Temporal Phenotype Matrix Engineering for Electronic Health Records – Enhancing Coronary Artery Disease Prediction

Kuan-Hui Liu, Cheng-Yu Chiang, Hsin-Yao Wang and Yi-Ju Tseng, *Member, IEEE*

***Abstract*—Nowadays, most studies still aggregate electronic health records (EHRs) into one record per patient for analysis and model development without considering temporal information, which is valuable for disease progression and outcome prediction. However, EHRs often exhibit sparsity and irregularity due to their inherent nature, and data preprocessing is needed to extract temporal information in EHRs. It is crucial to consider that imputation and aggregation techniques used during EHRs preprocessing can introduce artificial and unrealistic data, potentially leading to the loss of critical information. In this study, we proposed a temporal phenotype matrix engineering approach with auxiliary data layers (ADL) to extract important hidden information from EHRs. Our proposed approach was applied to the early prediction of coronary artery disease (CAD), one of the leading causes of death worldwide. We evaluated the performance of the long short term memory network (LSTM), convolutional neural network (CNN), and temporal convolution network (TCN) models on the CAD prediction task. Upon applying our proposed matrix engineering technique with ADL, we observed a substantial improvement, with an AUROC (area under the receiver operating characteristic) score of 0.919 ± 0.006 (a 10% increase, compared to when no ADL was included, 0.831 ± 0.011) in CNN model. In conclusion, this study highlights the benefits of the proposed temporal phenotype matrix engineering approach with ADL to address the sparsity and irregularity inherent in EHRs data.**

***Clinical Relevance*—**

Our findings underscore the potential of the proposed temporal phenotype matrix engineering approach with ADL for enhancing the early prediction of CAD, thereby contributing to improved patient outcomes and reduced mortality rates.

I. Introduction

Electronic health records (EHRs) store patient's medical history in a digital format. They contain various types of medical data, including patient demographics, diagnoses, laboratory and test results, medications, and radiology reports. Due to the richness of information, there is significant research potential for mining and exploring previously unknown correlations between diseases and heterogeneous data [1]. Records within EHRs are collected chronologically, making EHRs a valuable source of time series data. Nowadays, most studies still aggregate EHRs into one record per patient for analysis and model development without considering temporal information, which is valuable for disease progression and outcome prediction. Time series analysis allows for the detection of patterns and trends within EHR data, providing a better understanding of disease progression and treatment effectiveness over time. Additionally, time series models have been widely implemented for forecasting future health conditions [2]. The utilization of EHRs has the potential to greatly advance the diagnosis and forecasting tasks. However, several challenges arise due to the nature of EHRs [1]. One such challenge is the irregularity in data collection frequency, which depends on the patient's health condition. This irregularity can lead to sparse data.

To addressed these issues, data imputation and aggregation are commonly used methods. Various imputation techniques were applied to clinical data modeling [3, 4]. However. the impact of data imputation on model performance could be minimal [5], and its effectiveness depends on the nature of the data. In addition, there is a risk to lose critical information when applying aggregation.

We propose a novel temporal phenotype matrix engineering technique, which aims to extracts crucial information from EHRs. We applied this approach to predict coronary artery disease (CAD), a leading cause of death worldwide. Early prediction of CAD has shown to have a significant impact in reducing mortality rates.

II. Methods

A. Dataset

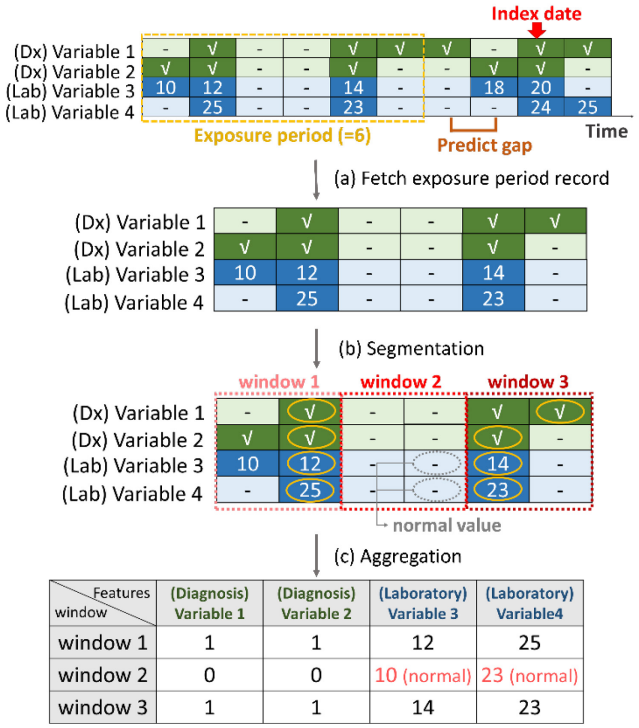
The study population comprised of patients who received laboratory tests for regular cardiac check-up in Chang Gung Memorial Hospitals, between January 1, 2001, and October 13, 2018. The laboratory tests for regular cardiac check-up included T-cholesterol, cholesterol, high density lipoprotein cholesterol (HDL-C), glycated Hemoglobin (Hb-A1c), low density lipoprotein cholesterol (LDL-C), and at least one C-Reactive protein (CRP) or high sensitivity CRP between 2001/01/01 and 2018/10/13. Demographic information (e.g., age and sex), diagnosis, and laboratory test results were obtained from Chang Gung Research Database (CGRD), which is the largest EHR research database in Taiwan. We categorized the original diagnosis codes into 283 Clinical Classification Software (CCS) diagnosis groups. Subsequently, we identified the laboratory tests that were conducted on more than 20% of the patients within the study population. A total of 80 laboratory tests were included in the analysis. The Chang Gung Medical Foundation Institutional Review Board

*Research supported by the National Science and Technology Council, Taiwan (NSTC 111-2628-E-A49-026-MY3).

Kuan-Hui Liu is with the Institute of Data Science and Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan (e-mail: nycu311554022.cs11@nycu.edu.tw). C.-Y Chiang was with the Department of Information Management, National Central University. H.-Y. Wang is with the Department of Laboratory Medicine, Chang Gung Memorial Hospital (CGMH) at Linkou, Taoyuan, Taiwan. Yi-Ju Tseng is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan (Corresponding author; phone: +886-3-5712121 #54713; fax: +886-3-5721490; email: yjtseng@nycu.edu.tw)

approved this study (IRB no. 201801771B0) and waived the requirement for patient consent.

B. Data Preprocessing



Features / window	(Diagnosis) Variable 1	(Diagnosis) Variable 2	(Laboratory) Variable 3	(Laboratory) Variable4
window 1	1	1	12	25
window 2	0	0	10 (normal)	23 (normal)
window 3	1	1	14	23

Figure 1. Data preprocessing workflow (a) For early prediction, we extracted data prior to the prediciton gap and within the exposure period. (b) Next, we segmented the data using a defined window period. (c) To tackle sparsity, we aggregated the records within each window.

To addressed the challenges of irregularity, sparsity, and high dimensionality in EHR analysis, we proposed a temporal phenotyping approach to generate time series data for further analysis.

We first defined the index data as the date of the first diagnosis of CAD. For patients who were never diagnosed with CAD during the study period, we use the last record date. To predict CAD in advance, we defined the prediction gap as 365 days and included the data prior the defined prediction gap for CAD prediction. The exposure period, for example 5 years, represents the total length of data included. With this information, we retrieved each patient's records within the exposure period (Figure 1 (a)).

Then, we segmented the exposure period into non-overlapping windows T of length L, which (T*L) is equal to the length of the exposure period shown in Figure 1 (b). In this case, T is the number of observations and L is the frequency. After segmentation, we aggregated the record data for each variable within the window period. For diagnosis data, we use '0' and '1' to indicate the absence or presence of a recorded diagnosis, respectively, within the window period (see Figure 1(c)). However, representing laboratory test results solely with '0' and '1' is insufficient, as these results contain valuable clinical information. Therefore, when multiple laboratory tests results are available within a window period, we choose the most recent results to represent the observation for that window. Conversely, if there is no record of a laboratory test within the window period, we imputed the missing value with the normal value for that specific laboratory test. This imputation method assumes that patients were in a healthy state at that time, thus not requiring the test. In this case, if we set T to 1 and L equal to exposure period, we could get the most recent record.

C. Classification neural network

After data pre-processing, we obtained multivariate time series data that captured temporal phenotype within EHR. Long-short-term memory (LSTM) is a popular approach of multivariate time series data classification task. Unlike the traditional Recurrent Neural Network (RNN), the LSTM model is designed to overcome the issues of vanishing and exploding gradients. This is achieved through the use of forget gate, which allows the LSTM to selectively remember important information and forget nonsignificant information.

Time series data can be further transformed into two-dimensional arrays, facilitating the application of convolution neural network (CNN) [6]. We leverage the CNN's inherent ability to identify the implicit features and patterns within the data. We adopted a basic CNN model architecture, which has two convolutional layers, two maximum pooling layers, one flatten layer, two fully connected layers, and two dropout layers for avoiding overfitting.

Building upon the concept of applying CNN to time series data, we also incorporate the Temporal Convolution Network (TCN). Because TCN is specifically designed to address the time series discovery problem and shares similarities with convolution network, TCN is outperformed RNN models in various datasets [7]. TCN consists of three key components. First, causal convolutions, enables TCN to process the time series data. Second, dilated convolutions are utilized to expand the receptive field to contain larger-scope information without pooling. Third, the residual connections were employed to help the network maintain a sufficient receptive field.

D. Auxiliary Data Layer (ADL)

During the data preprocessing step, particularly in the process of aggregation and imputation, there is a risk to lose critical information and potentially introducing artificial or unreal data. To mitigate this issue and ensure a more accurate representation of the patients' real health condition within the EHR, we designed an Auxiliary Data Layer (ADL).

One important task in data preprocessing is to impute missing values and ensure the data is complete for further analysis. However, it's worth noting that missing values can sometimes contain valuable information. Study proved that providing missing location to the model improve the performance of classification [8]. Due to the importance of missing location, we included the missing location in the proposed ADL. We denote a multivariate time series as X with V variables and T windows.

For $X = (x_1, x_2, \dots, x_T)^T \in \mathbb{R}^{T \times V}$, where for each $t \in \{1, 2, \dots, T\}$, $x_t \in \mathbb{R}^V$, x_t represent the all variables value in *the ith* window. And for each $t \in \{1, 2, \dots, T\}$; $v \in \{1, 2, \dots, V\}$, $x_t^v \in \mathbb{R}$, x_t^v means d-th variable value in t-th window. After denoting the time series data, we denote the missing locations as M, $M = (m_1, m_2, \dots, m_T)^T \in \{0, 1\}$,

where for each $t \in \{1, 2, \dots, T\}$, $m_t \in \{0, 1\}$, m_t represent the missing status in t-th window. And for $t \in \{1, 2, \dots, T\}$; $v \in \{1, 2, \dots, V\}$, $m_t^v \in \{0, 1\}$, m_t^v represents v-th variable in t-th window missing or not. The function is:

$$m_t^v = \begin{cases} 0, & \text{if } x_t^v \text{ is nan} \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

Apart from considering the missing locations, another aspect to address during the aggregation of multiple records within a window is the potential loss of valuable information and data trends. The conventional approach of discarding all records except the least one may result in the omission of important insights hidden within the discarded records. To solve this issue, similar to handling the missing locations, we include the number of records, maximum, minimum, and mean values within the window period during aggregation as part of the ADL [9].

By incorporating the ADLs with the original time series matrix, we can generate three-dimensional arrays as illustrated in Figure 2. Since the window size and number of variables are fixed, we appended data along the layer axis. As a result, we can get the data in shape of (number of layers, number of windows, number of variables).

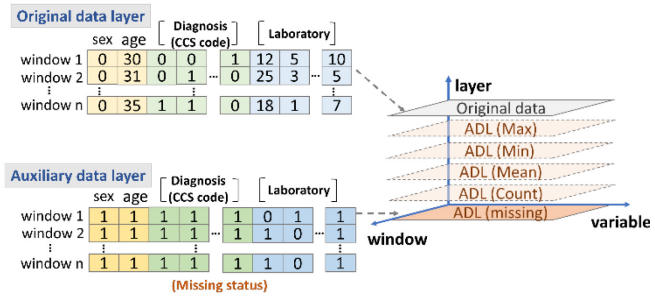


Figure 2. Process of Appending ADL

III. Experiment Design and Result

After pre-processing the data, we augmented the time-series matrix by incorporating sex and age information. The resulting matrix consists of 365 features. We have a total 8,188 patients diagnosed with CAD (case group) and 58,562 patients never diagnosed with CAD (control group), as shown in TABLE I.

TABLE I. Statistical Data in Demographic Data

		Case Group (N=8,818)	*Control group (N=58,562)*	*P-value*
Sex	**Female**	3,553 (40.3%)	28,074 (47.9%)	< 0.001
	Male	5,265 (59.7%)	30,488 (52.1%)	
Age	**Mean (SD)**	62.8 (11.9)	59.9 (15.6)	< 0.001
	Median [Q1, Q3]	63.0 [55.0, 71.2]	61.3 [48.8, 70.9]	< 0.001

A. Experimental Setting

The data was randomly divided into a training set, comprising 80% of the data, and a testing set, comprising the remaining 20%. To guarantee robustness, we repeat the evaluation process 20 times. The experiments were conducted using the NVIDIA GeForce RTX 3060 laptop GPU. The environment and hyperparameter are shown in Table II.

TABLE II. Setting and Environment of Hyperparameters in the Experiment

Environment	Keras 2.8.0 based on Tensorflow 2.8.0
Epoch (without ADL[a])	10
Epoch (with ADL)	8
Optimizer	Adam
Learning Rate	0.0005
Loss function	binary cross-entropy;

[a]ADL: Auxiliary data layer

During the data preprocessing stage, various parameters need to be determined, such as the exposure period, window period, and prediction gap. To identify the optimal combination of these parameters, we conducted experiments with different configurations. We found that the best performance was achieved using the following parameter values: an exposure period of 5-year, a window period of 365 days, and a prediction gap of 365 days.

B. Comparison of time series data and most recent record

The results showed that the performance of time series data (0.827 ± 0.009) is better than the model developed with the most recent record (0.820 ± 0.010). It proved that time series data can provide more efficient information in early prediction tasks.

C. Comparison among LSTM, CNN and TCN

The results indicate that both CNN (0.831 ± 0.011) and TCN (0.847 ± 0.009) outperform the LSTM (0.827 ± 0.009) models. These findings suggest that convolution neural network-based models, specifically CNN and TCN, are more suitable for the given task in this case.

D. Contribution of ADL

TABLE III. Comparative Results of the Original Time Series Data Without and With ADL.

Exposure period = 5-year Window period = 365 days Predict gap = 365 days			
Model	***Without ADL[a]***	***With ADL***	
		Missing	***Missing & Count***
TCN[b]	0.847 ± 0.009	0.897 ± 0.005	0.917 ± 0.007
CNN[c]	0.831 ± 0.011	0.892 ± 0.005	**0.919 ± 0.006**

[a]ADL: Auxiliary data layer [b]TCN: Temporal convolution network
[c]CNN: Convolution neural network

To further investigate the contribution of ADL in compensating for data loss during preprocessing, we incorporated the ADL into TCN and CNN models. Additionally, these models were modified to 3D format required by the inclusion of the ADL, as the data with ADL becomes three-dimensional arrays. The results, as shown in Table III, demonstrate that both the TCN and CNN models exhibited an improvement in performance of 6% and 7% when only the missing locations were included in ADL, respectively. Furthermore, when both missing locations and count of the

records are incorporated into the ADL, the models achieved the highest performance improvement of 10% compared to the models without ADL. The CNN model, incorporating missing locations and count of records in the ADL, achieved the best performance among the models evaluated. It attained an AUROC (area under the receiver operating characteristic curve) of 0.919± 0.006.

E. Comparison between different predict gaps

Figure 3 provides insights into the performance of the predictive models across different prediction gaps, with a focus on exploring how early the model can predict the onset of CAD. The primary objective is to identify the optimal prediction gap that allows for early prediction and subsequent treatment, leading to a reduction in the disease's mortality rate. Figure 3 shows a decrease in performance as the prediction gap extends. Specifically, there is a noticeable decrease in performance when the prediction gap reaches 730 days. This suggests that the models are less effective in predicting CAD at longer prediction gaps.

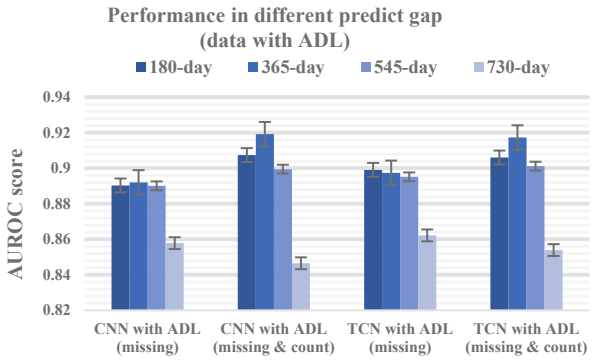


Figure 3. Performance between different models in different predict gaps among CNN (convolution neural network) and TCN (temporal convolution network) model with or without ADL.

IV. Discussion and Conclusion

To address the challenges of extracting temporal information posed by the irregularity and sparsity of EHR, we proposed a temporal phenotype matrix engineering approach to obtain multivariate time series matrix with ADL. By doing so, we can accurately capture and represent the temporal patterns of patients within the EHR. The inclusion of ADL in the multivariate time series matrix enables us to incorporate additional relevant information and compensate for data loss during the preprocessing step. This integration of ADL improved the performance of CAD prediction by providing a more comprehensive representation of patients' health conditions over time.

We observed that incorporating missing locations into the modeling process can significantly improve the performance of CAD prediction. That is, missing value within EHR data provide valuable information that allows the model to extract the most important features and patterns relevant to CAD prediction. Additionally, besides considering missing locations, we also incorporated other information [9] along with the original time-series matrix. We have noticed a distinct pattern where the performance of the CAD prediction model does not increase significantly when missing locations or counts are not incorporated. This observation suggests that missing values and the frequency of their occurrences within specific intervals hold substantial importance for accurate CAD prediction.

Our task is early prediction, so we also compare the performance of each model in different prediction gaps. The performance of the training with ADL decreases when the prediction gap length is extended. Although our approach can improve performance, while the prediction gap is longer than 730 days, performance will drop dramatically. So, we could try some models which can focus on the early data [10].

In conclusion, we proposed a temporal phenotype matrix engineering approach with ADL to tackle the challenges of irregularity and sparsity in the EHR analysis. By incorporating missing location and count as ADLs, we have demonstrated significant improvements in the performance of CAD prediction. Furthermore, in terms of early prediction, we can maintain excellent performance with a predict gap of less than one and a half years in our experiments. In CAD prediction, many researches also used computed tomography (CT) [11] or ECG data [12]. Based on it, we might develop toward multi-modality in the future.

References

[1] Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* 13.6 (2012): 395-405.

[2] Solares, Jose Roberto Ayala, et al. "Deep learning for electronic health records: A comparative review of multiple deep neural architectures." *Journal of biomedical informatics* 101 (2020): 103337.

[3] Bhattacharya, Moumita, et al. "Identifying ventricular arrhythmias and their predictors by applying machine learning methods to electronic health records in patients with hypertrophic cardiomyopathy (HCM-VAr-risk model)." *The American journal of cardiology* 123.10 (2019): 1681-1689.

[4] Beaulieu-Jones, Brett K., Jason H. Moore, and POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM. "Missing data imputation in the electronic health record using deeply learned autoencoders." *Pacific symposium on biocomputing* 2017.

[5] Bhaskaran, Krishnan, and Liam Smeeth. "What is the difference between missing completely at random and missing at random?." International journal of epidemiology 43.4 (2014): 1336-1339.

[6] Yeh, Marvin Chia-Han, et al. "Artificial intelligence–based prediction of lung cancer risk using nonimaging electronic medical records: Deep learning approach." *Journal of medical Internet research* 23.8 (2021): e26256.

[7] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).

[8] Fei, Ke, Qi Li, and Congcong Zhu. "Non-technical losses detection using missing values' pattern and neural architecture search." *International Journal of Electrical Power & Energy Systems* 134 (2022): 107410.

[9] Y.J. Tseng, X.O. Ping, J.D. Liang, P.M. Yang, G.T. Huang, and F. Lai. Multiple Time Series Clinical Data Processing for Classification with Merging Algorithm and Statistical Measures. *IEEE J Biomed Health Inform* 2015; 15(3):1036-43.

[10] Park, Jiheum, et al. "Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer." *Journal of Biomedical Informatics* 131 (2022): 104095.

[11] Zhang, Jia, et al. "Artificial intelligence in cardiovascular atherosclerosis imaging." *Journal of Personalized Medicine* 12.3 (2022): 420.

[12] Han Changho, et al. "Artificial intelligence-enabled ECG algorithm for the prediction of coronary artery calcification." *Frontiers in Cardiovascular Medicine* 9 (2022): 849223.