Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model

Yuesong Zou (master candidate in computational biology) Supervisor: Prof. Yue Li
yuesong.zou@mail.mcgill.ca
School of Computer Science

Introduction

- **Electronic health records (EHRs)** are systematic collections of longitudinal patient health information, including demographics, diagnosis, medications, etc. Precise analysis of EHRs help us to **understand human diseases** and to **design better healthcare systems**[1].
- However, EHR data are **heterogeneous**, **sparse**, and **noisy**. Deriving **robust** and **reliable** methods to analyze EHRs is **challenging**.

Contributions

To deal with the above challenges, we:

- incorporated existing medical knowledge and built a **multi-modal knowledge graph** of relations within and between diseases and drugs.
- built **GAT-ETM**, a graph-informed end-to-end multi-modal **topic model**, letting the existing medical knowledge guide our model learn informative and interpretable topics.
- applied our model on a mammoth EHR dataset which includes **1,200,000 patients** with up to **20 years** of follow-up, and thus got **100 meaningful disease and drug topics**.
- evaluated GAT-ETM on three tasks: held-out reconstruction, phenotype classification, drug recommendation, showed that compared to existing models, our model gives the most **informative patient embeddings**.

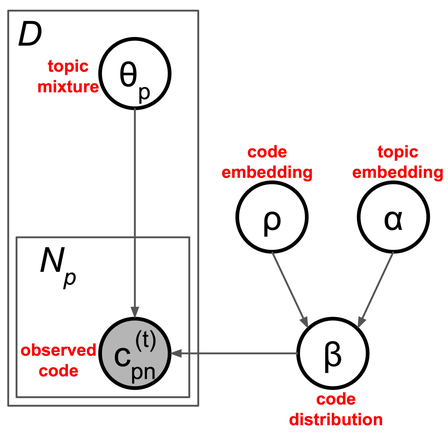
Methods

- **Topic model:**
  - Patients' topic mixture θ is assumed to fits to a logistic-normal distribution:

$$\boldsymbol{\delta}_p \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\theta}_p = \frac{\exp(\boldsymbol{\delta}_p)}{\sum_{k'} \exp(\boldsymbol{\delta}_{pk'})}$$

  - Topic-code distribution β is computed by dot product (considering both cosine similarity and frequency) between topic embeddings and medical code embeddings:

$$\boldsymbol{\beta}_k^{(t)} = \text{softmax}(\boldsymbol{\rho}^{(t)\mathsf{T}} \boldsymbol{\alpha}_{\cdot k}) = \frac{\exp(\boldsymbol{\rho}^{(t)\mathsf{T}} \boldsymbol{\alpha}_{\cdot k})}{\sum_v \exp(\boldsymbol{\rho}_{v\cdot}^{(t)\mathsf{T}} \boldsymbol{\alpha}_{\cdot k})}$$



  - Topic embeddings α is a matrix as learnable model parameters.
  - Medical code embedding β is obtained from a graph attention network (GAT) that extract semantic feature from the constructed medical knowledge graph.
  - finally, each EHR medical code (type t ∈ {disease, drug}) is sampled from:

$$c_{pn}^{(t)} \sim Cat(\boldsymbol{\beta}^{(t)} \boldsymbol{\theta}_p).$$

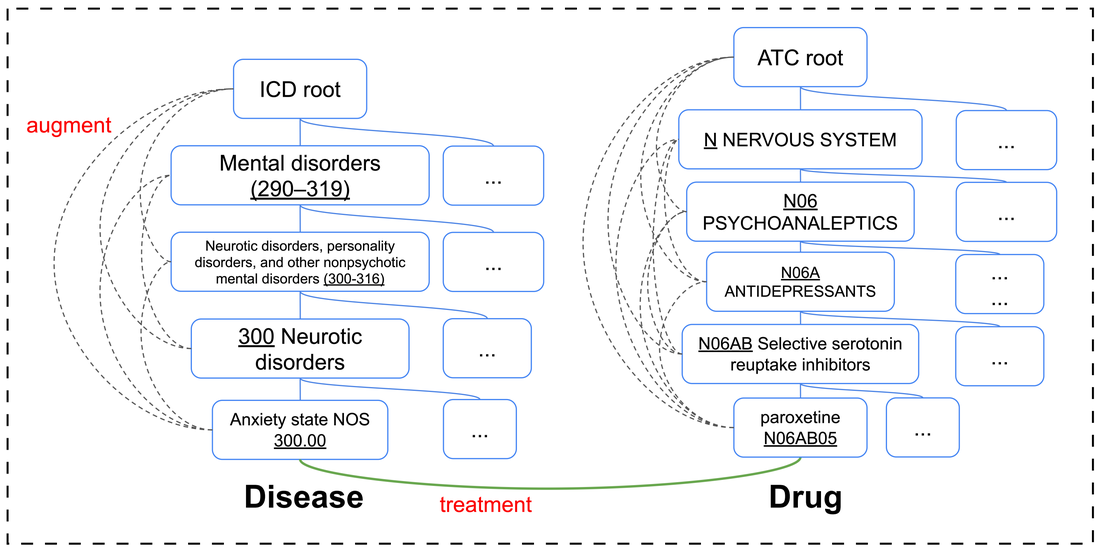
We infer true distribution of patient topic mixture $N(\mu_p, \sigma_p)$ by a two-layer feed forward neural network following a fusion layer that integrates multi-modal data. Evidence lower bound (ELBO) maximizes the likelihood while minimize the Kullback–Leibler divergence between the inferred distribution and the logistic-normal assumption. We maximize ELBO to train model parameters.

- **Medical Knowledge Graph**

Our graph consist of three parts:

- **taxonomy graph of disease**: ICD-9 disease classification system
- **taxonomy graph of drug**: ATC classification system
- **mapping between ICD-9 and ATC (drug - treat - disease)**

In both taxonomy graph, we linked each node with its all ancestors, because ancestral relationship means "including" (same as parental relationship).



Results

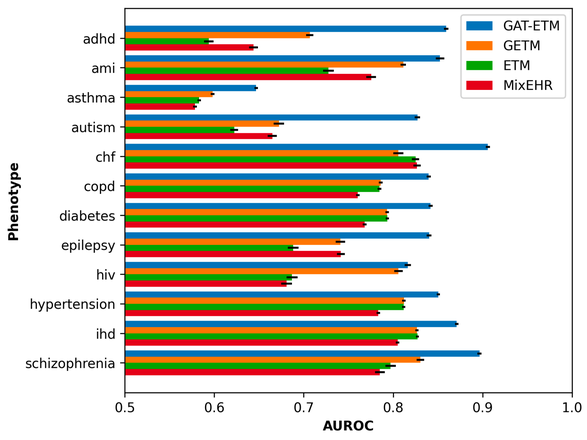
1. EHR Held-out Reconstruction

- compared our model with other topic-based state-of-the-art models: MixEHR[2], ETM[3], GETM[4].
- fed only half records of each patient to let models predict the remained half.
- our model **out-performed them on both reconstruction likelihood** (negative log-likelihood is the lower the better) **and topic quality** (considering topic coherence and divergence, the higher the better).

Model	Negative Log-likelihood	Topic Quality
MixEHR	203.97	0.0673
ETM	198.26	0.0704
GETM	184.32	0.1843
GAT-ETM	**172.69**	**0.1920**

2. Phenotype Classification

- used learned patient topic mixture θ as patient embeddings, trained a logistic regression model to classify diseases based on the 100-dimensional embeddings, and evaluated their ability of classification.
- conduct this classification task on 12 diseases and compared average Area Under the Receiver Operating Characteristic Curve (AUROC).
- the performance of our model has **a significant improvement over others on all 12 diseases** (following bar plot on the left).



Model	prec@5	recall@5	F1-score@5
ETM	0.1823	0.0833	0.1075
GETM	0.2378	0.1101	0.1418
GAT-ETM	**0.2600**	**0.1225**	**0.1569**

Model	Percentile of frequencies				
	20-40	40-60	60-80	80-100	avg.
ETM	0.0039	0.0188	0.0479	0.3847	0.3058
GETM	0.0213	0.0542	0.0934	0.4352	0.3597
GAT-ETM	**0.0345**	**0.0841**	**0.1239**	**0.4583**	**0.3815**

3. Drug Recommendation

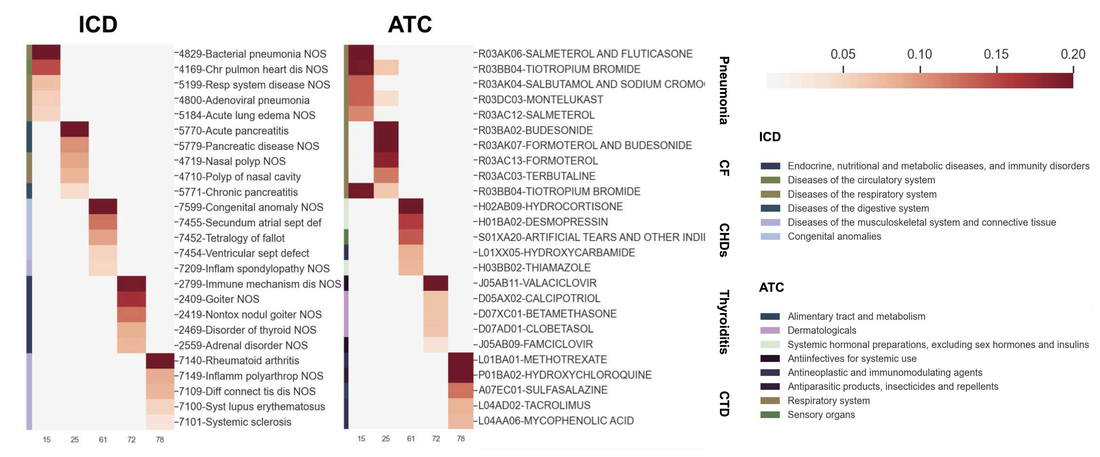
- fed only patients' disease codes to the models, let them impute drug codes.
- compared precision, recall, F1- score based on top-5 prediction given by the models; also compared imputation recall of drugs of different frequencies.
- ours has an **apparent improvement** (above tables on the right).

4. Case Study

Further investigation in several patients of the top worst imputation results showed that the drugs imputed by our model are mostly still close to and have similar usage as the ground truth drugs.

5. Learned Multi-modal Topics

Following is a heatmap visualization of top 5 codes of **five selected disease-drug topics**: pneumonia, cystic fibrosis, coronary heart diseases, thyroiditis, connective tissue disease.



References

1. Hodson R. Digital health[J]. Nature, 2019, 573(7775): S97-S97.
2. Li, Yue, et al. "Inferring multimodal latent topics from electronic health records." Nature communications 11.1 (2020): 1-17.
3. Dieng, Adji B., Francisco JR Ruiz, and David M. Blei. "Topic modeling in embedding spaces." Transactions of the Association for Computational Linguistics 8 (2020): 439-453.
4. Wang, Yuening, et al. "A graph-embedded topic model enables characterization of diverse pain phenotypes among UK Biobank individuals." iScience (2022): 104390.