Machine Learning for EHR (1)

- Electronic Health Record (EHR)
  - Electronic version of a patient medical history
- EHR includes the information of an encounter
  - Basic information of patient
  - Medical history
  - Lab results
  - Medicine record
  - Insurance details
  - …

Machine Learning for EHR (2)

- 1) Learning medical concept representations
  - There are over **3,100,000** concepts in the medical language system
    - Diagnosis code, medicine type, …
  - It will be very helpful if we can embed the concept into a representation
    - We can easily measure the patient similarity
    - Accelerate the clinical information retrieval
    - …
  - We can apply the word embedding/network embedding methods to learn the representation
    - AMIA2016-Learning low-dimensional representations of medical concepts
    - KDD2016-Multi-layer representation learning for medical concepts

Machine Learning for EHR (3)

- 2) Predictive healthcare
  - Predict the disease/medicine
    - Automatic diagnosis can assist the doctor to make decision
    - Online prescriptions
  - Representation learning + classification task
    - NeurIPS2016-Retain: An interpretable predictive model for healthcare using reverse time attention mechanism
    - NeurIPS2018-Mime: Multilevel medical embedding of electronic health records for predictive healthcare

Machine Learning for EHR (4)

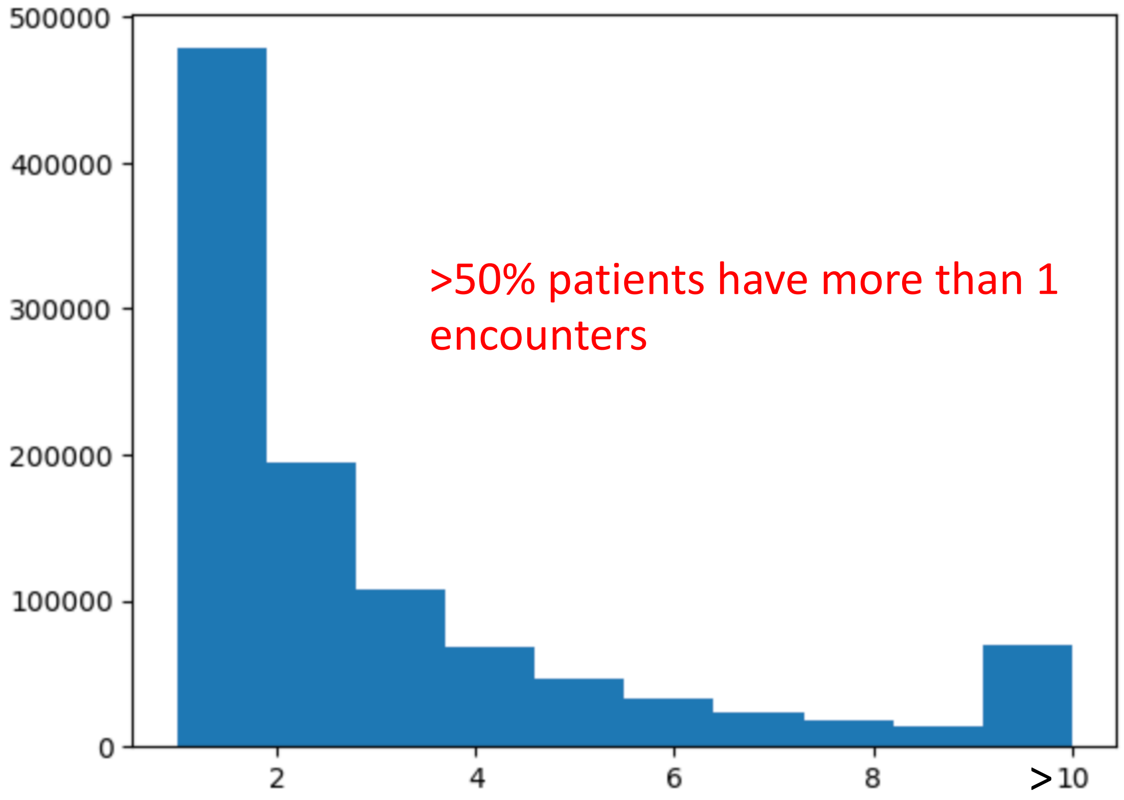
- 3) Anomaly detection
  - Detect the error diagnosis/medication
    - Incorrect data results in harm through suboptimal care delivery
  - Existing data quality assessment frameworks describe a set of dimensions (e.g., completeness, plausibility) evaluated using **basic rules**
    - Limitation: fail to consider patient-specific information and the correlation between different medical concept
  - In this project, we are going to model EHR data with **graph** structure to conduct more accurate prediction
    - Graph modality can help us to reveal the underlying relationship between different instances

Overview of dataset

- 3,640,261 encounters and 1,323 features
- Feature can be divided into 6 groups
  - 9 Basic features
  - 6 Time features
  - 283 past medication features
  - 578 Lab features
  - 104 Medication features
  - 21 Code features
  - 284 diagnosis features

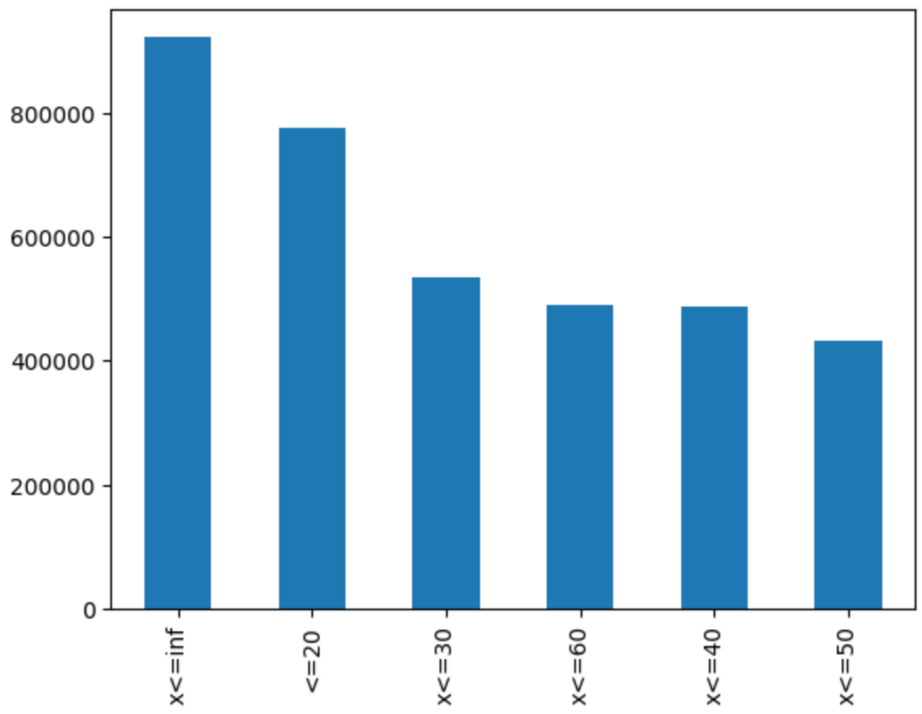
Basic features

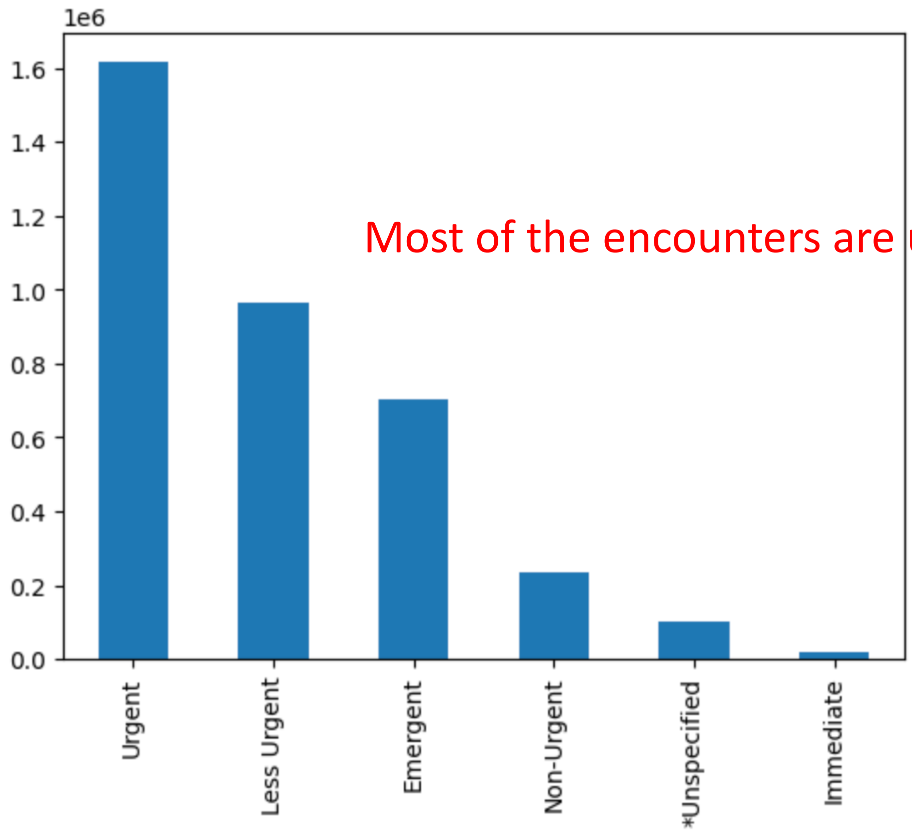
- 9 basic features including the information of patients
  - Age, sex, patient_id, Smokestatus, …
- Number of patient: 1,053,511
  - Number of patients with only one encc
  - Maximum number of encounters: 1,74
- Distribution of number of encounter



Basic features

- Distribution of age and esi (emergence level)

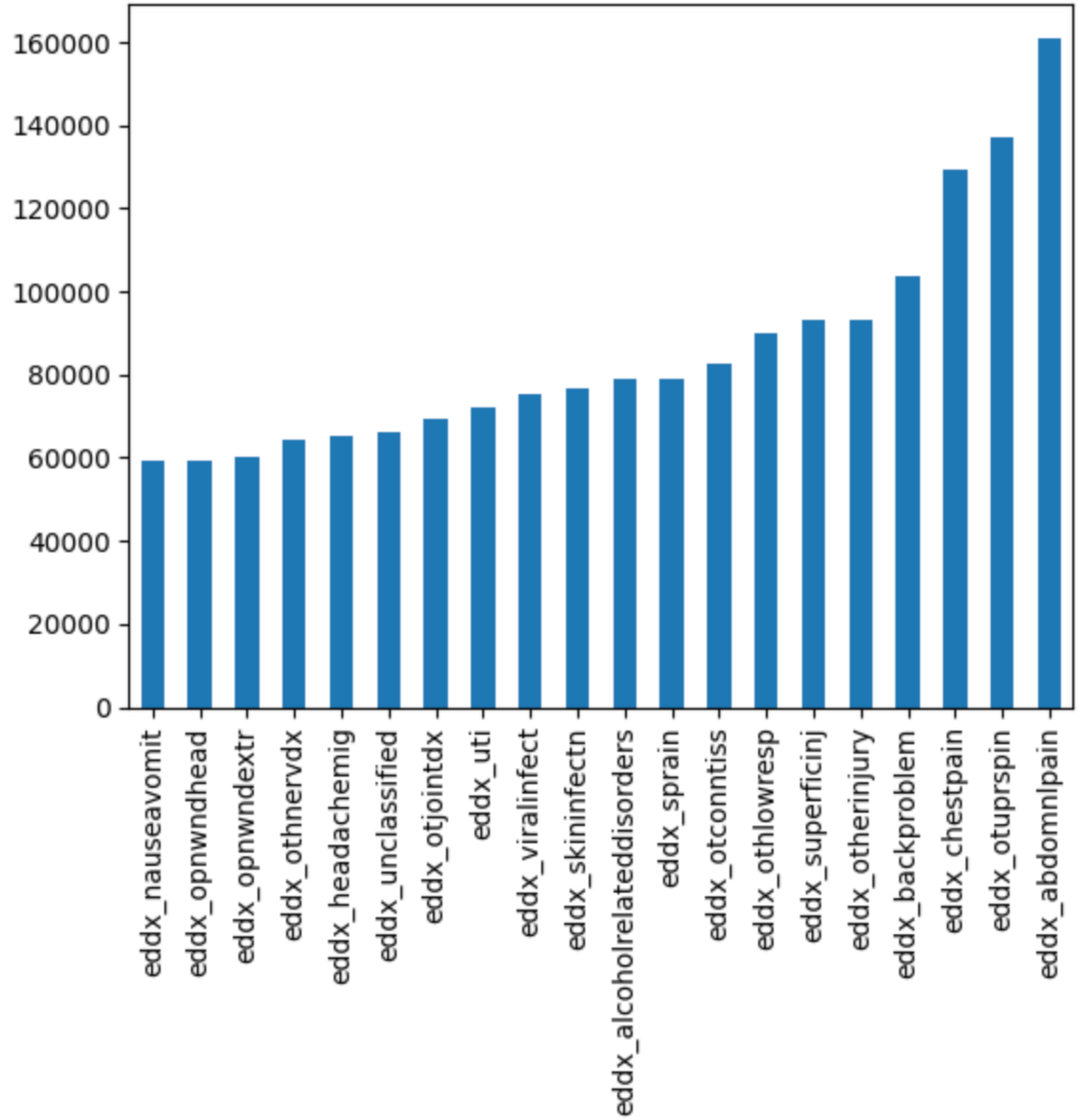




Most of the encounters are urgent

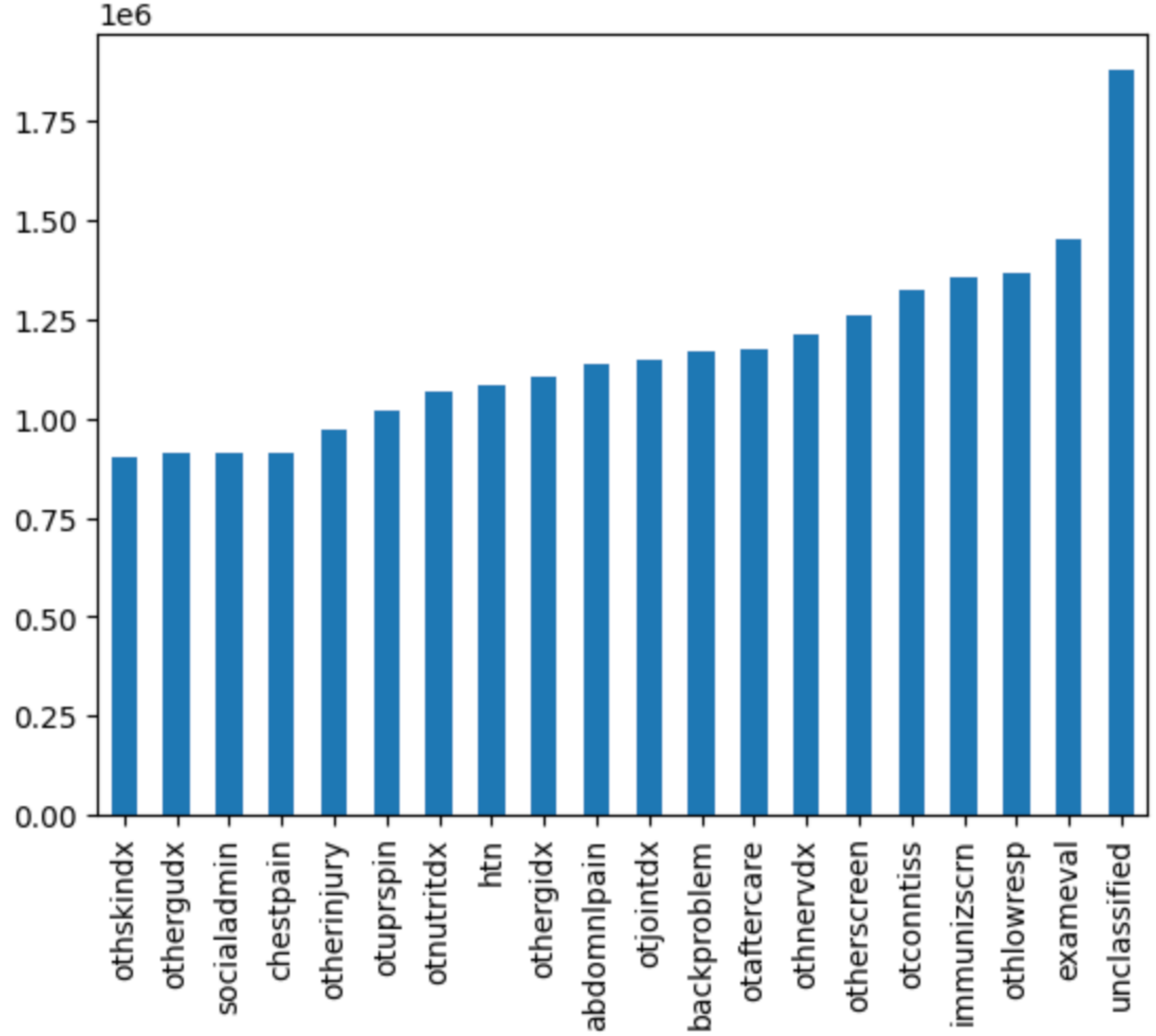
Diagnosis feature

- 284 diagnosis features are all binary
    - eddx_2ndarymalig, eddx_abdomhern
- 89.48% encounters have only one d
- 230,215/3,640,261 encounters don
- Top 30 diagnosis:



Past medication feature

- Binary features
  - Abdomhernia, abdomnlpain, …
- 385,664 encounters don't have pas
- Past medication features are differ
- Top 20 features



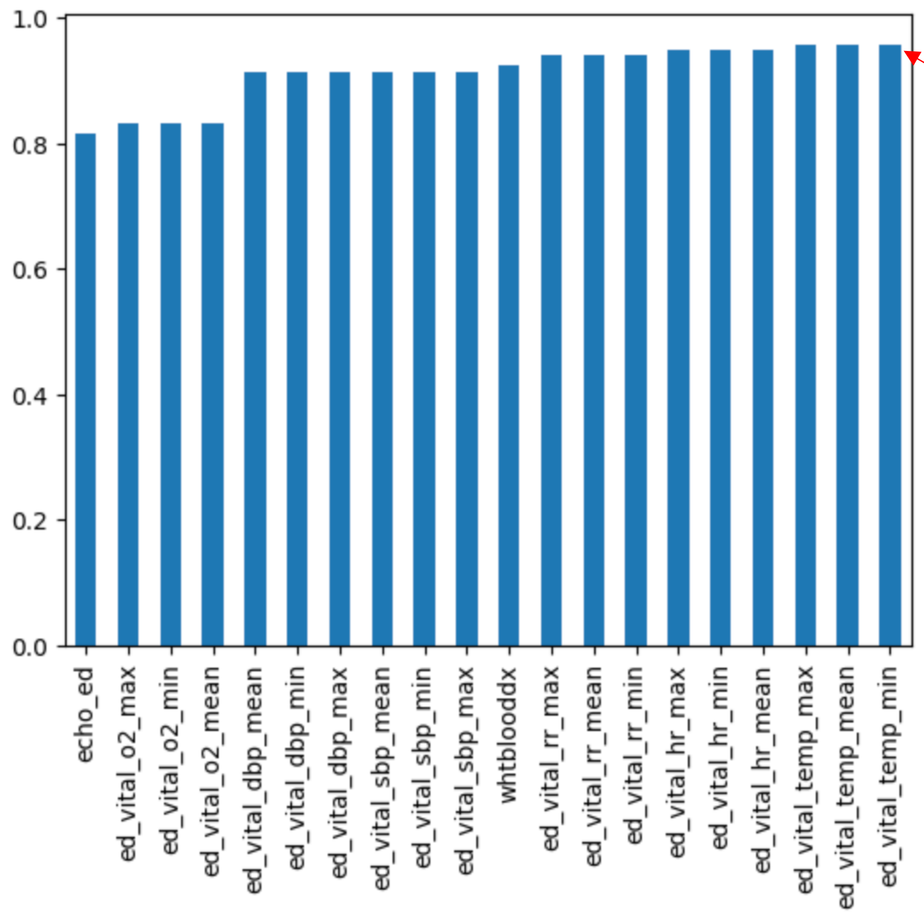
Lab features

- 578 lab features includes 6 kinds of features
  - Lab test, POC features, ED features, historical vital values, min and max of lab values, whether or not imaging was done
- Most of them are continuous features
- Average number of lab features: 110.057
- Some lab features are connected by most of the encounters

Lab features

- Some lab features are connected by most of the encounters

$$\frac{\text{Number of encounters}}{\text{total number of encounters}}$$



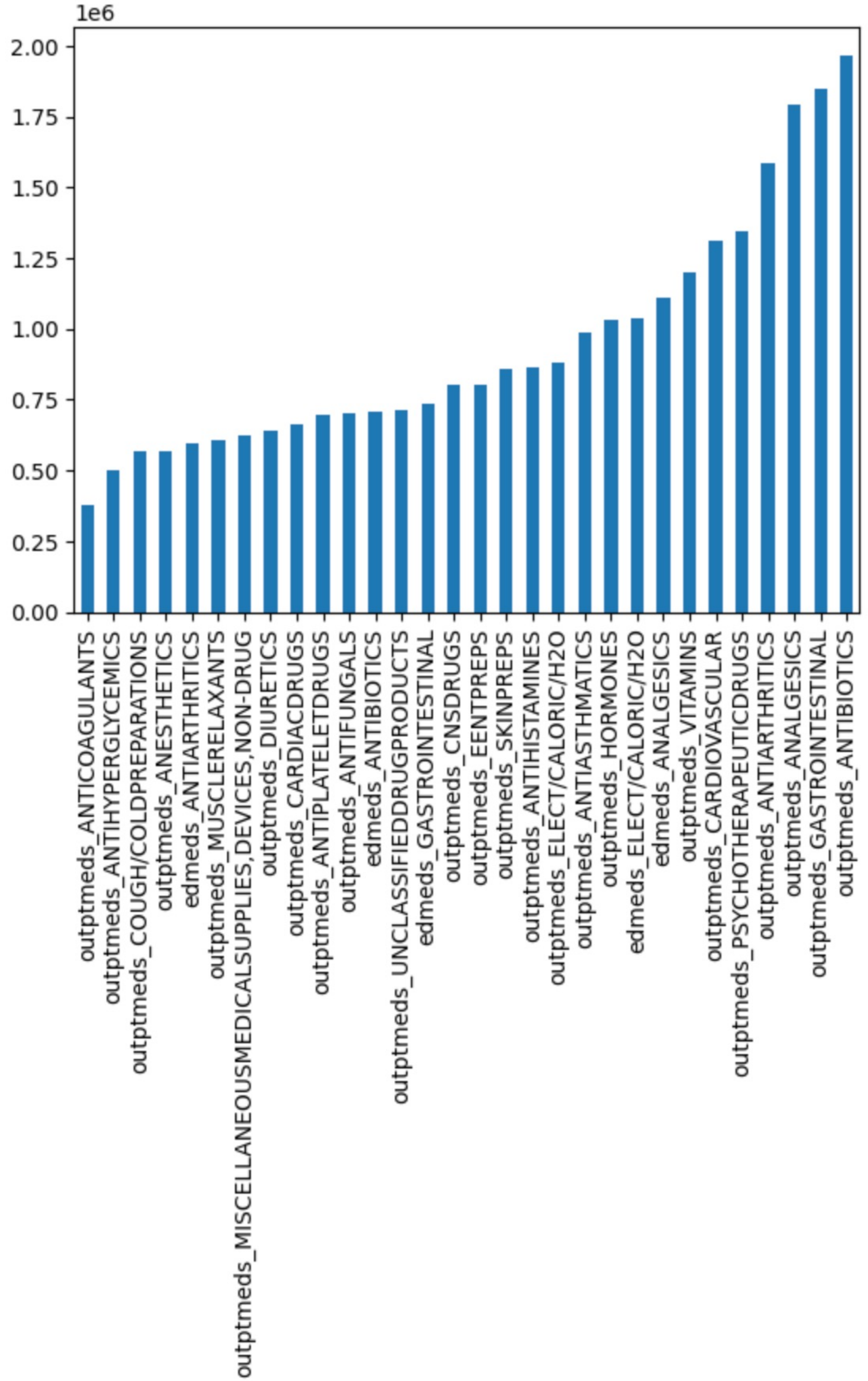
Almost connect with all the encounters!

Lab features

- Visualization of missing values
  - White part is the missing data

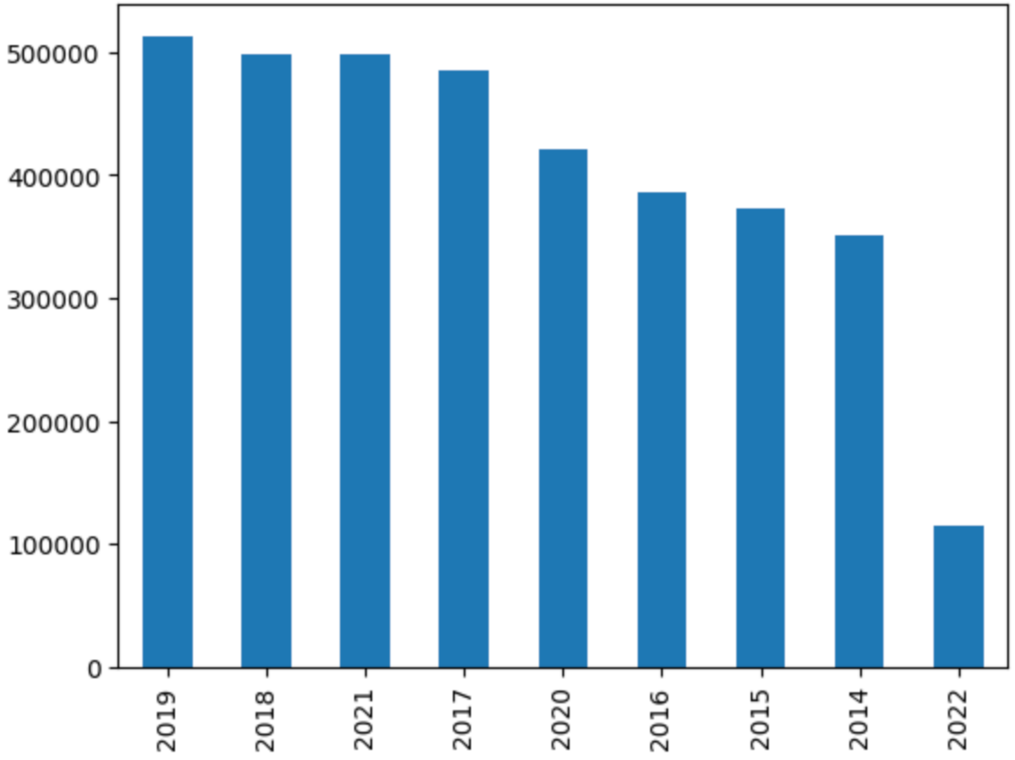
Medication features

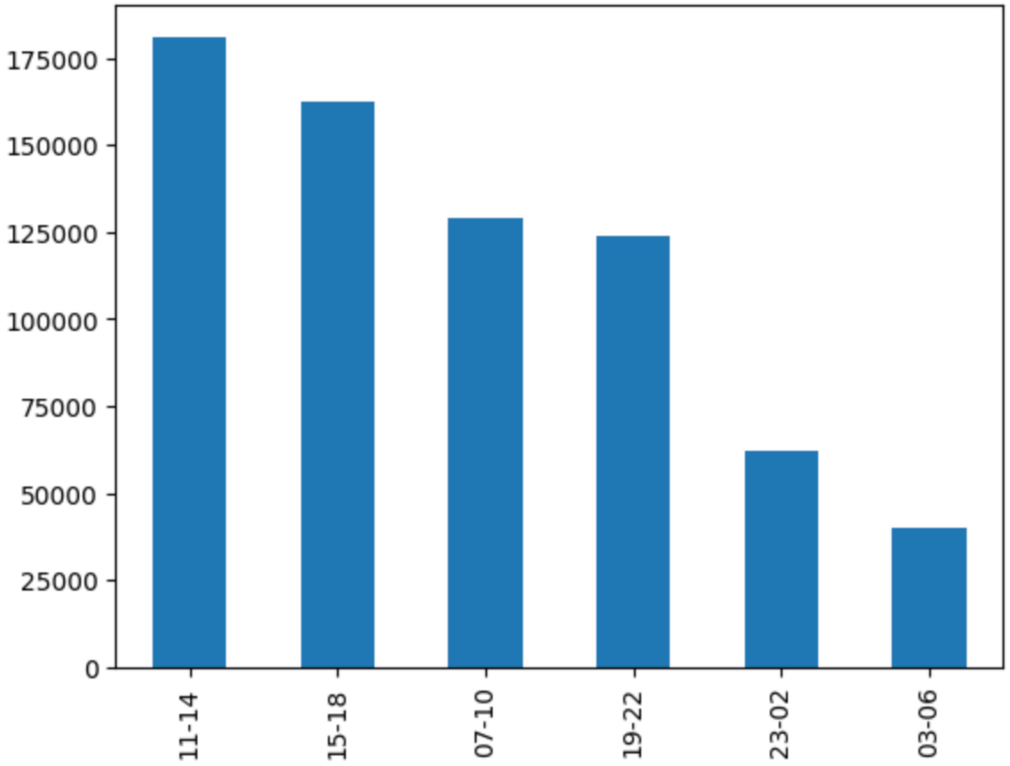
- 104 medication features
  - Medication prescribed outside of the ED
  - Medication prescribed in the ED
- Average number of medicine: 9.69
- 167,363/3,640,261 encounters don’t have me
- Top 30 medicine



Time features

- Time span: 2014-01-01 – 2022-03-01
- Distribution of the year and hour





Problem definition

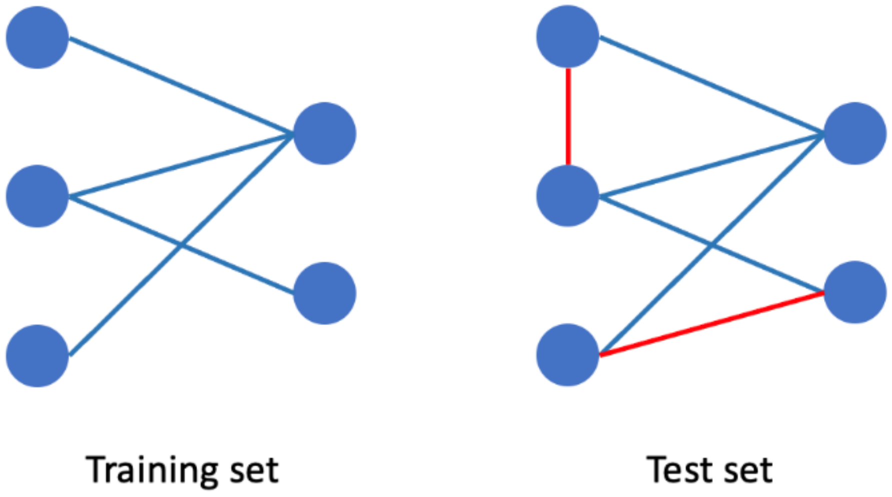
- Given an encounter record, predict the reliability of existing **diagnosis** and **medication,** as well as possible **diagnosis** and **medication**
- Graph can help us to reveal the underlying relationship between different instances
  - E.g., high correlation among some diagnoses
- Formulate the problem as a **link prediction** task
  - Encounter and diagnosis/medication are nodes in graph
  - Predict the probability that an edge between encounter and diagnosis/medication exists

Problem setting (1)

- Link prediction
  - Encounter, diagnosis, medication and lab test are nodes in graph
  - Predict the probability that an edge between encounter and diagnosis/medication exists
- Dataset split
  - Due to the large scale of the whole dataset, we now only consider the encounter during 2021-2022
  - Training set: 2021.01.01 – 2021.12.31
  - Validation set: 2022.01.01 – 2022.01.31
  - Test set: 2022.02.01 – 2022.03.31

Problem setting (2)

- Test set contains some edges which does not exist in training set to prevent data leakage



Ranking metric

- Given a query and some candidate keys, evaluate the quality of candidate list ranked by model
- Given an encounter, rank all the medication/diagnosis by model, and evaluate the quality of topK candidates by:
  - Recall: ratio of positive medication/diagnosis in the list to all positive ones
  - Ndcg: order of positive medication/diagnosis in the list
  - Precision: number of positive medication/diagnosis in the list
  - Hit: whether there is a positive medication/diagnosis in the list

Modeling (1)

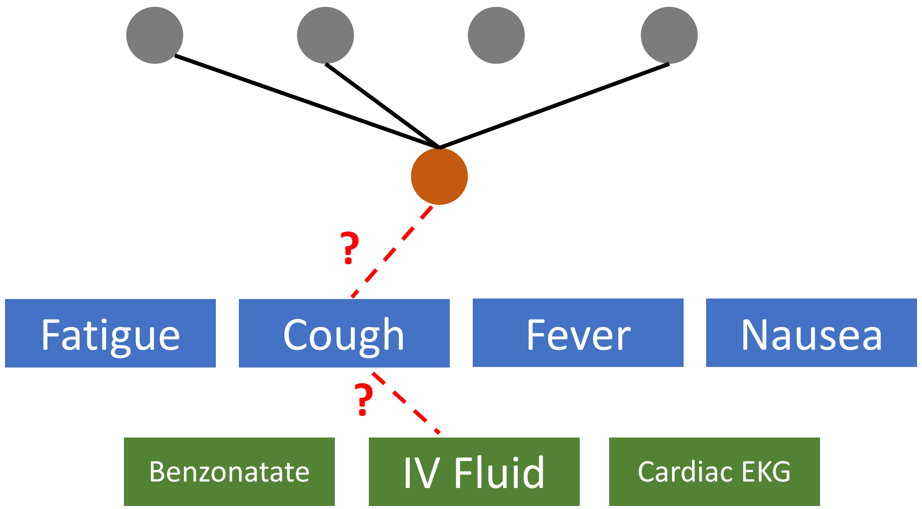
- Disease prognosis is determined by the results of lab tests
  - $X \rightarrow D$, where D is diagnosis and X is lab tests
- The diagnosis and the lab test results dictate which medication should be taken
  - $X, D \rightarrow M$, where M is medication
- Objective
  - Maximize the following probability distribution

$$\max p(M|X, D)P(D|X)$$

Probability of using this medication according to the diagnosis and lab test

Probability of disease prognosis according to lab test

Modeling (2)



Lab test

Step1: aggregate the information from lab nodes

encounter

Step2: predict the diagnosis

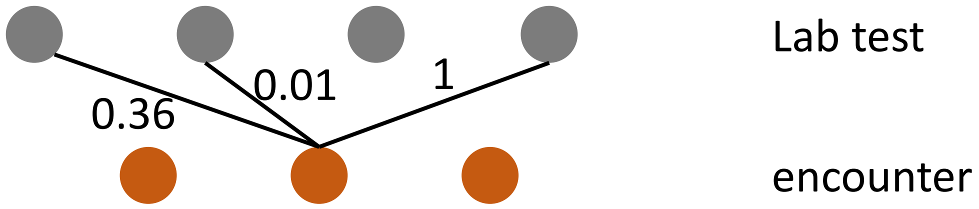
diagnosis

Step3: predict the medication based on diagnosis and lab information

medication

Modeling (3)

- Step1: aggregate the information from lab nodes
  - Lab test value is considered as the edge feature
- Apply the GNN framework
  - NeurIPS2020-Handling Missing Data with Graph Representation Learning
  - Update the edge and node representation at each message passing step



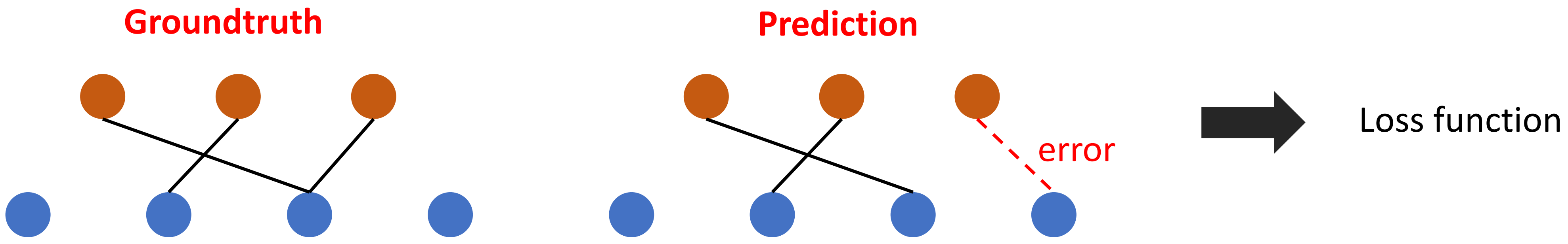
Modeling (4)

- Step2: predict the diagnosis
- Apply reconstruct loss
  - Optimize the model by reconstructing the graph structure among encounter and diagnosis nodes

Adjacent matrix

$$\min ||A - A'||$$

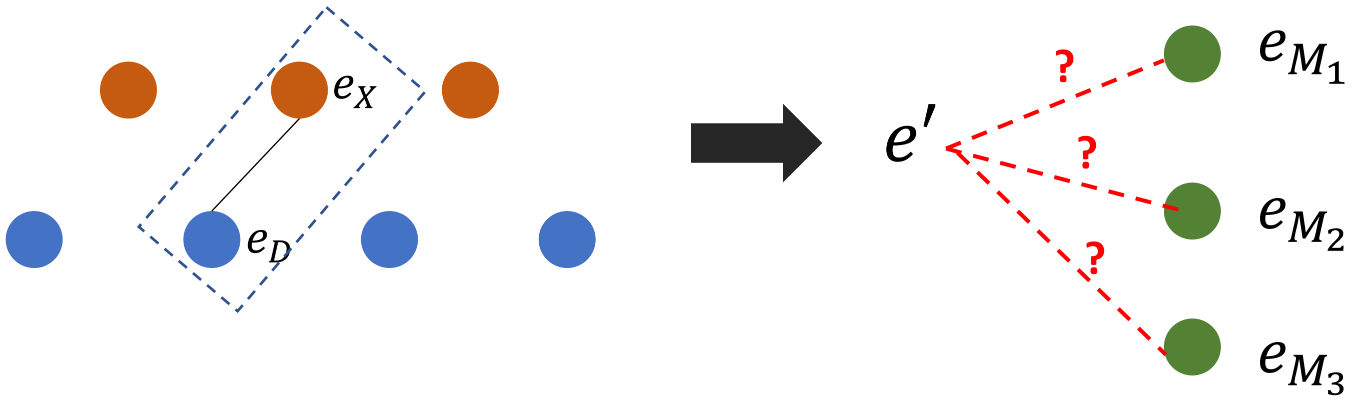
- Reconstruct the graph structure of sampled encounters



Modeling (4)

- Step3: predict the medication based on diagnosis and lab information
- Use both encounter and diagnosis embeddings to predict the medication
  - Apply a MLP to decode these two embeddings

$$e' = MLP(e_X || e_D)$$



Pretraining embedding

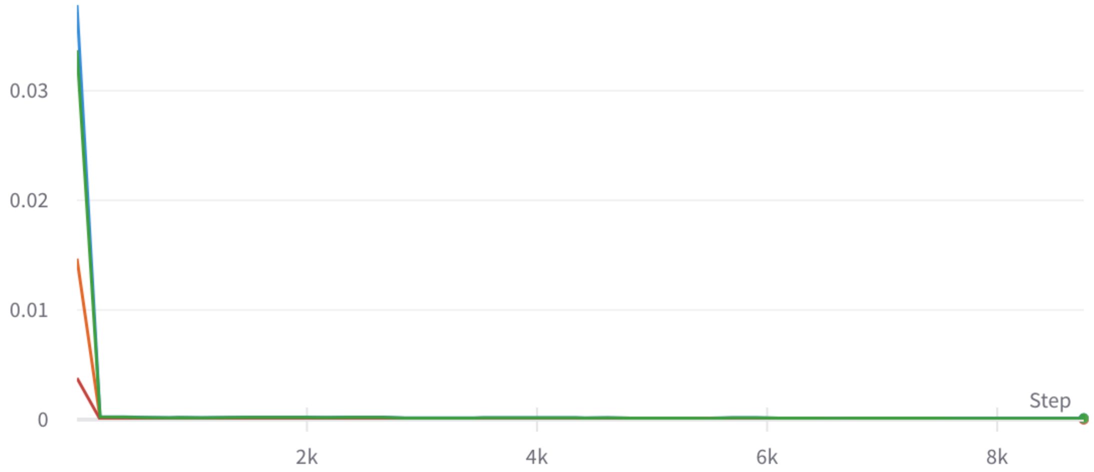
- Diagnosis has strong relationship with medication
- We can use **co-occurrence information** to learn the basic embeddings
- Organize the co-occurred diagnosis and medication as a ‘sentence’:

$$\underbrace{\text{diag1, diag2, med1, med2, med3, …}}_{\text{“sentence”}}$$

- Apply work embedding method to learn embeddings
  - GloVe: Global Vectors for Word Representation
  - The vectors of two “words” that occur together more frequently will be more similar in embedding space

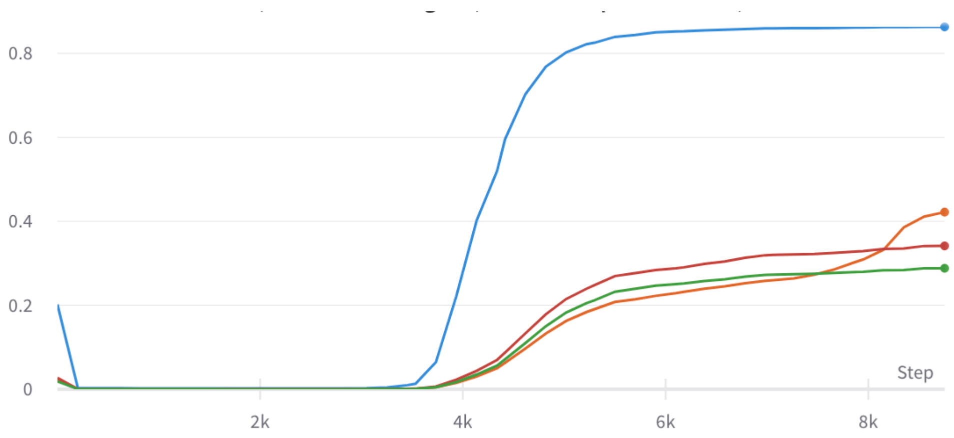
Result

diagnosis



Training step

medication



Hit:0.8631

ndcg:0.4218
precision:0.3414
recall:0.2881

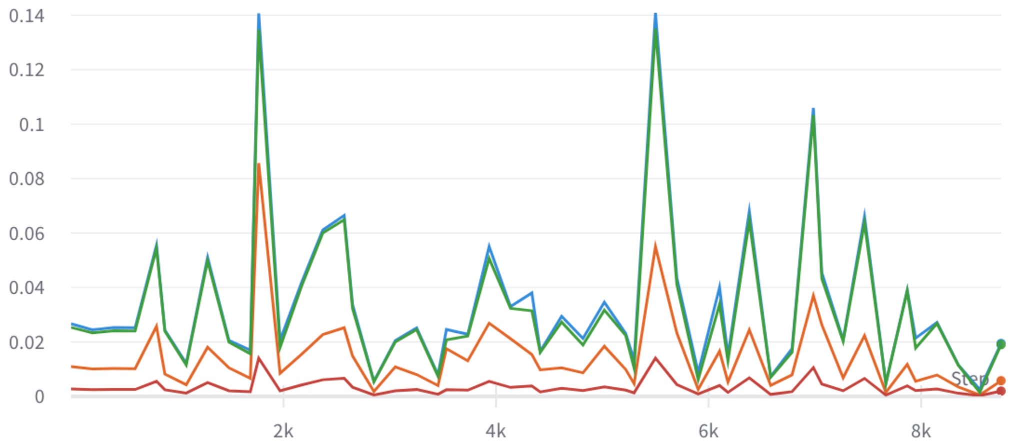
Training step

Why is performance on diagnosis so poor?

Result

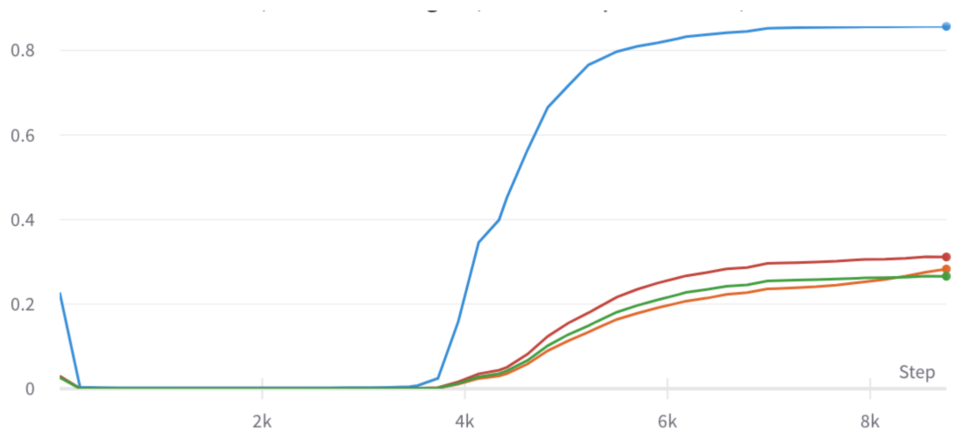
- Directly use encounter embeddings without GNN

diagnosis



Training step

medication



Training step

Hit:0.857

precision:0.3115

ndcg:0.283

recall:0.2658

Improvement

- How to use lab tests
  - Consider lab test as node is not good
    - Some lab nodes are almost fully connected by the encounters, resulting in an expensive aggregation process
    - Preliminary results show that performance on diagnosis is poor
  - Better way: Apply lab test results to connect the encounters by measuring the similarity
- Build a knowledge graph among all the medical concept in EHR
  - Identify different concept by node type and edge type
  - Connect the encounters with same patient id
  - Include the past medication entity
  - We can design auxiliary objective on the knowledge graph
    - Knowledge graph completion

Evaluation

- How to evaluate the performance on anomaly detection
  - We don’t have such annotated labels
- Generate some anomalous edges
  - Aane: Anomaly aware network embedding for anomalous link detection. *ICDM 2020*.
- Example
  - Randomly choose some source nodes in the graph
  - Randomly select some target nodes according to the predefined anomaly ratio
  - For each source node
    - sort the candidate target nodes by the distance in descending order
    - add the anomalous edge between source node and candidate target node until the number of anomalous links reaches the predefined anomaly ratio