Graph Kernel Prediction of Drug Prescription

Hao-Ren Yao*, Der-Chen Chang*, Ophir Frieder*, Wendy Huang§, and Tian-Shyug Lee¶
* Georgetown University, Washington, DC, USA
§ Meng Cheng Family Medicine Clinic, Kaohsiung City, Taiwan
¶ Fu Jen Catholic University, New Taipei City, Taiwan

***Abstract*—Predictive models for drug prescription exist; we propose an additional such model that uses a graphical representation of an electronic health record with the prescription process formulated as a kernelized binary (success or failure) classification problem. Our approach improves accuracy while maintaining a high level of interpretability. Results using the Taiwanese National Health Insurance Research Database (NHIRD) favorably compare our approach to other models.**

I. Introduction

Erroneous medication prescription is defined as a failure in the medication treatment process that results in an unsuccessful treatment or harmful outcome to patients [1]. Clinicians have the responsibility to accurately diagnose and adequately treat a patient's disease. For treatments that require medications, the ideal prescription is the one that is most effective and presents the least harmful side effects. Yet, this is not always achieved.

The rapid growth of patient Electronic Health Records (EHRs) provides opportunities to develop a data-driven analytical application on medical data [2]. Many approaches for many medical applications exist [3]–[7]. Our focus is the development of an accurate and efficient model to predict the success potential of a specific drug prescription for a given ailment.

Due to the complex nature of EHR data, implementing a predictive model is difficult. For example, electronic phenotyping is the process of extracting relevant features from EHRs, a major step before performing an analytical task [8]. However, as a feature extraction technique, electronic phenotyping may cause information loss on the discriminant features. Recently, the emergence of deep learning models pose other ways to analyze EHR data (e.g., EHR data embedding) which achieve better performance with significantly less feature engineering [9]. However, result interpretation of such systems is difficult.

We develop a framework to predict the success or failure of a drug used for disease treatment. The described approach is already in preliminary use and assists clinicians in identifying which drug prescription path to pursue. The method predicts success or failure of drug prescription by formulating it as a binary graph classification problem without the need of electronic phenotyping. First, we identify training data, namely success and failure patients for target disease treatment within a user-defined time period; we extract the set of medical events from patient EHRs that occur wihin this time quantum. Then, we perform a classification task on the graphical representation of the patient EHRs. Representing EHRs as a temporal graph is provenly effective, as shown in [10]. The graphical representation provides an opportunity to model the EHRs in a compact structure with high interpretability.

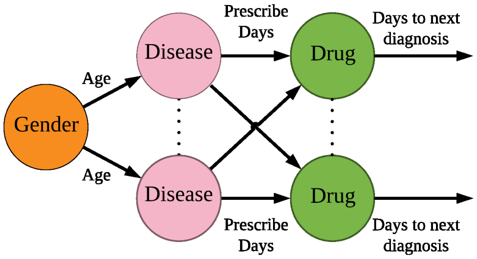


Fig. 1: An example of patient graph

Our contributions are as follows:

- We formulate a prediction task as a graph classification problem.
- We perform the classification task directly on the graph without the need of electronic phenotyping.
- We show that the graph kernel method is effective while providing high interpretability.

II. Methods

A. EHR Graph Representation

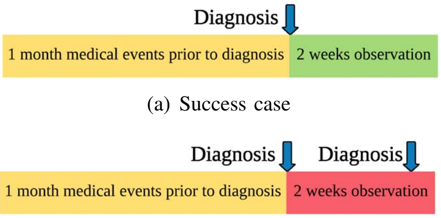
We formulate a patient's EHR or a subset of the EHR as a directed acyclic graph where each medical event represents a node, and two consecutive medical events form an edge with time difference (e.g., days) used as a weight between them. Patient demographic information, e.g., gender, connects to a first medical event with age as an edge weight.

Given n medical events, set $M = \{(m_1, t_1), \ldots, (m_n, t_1)\}$ represents a patient's EHR with m_i denoting a medical event such as diagnosis, and t_i denoting the time for m_i. We define the patient graph as follows:

Definition 1 (*Patient Graph*)**.** *The patient graph $P_g = (V, E)$ of events M is a weighted directed acyclic graph with its vertices V containing all events $m_i \in M$ and edges E containing all pairs of consecutive events (m_i, m_j). The edge weight from node i to node j is defined as $W_{ij} = t_j - t_i$ which defines the time interval between m_i, m_j.*

B. Success and Failure Prescription

Given a disease diagnosis and the clinical medical history of a patient, a drug prescription for the diagnosis is considered as


Diagnosis
1 month medical events prior to diagnosis
2 weeks observation
Diagnosis
Diagnosis
1 month medical events prior to diagnosis
2 weeks observation

(a) Success case

(b) Failure case

Fig. 2: Criteria for success and failure cases

a failure if there is an occurrence of clinical visit with a similar or identical type of disease diagnosis within a predefined time period. Otherwise, the prescription is considered a success. Figure 2 illustrates this criterion[1]. A failure is labelled as a positive outcome (as we manage to avoid a poor outcome), and a success as negative (as the prescription was valid and no corrective measure was necessary).

C. Predictive model formulation

Given a patient EHR, the patient's current diagnosis, and the drug prescription to the current diagnosis, we wish to predict the success or failure of the prescribed medication. We create a temporal graph G_i that consists of the current diagnosis, the drug prescription to the current diagnosis, and the medical events in the patient EHR prior to the current diagnosis. We then formulate a binary graph classification problem on the resulting temporal graph by considering the following dual optimization problem for a Support Vector Machine (SVM):

$$\begin{aligned}
&\underset{\alpha}{\text{maximize}} \ \sum_i \alpha_i - \frac{1}{2}\sum_{j,k} \alpha_j \alpha_k y_j y_k K(G_j, G_k) && \text{(1a)}\\
&\text{subject to} \ \ 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N && \text{(1b)}\\
&\qquad\qquad\quad \sum_i \alpha_i y_i = 0, \quad i = 1, \ldots, N && \text{(1c)}
\end{aligned}$$

where K is a positive definite graph kernel on input graphs G_j, G_k. C is a regularization parameter, and b is a bias term. Given the graph G_i, the bias term b can be computed by

$$b = y_i - \sum_{j=1}^{N} \alpha_j y_j K(G_i, G_j) \tag{2}$$

and the decision function is defined as:

$$f(G) = \sum_{i=1}^{N} \alpha_i y_i K(G_i, G) + b \tag{3}$$

Next, we introduce our developed graph kernel and differentiate it from existing models. The predictive framework is depicted in Figure 3.

III. EHR-based Graph Kernel

A. Temporal Topological Kernel

To provide an effective treatment, considering the temporal relationships between medical events is necessary. Existing graph kernels are discussed in [11]; primarily they focus on counting similar substructures, whereas the temporal relationships with each other are not considered. We, however, present a temporal topological kernel K_{tp}. Specifically, we transform input graphs to shortest path graphs by identical operations as discussed in [12] and define the kernel function as follows:

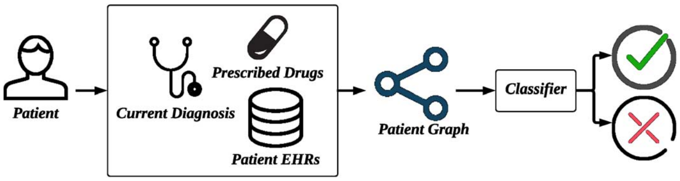


Fig. 3: Predictive framework

Definition 2 (*Temporal topological kernel*)**.** *Let $g_1 = (V_1, E_1)$ and $g_2 = (V_2, E_2)$ denote the shortest path graph of P_{g_1} and P_{g_2} by using the transformation discussed above, we define Temporal topological kernel K_{tp} as:*

$$K_{tp}(g_1, g_2) = \sum_{e_1 \in E_1, e_2 \in E_2} K_{ts}(e_1, e_2) \tag{4}$$

where K_{ts} is a temporal substructure kernel defined on edges $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ which calculates temporal similarity on substructures that connect to nodes in e_1, e_2.

The intuition of K_{tp} is based on calculating the similarity among temporal ordering on substructures (e.g., node neighborhoods) by K_{ts} between input graphs, recursively. If two graphs are similar, their temporal order for node neighborhood structures are similar. That is, for a given pair of nodes v_1, v_2 from two similar graphs g_1, g_2, the time difference from other nodes u_i, u_j in g_1, g_2 to v_1, v_2 where u_i, u_j lie in the subtrees that connect to v_1, v_2 must be similar.

Definition 3 (*Temporal substructure kernel*)**.** *Given a pair of edge $e_1 = (u_1, v_1), e_2 = (u_2, v_2)$, their associated edge weight function w_1, w_2 of g_1, g_2, and set of neighbor nodes N_1, N_2 of u_1, u_2, we define temporal substructure kernel K_{ts} as:*

$$\begin{aligned}
K_{ts}(e_1, e_2) = &\sum_{\substack{e_i=(n_i,u_1)\in E_1, n_i \in N_1 \\ e_j=(n_j,u_2)\in E_2, n_j \in N_2}} K_{ts}(e_i, e_j) + \\
&K_{time}(w_1(e_1), w_2(e_2)) \times \\
&K_{node}(u_1, u_2) \times K_{node}(v_1, v_2)
\end{aligned} \tag{5}$$

and base case definition for recursion part in Equation 5 when u_1 or u_2 is the root node:

$$\begin{aligned}
K_{ts}(e_1, e_2) = &K_{time}(w_1(e_1), w_2(e_2)) \times \\
&K_{node}(u_1, u_2) \times K_{node}(v_1, v_2),
\end{aligned} \tag{6}$$

where K_{time} is defined as:

$$K_{time}(w_1(e_1), w_2(e_2)) = e^{-1 \times |w_1(e_1) - w_2(e_2)|}, \tag{7}$$

and K_{node} is defined as:

$$K_{node}(u_1, u_2) = \begin{cases} 1, & \textit{if } label(u_1) = label(u_2) \\ 0, & \textit{otherwise} \end{cases} \tag{8}$$

[1] We follow the similar setting as in [10]

B. Kernel Validity

To show K_{ts} is a valid kernel, we must prove that it is positive definite.

Proof. K_{node} is a dirac delta function which is proven to be positive definite in [13]. K_{time} is positive definite since the transformation of exponential function is positive definite [14]. It is known that positive definiteness is closed under positive scalar linear combination and multiplication on positive definite kernels, and it is hold in base case definition in Equation 6. As a result, K_{ts} is positive definite, and K_{tp} is therefore positive definite. □

IV. Results and Discussion

A. Evaluation Data

We evaluate our approach using the Taiwanese National Health Insurance Research Database (NHIRD)[2] provided by the National Health Insurance Administration and the Ministry of Health and Welfare. Our collection contains a longer than 20-year complete medical history for one-million randomly sampled patients. Each record contains an ICD9-CM[3] code to indicate the disease diagnosed, and an ATC[4] code to indicate the drug prescription. Georgetown University Institutional Review Board (IRB) approvals were obtained.

We selected four diseases to study, ones whose treatments primarily rely on drug prescription (e.g., antibiotics). The objective is to predict whether the given prescription for disease diagnosis is a success or a failure as a binary graph classification task. We set an observation window for 1 month after the drug is prescribed with a 2 month medical history included prior to diagnosis for each patient. The statistics of these selected diseases are listed in Table I.

TABLE I: Disease Data Statistic

Disease	Pneumonia	Acute Otitis media	Acute cystitis	Urinary tract infection
ICD-9 Codes	481.*,482.* 483.*	381.0	595.0	599.0
# of patient	37,677	40,008	113,513	279,645
# of failure	12,439	14,999	35,728	94,105
# of success	25,238	25,009	77,785	185,540

B. Baseline Approaches

We compare our approach to the following baseline models:

- **Traditional models**. Logistic Regression (LR), Linear Support Vector Machine (L-SVM), and Random Forest (RF) with a bag of word model are used to build frequency vectors on all diagnoses and drug codes for each medical record. We also use word2vec [15] to embed all medical records into 256[5] dimensional vectors.
- **Deep learning models**. We use Med2Vec [16], Deepr [17], and recurrent neural network with long short term memory (RNN-LSTM) as our comparative deep learning models. For Med2Vec and deepr, we follow the same setting as in [16] and [17]. For RNN-LSTM, we embed each medical record in a low dimension vector as in [18] and feed into LSTM with fully connected softmax layer as prediction output.

C. Evaluation Metrics and Settings

We use accuracy (ACC), F1-score (F1), and the area under the receiver operating characteristic curve (AUROC) as our evaluation metrics. Datasets for each disease are divided into training, validation, and testing in an 80:10:10 ratio. All parameters in all models are fine tuned via 10-fold cross validation on the validation set. We repeat all experiments 10 times and report the average performance scores; we test them statistically by using pairwise t-test to validate the statistical significance of our proposed method. The p-value is set to 0.05 to reject the null-hypothesis, namely our solution insignificantly differs from previous efforts.

D. Evaluation Results

Results shown in Table II illustrate that our proposed method outperforms all baseline approaches. Surprisingly, the deep learning model fails to yield better performance than traditional methods. We surmise the reason may be due to characteristic differences in the EHR dataset. NHIRD is a nation-wide EHR database and one expects a high variance (difficulty in training) of patient features and demographics.

E. Classification Result Interpretation

A known issue for deep learning models is interpretability; that is, classification performance is often high but results are hard to interpret. To investigate the interpretability of our proposed model, we refer to the decision function in Equation 3. There, the class label for input graph G is determined by a linear combination of all training labels y_i and kernel values between G and all training examples G_i with associated dual coefficient α_i which can be interpreted as the overall importance of G_i during the label computation. We illustrate via an example, the training examples with top α values for failed prescription after training the Pneumonia dataset in Figure 4. Top influential G_i learned by SVM gives us an insight on how treatments look like in success or failure cases. Also, kernel values $K(G, G_i)$ define the similarity between G and G_i, providing clues to medical practitioners as to why the given treatment is a success or failure by looking into similar cases. All training examples are easily understood by medical practitioners due to their graphical representation.

V. Conclusion

Patient EHRs contain sparse, temporal, and heterogeneous data, complicating predictive model development. Our graph kernel based approach works directly on graph representations of patient EHR and has the following advantages:

1) We incorporate all medical event information including temporal relationships in a compact format.

[2] https://nhird.nhri.org.tw/en/

[3] The International Classification of Diseases, 9th Revision, Clinical Modification

[4] Anatomical Therapeutic Chemical

[5] We found this achieves the best performance after fine tuning

TABLE II: Performance comparison

	Pneumonia			Acute otitis media			Acute cystitis			Urinary tract infection		
Model	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
K-SVM	**0.6969**	**0.6671**	**0.6463**	**0.6860**	**0.6605**	**0.6208**	**0.7200**	**0.7172**	**0.6281**	**0.7402**	**0.6671**	**0.6463**
Med2Vec	0.5359	0.5393	0.4583	0.5644	0.5649	0.5929	0.5831	0.5865	0.5260	0.5725	0.5724	0.5612
Deepr	0.5673	0.6202	0.4137	0.5917	0.5916	0.5858	0.5923	0.5989	0.5309	0.6073	0.6133	0.5434
LSTM	0.5679	0.6035	0.4458	0.5607	0.6598	0.1862	0.5678	0.5706	0.5112	0.6448	0.6519	0.5779
LR-WB	0.6486	0.6023	0.5328	0.6152	0.5839	0.5791	0.5939	0.5720	0.4991	0.6486	0.6023	0.5328
SVM-WB	0.6463	0.6120	0.5369	0.6209	0.5955	0.5035	0.5950	0.6241	0.2809	0.6463	0.6120	0.5369
RF-WB	0.6603	0.6134	0.5874	0.6190	0.5772	0.5344	0.5887	0.6069	0.4372	0.6603	0.6134	0.5874
LR-BoW	0.6349	0.5950	0.6018	0.6103	0.5865	0.5850	0.6209	0.5806	0.5351	0.6349	0.5950	0.6018
SVM-BoW	0.6341	0.6013	0.5447	0.6008	0.5800	0.4249	0.6160	0.6242	0.3254	0.6341	0.6013	0.5447
RF-BoW	0.6500	0.6123	0.6258	0.6285	0.6146	0.6142	0.6083	0.6159	0.5387	0.6500	0.6125	0.6258

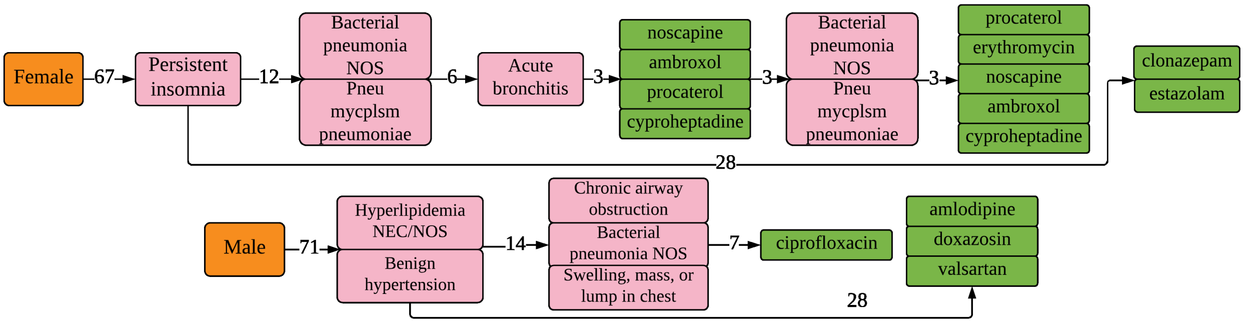


Fig. 4: Top two α patient graphs

2) We eliminate the need of electronic phenotyping by predicting directly from the graph structure.
3) Given the interpretability of the temporal graph representation of patient EHR, our classification results are self-explanatory and are easily understood by both the medical practitioner and patient.

The classification performance of our drug prescription success/failure surpasses all evaluated approaches; our approach is now in preliminary clinical use for a select set of common diseases.

Acknowledgment

We thank our collaboration partners at Fu Jen Catholic University to access to the NHIRD.

References

[1] J. Aronson, "Medication errors: what they are, how they happen, and how to avoid them." *QJM : monthly journal of the Association of Physicians*, vol. 102, no. 8, pp. 513–521, 2009.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, p. 395, 2012.

[3] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records: A survey (https://dl.acm.org/citation.cfm?id=3127881)," *ACM Computing Surveys*, vol. 50, 02 2017.

[4] H. Alphs-Jackson, J. Cashy, O. Frieder, and A. J. Schaeffer, "Data mining derived treatment algorithms from the electronic medical record improve theoretical empirical therapy for outpatient urinary tract infections," *The Journal of urology*, vol. 186, no. 6, pp. 2257–2262, 2011.

[5] H. C. Koh, G. Tan *et al.*, "Data mining applications in healthcare," *Journal of healthcare information management*, vol. 19, no. 2, p. 65, 2011.

[6] M. K. Obenshain, "Application of data mining techniques to healthcare data," *Infection Control & Hospital Epidemiology*, vol. 25, no. 8, pp. 690–695, 2004.

[7] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 2, no. 02, pp. 250–255, 2010.

[8] J. Pathak, A. N. Kho, and J. C. Denny, "Electronic health records-driven phenotyping: challenges, recent advances, and perspectives," 2013.

[9] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[10] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 705–714.

[11] S. Ghosh, N. Das, T. Gonçalves, P. Quaresma, and M. Kundu, "The journey of graph kernels through two decades," *Computer Science Review*, vol. 27, pp. 88–111, 2018.

[12] K. M. Borgwardt and H. P. Kriegel, "Shortest-path kernels on graphs," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 8 pp.–.

[13] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[14] D.-C. Chang, O. Frieder, and H.-R. Yao, "On bochner's theorem and its application to graph kernels," *Journal of Nonlinear and Convex Analysis*, vol. 19, no. 12, 2019.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[16] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1495–1504.

[17] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 22–30, Jan 2017.

[18] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1910–1919.