Training and Evaluating Interpretable Models on Electronic Health Records of COVID-19 Patients

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

A large cohort of Electronic Health Records (EHRs) collected by University of Washington School of Medicine since January, 2010 has recently become public. This includes more than 8.5k patients permitted for COVID-19 testing as per the official guidelines. Surprisingly, only a fraction $(\tilde{9}\%)$ of those allowed have been tested positive. A successful model may be used for a better prediction of potential candidates for testing to overcome the scarcity of testing resources. The features of such a model may provide new insights regarding the pathogenesis and epidemiology of COVID-19. We propose a modified novel interpretable model as an extension to Explainable Boosting Machine and showcase its results on our COVID dataset and further discuss the challenges in achieving this.

1 Introduction

Electronic Health Records (EHRs) are useful for temporal analysis of population health at a large scale. With the recent pandemic of COVID-19, we have witnessed how systematized collection of real-time patient data can help us combat such global crisis. However, there remains unsolved questions like could we predict the potential COVID-19 affected individuals based on their past EHR data. Recent attempts have shown that official medical guidelines of selecting COVID-19 susceptible are not efficient enough. We dig into this problem and put forward the challenges of training and evaluating prediction models on EHR data of COVID-19 Patients.

2 Related Work

Recent studies on EHR data have shown that one can identify the risk factors for hospitalization in confirmed COVID-19 patients [1]. However, the attempts to the prediction of potential COVID-19 targets from EHR data is limited. Another recent study by Chang et al. has shown that prior diagnoses and medications as risk factors for COVID-19 susceptibility and inpatient admission is possible with EHR data [2].

3 Problem Setting

The scientific problems that we have dealt with are discussed hereunder.

3.1 Research Questions

Most recent EHR data analysis consider bag-of-feature model which does not take into account other important information like dependency or relation among features. This can be denoted by

Submitted to NeurIPS 2021 AI for Science Workshop.

conditional dependency among features. For example, given *chest pain, fever* and *ECG*, without any structured information, we do not know which diagnosis is the reason for ordering ECG.

Secondly most EHR models are not interpretable in nature? For example can we check why a person was diagnosed with Covid Positive or was taken to the COVID ICU admission? These are critical questions that need to be answered and without interpretable models it is not possible understand clearly which feature attributes led to the corresponding diagnosis.

We try to answer the following questions

- **Question 1** : Among UW Medicine patients tested for SARS-CoV-2, predict the result of the initial test (positive or negative) using data available at the time the test is performed.
- **Question 2** : For patients with a positive RT-PCR for COVID-19 and who were tested at an outpatient visit, which patients were admitted to the hospital within 21 days of their RT-PCR test.
- **Question 3** : Among hospitalized patients, predict which patients will be admitted to ICU within 48 hours.

3.2 Explainable Boosting Machine

3.2.1 GAM and EBM

Generalized additive models (GAMs) are the gold standard for intelligibility in domains when only univariate terms are considered [6]. Standard GBM models can be represented as

$$g(E[y]) = \beta_c + \sum f_i(x_i) \tag{1}$$

where g is the link function that can be represented as either a regression or classification model and f_i denotes a feature function for the feature x_i.

Compared to previous GAM models, EBM can automatically detect and include pairwise dependencies in the form:

$$g(E[y]) = \beta_c + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j) \tag{2}$$

This helps in including more latent pairwise interaction information among features which further can improve accuracy predictions.

3.2.2 Interpretability

EBM is highly explainable because for the final prediction, each feature's contribution can be visualized. For instance for 3.1 we can visualize across the cohort of patients which symptoms or medical conditions had a higher degree of contribution towards positive diagnosis in RT-PCR test or in case of patients to be hospitalized in the next few weeks, which signs were more important relative to others.

Because EBM is an additive model, each feature contributes to predictions in a modular way that makes it easy to reason about the contribution of each feature to the prediction.

To make individual predictions, each function f_j acts as a lookup table per feature, and returns a term contribution. These term contributions are simply added up, and passed through the link function gg to compute the final prediction. Because of the modularity (additivity), term contributions can be sorted and visualized to show which features had the most impact on any individual prediction.

3.2.3 Proposed Model - EBM-Importance (EBM-I)

Instead of considering all pairwise interactions, we assign higher weights to those pairwise interactions for whom the two feature pairs' distribution has a higher correlation coefficient.

$$g(E[y]) = \beta_c + \sum f_i(x_i) + w_{i,j} * \sum f_{i,j}(x_i, x_j) \tag{3}$$

where $w_{i,j}$ is proportional in scale with $r(X_i, X_j)$, r being the correlation coefficient and all weights are normalised..

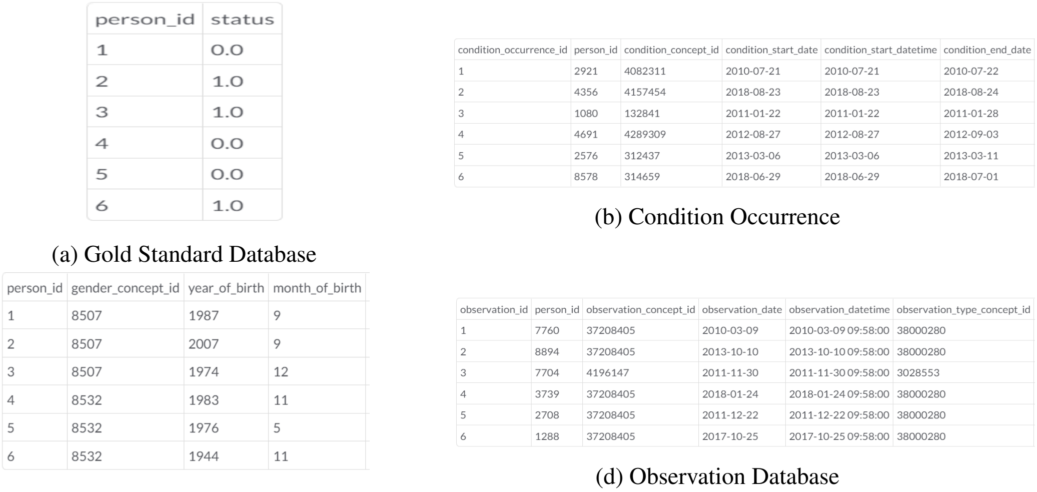
4 Experiments

4.1 Dataset Description

We perform our experimental evaluations on data collected from multiple large hospitals in the University of Washington Medical System. This represents 10 years of clinical records (2010-2020) from 9,500 patients, each of whom has at least one visit record and at least one test for COVID-19. These records include medications prescribed, conditions of patients, observations such as blood pressure and heart rate, demographic information, procedures, and laboratory measurements. Of the patients in the dataset, $\approx$ 800 patients have tested positive for COVID-19, with all of these patients having at least one visit.

Due to research purposes, we only consider a synthetic dataset provided as part of the Covid EHR Challenge [1] whose data distribution is made to mimic the real data distribution. We perform our analysis on 1251 patients as our training set and 536 patients as our test dataset.

- **Person Demographics Database** : The person demographics database contains demographic information who participated in the study like age and gender. Table 1c
- **Condition Occurrence Database** : Conditions are records of a patient suggesting the presence of a medical condition stated as a diagnosis, a sign or symptom, observed by a Provider or reported by the patient. 1b
- **Observation Database** : This table captures clinical facts about a Person obtained in the context of examination. It also includes data obtained from questionnaires during examination like lifestyle facts, medical history, family history. Table 1d
- **GoldStandard Database** : For evaluation purpose we train our model and evaluate on the goldstandard database. The file stores the true status of the patients w.r.t to the above research questions in 3.1. For example in question 1, the goldsandard database gives information for each patient whether they are tested positive/negative as labels $1.0/0.0$ respectively. Table 1a



person_id	status
1	0.0
2	1.0
3	1.0
4	0.0
5	0.0
6	1.0

(a) Gold Standard Database

condition_occurrence_id	person_id	condition_concept_id	condition_start_date	condition_start_datetime	condition_end_date
1	2921	4082311	2010-07-21	2010-07-21	2010-07-22
2	4356	4157454	2018-08-23	2018-08-23	2018-08-24
3	1080	132841	2011-01-22	2011-01-22	2011-01-28
4	4691	4289309	2012-08-27	2012-08-27	2012-09-03
5	2576	312437	2013-03-06	2013-03-06	2013-03-11
6	8578	314659	2018-06-29	2018-06-29	2018-07-01

(b) Condition Occurrence

person_id	gender_concept_id	year_of_birth	month_of_birth
1	8507	1987	9
2	8507	2007	9
3	8507	1974	12
4	8532	1983	11
5	8532	1976	5
6	8532	1944	11

(c) Person Demographics Information

observation_id	person_id	observation_concept_id	observation_date	observation_datetime	observation_type_concept_id
1	7760	37208405	2010-03-09	2010-03-09 09:58:00	38000280
2	8894	37208405	2013-10-10	2013-10-10 09:58:00	38000280
3	7704	4196147	2011-11-30	2011-11-30 09:58:00	3028553
4	3739	37208405	2018-01-24	2018-01-24 09:58:00	38000280
5	2708	37208405	2011-12-22	2011-12-22 09:58:00	38000280
6	1288	37208405	2017-10-25	2017-10-25 09:58:00	38000280

(d) Observation Database

Figure 1: Database collections

We combined the condition occurrence table with the person table to result in a single database for our computation. The concept id codes are from Athena directory which are basically SNOMED, ICD vocabulary.

Here, we compare our model's accuracy for the 3 questions w.r.t original EBM model in table 1

[1] Synapse Challenge

Question	EBM	EBM-I
Q1	0.81 ± 0.002	0.87 ± 0.0041
Q2	0.77 ± 0.008	0.81 ± ± 0.0012
Q3	0.67 ± 0.03	071 ± 0.004

Table 1: Accuracy of Predictions for our 3 questions

5 Challenges

The original proposal using Explainable Boosting Machine is a bit computationally inefficient when it is scaled to more number of patients > 5000. This is primarily due to the fact in the proposed model we consider pairwise interactions among the corresponding features (in our case ≈ 8619 features) and therefore for pairwise interactions calculations we need to consider total $17238 + 8619 = 25857$ features. Hence performance is compromised for EHR data which has large number of features.

We must explore the challenges that lie with the demography of EHR data. A recent study on EHRs showed that laboratory tests at the time of diagnosis reflect that hospital-level differences are in concurrence with country-level variations [3]. This is an important criteria which needs to be considered somehow in the model.

6 Future Directions

In future we aim to work on incorporating parallelism to the proposed EBM model when calculating the interactions among features. We aim to do so via a multi-threaded approach or via keeping atmost K such x_i, x_j feature pairs in different partitions and assign their computation in a corresponding thread.

Following the proposal of incorporating pairwise feature interactions, we also look forward to looking into recent advancements in EHR model analysis using Graph Neural Networks, where each node in the graph can be considered as the symptom feature and the connections can be considered as the interactions. Recent methods like [7] has proposed using transformer models that takes into account the concept of attention in neural networks to consider such pairwise interactions of features in EHR data.

References

[1] Oetjens, M.T., Luo, J.Z., Chang, A., Leader, J.B., Hartzel, D.N., Moore, B.S., Strande, N.T., Kirchner, H.L., Ledbetter, D.H., Justice, A.E. & Carey, D.J. (2020) Electronic health record analysis identifies kidney disease as the leading risk factor for hospitalization in confirmed COVID-19 patients. *PLoS One* **15**(11):e0242182.

[2] Chang, T.S., Ding, Y., Freund, M.K., Johnson, R., Schwarz, T., Yabu, J.M., Hazlett, C., Chiang, J.N., Wulf, A., Geschwind, D.H. & Butte, M.J. (2020) Prior diagnoses and medications as risk factors for COVID-19 in a Los Angeles Health System. *medRxiv*.

[3] Brat, G.A., Weber, G.M., Gehlenborg, N., Avillach, P., Palmer, N.P., Chiovato, L., Cimino, J., Waitman, L.R., Omenn, G.S., Malovini, A. & Moore, J.H. (2020) International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *Npj Digital Medicine* **3**(1):1-9.

[4] Hastie, Trevor, and Robert Tibshirani. "Generalized additive models: some applications." Journal of the American Statistical Association 82, no. 398 (1987): 371-386.

[5] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate intelligible models with pairwise interactions." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 623-631. 2013.

[6] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In KDD, 2012.

[7] Choi, Edward, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. "Learning the graphical structure of electronic health records with graph convolutional transformer." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 606-613. 2020.