CovRNN—A recurrent neural network model for predicting outcomes of COVID-19 patients: model development and validation using EHR data

Laila Rasmy, PhD[1], Masayuki Nigo, MD[2], Bijun Sai Kannadath, MBBS[3], Ziqian Xie, PhD[1], Bingyu Mao, MA[1], Khush Patel, MD[1], Yujia Zhou, MSc[1], Wanheng Zhang, M.S[4], Angela Ross, DNP[1], Hua Xu, PhD[1], Degui Zhi, PhD[1]

[1] School of Biomedical Informatics, University of Texas Health Science Center at Houston, U.S.
[2] McGovern Medical School, University of Texas Health Science Center at Houston, U.S.
[3] College of Medicine, University of Arizona - Phoenix, U.S.
[4] School of Public Health, University of Texas Health Science Center at Houston, U.S.

Corresponding Author:

Degui Zhi, Ph.D

UTHealth School of Biomedical Informatics
7000 Fannin Street
Houston, TX 77030
Degui.Zhi@uth.tmc.edu
(713) 500-3629

Keywords: COVID-19; predictive modeling; electronic health records; structured data; deep learning; recurrent neural network; health outcomes; mortality; ventilation; length of stay

Word Count: 4,408.

SUMMARY

Background

Predicting outcomes of COVID-19 patients at an early stage is critical for optimized clinical care and resource management, especially during a pandemic. Although multiple machine learning models have been proposed to address this issue, based on the need for extensive data pre-processing and feature engineering, these models have not been validated or implemented outside of the original study site.

Methods

In this study, we propose CovRNN, recurrent neural network (RNN)-based models to predict COVID-19 patients' outcomes, using their available electronic health record (EHR) data on admission, without the need for specific feature selection or missing data imputation. CovRNN is designed to predict three outcomes: in-hospital mortality, need for mechanical ventilation, and long length of stay (LOS >7 days). Predictions are made for time-to-event risk scores (survival prediction) and all-time risk scores (binary prediction). Our models were trained and validated using heterogeneous and de-identified data of 247,960 COVID-19 patients from 87 healthcare systems, derived from the Cerner® Real-World Dataset (CRWD). External validation was performed using three test sets (approximately 53,000 patients). Further, the transferability of CovRNN was validated using 36,140 de-identified patients' data derived from the Optum® de-identified COVID-19 Electronic Health Record v. 1015 dataset (2007–2020).

Findings

CovRNN shows higher performance than do traditional models. It achieved an area under the receiving operating characteristic (AUROC) of 93% for mortality and mechanical ventilation predictions on the CRWD test set (vs. 91·5% and 90% for light gradient boost machine (LGBM) and logistic regression (LR), respectively) and 86.5% for prediction of LOS > 7 days (vs. 81·7% and 80% for LGBM and LR, respectively). For survival prediction, CovRNN achieved a C-index of 86% for mortality and 92·6% for mechanical ventilation. External validation confirmed AUROCs in similar ranges.

Interpretation

Trained on a large heterogeneous real-world dataset, our CovRNN model showed high prediction accuracy, good calibration, and transferability through consistently good performance on multiple external datasets. Our results demonstrate the feasibility of a COVID-19 predictive model that delivers high accuracy without the need for complex feature engineering.

Funding

CPRIT Grants No. RP160015, RP170668, and RP210042.

Research in context

Evidence before this study

Although many methods for predicting COVID-19 outcomes have been developed, they have not been extensively externally validated due to their limited transferability. A key obstacle to the transferability of such methods is the need for laborious data preprocessing and feature engineering. An earlier systematic review article that critically assessed prediction models for diagnosis and prognosis of COVID-19 infection revealed that the majority of the 107 prognostic models published before July 2020 suffer a high risk of overfitting bias. Using the prediction model risk of bias assessment tool (PROBAST), the authors identified the common reasons for biased results, including model training on a small, locally sourced dataset, which leads to a high risk of model overfitting, as well as the lack of model calibration or external validation. To provide an updated survey of the literature, we conducted a Scopus and PubMed search for COVID-19 outcomes prediction articles published between July 2020 and the end of December 2021, using the keywords "COVID electronic health record ('mortality' or 'ventilator' or 'length of stay' or 'real-time') prediction." The literature search retrieved a total of 466 unique articles, and upon review, we found 53 studies that describe the development and validation of machine learning predictive models for COVID-19 patients' prognosis after admission. Out of the 53 studies, only four involved training and evaluating the models on a multi-sourced cohort of more than 20,000 COVID-19 patients. The proposed models in these studies, however, still require extensive data preprocessing and feature engineering, which limit the transferability, reliability, and sustainability of such models.

Added value of this study

We offer a "minimal preprocessing" machine learning model training framework that can flexibly adapt to the changing pandemic. Our framework is designed to consume electronic health record (EHR) data mapped to standard terminologies in common use without the need for specific feature selection or missing value imputation, for convenience and practicality. Our trained COVID-19 outcome prediction models achieve high prediction accuracy, comparable to the state-of-the-art in the literature, as well as good calibration and lower risk of bias, as they were trained and evaluated on large heterogeneous datasets collected from different health systems. In addition, we show that our models can be fine-tuned on new data for continuous improvement, as recommended by the FDA's good machine learning practice. Further, our framework includes a utility for model predictions explanation to facilitate clinical judgment of the model predictions.

Implications of the available evidence

We show that deep learning-based models can achieve state-of-the-art prediction accuracy while consuming the structured EHR categorical data in their standard format without the need for features selection or missing value imputations, which implies that the trained models can be easily validated on new data sources. Our trained models are validated across datasets from different sources, indicating the transferability of our models. Our model development framework can be further applied to train and evaluate predictive models for different types of clinical events. For clinicians who are fighting COVID-19 on the front lines, there are two potential actionable contributions of our work. Clinicians can (i) fine-tune our pre-trained model

on their local data, regardless of the size, establish utility, and then deploy; and (ii) use our comprehensive model development framework to train a predictive model, using their own data.

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which emerged in December 2019.[1] By the end of 2021, there were more than 295 million confirmed COVID-19 infections worldwide and more than 825,000 deaths in the United States alone.[2] Further, there have been around 3·7 million hospital admissions recorded since August 2020.[2] During the peaks of the pandemic waves, many states have reported near-capacity hospital and intensive care unit (ICU) utilization. Accurate prediction of the future clinical trajectories of COVID-19 patients at the time of admission is crucial for clinical decision making and enables efficient allocation of resources. Indeed, a number of models for the prediction of COVID-19 outcomes have been developed. Wynants et al.[3] reviewed 107 prognostic models as of July 1, 2020. The most common issue highlighted in this study is the high risk of bias associated with the reviewed models, due to either the small, locally sourced training dataset and the high risk of model overfitting or the lack of model calibration or external validation.[4,5] Through an updated survey of the literature, as of December 2021, we found that only four studies[6–9] involved training the proposed models on more than 20,000 patients. Moreover, all four models are based on a small set of specific features and need a laborious data preprocessing and feature engineering process that limits the transferability, reliability, and sustainability of the models.

In this study, we aim to develop an accurate and transferrable model for COVID-19 patients' outcomes on admission that include in-hospital mortality (iMort), need for mechanical ventilation during the stay (mVent), and hospital stay longer than one week (pLOS). Our model, CovRNN, utilizes a gated RNN architecture, proven to be effective in modeling patients' electronic health records (EHR) data.[10–14] To maximize transferability, CovRNN uses structured EHR data mapped to standard terminologies in common use without the need for specific feature selection or missing value imputation. For iMort and mVent prediction tasks, CovRNN predicts a time-to-event risk score that can be interpreted as a binary prediction with a time horizon (survival prediction) and an all-time risk score (binary prediction).

CovRNN was trained on a cloud-based, large heterogeneous dataset of 243,785 de-identified patients' data derived from 85 health systems available through the Cerner® Real-World COVID-19 Q3 Dataset (CRWD), hosted on the Cerner HealtheDataLab™. We evaluated our models on four test sets extracted from the CRWD and the Optum® de-identified COVID-19 Electronic Health Record v.1015 dataset (2007–2020). Each test set is different in size and represents a different use case so that we can evaluate the cross-hospital generalizability and model transferability between different EHR data sources. We also reported the results of subgroup analysis and ablation studies for a better understanding of the model's performance. In addition, we utilized the integrated gradient technique[15] to expose the factors of the CovRNN predictions that can enhance the interpretability.

To the best of our knowledge, CovRNN is the first COVID-19 outcome prediction model that can simultaneously achieve the following: (i) accurately predict different COVID-19 patients' outcomes on admission, and (ii) use readily available structured EHR in its categorical format without the need for specific feature selection or missing value imputation. In addition, the prospective compliance of CovRNN is evaluated against quality standards, including the transparent reporting of individual prognosis or diagnosis (TRIPOD) and the prediction model risk of bias assessment tool (PROBAST). We also showed the value added of the fine-tuning utility of CovRNN and how it can be used to improve the model's prediction accuracy. Such

utility can be further used for the continuous improvement of the model as per good machine learning practice (GMLP) recommendations to secure the model's reliability and sustainability. The source code of our model is publicly available at https://github.com/ZhiGroup/CovRNN to enable its applications and further evaluation by other researchers.

METHODS

Datasets and Cohort Description

We extracted our main training cohort from the CRWD hosted on the Cerner HealtheDataLab™, a cloud-based de-identified patients' dataset that included the clinical data for COVID-19 patients from 87 health systems as of the end of September 2020. The CRWD included only patients who had a minimum of one emergency or inpatient encounter with a diagnosis code that could be associated with COVID-19 exposure or infection, or a positive result for a COVID-19 laboratory test. CRWD included the patient's medical history for up to five years before their first COVID-19 infection. Further description of the CRWD is available in Supplementary Material A. In our study, we predefined our prediction point as the first COVID-19 hospitalization admission day to an emergency, observation, or in-patient unit, and we refer to this point as the *index date* (Figure 1). We thereby excluded all patients who had no recorded clinical information on or before the index date as well as patients who stayed for less than one day, as described in Supplementary Material B. We also excluded patients who had inconsistent dates, such as discharge dates before the hospitalization start date, as well as patients who were readmitted later and presented different outcomes. Our cohort included 247,960 patients, from which we held out two hospitals' data for external validation. The remaining 243,785 patients' data were split into training, validation, and test sets, with the ratio of 7:1:2. All of our reported CRWD results were on the held-out test set of 48,781 patients from the 85 health systems. For external validation, we evaluated the model on two randomly selected held-out hospitals from the CRWD, Hospital 1 from the south region, with 3,469 patients, and Hospital 2 from the west region, with 706 patients.

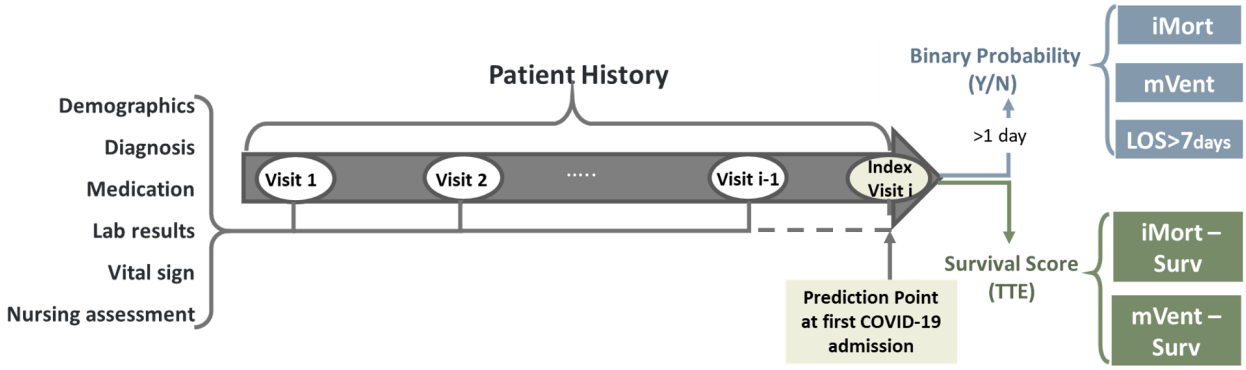

Figure 1. CovRNN prediction tasks.

For further external validation outside of the CRWD, we extracted a cohort of 36,140 de-identified patients' data derived from the Optum® de-identified COVID-19 Electronic Health Record dataset v.1015 dataset (2007-2020), which we refer to as "OPTUM" cohort. Further description of the Optum® dataset, along with differences and commonalities between the CRWD and OPTUM cohorts, are available in Supplementary Material A.

Data Preparation

We kept our data curation to the minimum level, as we describe below, to facilitate the transferability of our trained models among different datasets. We extracted all patient information on or before the date of their first hospital admission with COVID-19, including demographics, diagnosis, medication, procedures, laboratory results, and observations. To facilitate interoperability, we utilized standard terminologies in common use, such as ICD-9, ICD-10, and SNOMED-CT for diagnosis; LOINC and SNOMED-CT for laboratory results and observations; Multum drug identifiers and categories for medications; and CPT-4, HCPCS, ICD-9 PCs, and ICD-10 PCs for procedures. Such standard terminologies are readily available in the majority of EHR systems. In cases for which a coding system, such as Multum codes for medication, is not used, pre-existing mapping tools are available[16] that can be used to convert NDC medication codes to corresponding Multum information.

The majority of our features, such as diagnosis, medications, and procedures, were categorical. We converted numeric variables, such as laboratory results, to categorical variables as follows. For the CRWD, we used the "below normal low, normal, or above normal high" interpretation value provided in the CRWD rather than the actual numerical value; for OPTUM, we defined the result categories based on the corresponding normal result ranges. By doing so, we can further convert our input to either multi-hot or embedding matrices to feed to our models. Based on our previous study,[17] we decided to use the clinical information in the coding standards with which it was recorded, as the normalization of those codes to a more unified terminology provides minimal gain.[17] Further details of our data curation are available in Supplementary Material B. Our data curation pipeline is available at https://github.com/ZhiGroup/CovRNN.

Outcomes of Prediction Tasks

Our tasks include the prediction of COVID-19 patients' in-hospital mortality (iMort), need for mechanical ventilation (mVent), and prolonged length of stay (pLOS), on admission. For iMort event definition, we relied on the pre-assigned mortality flags on CRWD along with the "expired" encounter discharge disposition to confirm in-hospital mortality and identify the date of death. The iMort event definition was slightly different on the OPTUM data because there was no clear discharge disposition that indicated patient in-hospital death. We instead used the date of death and compared it against the hospitalization discharge date to assign the proper label. For mVent, we used mainly relevant mechanical ventilation procedure codes to define the outcome. In addition, with the CRWD, we used other relevant observations and recorded ventilator settings, not only to identify the instance of the event but also to identify the earliest time of the event (Supplementary Table 1). For iMort and mVent prediction tasks, we trained survival and binary classification-based prediction models. For survival analysis, we defined the time to event as the number of days between our index date and the earliest date that indicated the occurrence of the event, either a laboratory result or a recorded procedure for mVent or the discharge date of iMort. We used the hospitalization discharge date as the censoring time. We defined pLOS as a binary indicator for hospitalizations longer than 7 days, as the median length of stay (LOS) in the CRWD and OPTUM cohorts were 3 and 5, respectively, and we trained only a binary classification model for the pLOS task.

We used the area under the receiver operating characteristic (AUROC) as the main model discriminative performance metrics for the binary prediction models. We also reported other

clinically relevant metrics, including specificity at 95% sensitivity, the area under the precision-recall curve (AUPRC), and the sensitivity and specificity at optimum threshold, defined using the validation set. For survival analysis, we reported the concordance index[18] (c-index) as our main evaluation metric. In addition, we used the predicted score and calculated the AUROC at different periods.

Models

CovRNN models were based on a gated type of RNN, namely a gated recurrent unit (GRU), which is known for being an efficient sequential deep learning architecture for clinical event predictions.[12,19] Our models were designed to consume all demographics, diagnoses, medications, procedures, laboratory results, and other clinical event information readily available in the EHR before or on the index date to predict patient outcomes, without the need for specific feature selection or missing value imputation, for convenience and practicality. CovRNN also consumes the time difference between visits for a better temporal representation of patient history, which is known to slightly improve the prediction accuracy.[20,21] For binary classification tasks, we compared CovRNN against traditional machine learning algorithms, such as logistic regression (LR)[22] and light gradient boost machine (LGBM).[23] For survival prediction, we utilized the DeepSurv[24] architecture, while replacing the multiple layer perceptron (MLP) with GRU for better sequential information modeling. We were unable to adequately compare against machine learning survival models, such as random survival forest (RSF) or running proper factor analysis, for computational resource restrictions on the Cerner HealtheDataLab™, especially with the increased number of covariates and large training set size. Any version of the RSF model runnable on Cerner HealtheDataLab™ had a very small number of iterations/trees that led to poor and unreliable results; therefore, we decided not to report these results. Further implementation details are available in Supplementary Material C.

Experiments

For model development, we trained our models on 70% of the CRWD 85 hospitals' data (training set) and used 10% (validation set) to determine the best model trained, while controlling for overfitting. We reported the model performance on the remaining 20% held out for external validation (multi-hospital test set).

For further external validation, we used two additional levels of test sets (Figure 2). First, two hospitals' datasets were fully held out and used only to evaluate the cross-hospital generalizability: Hospital 1 (n = 3,469), from the south region, and Hospital 2 (n = 706), from the west region. Second, to evaluate the transferability of the CovRNN models across different EHR data sources, we used the OPTUM data set. Although the models can be directly used and evaluated on the OPTUM cohort, it is recommended to fine-tune the transferred model on a sample data of the destination, for two reasons: (a) Some clinical code distribution may vary or be newly presented at the destination data source; thus, during the model's fine-tuning, these codes would get introduced to the model and become embedded closer to codes of similar meaning; and (b) The definition of the outcome variables can be slightly different, given the limitations of each data source; for example, the mVent outcome was defined mainly in the OPTUM cohort, using only the procedure codes, whereas the CRWD used procedure codes and additional clinical event results. Therefore, to evaluate the value added of the models' fine-tuning, we transferred the best models trained on the CRWD and evaluated it on the

OPTUM cohort before and after fine-tuning. We also compared the performance of the fine-tuned models against new models that were trained only on the same OPTUM data used for fine-tuning.

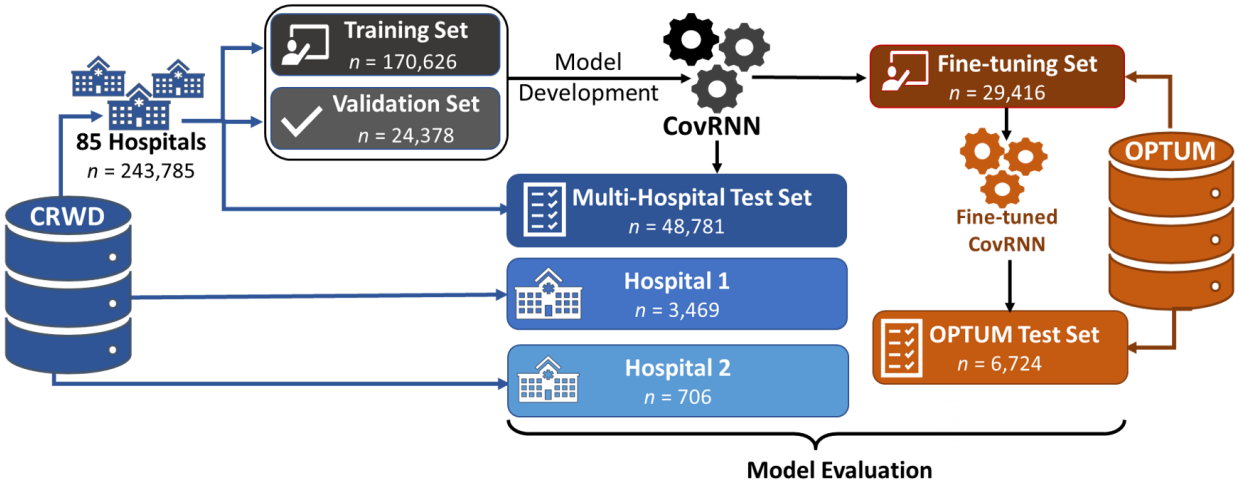

Figure 2. Model development and external validation datasets.

We conducted subgroup analysis for patients in regard to different age groups, races, baseline comorbidities, and geographical regions using the CRWD multi-hospital test set. Additional experimental designs including the evaluation of the added value for each clinical data category, the models' performance with single visit information using either the information provided on the index visit admission or the last visit with clinical information, and the impact of excluding intubated patients within the first 24 hours, referenced as "restricted" set, on mVent prediction, are presented in supplementary material D.

Models Interpretation

For CovRNN prediction interpretation, we used the integrated gradient technique[15] to expose the factors that contribute to the personalized model predictions. We used the integrated gradients technique due to its good theoretical properties, such as implementation invariance and completeness, and its implementation simplicity; as compared with methods such as layer-wise relevance propagation (LRP) or DeepLIFT, it does not require modification of the gradient backpropagation process and can be viewed as a deterministic and computationally efficient approximation of the gradient Shapley Additive Explanations (SHAP). This is unlike LR- and LGBM-based models, in which the existing interpretation utilities provide fixed feature-level importance, by using either the LR coefficients or the LGBM feature importance scores. For RNN-based models, we can achieve a more personalized explanation that shows the contribution scores for each code at each patient visit. For the preliminary evaluation, we extracted 20 random sample patients and presented their predicted risk scores as well as the contribution score assigned for each medical event and asked infectious disease specialists to evaluate its relevance.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

The description of the overall CRWD and OPTUM cohorts, presented in Table 1, shows that the prevalence of the outcome variables varies across the different data sources. The OPTUM cohort shows about twice the prevalence of in-hospital mortality and evidence of mechanical ventilation. In addition, the median length of stay is longer in the OPTUM cohort. Supplementary Table 2 includes further details of each subset.

Table 1. Descriptive statistics for CRWD and OPTUM extracted cohorts

Characteristics	CRWD *n* = 247,960	OPTUM *n* = 36,140
Age at the index visit		
Median (IQR)	57 (36–72)	60 (44–72)
Gender		
Female	130,540 (52·6%)	18,237 (50·5%)
Male	116,653 (47·0%)	17,885 (49·5%)
Race & Ethnicity		
Caucasian	168,606 (68·0%)	19,704 (54·5)
African American	36,762 (14·8%)	7,836 (21·7%)
Asian	5,494 (2·2%)	930 (2·6%)
American Indian / Alaska Native	4,285 (1·7%)	NA
Hispanic	72,068 (29·1%)	5,782 (16·0%)
Comorbidities		
Hypertension (HTN)	114,387 (47·7%)	22,035 (61·0%)
Diabetes (DM)	64,023 (26·7%)	12,942 (35·8%)
Congestive Heart Failure (CHF)	36,040 (15·0%)	6,568 (18·2%)
Chronic Kidney Disease (CKD)	34,789 (14·5%)	7,517 (20·8%)
Cancer	19,145 (8·0%)	5,094 (14·1%)
Outcomes		
Mortality	13,607 (5·5%)	4,831 (13·4%)
Median TTE	8 (4–16)	5 (3– 10)
Mechanical ventilation	33,505 (13·5%)	9,582 (26·5%)
Intubated on first day	17,811 (7·2%)	4,466 (12·4%)
Median TTE	2 (1–5)	3 (2 –7)
Length of stay		
Median (IQR)	3 (1– 6)	5 (3–10)
Total number of unique features	123,642	67,128
Number of Healthcare Systems	87	197

IQR: interquartile range, TTE: time to event in days

CovRNN binary classification models achieved an AUROC of 93% for iMort and mVent tasks on the CRWD test set (vs. 91·5% and 90% for LGBM and LR, respectively) and 86·5% for

pLOS (vs. 81·7% and 80% for LGBM and LR, respectively) (Table 2). External validation on held-out hospitals data showed AUROC ranges of 91·5%–97% for iMort and mVent binary predictions. pLOS prediction task showed AUROC 87·2% and 88·3% for Hospital 1 and Hospital 2, respectively (Table 2). External validation on the OPTUM cohort showed an AUROC after fine-tuning of 91·3%, 91·5%, and 81·0% for iMort, mVent, and pLOS, respectively (Table 3). Additional metrics, including specificity at 95% sensitivity, AUPRC, and sensitivity and specificity at the optimum threshold, are presented in Supplementary Table 2.

Table 2. Model performance on different CRWD test sets.

		In-hospital Mortality (iMort)				Mechanical Ventilation (mVent)				Stay > 7 days (pLOS)		
	n	LR	LGBM	Cov-RNN	CovRNN–SURV*	LR	LGBM	Cov-RNN	CovRNN–SURV*	LR	LGBM	Cov-RNN
Multi-hospital Test Set	48,781	90·3 (89·8–90·8)	91·5 (91·1–92·0)	93·0 (92·6–93·4)	86·0 (85·1–86·9)	89·5 (89·1–89·9)	91·2 (90·8–91·5)	92·9 (92·6–93·2)	92.6 (92·2–93·0)	80·0 (79·5–80·4)	81·7 (81·3–82·2)	86·5 (86·2–86·9)
Hospital 1	3,469	88·8 (86·9–90·5)	91·0 (89·5–92·4)	91·8 (90·3–93·2)	86·0 (83·2–88·5)	86·7 (85·1–88·4)	88·4 (87·0–89·9)	91·5 (90·2–92·8)	90.8 (89·4–92·2)	77·3 (75·5–79·1)	78·5 (76·7–80·2)	87·2 (85·8–88·4)
Hospital 2	706	94·6 (91·9–96·9)	95·1 (92·7–97·2)	97·0 (95·2–98·6)	91·6 (87·5–94·8)	93·5 (90·7–95·8)	95·6 (93·8–97·1)	96·0 (94·2–97·7)	93.8 (91·4–96·0)	80·9 (76·9–84·7)	84·3 (80·5–87·7)	88·3 (85·6–90·9)

** CovRNN–SURV uses time-to-event outcomes and the c-index are shown, other models use binary outcomes and AUROC are shown.*

Table 3. CovRNN models trained on CRWD performance on OPTUM before and after fine-tuning

Outcome	Newly trained model on OPTUM training cohort	Directly using CRWD transferred model	After Fine-tuning
In-hospital Mortality	88·6	87·0	91·3
Mechanical Ventilation	90·4	72·5	91·5
Stay > 7 days	78·1	68·0	81·0
In-hospital Mortality – Survival*	86·1	77·1	88·9
Mechanical Ventilation	90·2	69·2	93·7

– **Survival***

** values corresponding to survival models are in C-index, other models mainly using binary classifications are showing AUROC*

CovRNN survival models' evaluation on the CRWD test set achieved a c-index of 86·0 for iMort and 92·6 for mVent (Table 2). Using the survival models to predict patient risk to develop the event at a certain time point within the period between Day 1 and Day 60 showed an AUROC range from 93·6% to 88·8% for iMort and from 95·5% to 91·4% for mVent. Similarly, the survival models showed a c-index range from 86·5 to 93·8 for iMort and mVent tasks on the held-out hospitals data (Table 2, Figure 3a) and the OPTUM test set after fine-tuning (Table 3, Figure 3b).

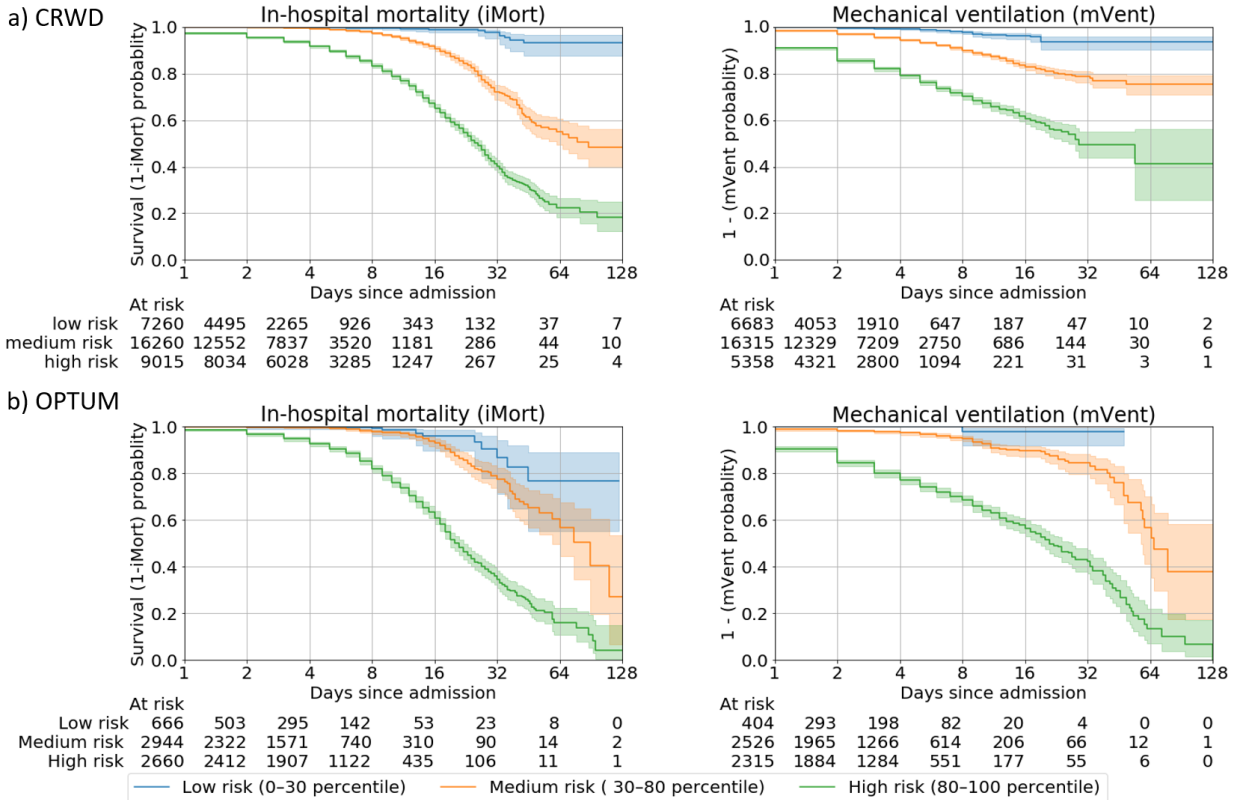

Figure 3. Kaplan-Meier (KM) curves that show the stratified survival analysis. Stratification of patients is according to their predicted survival score over time in days since admission. Shaded areas indicate 95% CIs calculated on the logarithmic scale from the standard errors of the Kaplan-Meier estimator with the center values as corresponding to the Kaplan–Meier estimate

For CovRNN binary classification and survival models, the transferred models consistently achieved better performance after fine-tuning compared to training new models on the OPTUM data. Our transferred binary classification models after fine-tuning showed AUROCs of 91·3%, 91·2%, and 81% for iMort, mVent, and pLOS prediction tasks, respectively, compared to 88·7%, 90·4%, and 78·1% for newly trained models (Table 3). Similarly, transferred survival models showed a c-index range from 88·9% to 93·7% after fine-tuning for iMort and mVent versus 86·1% to 90·2% for newly trained models (Table 3).

We conducted three experiments. The first experiment is an ablation study that showed that each clinical data category contributes to an increase in the model prediction accuracy (Supplementary Table 4). The second experiment results showed that the use of the full patient history continuously had a better performance than using the last (index) visit information only (Supplementary Table 5). In the third experiment, we found that CovRNN's performance remained unchanged on the "restricted" test set, regardless of whether we kept or excluded the patients who got intubated within the first 24 hours from the training cohort (Supplementary Table 6).

The subgroup analysis showed that prediction accuracy remains mainly consistent among different comorbidities, age groups, races, and regions. The most notable trend is that the prediction accuracy is better among the younger population (Figure 4). In addition, CovRNN binary classification models showed good calibration without sacrificing high prediction accuracy, as shown in the calibration plots (Figure 5). In Supplementary Figure 4, we present a sample visualization that shows the integrated gradient-based explanation of the CovRNN model's true positive prediction for a pLOS case. The information in the figure, however, is based on a sample and not the full patient data; the full data for a subset of more than 20 patients for each prediction task were presented to an infectious disease specialist, who found it informative. Further evaluation of the model explanation is warranted, taking into consideration that the evaluation of such personalized explanations, whereby the same clinical code can have different contribution scores at each patient and visit level, given the different context, is laborious. To demonstrate our efforts to abide by transparent reporting standards, we provide the TRIPOD and PROBAST assessments as Supplementary Material D and Supplementary Material E, respectively.

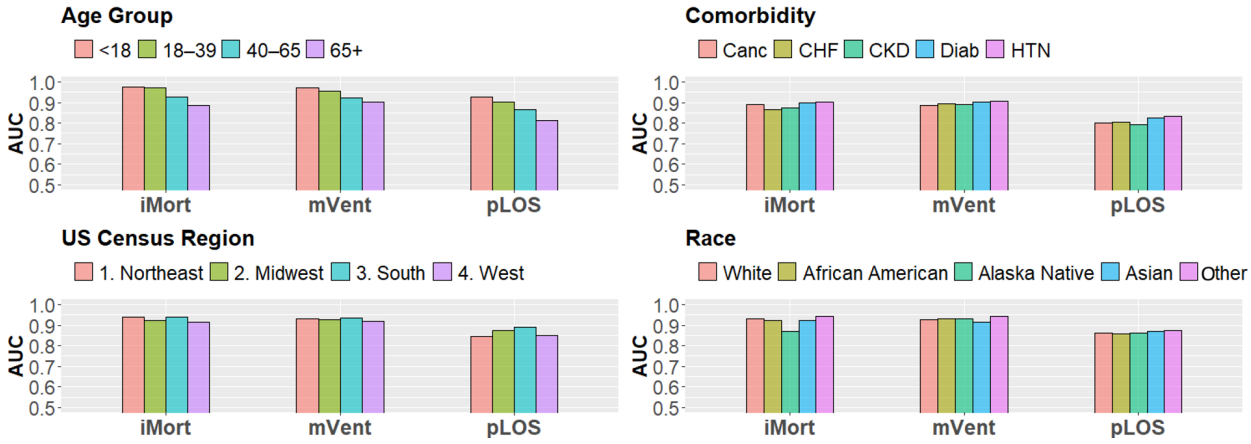

Figure 4. Subgroup analysis using the CRWD multi-hospital test set

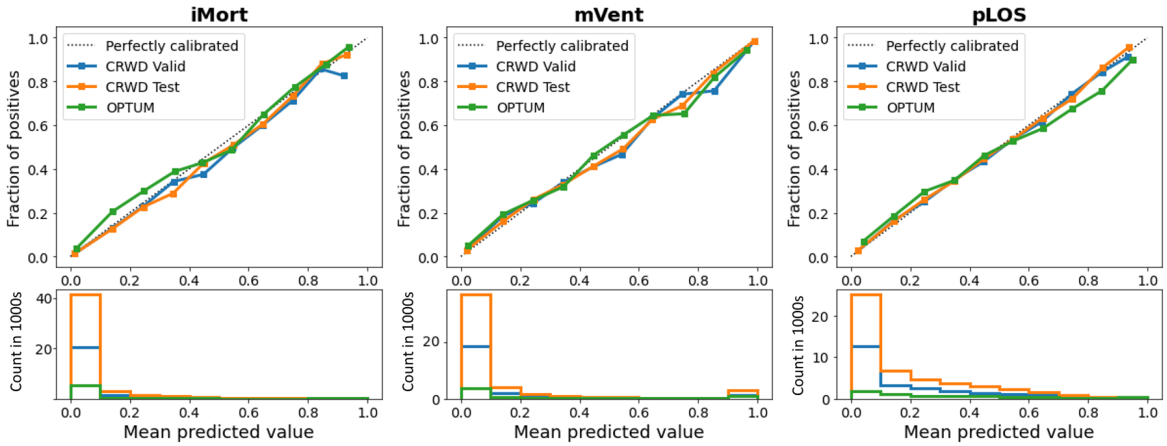

Figure 5. Calibration plots for the CRWD validation set, CRWD multi-hospital test set, and OPTUM test set.

DISCUSSION

Our experiments showed that CovRNN models trained on a large heterogeneous dataset of approximately 200,000 COVID-19 patients required minimum data curation to achieve high prediction accuracy (AUROC: 97–86%) for different patient clinical outcomes, namely iMort, mVent, and pLOS. CovRNN not only showed high prediction accuracy but also demonstrated good transferability between two large de-identified EHR databases with different structures, good external validity, proper model calibration, and utility of fine-tuning for continuous improvement. In addition, we used integrated gradients as a utility to expose the factors that contribute to the model predicted scores.

CovRNN models were predominant, with consistently higher performance as compared to other baseline methods. Interestingly, we found that the maximum difference between LR, LGBM, and CovRNN models' AUROC for iMort and mVent was around 3%, whereas it exceeds 6% for pLOS. Similarly, we observed that the pLOS prediction accuracy was highly affected by the inclusion of patient history. Therefore, we believe that considering the sequence of events that occurred in the past is of higher importance for the pLOS prediction task versus the iMort and mVent prediction tasks, for which the most recent events are of higher importance.

Feature ablation experiment demonstrated that the use of the patient's known comorbidities alone can achieve an AUROC of 86% for iMort and 84% for mVent (Supplementary Material D Experiment 1), which is expected as it summarises the patient's health condition. The addition of medication history and laboratory results increased the model performance by 3–4% for each category, as it filled the gap and introduced useful information to the model which were not captured by the recorded diagnoses. Afterwards, a relatively small degree of optimisation (<1%) was obtained from adding procedures, and nursing assessments, as they mostly introduce redundant information already captured from the previously fed data.

Although several studies reported predictive machine learning models with prediction accuracy comparable to our,[6,9,25–27] our model was trained and evaluated on larger, multicenter cohorts from two large, well-known de-identified EHR databases in the United States (a total of more than 300,000 patients). CovRNN outperforms other COVID-19 outcomes prediction models trained and evaluated on more than 50,000 COVID-19 patients which commonly rely on boosting-based algorithms[7,8]. The N3C study[9] had a similar number of COVID-19 patients (160,000) included in their training set; however, their reported prediction accuracy (AUROC)

against iMort and mVent (combined as severity indicator) was only 86%. In addition, the majority of published studies with machine learning models predict the outcome in a very short follow-up window, such as 1 hour to 1 day from the index time-point.[6,27] Further, some studies did not specify the time window of prediction or used limited historical data.[28] As the window periods become shorter, the prediction task is easier, and, thus, the accuracy is higher, but the results are less valuable, as physicians can predict those clinical outcomes better without using models. We reported CovRNN survival model results to demonstrate the flexibility of our approach. We believe, however, that predicting the probability of the adverse events' occurring within the hospital stay should be informative enough for clinicians to make appropriate decisions on admission and may not be limited to a specific time range. Therefore, we also focused on the evaluation and calibration of the binary classification models.

In our study, we included data available on or before the index admission to predict clinical outcomes throughout the hospital stay. Of note, our cohorts also included patients who stayed in observation units without hospitalization. Many physicians often encounter significant dilemmas when deciding the patient's disposition, such as discharge or higher level of care, in the observation unit or at the time of hospitalization. Further, patients with COVID-19 often progress rapidly, especially after about 7 days from the onset of symptoms, even when they initially present with mild symptoms.[29] This characteristic clinical course in patients with COVID-19 makes it substantially difficult for clinicians to predict future outcomes on the first day of hospital encounters. Our model is particularly helpful in those clinical scenarios, as the trajectory of the most important clinical outcomes, such as in-hospital mortality, was predicted with specificity of 71% at 95% sensitivity (Supplementary Table 3). As indicated in Supplementary Table 3, the threshold can be easily adjusted to prioritize the sensitivity or specificity to meet the clinicians' needs. For example, in the situation in which our model predicts the patient's death with high specificity, physicians could initiate an early discussion of poor outcomes with the patient and/or goals of care in appropriate cases. As the risk of further COVID-19 surges still cannot be ruled out and scenarios of healthcare systems being overwhelmed with a large number of infected patients is still a distinct possibility, CovRNN can be a useful tool while triaging patients. The score provided by CovRNN can be utilized to risk stratify large numbers of patients based on their readily available data in a few seconds. The minimal need for data curation and reliance on the power of the deep learning model architecture for learning proper feature representations from large data are key advantages of our CovRNN model. We were able to transfer the model between two completely different datasets that have some differences, particularly in clinical codes distributions. With a simple model fine-tuning step on sample data from the destination dataset, the model consistently achieved high prediction accuracy. Although we focused on COVID-19 patient outcomes, this is a good example for a proof of concept, and we can apply the same methodology to predict different clinical circumstances.

Limitations

Our study has several limitations. First, our data analysis includes only retrospective data. Despite our efforts to avoid potential bias by separating training, validation, and test datasets as well as external validation on a different data source, potential biases are inevitable. A prospective validation study is warranted, ideally, in hospitals that did not participate in data sharing with the database that we used to secure the validation of transferability. Second, our model focused only on predicting clinical outcomes at the time of hospital admission . It is possible to use multiple time points during the hospital stay to update models to achieve "real-time" predictions. Because minimal data preprocessing is required, our model can be easily modified to use different data points to predict future clinical outcomes. Third, real-world EHR structured data are not always associated with standard codes. For example, data from Cerner Millennium may not be codified at all in the source system or can only be associated with clients' proprietary event codes. Hospitals commonly have access to utilities to map their structured EHR data into industry-standard codes, which we utilized in our model, to facilitate interoperability, FHIR queries, data-sharing, billing, and public health reporting tasks. Such utilities are sometimes provided natively by their EHR vendors. Such mappings are required to get the benefit of our pre-trained CovRNN models, otherwise, we recommend using our CovRNN training framework to train compatible models utilizing the system's proprietary event codes. However, these codes or representations are only meaningful in the context of the originating system, and they are not helpful to train transferable models.

Fourth, we performed only a preliminary evaluation for the model predictions explanations, whereby we extracted data from 20 random sample patients and presented their predicted risk scores as well as the contribution score assigned for each medical event and asked infectious disease specialists to evaluate its relevance. Although we acknowledge that this is not a rigorous evaluation method, it demonstrated that our proposed model provides the tool that allows model transparency and helps to engage clinicians and facilitate their judgment on the model predictions. Future work is warranted to systematically evaluate the model's transparency. Fifth, the dynamics of COVID-19 management in hospitals and patient surges from pandemic waves have changed over time, which would modify the patient outcomes over time. Thus, the accuracy of our model reported for mostly the first wave (up to September 2020) may not be the same for future datasets. For example, the patients who received COVID-19 vaccines likely have different clinical outcomes.[30] Because our model is trained on historical data, the model can be easily fine-tuned on more current data to improve its prediction accuracy, which is one of the major advantages of deep learning models. Future work is warranted to fine-tune and evaluate our models on later waves data.

CONCLUSION

Through benchmarking, we found that our CovRNN can provide accurate and transferable predictive models for a wide range of outcomes and that we can continuously improve upon the model through periodic fine-tuning. Further, our data preparation pipeline is kept to a minimum to facilitate the transferability of the models and facilitate further validation on new data sources. Our model development framework can be further applied to train and evaluate predictive models for different types of clinical events. For clinicians who are fighting COVID-19 on the front lines, there are two potential actionable contributions of our work. Clinicians can (i) fine-tune our pre-trained model on their local data, regardless of the size, establish utility, and then deploy; and (ii) use our comprehensive model development framework to train a predictive model, using their own data.

DATA AVAILABILITY

The data that support the findings of this study, the Cerner® Real-World COVID-19 Q3 Dataset and Optum® de-identified COVID-19 Electronic Health Record v. 1015 dataset (2007–2020), are available for licensing at Cerner Corporation and Optum, Inc., respectively. Data access may require a data sharing agreement and may incur data access fees.

AUTHORS' CONTRIBUTION

L.R. and D.Z. conceived the idea. L.R. led the design and the implementation of experiments. L.R., K.P., M.N., and B.K. reviewed the evidence before the study. M.N. contributed to the discussion and the model explanation evaluation. Z.X. contributed to the model explanation. L.R., B.M., and K.P. ran the experiments on the OPTUM data. L.R. and Y.Z. extracted the EHR data. W.Z. added the visualizations. L.R. led the writing. B.K., M.N., H.X., and D.Z. also contributed to the writing. AR assessed the study against TRIPOD and PROBAST standards. H.X. and D.Z. supervised the project. L.R. and D.Z. finalized the manuscript. L.R. B.K., and M.N. had access to and verified the CRWD data. L.R, Y.Z., B.M., and K.P had access to and verified the OPTUM data. All co-authors reviewed and approved the manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

ETHICS COMMITTEE APPROVAL

The Committee for the Protection of Human Subjects at the University of Texas Health Science Center in Houston reviewed the IRB # HSC-SBMI-20-0836 for the "Analysis of COVID-19 related data in Cerner's HealtheDataLab" project. The committee determined the project to qualify for exempt status according to 45 CFR 46.101(b)

ACKNOWLEDGMENTS

We would like to acknowledge the use of the Cerner Real World® dataset access and the support from Cerner Research Team especially Cheryl Akridge. We also acknowledge the assistance provided by the University of Texas Health Science Center in Houston (UTHealth) School of Biomedical Informatics (SBMI) Data Service team and the UTHealth Center of Healthcare Data.

FUNDING

This work was supported by the Cancer Prevention and Research Institute of Texas (CPRIT) Grant No. RP170668 and the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015 and CPRIT Grant No. RP210042).

References

1 Coronavirus disease (COVID-19) – World Health Organization.

https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed May 29, 2021).

2 CDC. COVID Data Tracker. 2020; published online March 28. https://covid.cdc.gov/covid-data-tracker (accessed March 28, 2021).

3 Wynants L, Van Calster B, Collins GS, *et al.* Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369**: m1328.

4 Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in Covid-19. BMJ. 2020; **369**: m1464.

5 Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. Lancet Digit Health. 2020; **2**: e496–7.

6 Schwab P, Mehrjou A, Parbhoo S, *et al.* Real-time prediction of COVID-19 related mortality using electronic health records. *Nat Commun* 2021; **12**: 1058.

7 He F, Page JH, Weinberg KR, Mishra A. Development and validation of simplified machine learning algorithms to predict prognosis of hospitalized COVID-19 patients: a multi-center, retrospective study. *J Med Internet Res* 2021; published online Dec 19. DOI:10.2196/31549.

8 Feng A. Using electronic health records to accurately predict COVID-19 health outcomes through a novel machine learning pipeline. Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. 2021. DOI:10.1145/3459930.3469490.

9 Bennett TD, Moffitt RA, Hajagos JG, *et al.* Clinical Characterization and Prediction of Clinical Severity of SARS-CoV-2 Infection Among US Adults Using Data From the US National COVID Cohort Collaborative. *JAMA Netw Open* 2021; **4**: e2116901–e2116901.

10 Rasmy L, Wu Y, Wang N, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018; **84**: 11–6.

11 Xiang Y, Ji H, Zhou Y, *et al.* Asthma Exacerbation Prediction and Risk Factor Analysis Based on a Time-Sensitive, Attentive Neural Network: Retrospective Cohort Study. *J Med Internet Res* 2020; **22**: e16981.

12 Rasmy L, Zhu J, Li Z, Hao X, Zhi D. Simple Recurrent Neural Networks is all we need for clinical events predictions using EHR data. 2019; published online Aug 27. DOI:10.13140/RG.2.2.13199.51368.

13 Wanyan T, Honarvar H, Jaladanki SK, *et al.* Contrastive Learning Improves Critical Event Prediction in COVID-19 Patients. *ArXiv* 2021; published online Jan 11. https://www.ncbi.nlm.nih.gov/pubmed/33442560.

14 Rajkomar A, Oren E, Chen K, *et al.* Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 2018; **1**: 1–10.

15 Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: International Conference on Machine Learning. PMLR, 2017: 3319–28.

16 UMLS Metathesaurus - MMSL (Multum) - Synopsis. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MMSL/index.html (accessed May 27, 2021).

17 Rasmy L, Tiryaki F, Zhou Y, *et al.* Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. *J Am Med Inform Assoc* 2020; **27**: 1593–9.

18 Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA* 1982; **247**: 2543–6.

19 Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; **24**: 361.

20 Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Adv Neural Inf Process Syst* 2016; **29**. https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf (accessed Sept 2, 2021).

21 Modeling asynchronous event sequences with RNNs. *J Biomed Inform* 2018; **83**: 167–77.

22 sklearn.linear_model.LogisticRegression — scikit-learn 0.24.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed May 27, 2021).

23 Welcome to LightGBM's documentation! — LightGBM 3.2.1.99 documentation. https://lightgbm.readthedocs.io/ (accessed May 27, 2021).

24 Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018; **18**: 1–12.

25 Villegas M, Gonzalez-Agirre A, Gutiérrez-Fandiño A, *et al.* Predicting the Evolution of COVID-19 Mortality Risk: a Recurrent Neural Network Approach. DOI:10.1101/2020.12.22.20244061.

26 Razavian N, Major VJ, Sudarshan M, *et al.* A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. *NPJ Digit Med* 2020; **3**: 130.

27 Yadaw AS, Li Y-C, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. *Lancet Digit Health* 2020; **2**: e516–25.

28 Estiri H, Strasser ZH, Murphy SN. Individualized prediction of COVID-19 adverse outcomes with MLHO. *Sci Rep* 2021; **11**: 5322.

29 CDC. Healthcare Workers. 2021; published online May 27. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html (accessed June 7, 2021).

30 COVID-19 Vaccine Breakthrough Infections Reported to CDC — United States, January 1–April 30, 2021. *MMWR Morb Mortal Wkly Rep* 2021; **70**. DOI:10.15585/mmwr.mm7021e3.