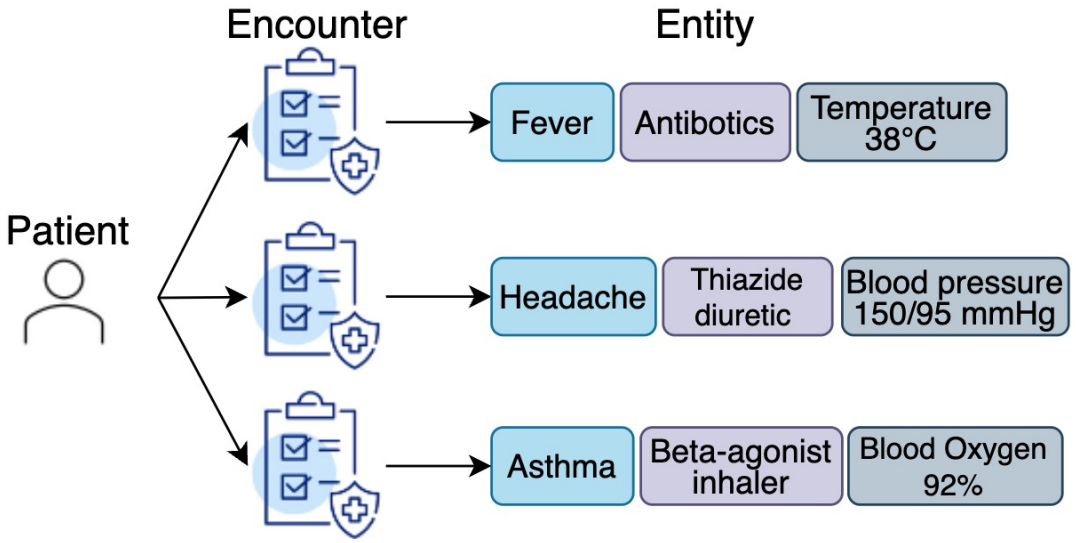
HEART: Heterogeneous Relation-Aware Transformer for EHR

Tinglin Huang, Rohan Krishna Thakur, Meili Gupta, Vimig Socrates,

David van Dijk, R. Andrew Taylor, Rex Ying

Electronic Health Record (EHR) Data

- EHR includes the information of an encounter by digitizing the medical information
  - Demography, Diagnosis, Medication, Lab results, Procedures
- It is a hierarchical structures
  - One patient has multiple encounters, and one encounter includes a range of medical entities

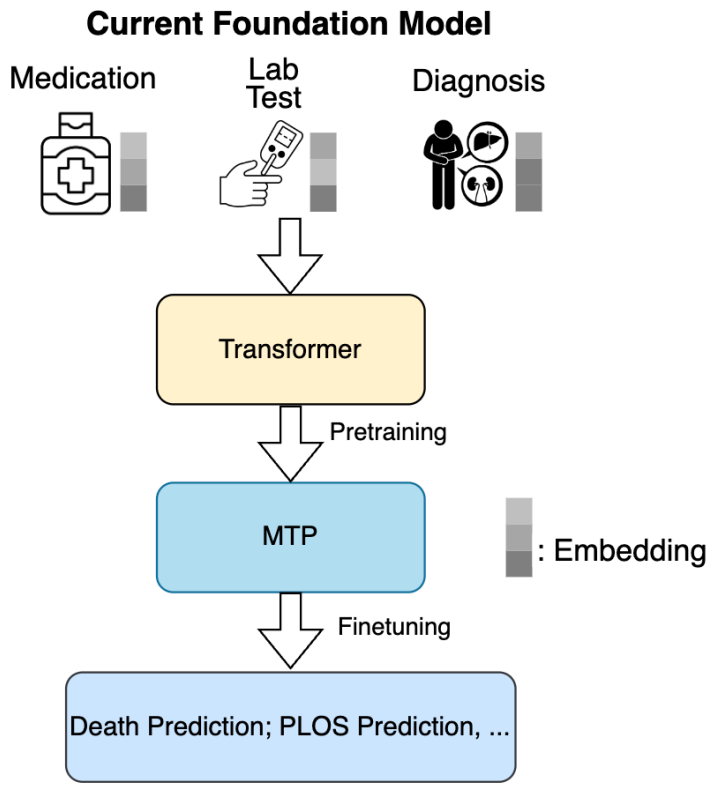


Foundation Models for EHR (1)

- EHR allows us to develop healthcare applications with computational methods
  - Such as automatic diagnosis and health outcome prediction
- 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
- **Foundation models** have demonstrated their effectiveness in medical domains

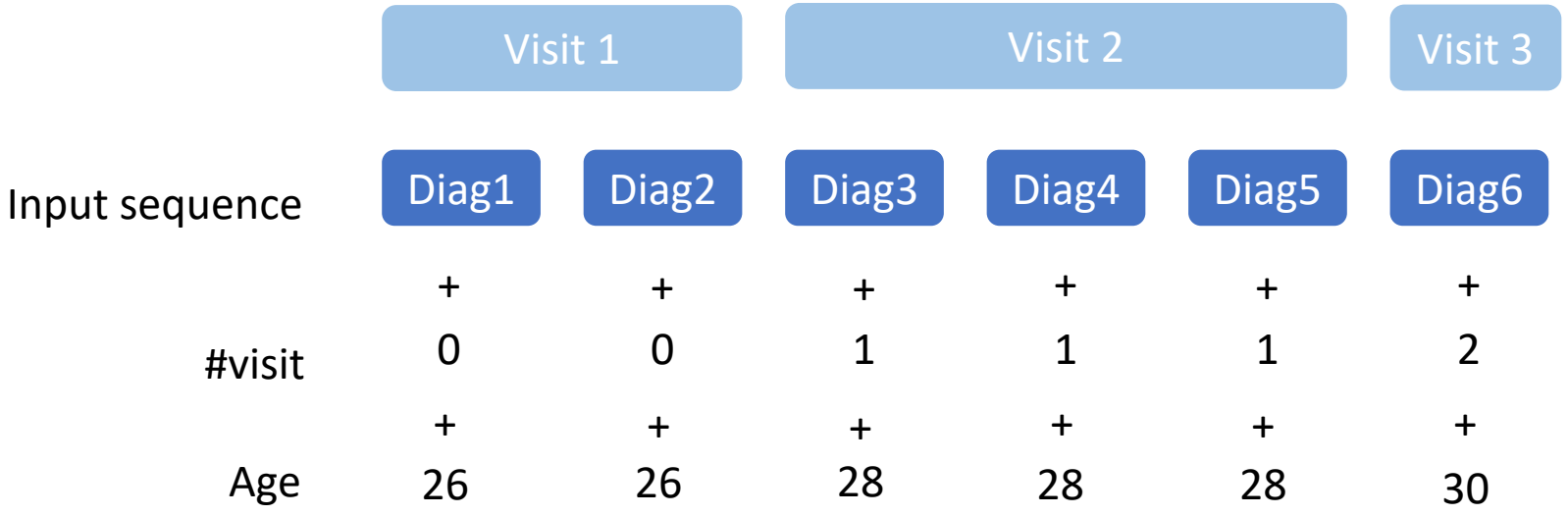
Foundation Models for EHR (2)

- Foundation model treats medical entities as tokens and organize the entities included in the encounters as sentences
- Pretrain Transformer with self-supervised objective
  - Masked token prediction (MTP)
- Fine-tune the model for different downstream tasks



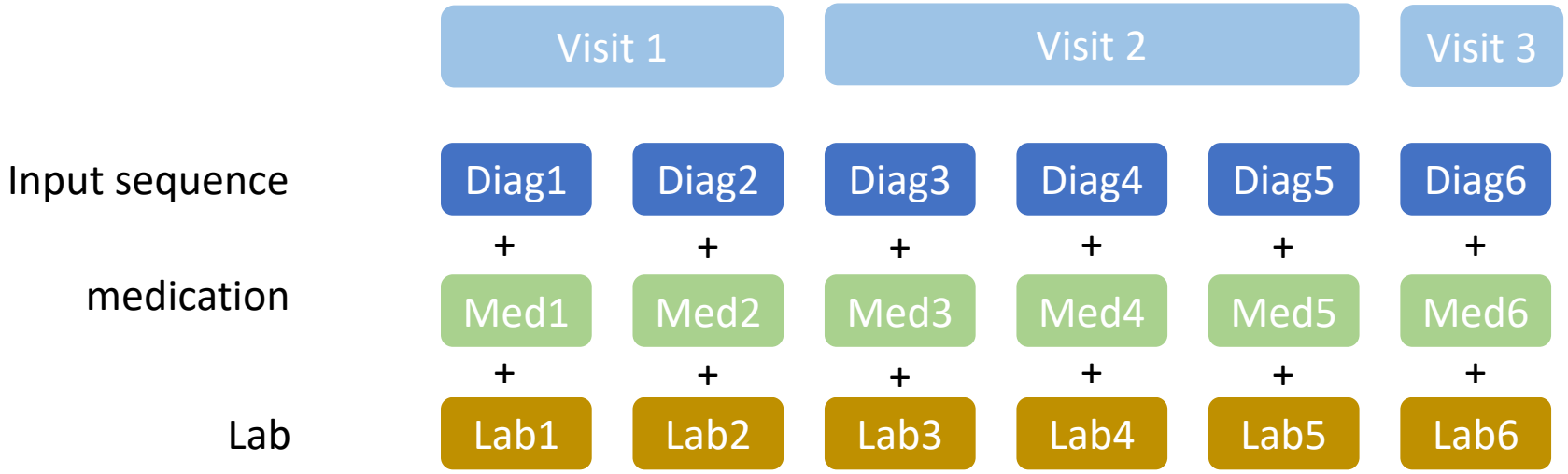
BEHRT and Med-BERT: Related Work

- For example, BEHRT and Med-BERT
  - They organize the diagnosis entities as a sentence and feed it to Transformer
  - The number of visits and demography is used as the positional encoding



ExBEHRT: Related Work

- A recent work ExBEHRT utilizes different modalities within the modeling
  - Lab test, Medication, …
  - By adding representations of other modalities to diagnosis representations

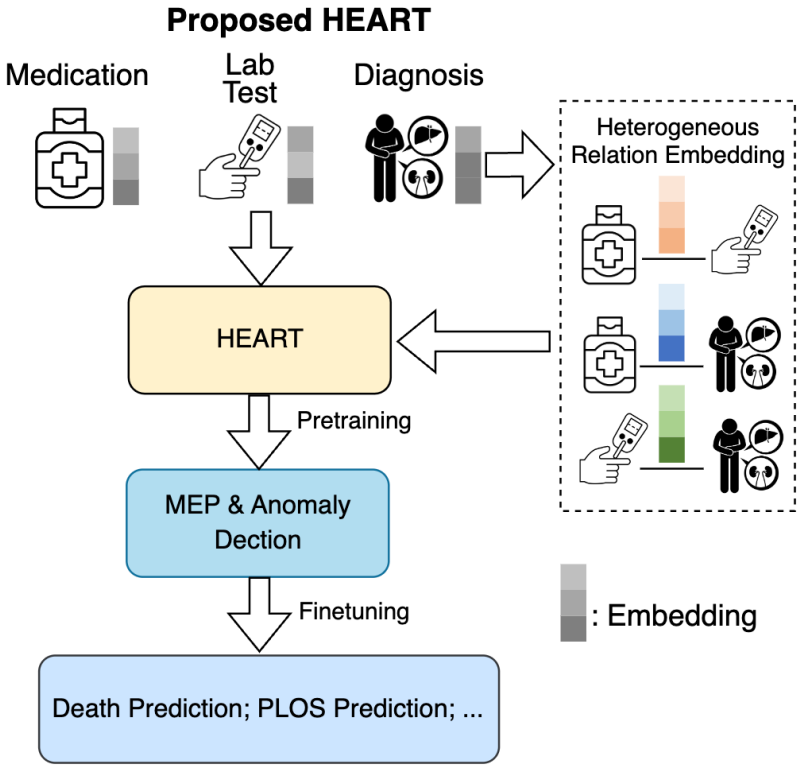


Challenges with Current Models

- However, current models fall short of utilizing the **heterogeneous correlations** inherent between medical entities
  - BEHRT, Med-BERT: focus merely on encoding diagnoses
  - ExBEHRT: encode different modalities in the same representation space
- Exploiting how entities interact with each other in the heterogeneous space enables the model to perform complex reasoning
  - E.g., the relationship between "**Antibiotics**" (medication) and both "**Fever**" (diagnosis) and "**Antibody Tests: Positive**" (lab test)

Our Proposed Solution: HEART

- HEART, a heterogeneous relation-aware Transforme for EHR
  - Code Structure Embedding
  - Heterogeneous Relation Embedding
    - The pairwise relation between each pair of entities is parameterized as an embedding
  - Multi-Level Attention Scheme
    - Alternatively perform encoding across entity level and encounter level
  - Pretraining Task
    - Adapt the masked token prediction objective to the position-agnostic missing entity prediction task
    - Introduce anomaly detection as an additional objective



HEART: Input representation (1)

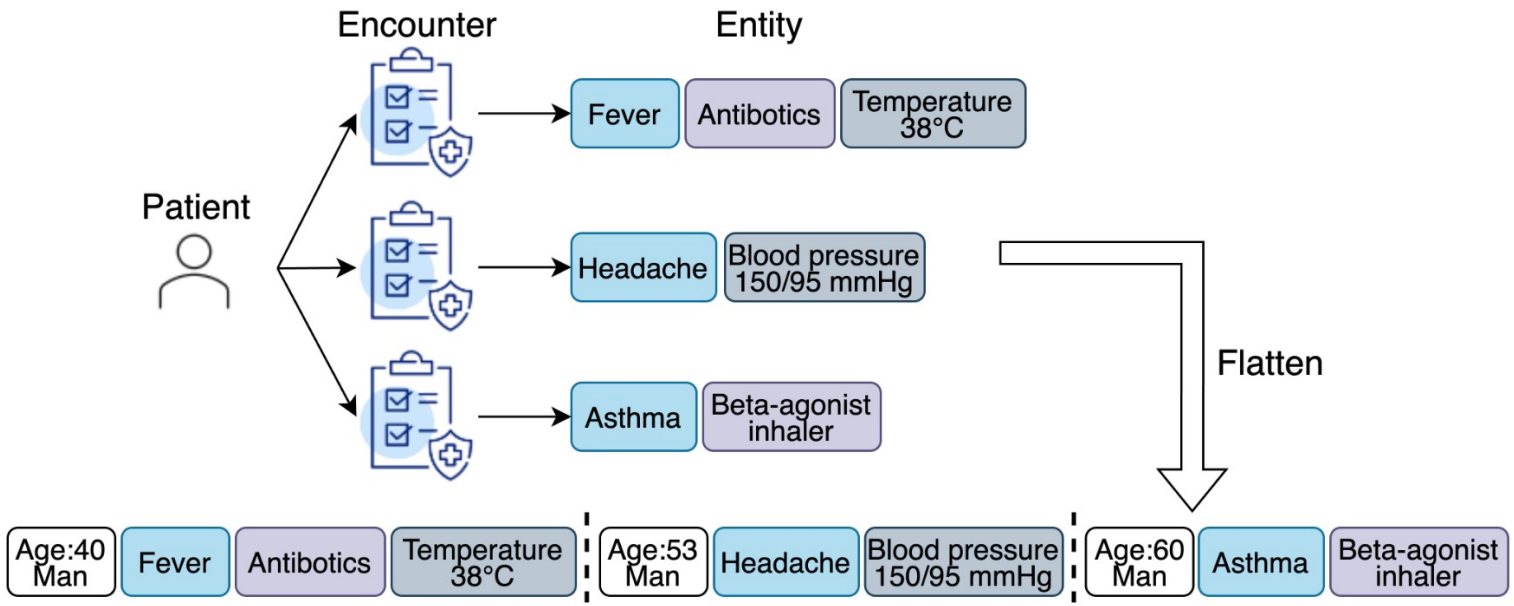
- Given a patient, we flatten the corresponding historical visits into several sequences of entities V
  - Apply patient's demography D as [CLS] token and add it to each sequence

$$\left[\underbrace{[D_1, V_{1,1}, \cdots, V_{1,N_1}]}_{\text{1-th encounter}}, \cdots, \underbrace{[D_S, V_{S,1}, \cdots, V_{S,N_S}]}_{S\text{-th encounter}}\right]$$

  - $V_{i,n}$ is n-th token at i-th encounter
  - The combination of age and gender, e.g., "<20, man>", is treated as demography token

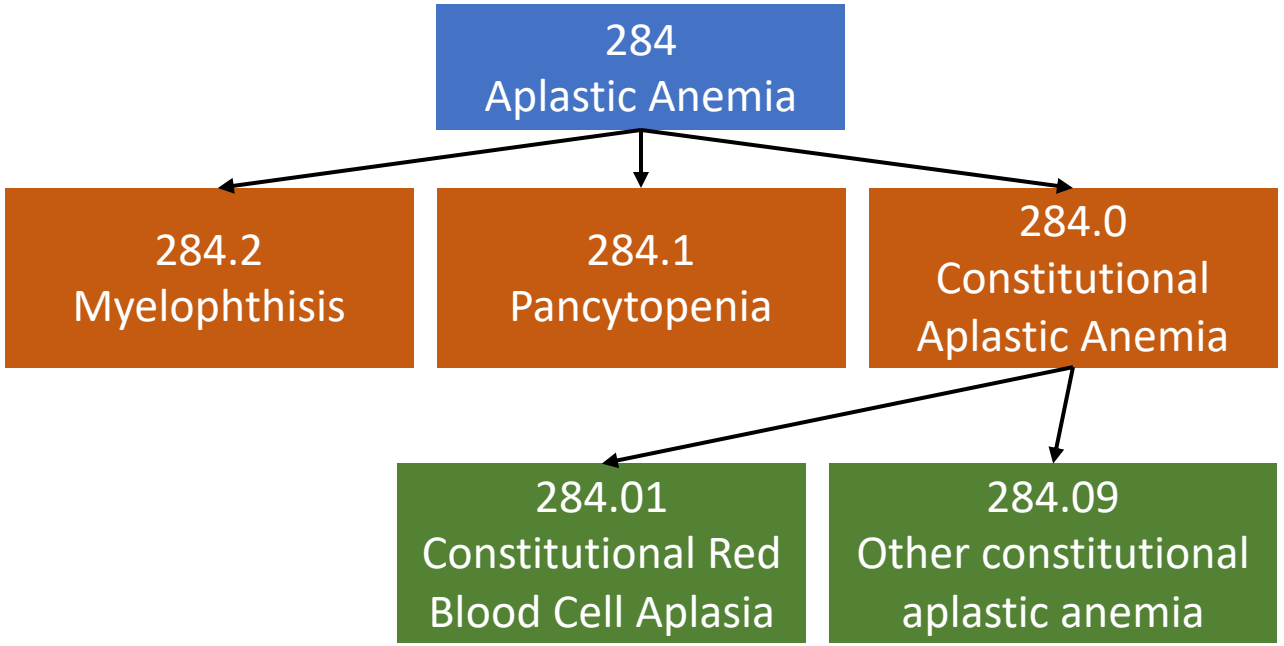
HEART: Input representation (2)

- An illustrative example



HEART: Code Structure Embedding (1)

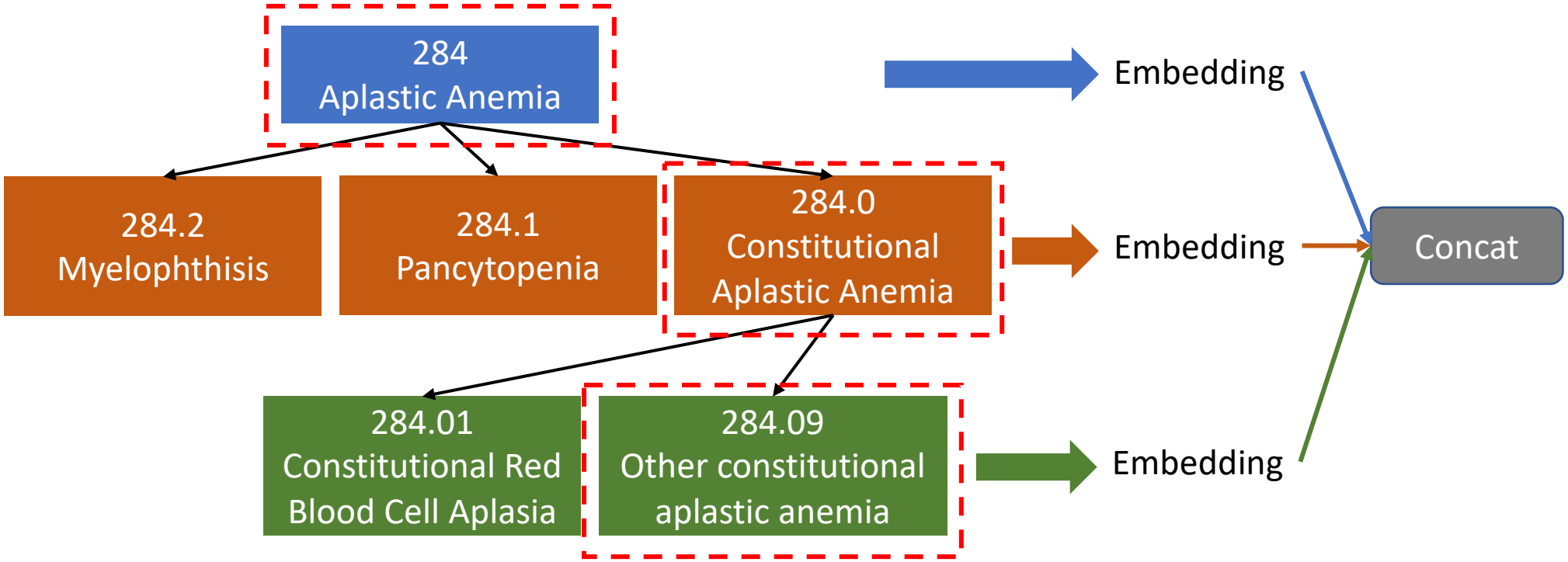
- Medical concepts often function as tree-like classification systems:



- Such code underscores the interconnectedness of medical concepts

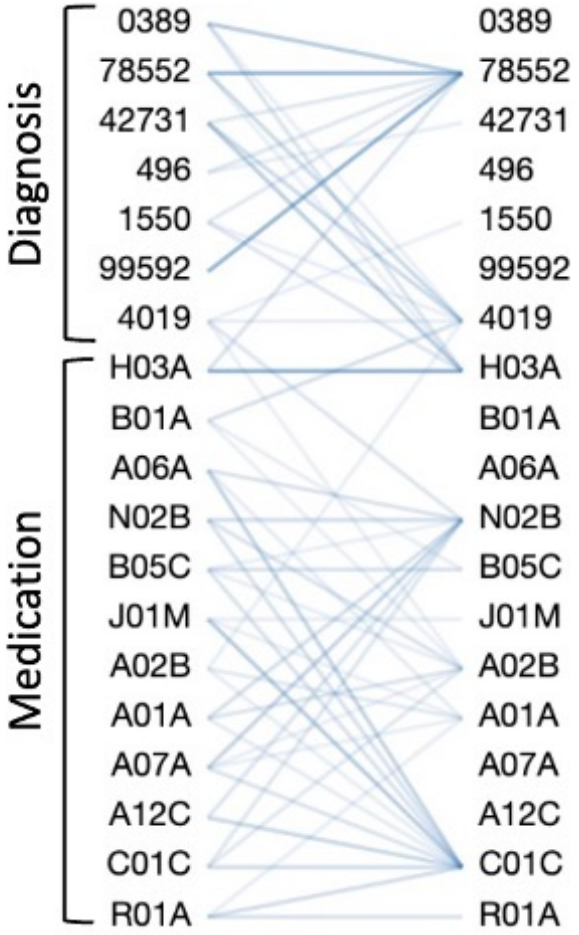
HEART: Code Structure Embedding (2)

- Given a diagnosis or medication entity, we divide its embedding into three segments, each associated with one of these code levels
  - Diagnosis: ICD-9, Medicine: ATC



HEART: Heterogeneous Relation Embedding (1)

- Encode relationship between entities with pairwise representation
- The input includes token embedding and pairwise representations
  - $N \times D$ and $N \times N \times D$



HEART: Heterogeneous Relation Embedding (2)

- We incorporate relation embeddings within model, which captures and encodes the heterogeneous correlation between entities
- For an entity pair (V_n, V_m) in the same encounter, we apply **type-specific** transformation:

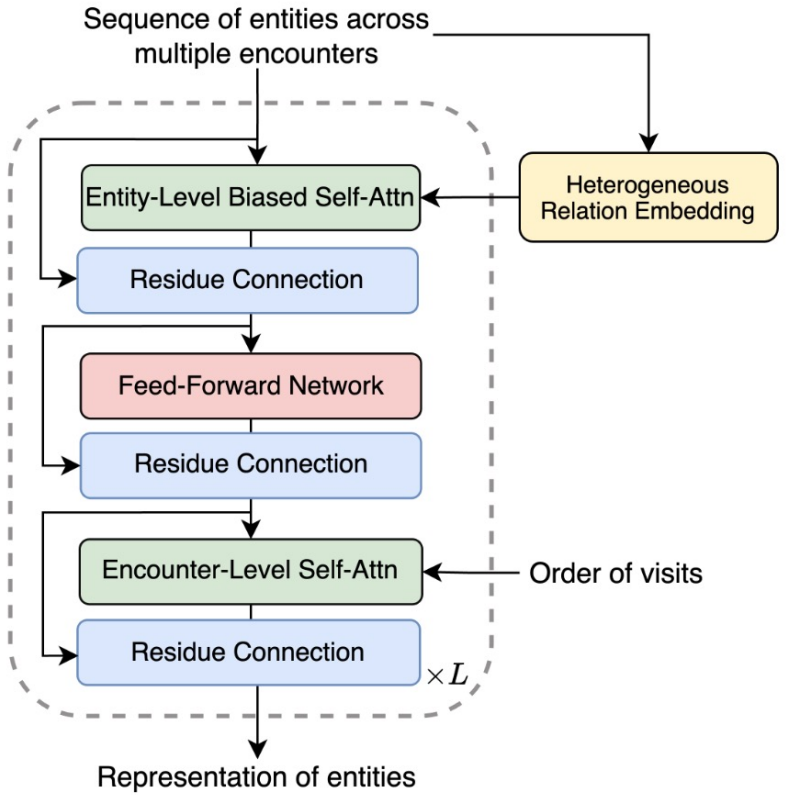
$$\boldsymbol{R}_n = \text{Linear}_{\tau(V_n)}(\boldsymbol{V}_n),\ \boldsymbol{R}_m = \text{Linear}_{\tau(V_m)}(\boldsymbol{V}_m)$$

  - Differentiating between the heterogeneous semantics of different type pairs
- Then we combine these two embeddings to generate the relational embedding for the entity pair

$$\boldsymbol{R}_{n \leftarrow m} = \text{Linear}(\boldsymbol{R}_n || \boldsymbol{R}_m)$$

HEART: Multi-level attention scheme

- Although more heterogeneous entities introduce rich semantic information, the **computational cost** increases quadratically
- Self-attention is alternately applied across these two levels, effectively increasing the receptive fields for each entity
  - Entity-level: Biased Self-Attention Module
  - Encounter-level: Self-Attention solely on Demography tokens



HEART: Entity-Level Biased Self-Attention Module

- As for entity-level context, we focus on encoding the entity within the same visit

$$\left[D', V'_1, \cdots, V'_N\right] = \text{Entity-Attn}\left([D, V_1, \cdots, V_N]\right)$$

- We incorporate the heterogeneous information from the relation embeddings, which serves as a bias term:

$$b_{nm} = \text{Linear}(\text{LN}(R_{n\leftarrow m}))$$

Attention score between n-th and m-th entities

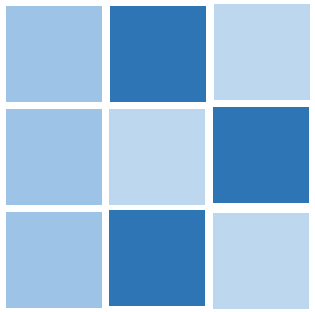
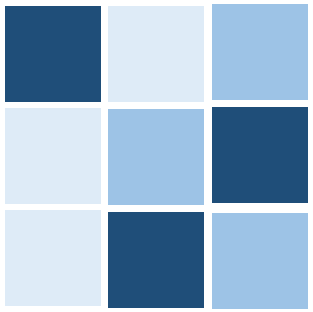
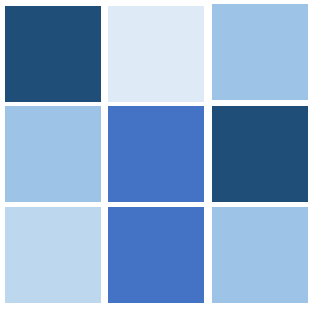
$$A_{nm} = \text{Softmax}_n\left(\frac{1}{\sqrt{d}} V_n^{qry^T} V_m^{key} + b_{nm}\right)$$

- Relation embedding will also serve as context

$$V'_n = \text{Linear}\left(\text{LN}\left(\sum_m A_{nm} V_m^{val} || \sum_m A_{nm} R_{n\leftarrow m}\right)\right) + V_n$$

HEART: Entity-Level Biased Self-Attention Module

- Applying relation embedding as bias can directly prioritize or deprioritize certain connections
  - Illustrative example:

Softmax( + ) = 

Algorithm 2: Entity-Level Biased Self-Attention Module

Input: Entity embeddings $[D, V_1, \cdots, V_N]$ from the same encounter, Relation embeddings for each entity pair $\{R_{n \leftarrow m} \mid n, m \leq N\}$.

Hyper-parameters: Hidden size d

Output: Updated entity embeddings $\left[D, V'_1, \cdots, V'_N\right]$

for *every entity* n **do**

  // QKV Linear transformer over all the entities.

  $V_n^{qry} = \text{Linear}^{qry}(V_n)$

  $V_n^{key} = \text{Linear}^{key}(V_n)$

  $V_n^{val} = \text{Linear}^{val}(V_n)$

end

for *every entity pair* (n, m) **do**

  // Calculate bias based on the relation embedding.

  $b_{nm} = \text{Linear}(R_{n \leftarrow m})$

  // Calculate the biased attention map.

  $A_{nm} = \left(\frac{1}{\sqrt{d}} V_n^{qryT} V_m^{key} + b_{nm}\right)$

end

$A = \text{Softmax}(A)$

for *every entity* n **do**

  // Aggregation.

  $V'_n = \text{Linear}\left(\text{LN}(A_{nm} V_n^{val} || A_{nm} R_{n \leftarrow m})\right) + V_n$

end

return $\left[D, V'_1, \cdots, V'_N\right]$

HEART: Encounter-Level Self-Attention Module

- We limit the attention to demography tokens across encounters

$$\left[\boldsymbol{D}'_1, \cdots, \boldsymbol{D}'_S\right] = \text{Enc-Attn}\left([\boldsymbol{D}_1, \cdots, \boldsymbol{D}_S]\right)$$

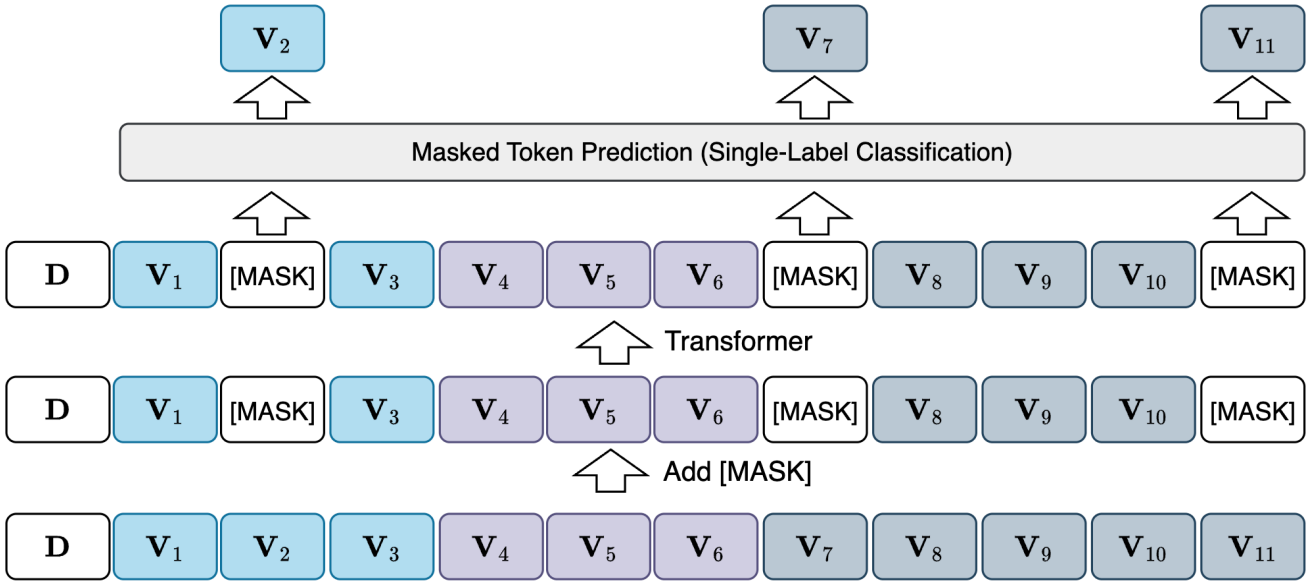
- Besides standard self-attention mechanism, we incorporate order of visits as a position encoding in the query and key transformations:

$$\boldsymbol{D}_i^{qry} = \text{Linear}(\boldsymbol{D}_i + \boldsymbol{t}_i), \boldsymbol{D}_i^{key} = \text{Linear}(\boldsymbol{D}_i + \boldsymbol{t}_i)$$

  - $\boldsymbol{t}_i$ is the embedding for i-th visit

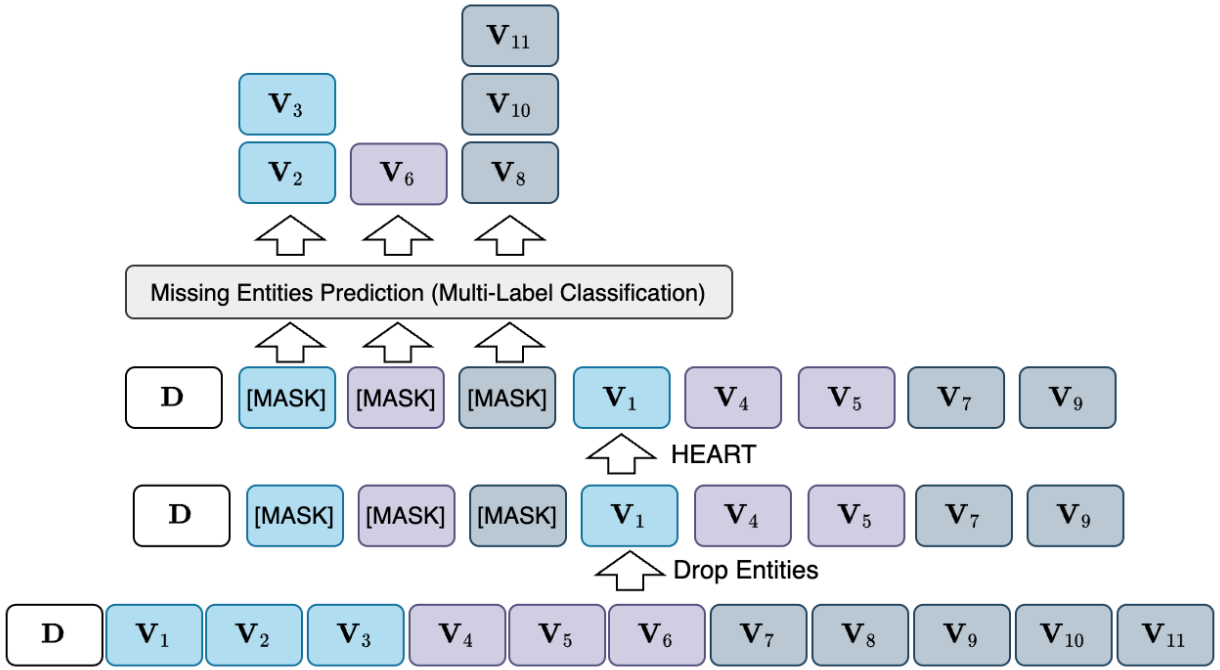
HEART: Missing Entity Prediction (1)

- Masked token prediction (MTP)
  - Replaces actual tokens with [MASK] and performs single-label classification at each masked position to predict the original token
  - **Position-dependent** and thus not suitable for EHR due to the unordered nature of medical entities



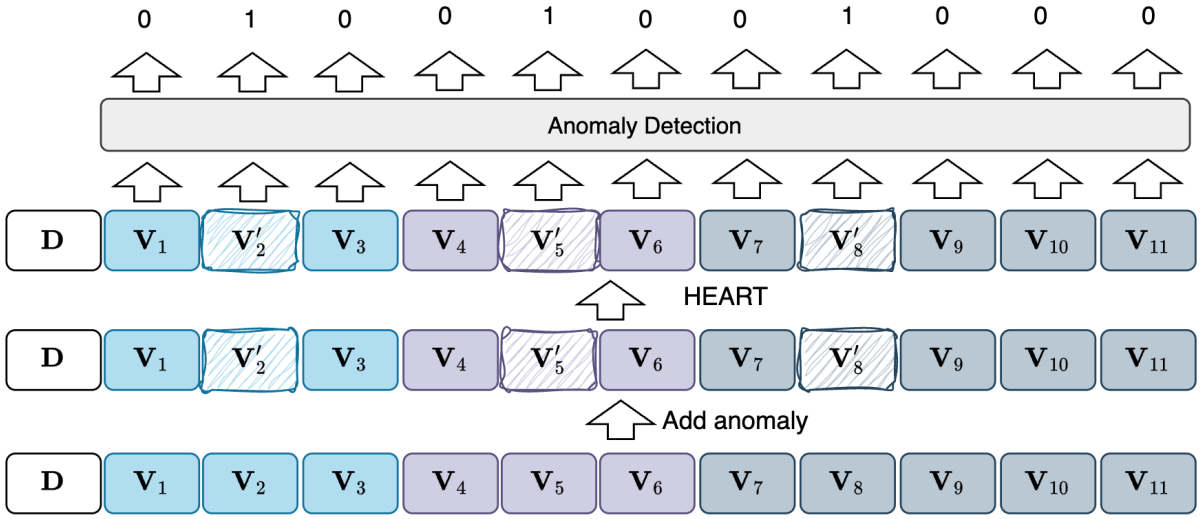
HEART: Missing Entity Prediction (2)

- Missing entity prediction (MEP)
  - Randomly drop some entities and let the model perform multi-label classification based on one [MASK] for each entity type
  - **Position-agnostic** and heterogeneity-aware



HEART: Anomaly Detection Task

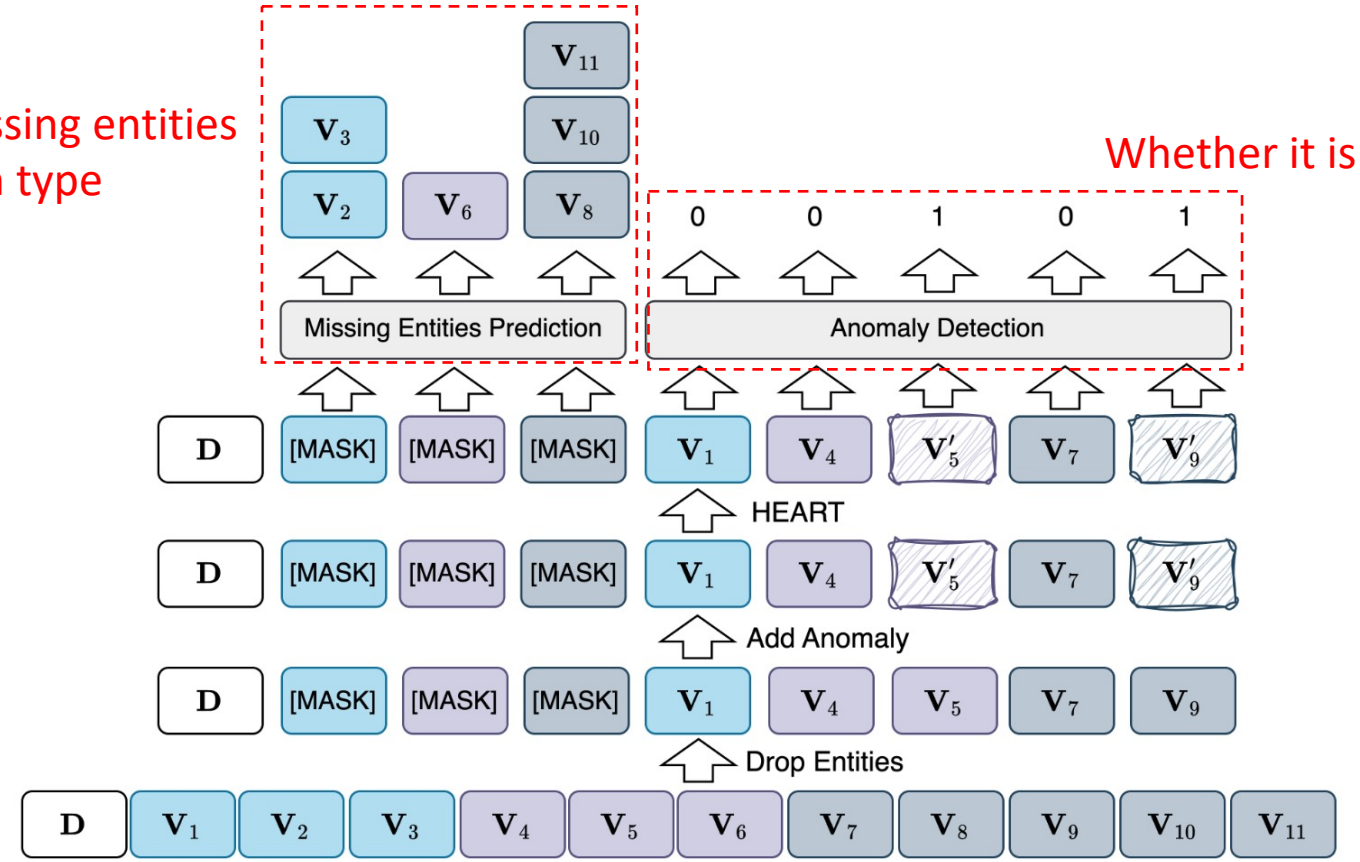
- EHR data often contain noise due to human input errors
  - Incorrect data can significantly impact billing and coverage decisions
- We incorporate anomaly detection as our pretrained task to encourage the model to identify unrelated entities given a context
  - Randomly replace some entities with the entities of the same type
  - Let the model predict whether the tokens are anomalies



HEART: Pretrained Tasks Overview

Predict the missing entities for each type

Whether it is anomaly data



Different colors represent different modalities

Experiments

- Two EHR datasets
  - MIMIC-III and eICU
- Pipeline
  - Pretrain the model with 70% patients
  - Finetune the model with only 20% of the remaining datasets and conduct evaluation
- Downstream tasks
  - Death prediction
  - Prolonged length of stay (PLOS) prediction
  - Readmission prediction
  - Next Diagnosis Prediction in 6/12 months

	MIMIC-III	eICU
#patients	33,582	86,290
#encounter	40,770	100,839
#diagnosis	1,998	903
#medication	149	2044
#procedure	801	n/a
#lab	1,500	756
avg # of visits	1.214	1.169
max # of visits	15	24
avg # of diagnosis	10.765	3.344
avg # of medication	11.248	24.036
avg # of procedure	4.459	n/a
avg # of lab	41.592	39.779

Experiments

- Comparison results
  - On the eICU dataset, the model can only perform death and PLOS prediction since it lacks sequential information

Dataset	Task	G-BERT		BEHRT		Med-BERT		ExBEHRT		HEART		
		F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	
MIMIC-III	Death	0.6452	0.8546	0.6491	0.8449	0.6606	0.8543	0.6338	0.8420	**0.7238**	**0.9029**	↑ 9.5%
	PLOS	0.6957	0.6988	0.6852	0.7081	0.7240	0.7617	0.7410	0.8036	**0.7572**	**0.8222**	↑ 2.2%
	Readmission	0.6043	0.6473	0.6475	0.7563	0.6631	0.7559	0.6511	0.7514	**0.6968**	**0.7881**	↑ 5.0%
eICU	Death	0.7580	0.9059	0.6287	0.8039	0.6069	0.7950	0.6286	0.8049	**0.8049**	**0.9333**	↑ 6.2%
	PLOS	0.6997	0.8240	0.5060	0.6356	0.4423	0.6355	0.5261	0.6509	**0.7125**	**0.8373**	↑ 1.8%

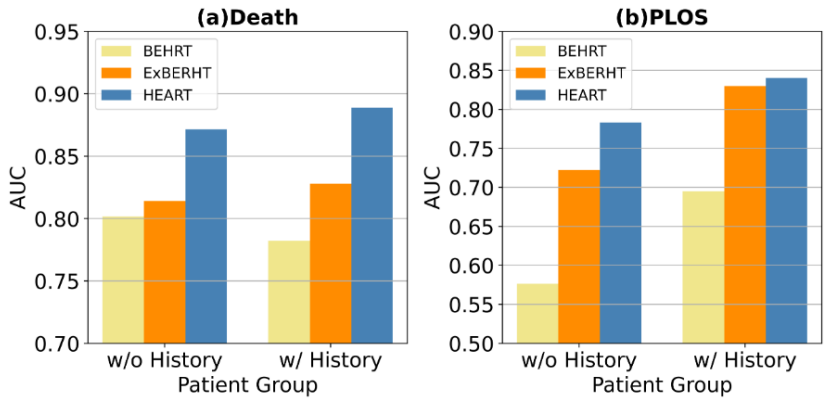
Dataset	Task	G-BERT		BEHRT		Med-BERT		ExBEHRT		HEART		
		Precision	PRAUC	Precision	PRAUC	Precision	PRAUC	Precision	PRAUC	Precision	PRAUC	
MIMIC-III	6 Months	0.2668	0.1490	0.3637	0.1733	0.4225	0.1854	0.2006	0.1541	**0.4867**	**0.1911**	↑ 15.1%
	12 Months	0.3382	0.1666	0.4019	0.1855	0.3803	0.1906	0.3015	0.1682	**0.4546**	**0.2004**	↑ 13.1%

Further Study

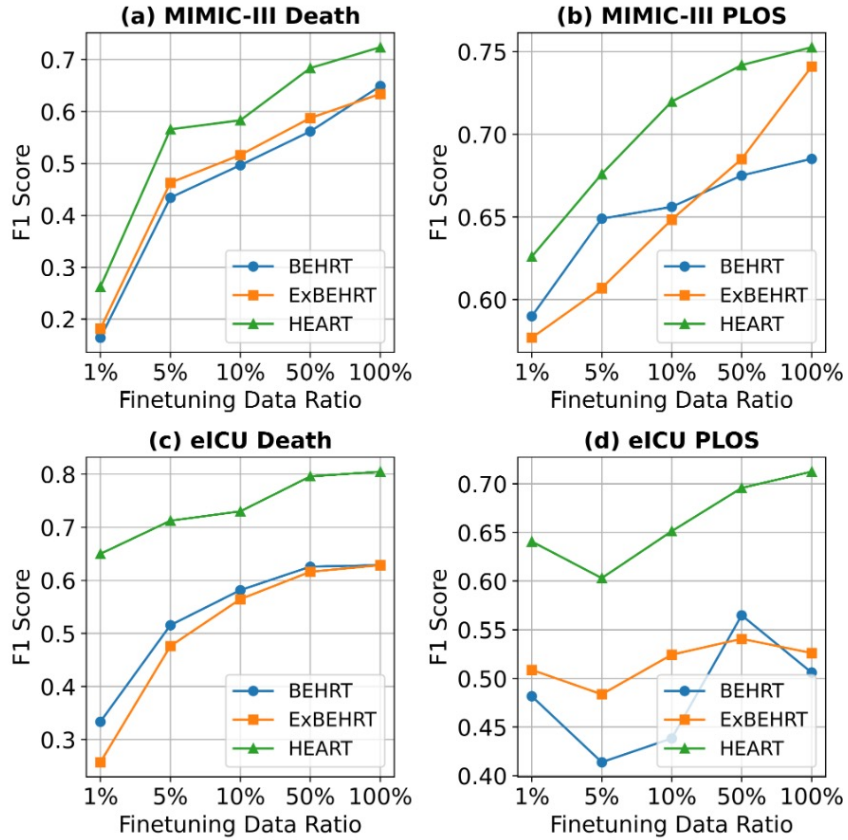
- Ablation study

	Death		PLOS		Readmission		Diagnosis (6M)	
	F1	AUC	F1	AUC	F1	AUC	Precision	PRAUC
HEART	0.7238	0.9029	0.7527	0.8222	0.6968	0.7881	0.4867	0.1911
w/o Relation Embedding	0.6707	0.8510	0.7238	0.7744	0.6758	0.7791	0.4455	0.1867
w/o Multi-Level Attention	0.7217	0.9014	0.7518	0.8185	0.6079	0.7492	0.4219	0.1849
w/o Anomaly Detection	0.7138	0.8965	0.7439	0.7879	0.6889	0.7813	0.4413	0.1830

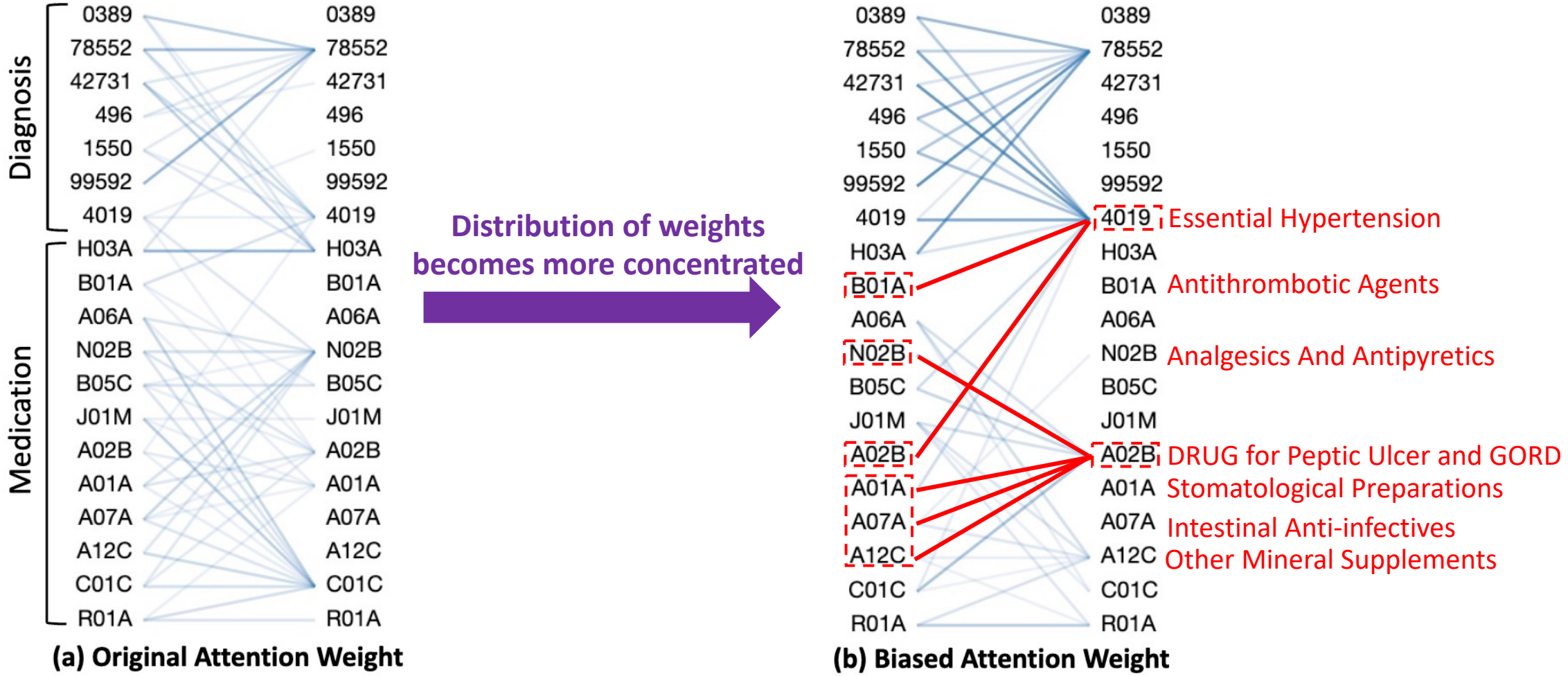
- Performance on different patient groups
  - Patients w/ and w/o past encounters



- Performance with different finetuning data ratio



Case Study on Relation Embedding



Conclusion

- We propose HEART, a pretrained language model for structured EHR data
  - 1) HEART seamlessly integrates heterogeneous medical entity information and captures their pairwise relationships through relation embedding
  - 2) An efficient encoding scheme is proposed
  - 3) The model is pretrained with two tasks, which is more suitable to EHR
  - 4) Results on 5 downstream tasks using 2 datasets demonstrate its effectiveness