -1.41 |
| CURB-65 | 1.22 | 3.02 | 2.73 |
| ARDS Mortality | 2.63 | 2.96 | 2.19 |
| Charlson Comorbidity Index | 4.64 | 9.21 | 4.18 |
| Novel | 2.20 | 3.74 | 2.47 |

Coefficients of the models in log-odds

**eTable 7: Model Performance Metrics**

The purpose of the main stacked model was to create a ranked patient list by probability of mortality. If the model was to be used to as part of a clinical decision support alert, then a threshold for the estimated probability would need to be used to define when an alert fires.

For example, if the threshold was set at 0.40, then the reported models have the following performance statistics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Positive Predictive Value | Sensitivity | Specificity | Accuracy | F1 |
| Retrospective | 0.72 | 0.49 | 0.99 | 0.95 | 0.59 |
| Prospective | 0.65 | 0.41 | 0.99 | 0.98 | 0.50 |
| COVID-19 Positive | 0.79 | 0.43 | 0.98 | 0.91 | 0.55 |

**eTable 8: Calibration assessments of the component and stacked models.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Retrospective | | | Prospective | | | COVID-19 Positive | | |
|  | Brier | Intercept | Slope | Brier | Intercept | Slope | Brier | Intercept | Slope |
| SOFA | 0.04 | -0.06 (-0.16, 0.05); \*P\* = 0.3119 | 0.98 (0.94, 1.02); \*P\* = 0.3146 | 0.02 | -0.42 (-0.56, -0.29); \*P\* < 0.0001 | 0.99 (0.94, 1.03); \*P\* = 0.5819 | 0.08 | 0.01 (-0.27, 0.30); \*P\* = 0.9254 | 0.68 (0.59, 0.77); \*P\* < 0.0001 |
| qSOFA | 0.05 | -0.02 (-0.14, 0.10); \*P\* = 0.7735 | 0.99 (0.94, 1.04); \*P\* = 0.6637 | 0.02 | 0.04 (-0.13, 0.21); \*P\* = 0.6330 | 1.16 (1.10, 1.22); \*P\* < 0.0001 | 0.09 | 0.51 (0.15, 0.88); \*P\* = 0.0056 | 0.82 (0.70, 0.94); \*P\* = 0.0039 |
| CURB-65 | 0.06 | -0.01 (-0.14, 0.12); \*P\* = 0.8886 | 0.99 (0.94, 1.05); \*P\* = 0.7617 | 0.02 | -0.05 (-0.21, 0.11); \*P\* = 0.5583 | 1.21 (1.15, 1.28); \*P\* < 0.0001 | 0.08 | 1.61 (1.17, 2.04); \*P\* < 0.0001 | 1.27 (1.10, 1.44); \*P\* = 0.0018 |
| ARDS Mortality | 0.05 | -0.21 (-0.33, -0.09); \*P\* = 0.0004 | 0.91 (0.87, 0.95); \*P\* < 0.0001 | 0.02 | -0.14 (-0.30, 0.02); \*P\* = 0.0919 | 1.18 (1.12, 1.24); \*P\* < 0.0001 | 0.08 | 1.61 (1.09, 2.13); \*P\* < 0.0001 | 1.31 (1.12, 1.50); \*P\* = 0.0013 |
| Charlson Comorbidity Index | 0.06 | -0.32 (-0.61, -0.03); \*P\* = 0.0315 | 0.87 (0.76, 0.98); \*P\* = 0.0253 | 0.03 | -0.25 (-0.58, 0.09); \*P\* = 0.1443 | 1.25 (1.12, 1.38); \*P\* = 0.0001 | 0.11 | 3.29 (2.21, 4.36); \*P\* < 0.0001 | 1.87 (1.48, 2.26); \*P\* < 0.0001 |
| Novel | 0.05 | 0.01 (-0.13, 0.15); \*P\* = 0.9228 | 1.00 (0.95, 1.05); \*P\* = 0.9269 | 0.02 | -0.24 (-0.41, -0.06); \*P\* = 0.0083 | 1.19 (1.13, 1.26); \*P\* < 0.0001 | 0.09 | 2.59 (1.95, 3.22); \*P\* < 0.0001 | 1.62 (1.39, 1.85); \*P\* < 0.0001 |
| Ensemble | 0.04 | 0.04 (-0.09, 0.16); \*P\* = 0.5855 | 1.02 (0.97, 1.07); \*P\* = 0.3834 | 0.02 | -0.14 (-0.30, 0.03); \*P\* = 0.1059 | 1.12 (1.06, 1.17); \*P\* < 0.0001 | 0.07 | 0.32 (-0.04, 0.67); \*P\* = 0.0779 | 0.81 (0.70, 0.92); \*P\* = 0.0009 |

The Brier Score is the average squared distance between the estimated probability of mortality and the actual mortality status of the patients. The intercept and slope columns report the point estimate, 95% confidence intervals and p-values for Cox calibration regression.[7] The p-value for the intercept results from testing the null hypothesis of the intercept being 0; the p-value for the slope results from testing the null hypothesis of the slope being 1. Interpretations for the slope: slope > 1 suggests estimated mortality probabilities do not vary enough; a positive slope less than 1 suggest the estimated mortality probabilities vary too much.