CovRNN: predicting outcomes of COVID-19 patients on admission using their electronic health records with minimal data processing

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

[1]School of Biomedical Informatics, UTHealth, Houston, TX; [2]McGovern Medical School, UTHealth, Houston, TX; [3]College of Medicine, University of Arizona – Phoenix, AZ; [4]School of Public Health, UTHealth, Houston, TX.

Introduction

Coronavirus disease (COVID-19) is an infectious disease that emerged in December 2019 and it leads to the death of more than four million patients worldwide by the mid of July 2021. During the peaks of the pandemic waves, many places 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. A survey of the literature till the end of April 2021 showed only two studies[1,2] trained their proposed models on more than 20,000 patients. Both of which are based on specific features and needed laborious data preprocessing and feature engineering processes that limit the transferability, reliability, and sustainability of such models.

We developed CovRNN, a collection of deep learning based predictive models for the clinical outcomes of COVID-19 patients including: in-hospital mortality (iMort), need for mechanical ventilation during the stay (mVent), and hospital stay longer than one week (pLOS). CovRNN models were trained on more than 200,000 hospitalized COVID-19 patients to accurately predict patients' outcomes on admission using readily available structured electronic health records data (EHR) without the need for specific feature selection, feature engineering, or missing values imputation.

Methods

CovRNN utilized a gated recurrent neural networks (RNN) architecture namely gated recurrent unit (GRU) that encodes the temporal nature of patient history including the most recent COVID-19 visit as well as distant events that happened years ago. For iMort and mVent prediction tasks, CovRNN was trained to predict both time-to-event risk scores (survival prediction) as well as the all-time risk scores (binary prediction), to fit different clinical needs for healthcare workers confronting COVID-19.

We kept our data processing to the minimum 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. We utilized standard terminologies that are readily available in the majority of EHR systems to facilitate interoperability: ICD 9, ICD10, and SNOMED-CT for diagnosis; LOINC and SNOMED-CT for laboratory tests and observations; Multum drug identifiers and categories for medications; and CPT-4, HCPCS, ICD-9 PCs, and ICD-10 PCs for procedures. In cases where a coding system is not used such as Multum codes for medication, pre-existing mapping tools are available that can be used to convert NDC medication codes to corresponding Multum information. For laboratory results, we used clinical interpretations such as "below normal low", "normal", or "above normal high", instead of using the actual numerical value. By doing so all our input data became in a categorical format so we can further convert our input into trainable embedding matrices which will learn the features representation that mediates our model robustness to input errors and generalizability.

CovRNN was trained on a large heterogeneous dataset of 243,785 de-identified patients data derived from 85 different health systems and available through the Cerner® Real-World Dataset (CRWD) hosted on the Cerner HealthDataLab. We reported our prediction accuracy and model calibration results on a large heterogeneous held-out test set of 48,781 patients. For external validation, we evaluated the CovRNN on two held-out hospitals' data and compared the model performance against traditional machine learning models logistic regression (LR) and light gradient boosting machine (LGBM) for baseline comparison. We reported the results of subgroup analysis to understand CovRNN performance in different populations and utilized the integrated gradient technique to explain the model predictions.

Results

CovRNN achieved AUC of around 93% for both in-hospital mortality and need for mechanical ventilation binary predictions (vs. around 91.5% and 90% for LGBM and 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, showed a consistently high prediction accuracy for CovRNN models, as shown in Table 1 and Table 2. Subgroup analysis showed consistent results among different populations based on race, geographical region, and comorbidities as appear in figure 2, while a slight decrease in the model prediction accuracy was observed for older age (65 years or older).

Table 1. Binary classifications models Performance on different CRWD test sets.

	n	In-hospital Mortality (iMort)			Mechanical Ventilation (mVent)			Stay > 7 days (pLOS)		
		LR	LGBM	CovRNN	LR	LGBM	CovRNN	LR	LGBM	CovRNN
Multi- hospital Test Set	48,781	90·3	91·5	93·0	89·5	91·2	92·9	80·0	81·7	86·5
Hospital 1	3,469	88·8	91·0	91·8	86·7	88·4	91·5	77·3	78·5	87·2
Hospital 2	706	94·6	95·1	97·0	93·5	95·6	96·0	80·9	84·3	88·3

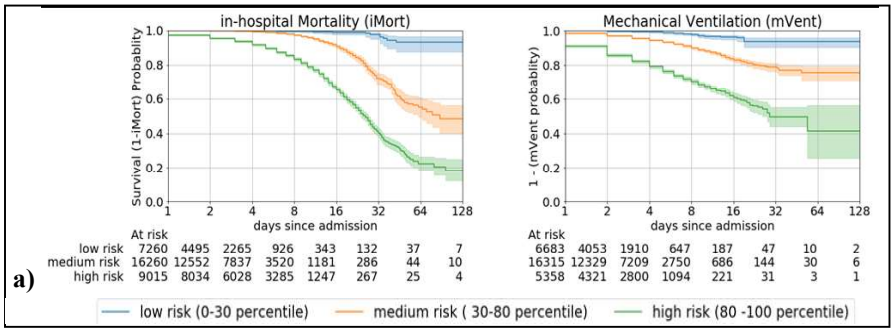

Figure 1. Kaplan–Meier (KM) curves showing the stratified survival analysis.

Table 2. Survival Analysis C-index

b)	iMort	mVent
Multi- hospital Test Set	86·0	92.6
Hospital 1	86·0	90.8
Hospital 2	91·6	93.8

Discussion and Conclusion

To the best of our knowledge, CovRNN is the first RNN based model that can accurately predict different COVID-19 patient's outcomes on admission, using readily available structured EHR in its raw categorical format without any need for specific feature selection, or missing values imputation.

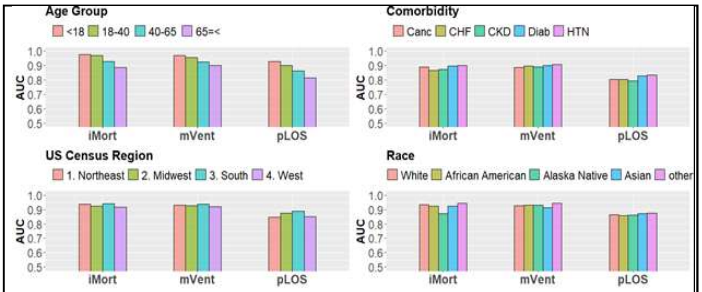

Figure 2. Subgroup Analysis

Our trained COVID-19 outcomes 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 it was trained and evaluated on large heterogeneous datasets collected from different health systems. Moving forward, our models can be fine-tuned on new data for continuous improvement as recommended by the FDA's good machine learning practice. Furthermore, our framework includes a utility for model predictions explanation to facilitate clinical judgment on the model predictions. For clinicians fighting COVID-19 in the frontline, there are two potential actionable contributions of our work: (i) They can fine-tune our pre-trained model on their local data regardless of the size; (ii) Use our comprehensive model development framework to train a predictive model using their own data. In our work we considered the major factors to establish the feasibility of our model, however, further clinical validation is required.

Acknowledgments L.R. is supported by the UTHealth Innovation for Cancer Prevention Research Training Program Pre-Doctoral Fellowship (CPRIT Grant No. RP160015)

References

1. Schwab P, Mehrjou A, Parbhoo S, Celi LA, Hetzel J, Hofer M, et al. Real-time prediction of COVID-19 related mortality using electronic health records. Nat Commun 2021;12(1):1–16.
2. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM, 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(7):e2116901–e2116901.