Graph-based Attention Model for Healthcare Representation Learning

Author: Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, Jimeng Sun

Source: KDD '17

Advisor: Jia-Ling Koh

Speaker: Shih-Han Lo

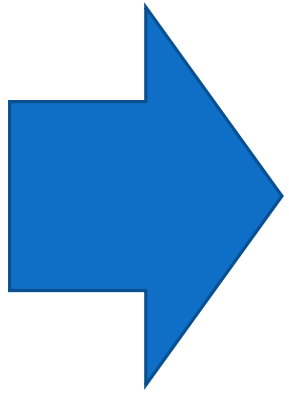
Date: 2018/03/27

Outline

- **Introduction**
- Method
- Experiment
- Conclusion

Motivation


e-health
Search
A-F
G-L
M-R
T-Z
Medical Record
HN
Date :
Patient name :
D.O.B./Age :
GP Name :
Address :
Tel :
Blood Group :
Weight :
Height :
Home
Schedule



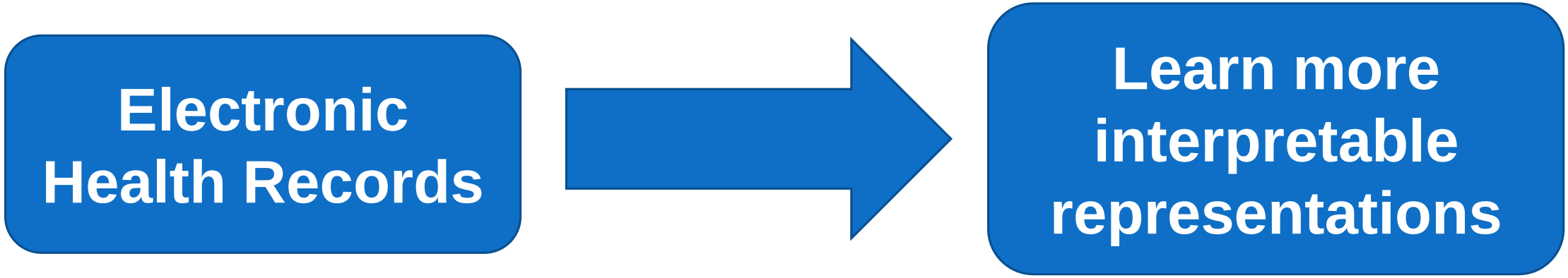

Data
Collaborative
Medical
DiseaseManagement
Risk Assessment
ProviderReimbursement
QualityImprovement
PrimaryCareDelivery
Predictive Modeling
Preventive
CaseManagement
Analytics
Needs Assessment
Coordinated
Comprehensive
Partnership
Healthcare
Integrated

Motivation

- Two important challenges remain for learning in healthcare:
  - Data insufficiency
  - Interpretation

Goal

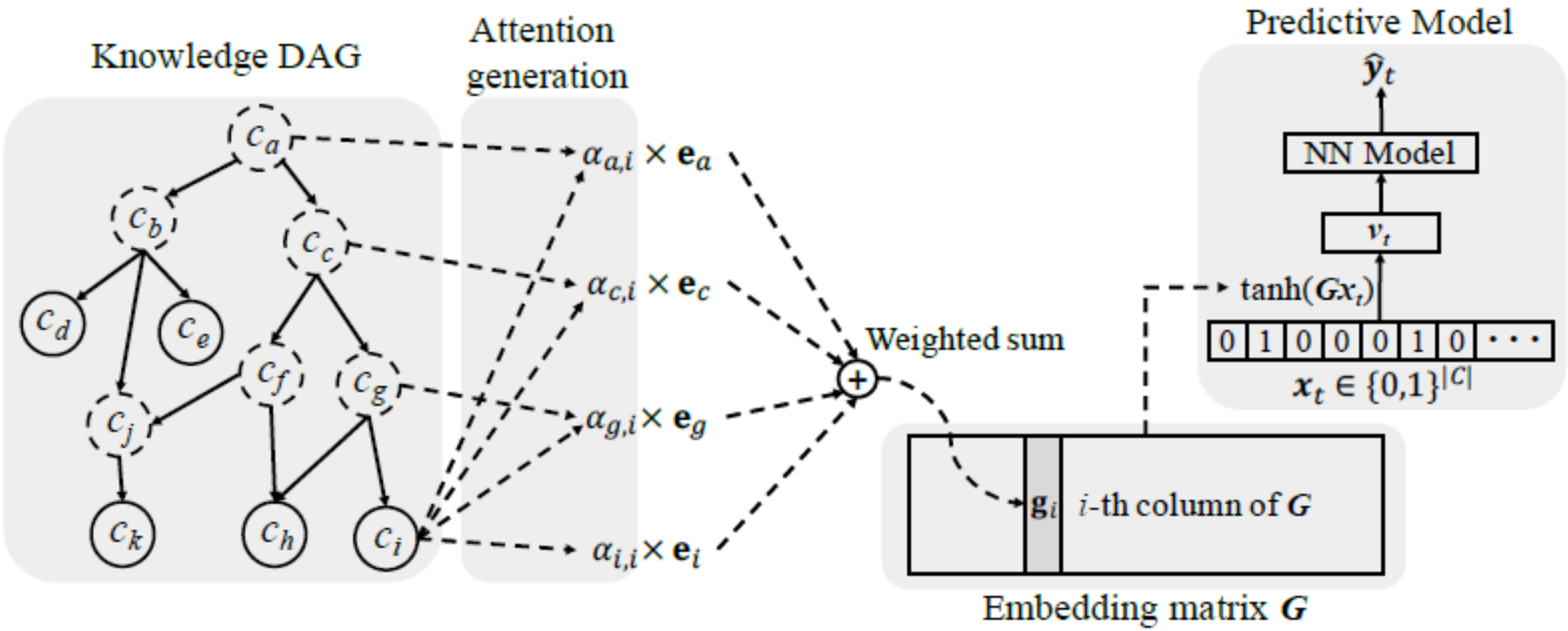
(Using medical ontologies)



Outline

- Introduction
- **Method**
- Experiment
- Conclusion

Framework


Knowledge DAG
Attention generation
Predictive Model
$\hat{y}_t$
NN Model
v_t
$\tanh(Gx_t)$
0 1 0 0 0 1 0 · · ·
$x_t \in \{0,1\}^{|C|}$
c_a
c_b
c_c
c_d
c_e
c_f
c_g
c_j
c_k
c_h
c_i
$\alpha_{a,i} \times \mathbf{e}_a$
$\alpha_{c,i} \times \mathbf{e}_c$
$\alpha_{g,i} \times \mathbf{e}_g$
$\alpha_{i,i} \times \mathbf{e}_i$
Weighted sum
$\mathbf{g}_i$ i-th column of G
Embedding matrix G

Definition

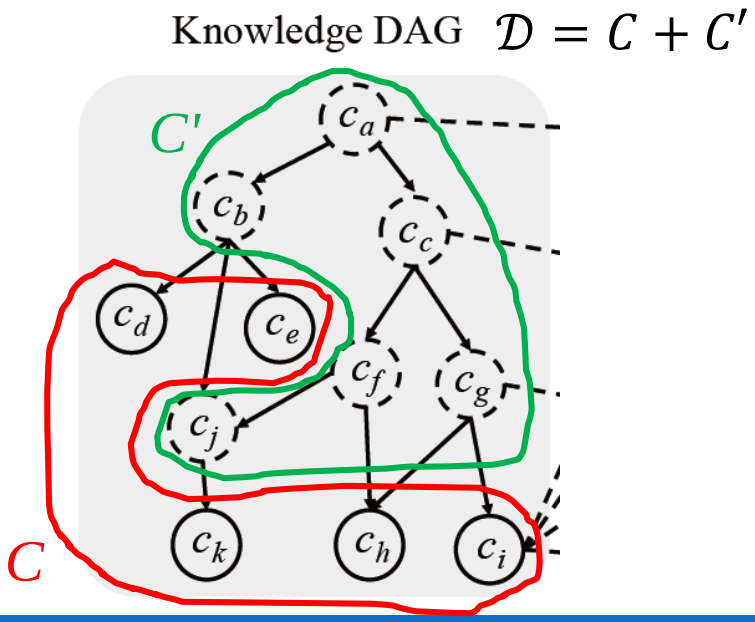
- Medical codes (concepts) from EHR
  - $C = \{c_1, c_2, \dots, c_{|C|}\}$

V_t: 0 1 0 0 0 1 0 • • •

$\boldsymbol{x}_t \in \{0,1\}^{|C|}$

- Sequence of visits from each patient
  - $V_1, V_2, \dots, V_T$

- Non-leaf nodes (ancestors)
  - $C' = \{c_{|C|+1}, c_{|C|+2}, \dots, c_{|C|+|C'|}\}$



Initializing Basic Embeddings

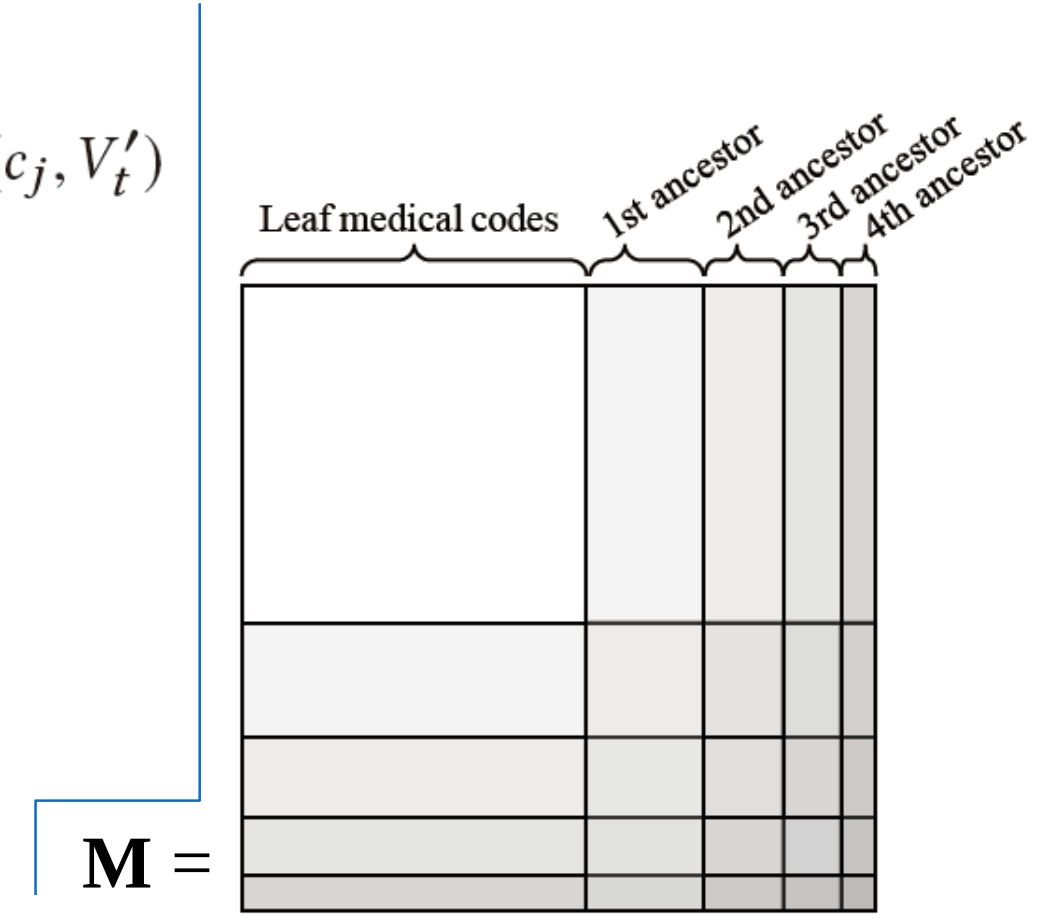
- Co-occurrence of codes c_i and c_j:

$$co\text{-}occurrence(c_i, c_j, V_t') = count(c_i, V_t') \times count(c_j, V_t')$$

- Co-occurrence matrix: $\mathbf{M} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$
- Loss function [31]:

$$J = \sum_{i,j=1}^{|\mathcal{D}|} f(\mathbf{M}_{ij})(\mathbf{e}_i^\top \mathbf{e}_j + b_i + b_j - \log \mathbf{M}_{ij})^2$$

$$\text{where } f(x) = \begin{cases} (x \ / \ x_{max})^\alpha & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$$



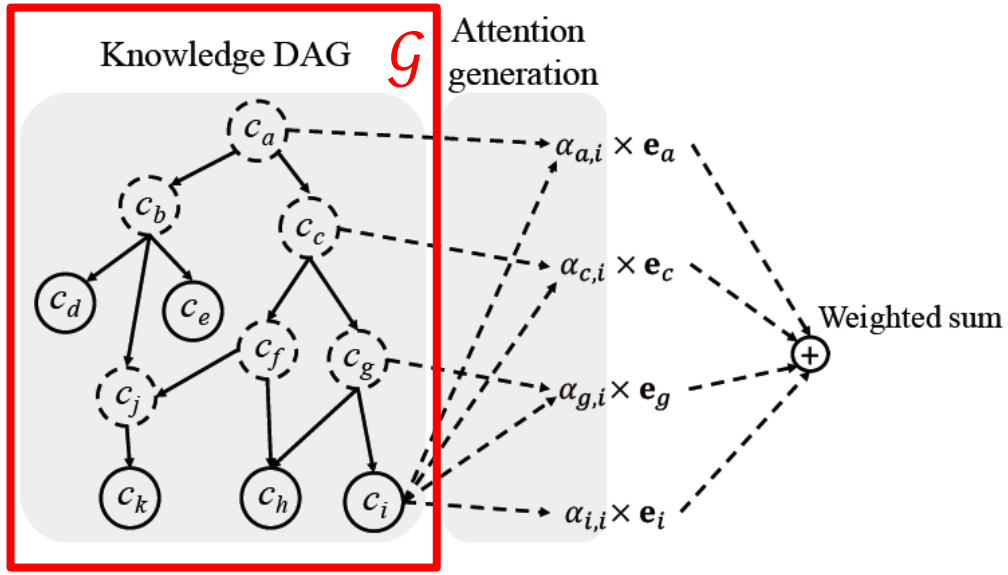
Attention Mechanism

- GRAM leverages the parent-child relationship of $\mathcal{G}$ to learn robust representations.

- Leaf node's final representation:

$$\mathbf{g}_i = \sum_{j \in \mathcal{A}(i)} \alpha_{ij} \mathbf{e}_j, \qquad \sum_{j \in \mathcal{A}(i)} \alpha_{ij} = 1, \ \alpha_{ij} \geq 0 \ \text{ for } j \in \mathcal{A}(i)$$

where $$\alpha_{ij} = \frac{\exp(f(\mathbf{e}_i, \mathbf{e}_j))}{\sum_{k \in \mathcal{A}(i)} \exp(f(\mathbf{e}_i, \mathbf{e}_k))}$$



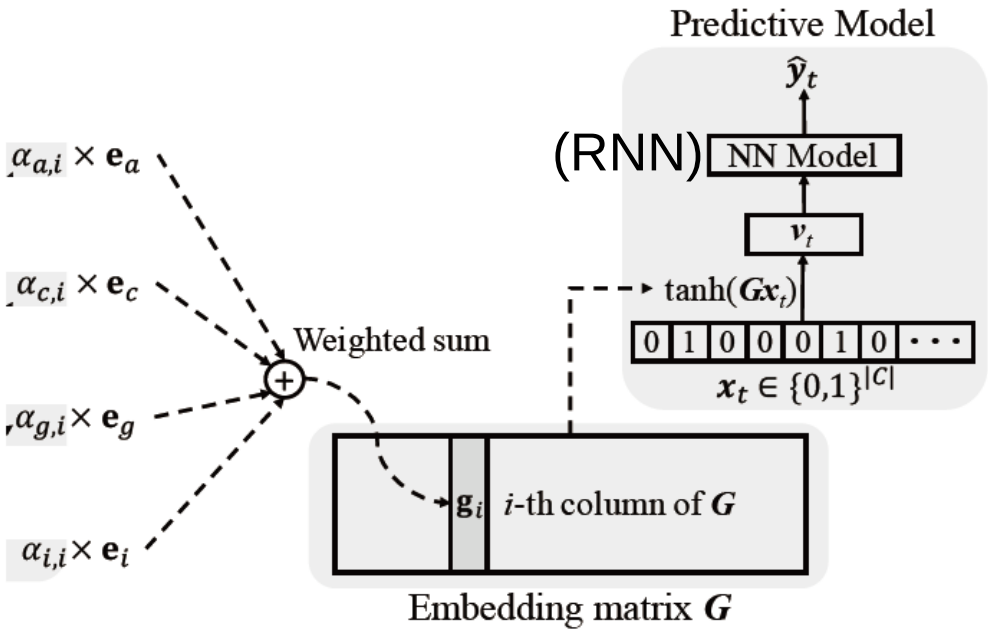
Training with a Predictive Model

- Predict the codes of the next visit:

$$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t = \tanh(\mathbf{G}[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t]),$$

$$\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_t = \mathrm{RNN}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t, \theta_r),$$

$$\widehat{\mathbf{y}}_t = \widehat{\mathbf{x}}_{t+1} = \mathrm{Softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})$$



- Loss function:

$$\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_T) = -\frac{1}{T-1}\sum_{t=1}^{T-1}\left(\mathbf{y}_t{}^{\top}\log(\widehat{\mathbf{y}}_t) + (1-\mathbf{y}_t)^{\top}\log(1-\widehat{\mathbf{y}}_t)\right)$$

Outline

- Introduction
- Method
- **Experiment**
- Conclusion

Datasets

Dataset	Sutter PAMF	MIMIC-III	Sutter HF cohort
# of patients	258,555†	7,499†	30,727† (3,408 cases)
# of visits	13,920,759	19,911	572,551
Avg. # of visits per patient	53.8	2.66	38.38
# of unique ICD9 codes	10,437	4,893	5,689
Avg. # of codes per visit	1.98	13.1	2.06
Max # of codes per visit	54	39	29

† For all datasets, we chose patients who made at least two visits.

Prediction Performance

Model	0-20	20-40	40-60	60-80	80-100
GRAM+	**0.0150**	**0.3242**	0.4325	0.4238	0.4903
GRAM	0.0042	0.2987	0.4224	0.4193	0.4895
RandomDAG	0.0050	0.2700	0.4010	0.4059	0.4853
RNN+	0.0069	0.2742	0.4140	0.4212	**0.4959**
RNN	0.0080	0.2691	0.4134	0.4227	0.4951
SimpleRollUp	0.0085	0.3078	**0.4369**	**0.4330**	0.4924
RollUpRare	0.0062	0.2768	0.4176	0.4226	0.4956

(a) *Accuracy@5* of sequential diagnoses prediction on Sutter data

Model	0-20	20-40	40-60	60-80	80-100
GRAM+	**0.0672**	**0.1787**	**0.2644**	0.2490	0.6267
GRAM	0.0556	0.1016	0.1935	0.2296	0.6363
RandomDAG	0.0329	0.0708	0.1346	0.1512	0.4494
RNN+	0.0454	0.0843	0.2080	0.2494	0.6239
RNN	0.0454	0.0731	0.1804	0.2371	0.6243
SimpleRollUp	0.0578	0.1328	0.2455	**0.2667**	**0.6387**
RollUpRare	0.0454	0.0653	0.1843	0.2364	0.6277

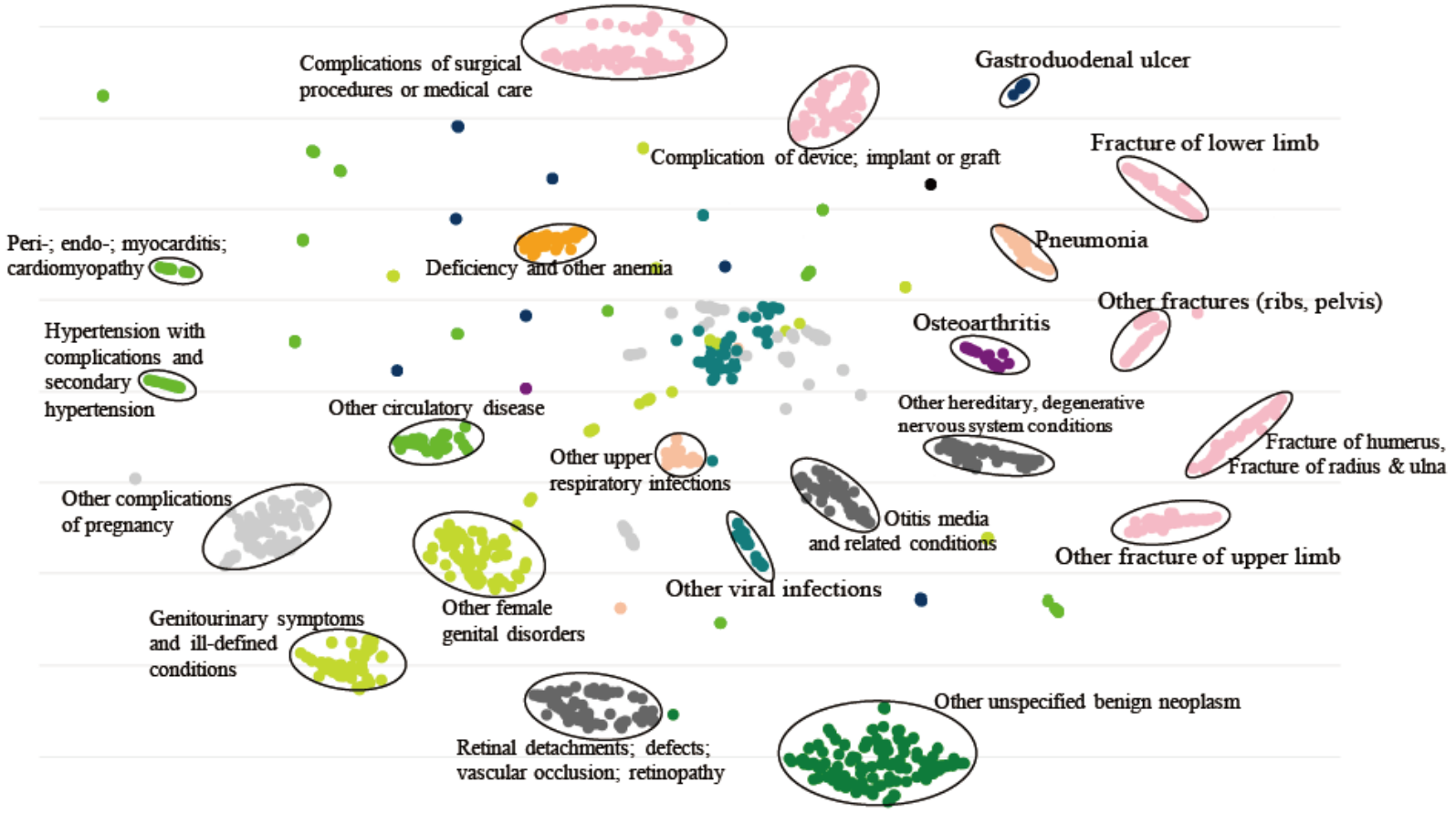
(b) *Accuracy@20* of sequential diagnoses prediction on MIMIC-III

Prediction Performance

Model	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
GRAM+	0.7970	**0.8223**	0.8307	**0.8332**	**0.8389**	**0.8404**	**0.8452**	**0.8456**	**0.8447**	**0.8448**
GRAM	**0.7981**	0.8217	**0.8340**	**0.8332**	0.8372	0.8377	0.8440	0.8431	0.8430	0.8447
RandomDAG	0.7644	0.7882	0.7986	0.8070	0.8143	0.8185	0.8274	0.8312	0.8254	0.8226
RNN+	0.7930	0.8117	0.8162	0.8215	0.8261	0.8333	0.8343	0.8353	0.8345	0.8335
RNN	0.7811	0.7942	0.8066	0.8111	0.8156	0.8207	0.8258	0.8278	0.8297	0.8314
SimpleRollUp	0.7799	0.8022	0.8108	0.8133	0.8177	0.8207	0.8223	0.8272	0.8269	0.8258
RollUpRare	0.7830	0.8067	0.8064	0.8119	0.8211	0.8202	0.8262	0.8296	0.8307	0.8291

(c) AUC of HF onset prediction on Sutter HF cohort

Qualitative Evaluation



(a) Scatterplot of the final representations g_i's of GRAM+

Qualitative Evaluation

(b) Scatterplot of the final representations g_i's of GRAM

(c) Scatterplot of the trained embedding matrix W_{emb} of RNN+

Outline

- Introduction
- Method
- Experiment
- **Conclusion**

Conclusion

- GRAM uses both a knowledge DAG and EHR to learn an accurate and interpretable representations of medical concepts.

- We showed improvement in the prediction performance:
  - Low-frequency diseases.
  - Small datasets.