A Comprehensive Benchmark For COVID-19 Predictive Modeling Using Electronic Health Records in Intensive Care

Junyi Gao*, MS[1,2,3], Yinghao Zhu*, BE[1], Wenqing Wang*, BE[1], Yasha Wang, PhD[1], Wen Tang, MD[4], Liantao Ma, PhD[1]

[1]Peking University, Beijing, China; [2]University of Edinburgh, Edinburgh, UK; [3]Health Data Research UK, UK; [4]Peking University Third Hospital, Beijing, China

Abstract

The COVID-19 pandemic has posed a heavy burden to the healthcare system worldwide. Despite many deep learning models have been proposed to conduct COVID-19 clinical predictive tasks in intensive care, we are the first to provide a comprehensive benchmark for a fair comparison. We propose the Outcome-specific length-of-stay and Early mortality prediction task for patients in intensive care units. Our benchmark and online platform can further facilitate healthcare researchers and clinicians for COVID-19 predictive modeling.

Introduction

The COVID-19 pandemic needs no introduction. As of May 2022, the virus has caused over 500 million infected cases and over 6 million deaths. Though research shows that new variants of COVID-19 are less deadly, they are more spreadable and cause the number of cases still be surging globally. Under current circumstances, achieving early risk prediction and estimating the disease progression, especially for COVID-19 patients in intensive care units, have been an important topic to allocate limited medical resources and relieve the burdens of our healthcare system. Electronic health record (EHR) data and intelligent models have been viable solutions to solve this challenge. Many machine learning and deep learning models have been proposed to utilize COVID-19 patients' EHR data to conduct clinical prediction tasks, including severity[1], length-of-stay (LOS)[2], etc. There are more previous general EHR predictive models, which can also be applied to COVID-19 prediction tasks. These works achieve better prediction performances compared with simple statistical models. However, they still face two challenges in terms of task practicality and clinical versatility: (1). How to adapt traditional prediction tasks for COVID-19 patients in intensive care units; (2). How to choose the best one among various options of predictive models.

In this work, we propose a comprehensive benchmark for COVID-19 predictive modeling with EHR data in intensive care. The proposed pipeline of this benchmark is shown in **Figure 1**.

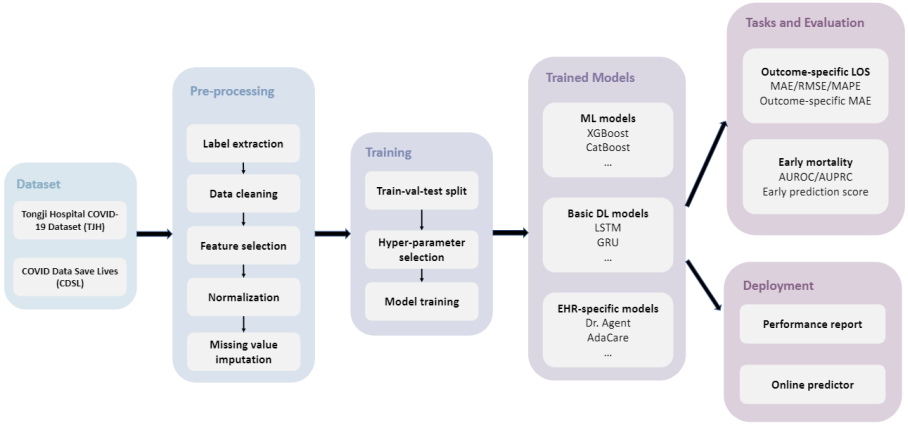

Figure 1. The proposed benchmark pipeline.

We propose fair, detailed, open-source data-preprocessing pipelines for two publicly available real-world COVID-19 EHR dataset (*Tongji Hospital TJH*[1] and *Covid Data Save Lives CDSL*[2]). We propose two clinical prediction tasks, *Outcome-specific length-of-stay prediction* and *Early mortality prediction* for COVID-19 patients in intensive care units. The two tasks are adapted from the naive length-of-stay and mortality prediction tasks to accommodate the clinical practice for COVID-19 patients. We also propose corresponding metrics to evaluate model performances on two tasks (i.e., *outcome-specific MAE, OSAME* and *early mortality prediction score, ES*). We evaluate 17 state-of-the-art predictive models on two tasks, including 5 machine learning models, 6 basic deep learning models and 6 deep

[1] https://github.io/HAIRLAB/Pre_Surv_COVID_19

[2] https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version

learning predictive models specifically designed for EHR data. We build an online platform to deploy benchmark experiment results, including model performances with all hyper-parameter combinations on both tasks. We also deploy all trained models online, which allows clinicians and researchers to easily use these models to directly get prediction results using their own data. The code is publicly available at https://github.com/yhzhu99/covid-ehr-benchmarks and the system is available at http://106.15.42.91:8000/.

Results

The benchmark prediction performances of two tasks on two datasets are shown in **Table 1** and **Table 2**. We do not show the performances of all 17 models due to space limitations. The entire performance table is available in the online system. The two-stage and multi-task in **Table 2** denote two training strategies. In the two-stage setting, we train the two models to predict outcome and LOS, respectively, and evaluate the performance together. In the multi-task setting, we train one single model with the same backbone but two different prediction heads. We found that deep learning models generally outperform traditional machine learning models such as random forest or XGBoost since they can utilize temporal disease progression information.

Table 1 Early mortality prediction performance on two datasets

Dataset	TJH			CDSL		
Metric	AUPRC	AUROC	ES	AUPRC	AUROC	ES
Random Forest	94.59 ± 3.25	92.95 ± 3.96	82.91 ± 7.29	58.33 ± 0.94	68.44 ± 0.37	10.93 ± 2.30
GBDT	95.28 ± 2.97	94.15 ± 3.32	89.23 ± 4.81	58.93 ± 1.57	70.82 ± 0.39	22.86 ± 1.04
LSTM	98.66 ± 1.69	98.78 ± 1.43	88.03 ± 6.47	66.22 ± 3.49	91.33 ± 0.82	38.51 ± 4.11
TCN	99.21 ± 0.99	99.21 ± 1.04	**93.19 ± 5.08**	69.96 ± 7.90	91.33 ± 2.19	**41.23 ± 11.74**
RETAIN	99.24 ± 0.72	99.21 ± 0.86	88.43 ± 8.03	68.77 ± 3.59	91.02 ± 1.09	36.75 ± 3.94
AdaCare	**99.29 ± 0.89**	**99.29 ± 0.93**	91.16 ± 5.25	69.55 ± 0.53	91.70 ± 0.14	39.83 ± 2.19
ConCare	98.90 ± 0.97	98.86 ± 0.97	91.14 ± 6.99	**75.33 ± 2.99**	**93.35 ± 0.83**	38.09 ± 9.56

Table 2 Outcome-specific length-of-stay prediction performance on two datasets

Dataset	TJH			CDSL		
Metric	MAE (↓)	RMSE (↓)	OSMAE (↓)	MAE (↓)	RMSE (↓)	OSMAE (↓)
Random Forest (two-stage)	5.31 ± 0.74	7.66 ± 1.12	4.73 ± 1.11	4.38 ± 0.00	7.14 ± 0.00	4.46 ± 0.05
GBDT (two-stage)	5.40 ± 0.75	7.75 ± 1.15	4.56 ± 0.91	4.34 ± 0.01	7.08 ± 0.01	4.40 ± 0.08
LSTM (multi-task)	4.05 ± 0.67	5.81 ± 1.00	4.24 ± 1.35	3.68 ± 0.08	6.16 ± 0.07	3.65 ± 0.13
LSTM (two-stage)	4.56 ± 0.56	6.32 ± 0.85	4.09 ± 1.09	3.21 ± 0.24	4.98 ± 0.66	3.84 ± 0.22
TCN (multi-task)	3.94 ± 0.77	5.59 ± 1.19	4.02 ± 1.05	3.43 ± 0.05	5.86 ± 0.07	**3.26 ± 0.15**
TCN (two-stage)	4.19 ± 0.65	5.89 ± 0.88	**3.52 ± 0.94**	**3.12 ± 0.11**	**4.76 ± 0.48**	3.66 ± 0.32
RETAIN (multi-task)	4.12 ± 0.63	5.61 ± 1.02	4.37 ± 1.09	3.58 ± 0.07	6.05 ± 0.08	3.45 ± 0.10
RETAIN (two-stage)	**3.83 ± 0.99**	**5.37 ± 1.24**	3.57 ± 1.12	3.27 ± 0.16	4.99 ± 0.58	3.97 ± 0.31
ConCare (multi-task)	4.50 ± 0.56	6.50 ± 0.81	5.01 ± 1.49	3.25 ± 0.15	5.35 ± 0.26	3.32 ± 0.16
ConCare (two-stage)	4.52 ± 0.60	6.38 ± 0.86	3.81 ± 0.71	3.38 ± 0.14	5.07 ± 0.44	3.71 ± 0.34

Conclusions

In this work, we propose a comprehensive benchmark for COVID-19 predictive modeling with EHR data in critical care. We set up the data pre-processing pipeline and propose two tasks based on the clinical practice of COVID-19 patients and implement 17 state-of-the-art machine learning and deep learning prediction models from existing COVID-19 as well as more general EHR predictive works. We also deploy all trained models online, which allows clinicians to easily use these models and get prediction results using their own data. We hope our efforts can further facilitate deep learning and machine learning research for COVID-19 predictive modeling.

References

1. Gao J, Yang C, Heintz J, Barrows S, Albers E, Stapel M, et al. MedML: Fusing Medical Knowledge and Machine Learning Models for Early Pediatric COVID-19 Hospitalization and Severity Prediction. Iscience. 2022:104970.
2. Ma L, Ma X, Gao J, Jiao X, Yu Z, Zhang C, et al., editors. Distilling knowledge from publicly available online EMR data to emerging epidemic for prognosis. Proceedings of the Web Conference 2021; 2021.