Time-aware Embeddings of Clinical Data using a Knowledge Graph

Karthik Soman, Charlotte A. Nelson, Gabriel Cerono and Sergio E. Baranzini*
Weill Institute for Neuroscience, Department of Neurology, University of California San Francisco, San Francisco, California, United States of America
**Email: sergio.baranzini@ucsf.edu*

Meaningful representations of clinical data using embedding vectors is a pivotal step to invoke any machine learning (ML) algorithm for data inference. In this article, we propose a time-aware embedding approach of electronic health records onto a biomedical knowledge graph for creating machine readable patient representations. This approach not only captures the temporal dynamics of patient clinical trajectories, but also enriches it with additional biological information from the knowledge graph. To gauge the predictivity of this approach, we propose an ML pipeline called *TANDEM* (Temporal and Non-temporal Dynamics Embedded Model) and apply it on the early detection of Parkinson's disease. TANDEM results in a classification AUC score of 0.85 on unseen test dataset. These predictions are further explained by providing a biological insight using the knowledge graph. Taken together, we show that temporal embeddings of clinical data could be a meaningful predictive representation for downstream ML pipelines in clinical decision-making.

Keywords: temporal embedding; knowledge graph; electronic health record; machine learning.

1. Introduction

Clinical data comes from multiple modalities and encompasses heterogeneous information related to patient health. Electronic health records (EHR), a structured clinical data, encompasses different health variables of a patient such as diagnosis, medications, lab tests, clinical visit encounters, etc. Machine learning (ML) algorithms, owing to their ability to decipher patterns in large scale heterogeneous data, could be used to tap the invaluable information embedded in the EHR data for insightful clinical predictions[1]. There have been previous efforts along this line such as clinical concept embeddings, disease phenotyping/diagnosis and EHR de-identification[2,3].

Patient representation learning is an important aspect for running ML pipelines. Such representations are generally lower-dimensional latent vectors with predictive value for patient's health status[3]. This predictive value is further capitalized for downstream clinical predictive modeling. There have been predictive analyses that utilized the longitudinal aspect of EHR data such as measurements of lab tests[4], temporal history of diagnosis, medication and procedure codes[5] and long term temporal dependencies in patient medical records[6]. These modeling approaches utilized sequence models like Recurrent Neural Network (RNN) to capture the temporal dynamics in the longitudinal EHR data and embed patients' health state trajectories as internal latent

representation[2]. Although such approaches have proven to be useful in predictive medicine, the abstract nature of patient representation affects their clinical interpretability.

There have been interpretable modeling approaches using knowledge networks for clinically relevant problems[7–9]. The major aspect of such an approach is the existence of biologically relevant edges in a knowledge network that could connect entities from molecular to phenotypic level[10]. Such a network level approach helps to understand the relationship between disease and underlying molecular/genetic pathways, thereby providing an insightful knowledge that transcends multiple levels of biology. There have been recent efforts to integrate EHR data with knowledge networks for a network level concept embedding and disease prediction[11,12], but without considering the longitudinal aspect of clinical data.

In this paper, we try to achieve the best of both worlds, i.e. embedding longitudinal EHR data on a biomedical knowledge graph to capture the temporal dynamics of patient clinical trajectory at a network level. We hypothesize that such an embedding approach could represent the health status of a patient with enriched biological information at a higher temporal resolution which could ultimately improve the predictability of disease diagnosis. With this objective, we introduce the concept of knowledge graph based temporal embeddings, and use them in an explainable modeling approach called *TANDEM* for the diagnosis of chronic diseases, in this study - Parkinson's Disease (PD).

2. Methods

2.1. *Scalable Precision medicine Open Knowledge Engine (SPOKE)*

SPOKE is a heterogeneous biomedical knowledge network with more than 3 million nodes of 16 types (such as genes, proteins, disease, symptoms etc.) and more than 16 million edges of 32 types between those nodes[11]. SPOKE integrates over 40 publicly available databases that are biologically relevant (such as GWAS, DOID, Uniprot, ChEMBL, DrugBank, SIDER, MESH). Graphical user interface of SPOKE network is made publicly available (https://spoke.rbvi.ucsf.edu/). In this study, we utilized the biological associations present in this large scale network to create meaningful patient representations for downstream ML analysis.

2.2. *Creating temporal embeddings of patients*

In the previous study[11], SPOKE knowledge graph was connected to EHR data using Observational Medical Outcomes Partnership (OMOP) common data model and Unified Medical Language System's (UMLS) Metathesaurus mappings. Then an embedding vector, called Propagate SPOKE Entry Vectors (PSEVs), for a clinical concept was created by using a modified version of topic-sensitive PageRank[11,13]. PSEVs can be created for any code in the EHR that has been recorded for

a cohort of patients (e.g. Parkinson's Disease). A PSEV vector of a clinical concept stores how important each node in SPOKE is for that particular concept, which hence gives a network level representation of an EHR concept.

In this study, to produce temporal embeddings for an individual patient, PSEVs corresponding to the EHR codes (taken from the de-identified EHR database of UCSF medical center) from a specified time range (frame width = 1 year) in a patient's timeline were aggregated and normalized to create a patient specific embedding vector (Figure 1A). Stacking such embedding vectors from each time frame gave rise to a two-dimensional array whose rows represented time and columns represented SPOKE nodes (Figure 1A). We named this as temporal SPOKEsig since it holds the temporal dynamics of SPOKE nodes as a function of a patient's clinical data. We also created non-temporal SPOKEsig i.e. patient embedding without considering the temporal order of EHR concepts, hence generating a one-dimensional array of vector (i.e. no time axis, Figure 1A).

In this study we created embeddings for two patient cohorts (i.e. PD and non-PD). Patients were included in the PD cohorts if a PD diagnosis code was present in their EHR *diagnosis* table. We selected only those patients with enough temporal history (i.e. having clinical data in more than one year of time frame in their timeline). In the interest of analyzing disease dynamics and classifying patients into PD or non-PD classes before the clinical diagnosis, we created embeddings starting from one year before their actual clinical diagnosis and going further back in time (i.e. early detection of PD, Figure 1A). We created two sets of such embedding vectors for each cohort where one set was used for feature selection and training the downstream ML model and the other set was used to evaluate the performance of the model.

Considering *M* number of nodes in SPOKE, a patient cohort with *N* patients can be represented by a two-dimensional array of size NxM using the non-temporal approach (Figure 1B). The same patient cohort can be represented by a three-dimensional array of size NxTxM using the temporal approach, where *T* denotes the time axis of the embedding vector (Figure 1B). *T* corresponds to the largest visiting time of a patient in the cohort of interest, in this study the PD cohort.

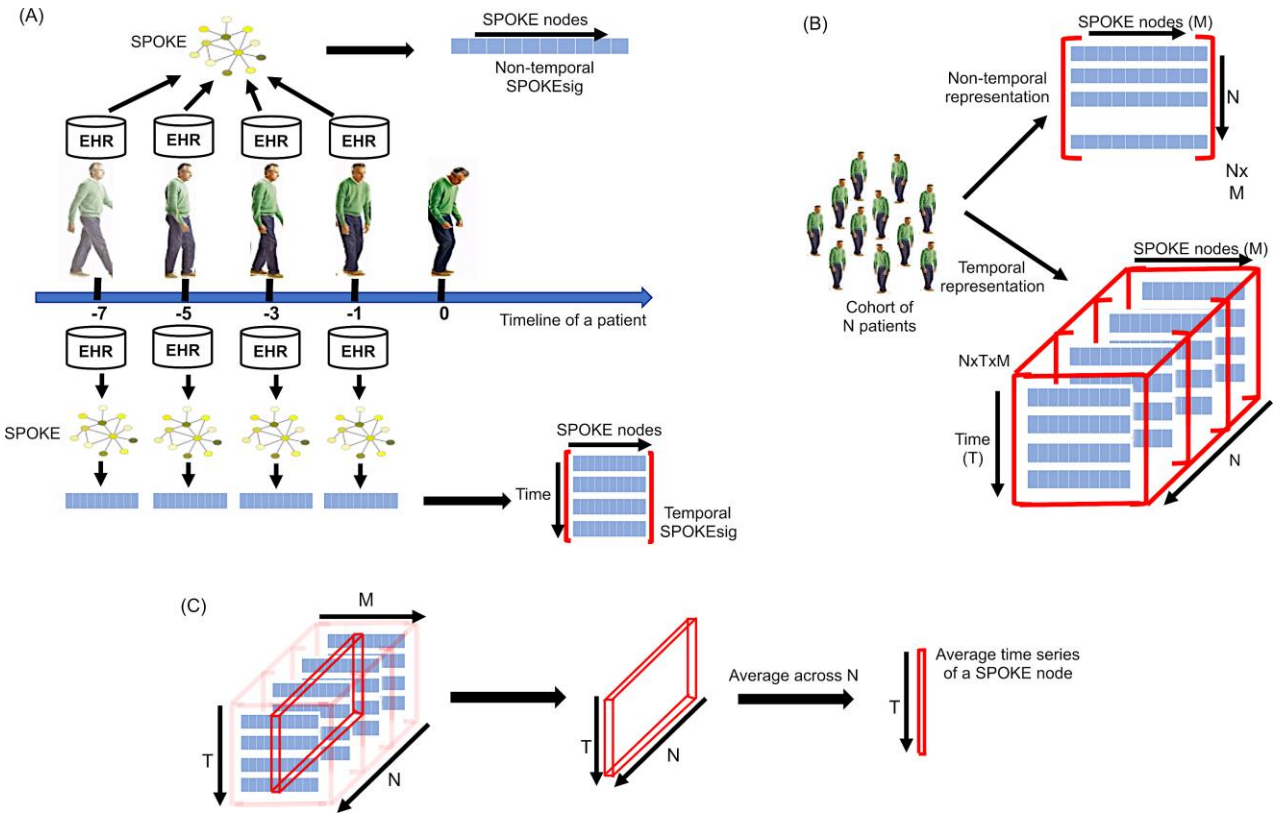

Fig. 1. (A) shows the schematic for the generation of temporal and non-temporal patient embeddings. The middle arrow shows the patient timeline where 0 represents the time when the diagnosis was made for the first time. -1 represents one year before the clinical diagnosis and a similar explanation holds for other tick labels shown on the timeline. (B) shows the way in which a patient cohort can be represented using non-temporal and temporal SPOKEsig approaches. (C) Schematic for the computation of the average time series of a SPOKE node. Starting from the left, it shows the time series of a SPOKE node as a strip in the three-dimensional array of temporal SPOKEsig. Averaging that strip across the depth (i.e. number of patient samples *N*) gives the average time series of that SPOKE node.

2.3. *Knowledge graph time series and feature selection*

For any useful data inference using an ML algorithm, the first step is to select predictive features from the embedding vectors that are used as training data for downstream ML pipeline. In a three-dimensional temporal SPOKEsig, each feature is a time series corresponding to nodes in the SPOKE knowledge graph. To evaluate how these nodes evolve in time with respect to disease progression, we first computed the average time series of each SPOKE node across all patients in the training data of each cohort (Figure 1C).

We then applied a non-parametric statistical test (*Mann-Kendall Trend Test, MKTT* [14]) on each average time series to identify a trend[15]. Trend can be treated as a feature that gives a measure of how time series evolve. MKTT only tests for linear monotonic trends in a time series[15]. Hence, a

time-series can be classified as an increasing, decreasing or no trend. In addition to the trend type, the test also returns a trend value (slope) present in the time series and a p-value associated with it. Since we are looking at a classification problem, we wanted to retain predictive temporal features that show disparate temporal dynamics between the cohorts. Hence, we selected those features that satisfied at least one of the following three criteria:

1. A node has a trend in one cohort and no trend in the other cohort
2. A node has opposite trends in two cohorts
3. A node has the same trend in two cohorts, then select only if its slope in one cohort is more than double than in the other.

2.4. *Transformation of temporal embeddings of a patient cohort*

After feature selection, the next step is to train an ML classifier to identify if a patient has PD or not (two-class problem). Since temporal embeddings are sequential data (because of the time dimension), state-of-the-art models to learn such data are recurrent neural networks (RNN) like Long short-term memory (LSTM) networks[16], Gated recurrent unit (GRU) networks[17]. However, the patient cohort size used in this study was not large enough to train such deep neural networks with trainable parameters in the order of millions. This situation (less data and more parameters) could lead to data overfitting and that could affect the generalizability of the trained model. In such situations, previous studies have chosen models like random forest (RF) owing to their ensemble architecture [18–20] and we chose the same in our case.

To train a RF classifier, we transformed the temporal SPOKEsig from a three-dimensional array (NxTxM') to a two-dimensional array (NxM') where *N* corresponds to total number of patients in a cohort, *T* represents time and *M'* represents the selected features from an initial *M* features (after feature selection, $M' < M$). To retain the embedded temporal information in the transformed two-dimensional representation, we performed a linear approximation of temporal SPOKEsig by computing the trend value present in each time series of SPOKE nodes across all patients. Figure 2 shows the steps involved in this transformation process.

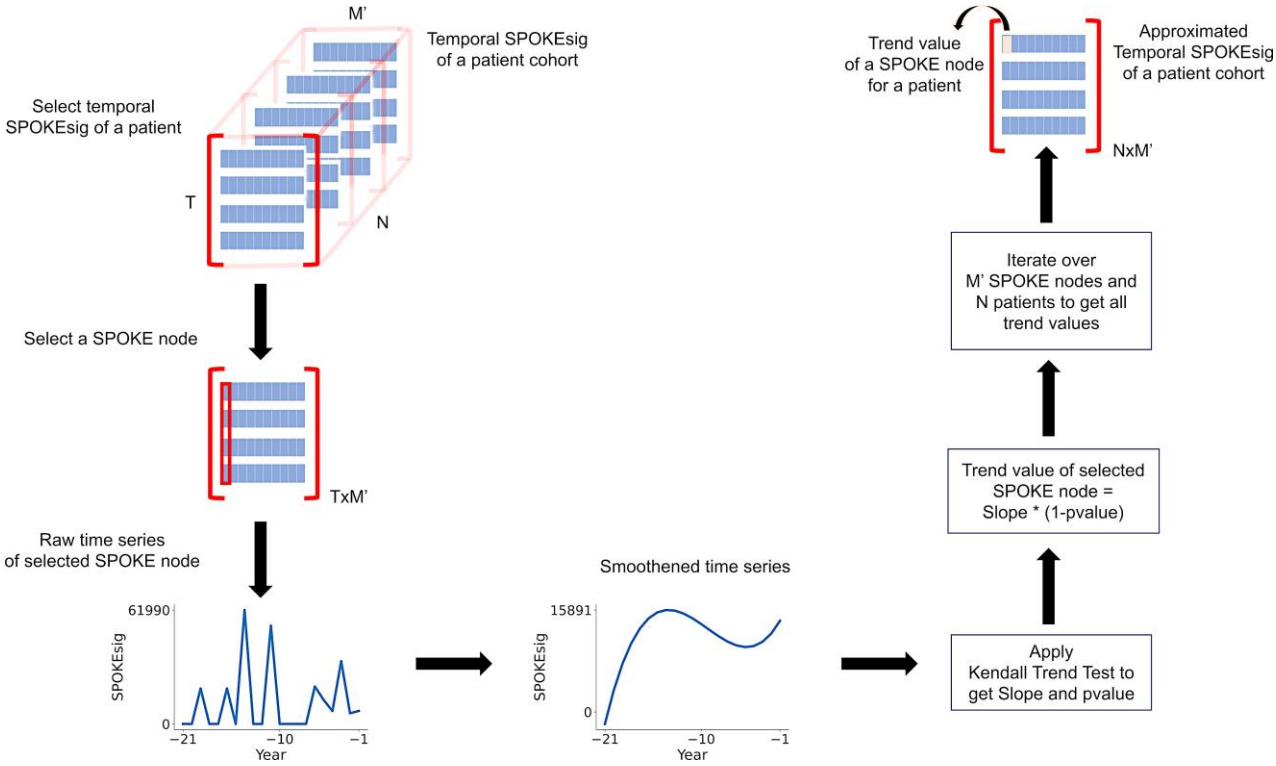

Fig. 2. Steps involved in the linear approximation of three-dimensional temporal SPOKEsig. Following the direction of arrows, it starts with selecting a temporal SPOKEsig of a patient, followed by selecting a time series of a SPOKE node. To prevent any false trend value estimates (because of the zero elements in the series coming from the sporadic hospital visits made by the patient), the raw time series was smoothened using Savitzky-Golay filter (window size = 21 and polynomial order = 3). We then applied Kendall trend test on the smoothed time series to get the trend (slope) and p-value. Final trend value was considered as the estimated slope multiplied by the probability for the presence of trend in that time series (which is 1-p-value). These steps were iterated for all SPOKE nodes across all patients in a cohort to get the approximated temporal SPOKEsig of a patient cohort which is a two-dimensional array.

To compensate for this linear approximation transformation, a second feature selection was done on the transformed two-dimensional array (of training data) such that we selected only those features whose absolute difference in their average slope values between PD and non-PD cohort is greater than a threshold value of 406 (chosen empirically).

2.5. *Temporal and non-temporal dynamics embedded model (TANDEM) for disease classification*

TANDEM includes both temporal and non-temporal embeddings of patients for disease classification. Specifically, we trained two separate RF models, one using approximated temporal SPOKEsig and the other one using non-temporal SPOKEsig. One model evaluated the linear trend

and the other model evaluated the area swept by the time series of SPOKE nodes. Hence, both classifiers looked at two fundamentally different aspects of the time series data. Each model was trained using their respective training data. Since there existed less PD samples than non-PD samples in the data, training data was imbalanced. Hence, while training the classifiers, proper weights were assigned to patient samples in the training data based on their class distribution (hence more weightage was given to PD samples while training). Individual prediction scores of these two models were further normalized by their percentile scores. Finally, a logistic classifier was trained (using binary cross-entropy as the loss function) using the normalized prediction scores from temporal and non-temporal RF models to compute the final disease prediction score.

Classification performance was evaluated using an unseen test dataset. Model performance was quantified by computing the Area Under the Curve (AUC) of Receiver Operator Characteristic (ROC) curve. Bootstrap analysis was done by randomly sampling prediction scores (corresponding to both classes) with replacement and then computing AUC score for that sample. This process was repeated for 1000 times which generated a distribution of AUC scores for the model. In addition to AUC, we also computed F1 score and Average Precision score of each model for comparison.

3. Results

3.1. *Patient temporal embedding*

We selected a total of 283 PD and 74,059 non-PD patients respectively as training dataset. We had a separate test dataset (for model evaluation) with 1994 patients (17 PD and 1977 non-PD). EHR history of both cohorts spanned a maximum of 21 years from one year prior to the clinical diagnosis. There were a total of 389,297 SPOKE nodes in the embedding vector (i.e. dimension of the vector).

3.2. *Feature selection and PCA visualization*

Following the feature selection method using the MKT test (mentioned in the Methods section), we were able to reduce the features of temporal SPOKEsig from 389,297 to 109,256 (28.1% of initial features). Next, temporal dynamics of the selected and non-selected features were visualized by projecting them onto the first three principal components (Figure 3). A second feature selection on the linear approximated temporal SPOKEsigs (see Methods) reduced features from 109256 to 42012 (38.5% of initial features).

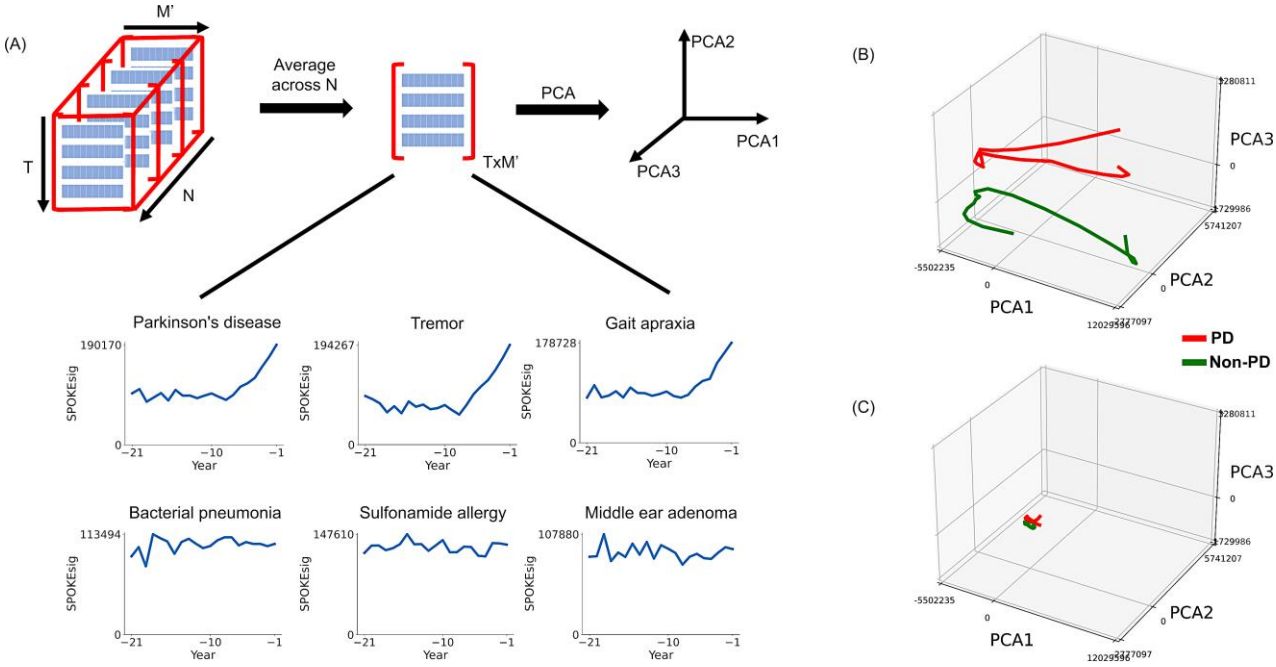

Fig. 3. (A) shows the steps in applying PCA on feature selected temporal SPOKEsig. The insight shows six examples of SPOKE node time series corresponding to PD cohort (averaged across patient samples). Upper row corresponds to SPOKE nodes that are closely related to PD and lower row corresponds to nodes that are less related to PD. (B) Temporal trajectory of selected features in PCA space. Two distinct trajectories are evident in the PCA space and the color code is shown in the legend. (C) Temporal trajectory of non-selected features in PCA space. For the sake of visual comparison, we included only those non-selected features that showed no trend in both PD and non-PD cohorts and had a p value > 0.5.

3.3. *Disease classification using TANDEM architecture*

AUC bootstrap analysis on the test data showed that temporal model showed higher performance than the non-temporal model (Figure 4A, Table 1, p-value=$4.5*10^{-52}$, N=1000, Mann Whitney U test). However, TANDEM architecture outperformed these two models significantly (Figure 4A, Table 1). We also compared these models using their F1-score and average precision score on the test data and it showed that in both cases TANDEM model held the highest score (Figure 4B-C).

For the explainability of TANDEM predictions from a biological perspective, we estimated the temporal slope (rate of growth) of PD related gene nodes' time series (i.e. gene nodes connected to PD node in SPOKE) for all patients that were correctly predicted by the TANDEM model. 13 PD (out of 17) and 1659 non-PD (out of 1977) test patients were correctly predicted by the TANDEM model. PD genes showed higher rate of temporal evolution in these PD patient group than the non-PD group (p-value = $1.4*10^{-06}$, N = 141, Mann Whitney U test, Figure 4D). We also showed the temporal evolution of PD-gene network for a single patient across three discrete time points in a patient's timeline (Figure 4E for PD patient and Figure 4F for non-PD patient).

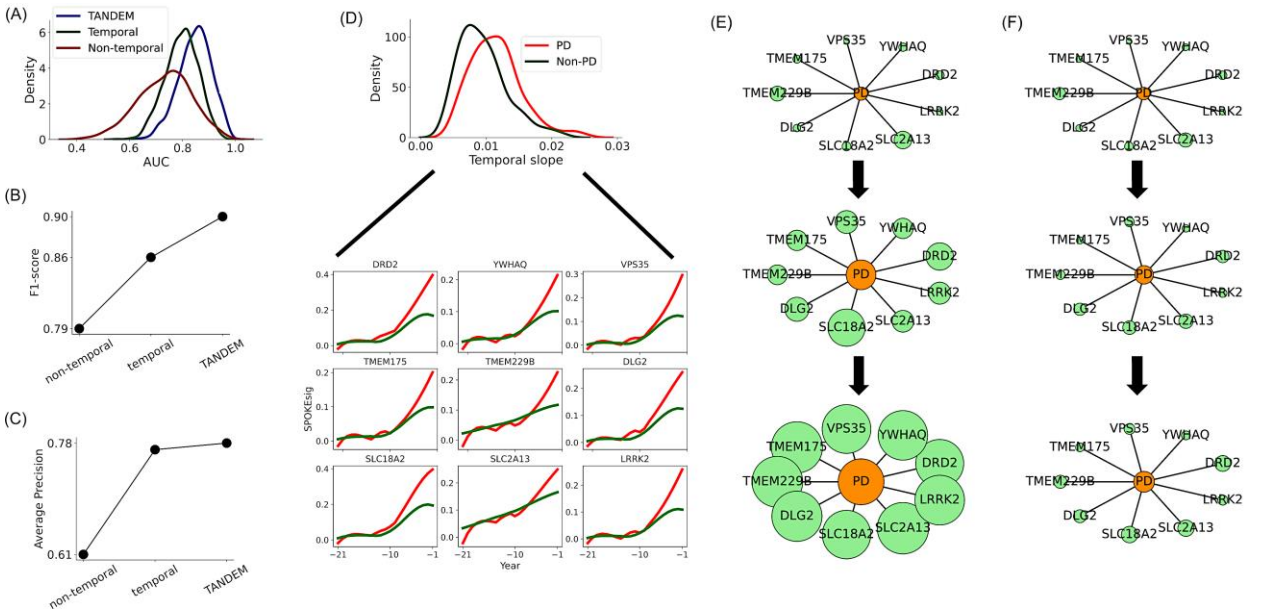

Fig. 4. (A) AUC distributions of three models in PD classification. (B)-(C) F1-score and Average Precision score of three models respectively (D) Distribution of temporal slope of PD related genes averaged across test patients correctly predicted by TANDEM. Insight shows the average time series of 9 PD related genes for PD (red) and non-PD (green) cohorts from the above distribution. (E)-(F) show the temporal evolution of PD-gene network for a PD patient (E) and a non-PD patient (F) across -15 (top), -4 (middle) and -1 year (bottom) before their clinical diagnosis. Green color nodes represent genes and the orange color node represents disease (PD). Size of a node at a specific time is proportional to the relevance of that node for an individual patient in that time.

Table 1. Comparison of model performances

Model	AUC ($\mu \pm \sigma$)	95% CI	Comparison with TANDEM (p-value, Mann Whitney U test, N = 1000)
Temporal	0.8± 0.06	(0.67, 0.91)	$3.1*10^{-64}$
Non-temporal	0.73±0.1	(0.52, 0.92)	$4.5*10^{-145}$
TANDEM	0.85±0.06	(0.71, 0.96)	-

4. Discussion

If we consider clinical events of a patient in the order in which they occurred, they naturally form a time series. By embedding this longitudinal EHR data on a knowledge network, we tried to achieve a network level interpretation of the temporal dynamics of disease (in this case PD). This approach could possibly bridge the two EHR modeling approaches i.e. knowledge network approach[10] and longitudinal data approach[2].

TANDEM model underlines the complementary nature of temporal and non-temporal features of clinical data in disease diagnosis. These two aspects of TANDEM worked in tandem and enhanced the overall prediction performance. Since the temporal SPOKEsig enriches a patient's clinical trajectory with additional biological information, this approach could give a biological perspective to the model predictions and thereby making it an explainable approach. For example, there was an increased temporal slope associated with the gene LRRK2 among PD patients correctly predicted by the model. There have been previous studies that pointed out the criticality of mutations in the LRRK2 gene and PD pathogenesis, thus making it a predominant genetic risk factor for PD[21,22]. This followed by the visualization of temporal evolution of PD-gene network at individual patient level brings an intuitive biological insight into the model's prediction. As a future work, we plan to apply this modeling architecture to other complex diseases to test its generalizability.

A major challenge in this study was the mapping of clinical data to SPOKE graph for creating embedding vectors. Not all EHR variables map to SPOKE nodes and hence that transformation was lossy. However, additional biological information from SPOKE knowledge graph could be considered as a compensatory factor for this loss. Another challenge is the limitation of patient data. Since this study relied on the temporal history of EHR data, we had to drop patients with fewer temporal information to analyze (~20% patients were dropped). This could be a bottleneck for a data driven pipeline. Lastly, linear approximation of temporal SPOKEsig could have compromised its predictive power. Hence, as a future work, we plan to use the three-dimensional temporal SPOKEsig in its entirety for disease prediction using deep learning sequence models.

Availability of Code and Data

We have made available patient graph representations and the python code for TANDEM in the github repository (https://github.com/BaranziniLab/TANDEM).

Acknowledgments

The development of SPOKE and its applications are being funded by grants from the National Science Foundation (NSF_2033569), NIH/NCATS (NIH_NOA_1OT2TR003450), and the UCSF Marcus Program in Precision Medicine Innovation. SEB holds the Heidrich Family and Friends Endowed Chair of Neurology at UCSF. SEB holds the Distinguished Professorship in Neurology I at UCSF.

References

1. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. https://doi.org/10.1109/JBHI.2017.2767063.

2. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* vol. 25 1419–1428 Preprint at https://doi.org/10.1093/jamia/ocy068 (2018).

3. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **6**, 26094 (2016).

4. Razavian, N., Marcus, J. & Sontag, D. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. (2016) doi:10.48550/arXiv.1608.00647.

5. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. (2015) doi:10.48550/arXiv.1511.05942.

6. Pham, T., Tran, T., Phung, D. & Venkatesh, S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. (2016) doi:10.48550/arXiv.1602.00357.

7. Bean, D. M. *et al.* Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.* **7**, 1–11 (2017).

8. Himmelstein, D. S. & Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput. Biol.* **11**, e1004259 (2015).

9. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs for repurposing. (2017) doi:10.7554/eLife.26726.

10. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2010).

11. Nelson, C. A., Butte, A. J. & Baranzini, S. E. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nat. Commun.* **10**, 1–10 (2019).

12. Nelson, C. A., Bove, R., Butte, A. J. & Baranzini, S. E. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *J. Am. Med. Inform. Assoc.* **29**, 424–434 (2021).

13. Haveliwala, T. H. Topic-sensitive PageRank. *Proceedings of the eleventh international conference on World Wide Web - WWW '02* Preprint at https://doi.org/10.1145/511446.511513 (2002).

14. Mann, H. B. Nonparametric Tests Against Trend. *Econometrica* vol. 13 245 Preprint at https://doi.org/10.2307/1907187 (1945).

15. Wang, F. *et al.* Re-evaluation of the Power of the Mann-Kendall Test for Detecting Monotonic Trends in Hydrometeorological Time Series. *Front. Earth Sci.* **0**, (2020).

16. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

17. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. (2014) doi:10.48550/arXiv.1412.3555.

18. Dietterich, T. G. Ensemble Methods in Machine Learning. in *Multiple Classifier Systems* 1–15 (Springer Berlin Heidelberg, 2000).

19. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

20. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 1–13 (2006).

21. Tolosa, E., Vila, M., Klein, C. & Rascol, O. LRRK2 in Parkinson disease: challenges of clinical trials. *Nat. Rev. Neurol.* **16**, 97–107 (2020).

22. Dächsel, J. C. & Farrer, M. J. LRRK2 and Parkinson Disease. *Arch. Neurol.* **67**, 542–547 (2010).