FROM PREDICTION TO FOUNDATION: DEEP LEARNING MODELS FOR PATIENT CARE OPTIMIZATION



Jon Sege & Vincent Pan

White Plains Hospital

WHO WE ARE



Jon Sege

- AVP, Data Management & Analytics
- M.S. Environmental Engineering, UC Berkeley
- Environmental Science, Digital Media, Healthcare
- Passionate about data for public health impact



Vincent Pan, CFA

- Data Scientist
- M.S. Data Science, Columbia University
- Financial, Reinsurance, Private Equity, Healthcare
- Passionate about using data for good

WHAT WE DO

Our scope of practice

- Advanced analytics
- Data science
- Infrastructure
- Deployment

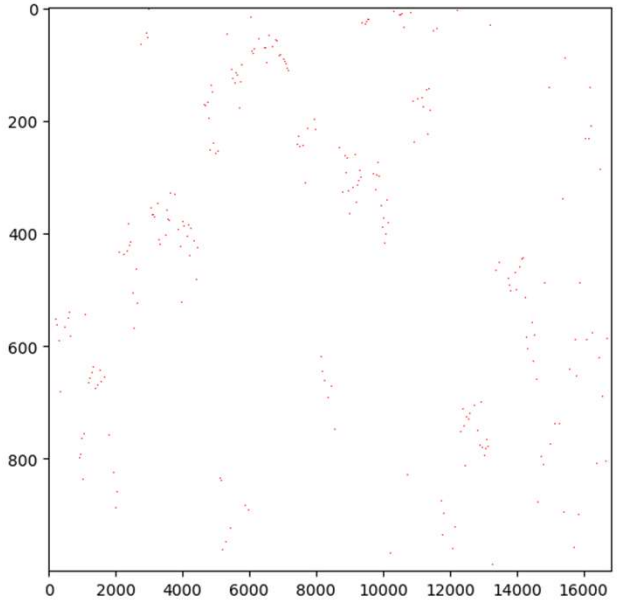
Our stakeholders

- Executive leadership
- Clinical leadership
- Quality & PI teams

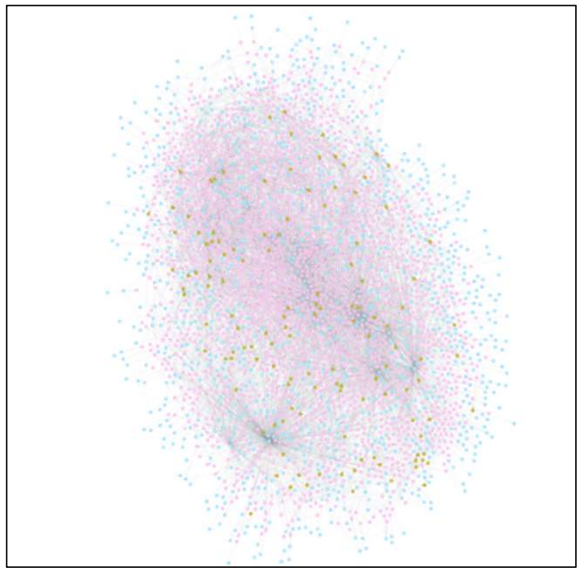
What? ➡ Why? ➡ What Next?

WHY DEEP LEARNING IN HEALTHCARE

Feature Density
Diagnoses



Relationships
Diagnoses & Procedures

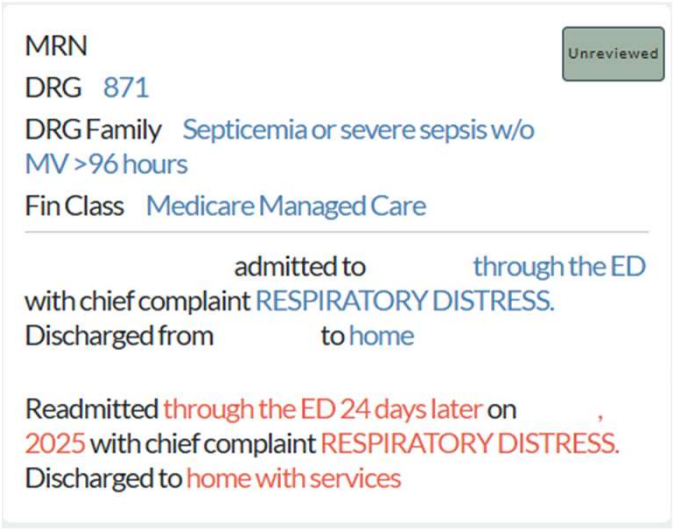


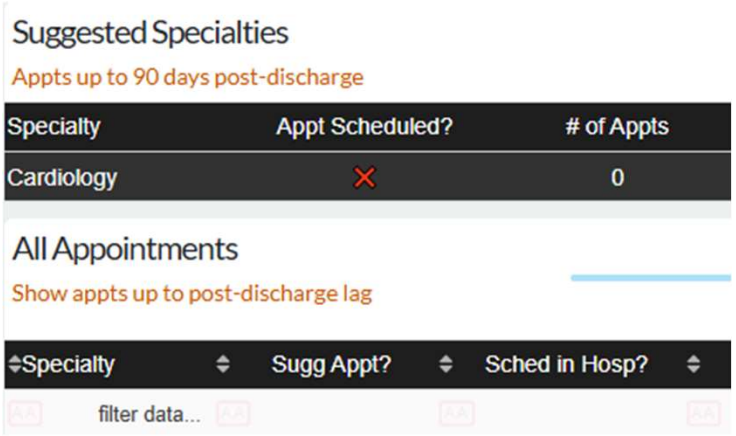
- Handles High-Dimensional Data
- Automates Feature Engineering
- Scales with Big Data
- Learns from Temporal Sequences

USE CASE EXAMPLE: PREDICT SPECIALISTS

- Anticipate Specialist Needs
- Enhance Care Coordination
- Improve Clinical Decision Support
- Optimize Workflow and Staffing
- Right Care, Right Time

USE CASE EXAMPLE: PREDICT SPECIALISTS

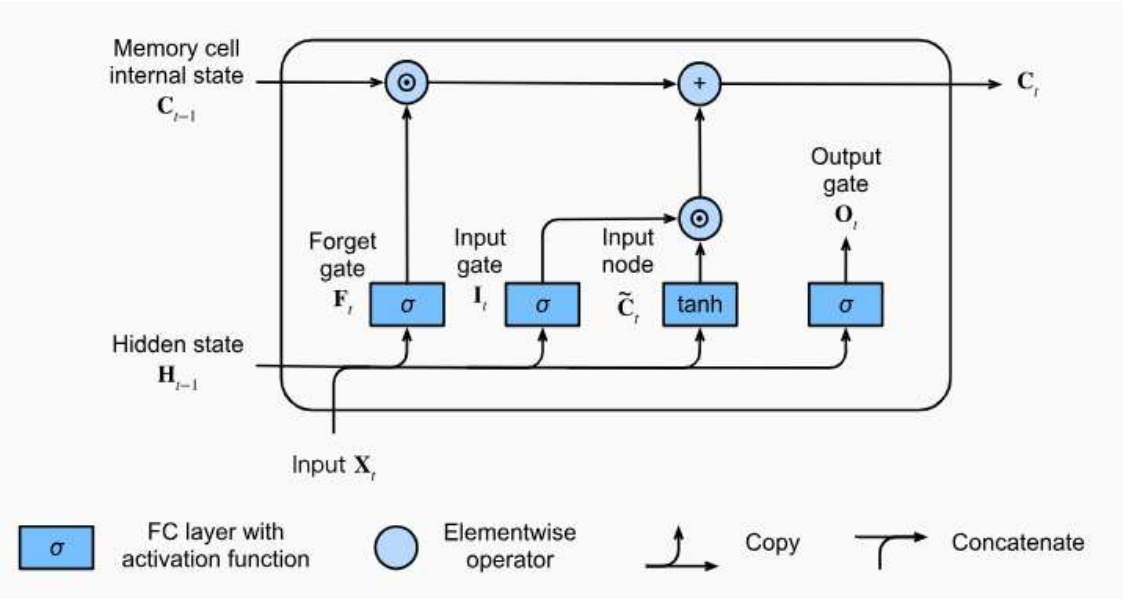




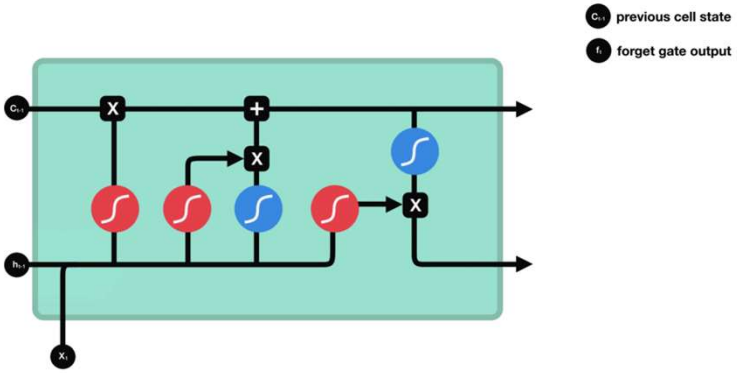
- Pt admitted for sepsis. LSTM model suggested Cardiology follow-up
- Subsequent case 24 days later with primary diagnosis NSTEMI heart attack
- Between the patient's discharge date and readmission date 257 Cardiology New Patient appointments were scheduled within the service area, incl. 63 no-shows / cancellations

Why LSTM (Long Short-Term Memory)

- Supports Multi-Label Output
- Handles Variable-Length Sequences
- Models Clinical Code Sequences
- Captures Long-Term Dependencies
- Learns Contextual Patterns in Codes



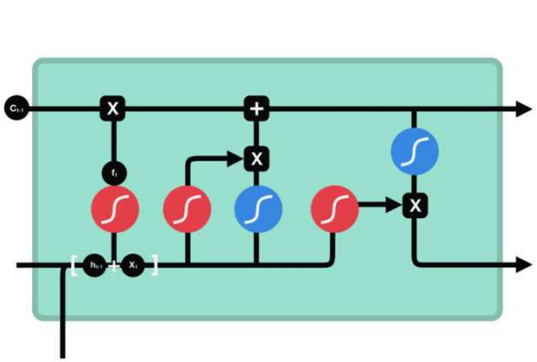
Forget gate



previous cell state

forget gate output

Input Gate



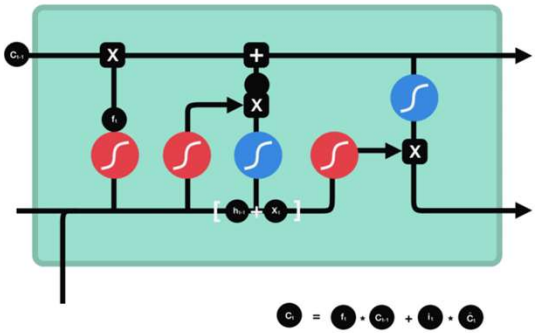
previous cell state

forget gate output

input gate output

candidate

Cell State



$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

previous cell state

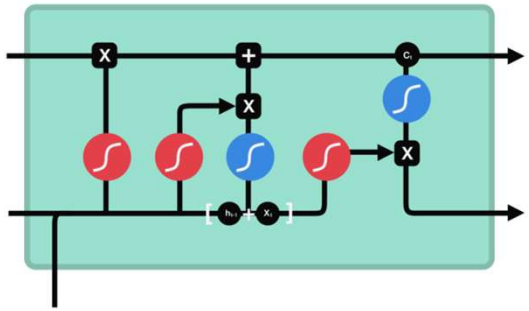
forget gate output

input gate output

candidate

new cell state

Output Gate



previous cell state

forget gate output

input gate output

candidate

new cell state

output gate output

hidden state

MODEL TRAINING & PERFORMANCE

➢ Input vs Output

icd10_codes	specialties
[A41.9, C25.9, J18.9, E43, C77.2, C78.00, D63....	[Infectious Diseases, Surgery, Neurology, Pall...
[K35.80, Q21.1, J45.909, I27.20, I10, M32.9, Z...	[Rheumatology, Surgery]
[C79.89, R18.0, J91.0, K70.30, C78.7, C80.1, R...	[Hematology]
[S82.852A, D68.51, D68.59, F41.9, Z86.718, M51...	[Orthopedic Surgery, Cardiology, Hematology an...
[N40.1, R33.8, M10.9, N18.30, Z86.718, M19.90,...	[Urology, Cardiology]
...	...
[A41.89, B27.89, R65.20, J03.80, B27.90]	[Surgery]
[I48.0, I24.89, I10, I48.92, Z96.612, E78.5, Z...	[Cardiology]
[M54.12, D50.9, F41.9, K21.9, R79.89]	[Neurology, Cardiology]
[I63.81, Q21.12, I66.01, E86.1, E66.9, I10, G4...	[Neurology]
[T85.590A, E87.20, K83.01, K83.1, E83.52, H90....	[Hepatology]

➢ Framework



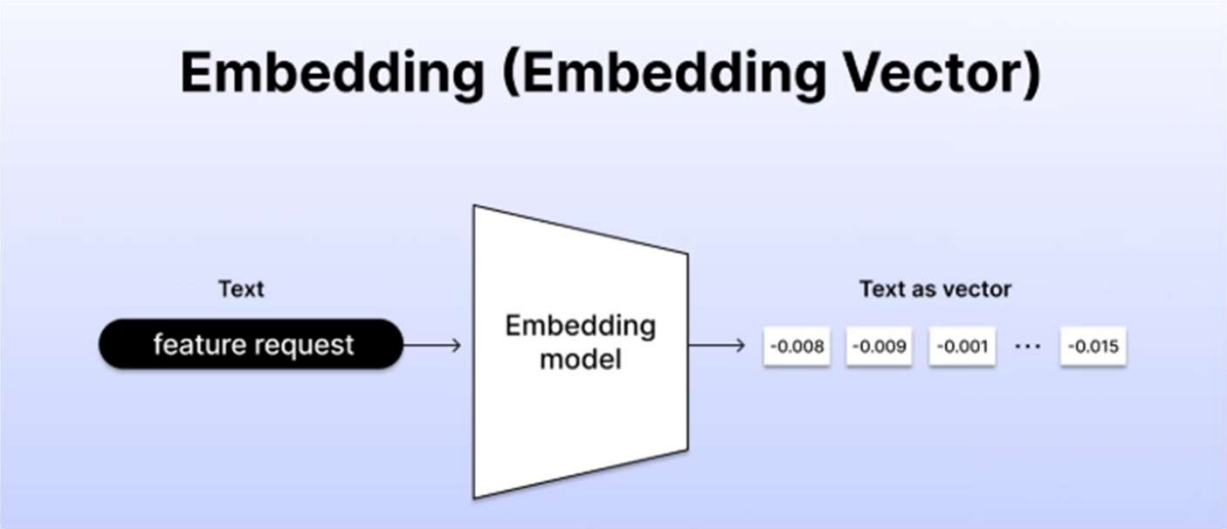
➢ Evaluation Metric

- Jaccard Index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- Training set: 0.65
- Validation set: 0.64

WHY EMBEDDINGS



- Capture Semantic Relationships
- Reduce Dimensionality
- Enable Transfer Learning
- Improve Generalization

MASKED CODE MODELING (MCM) (*GCT-STYLE BERT)*

Reference

Choi, E., *et al.* (2018)

Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019)

BERT: Pre-training of Bidirectional Transformers for Language Understanding

Concept

Graphical Structure in Electronic Health Records (EHR)

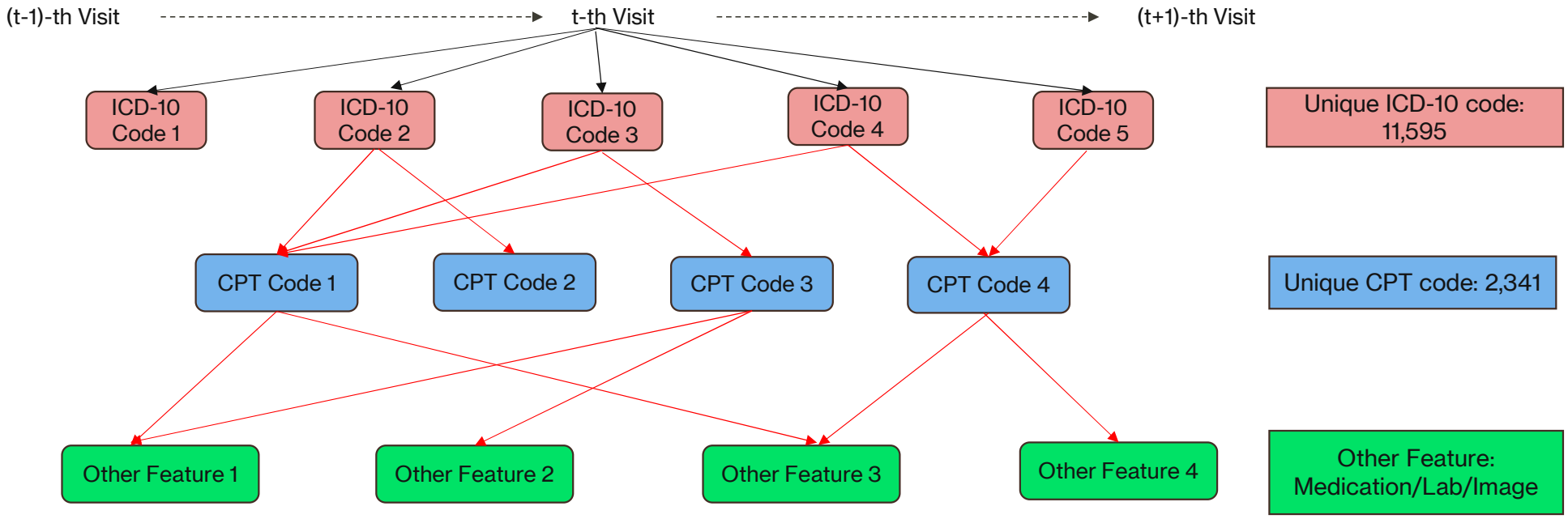
- Captures complex interactions between diagnoses, procedures and treatments.
- Helps model temporal and relational dependencies across visits.

Model

Graph Convolutional Transformer (GCT)

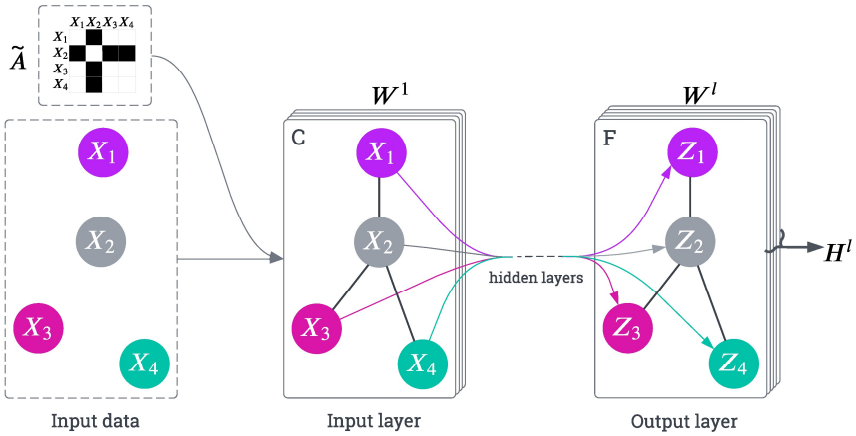
- Learns latent graph structure among clinical codes.
- Enhances contextual representation using relational attention.

GRAPHICAL STRUCTURE OF A VISIT

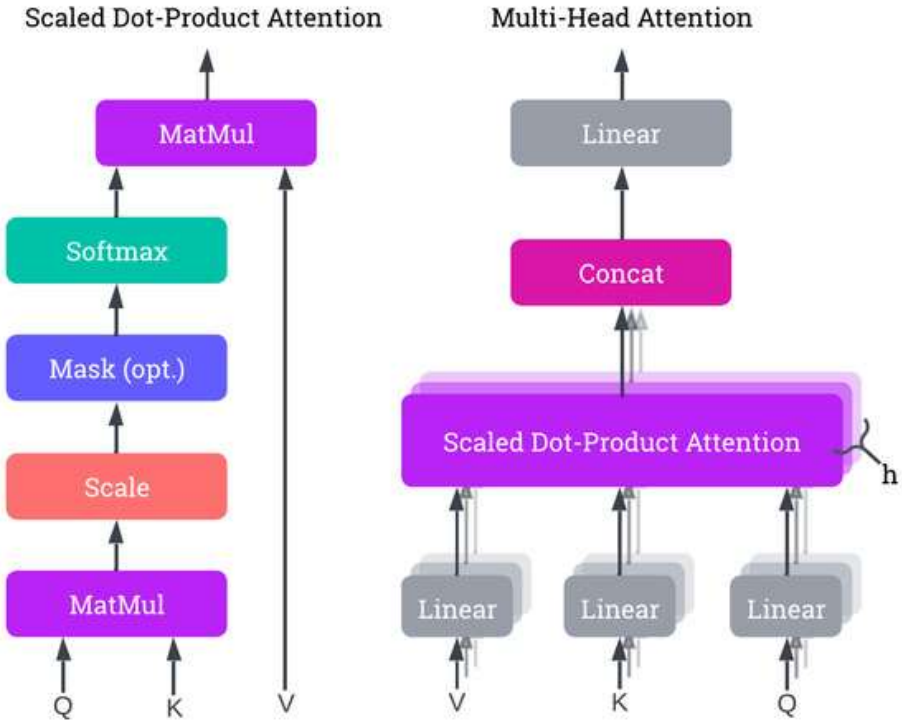


GCT ARCHITECTURE

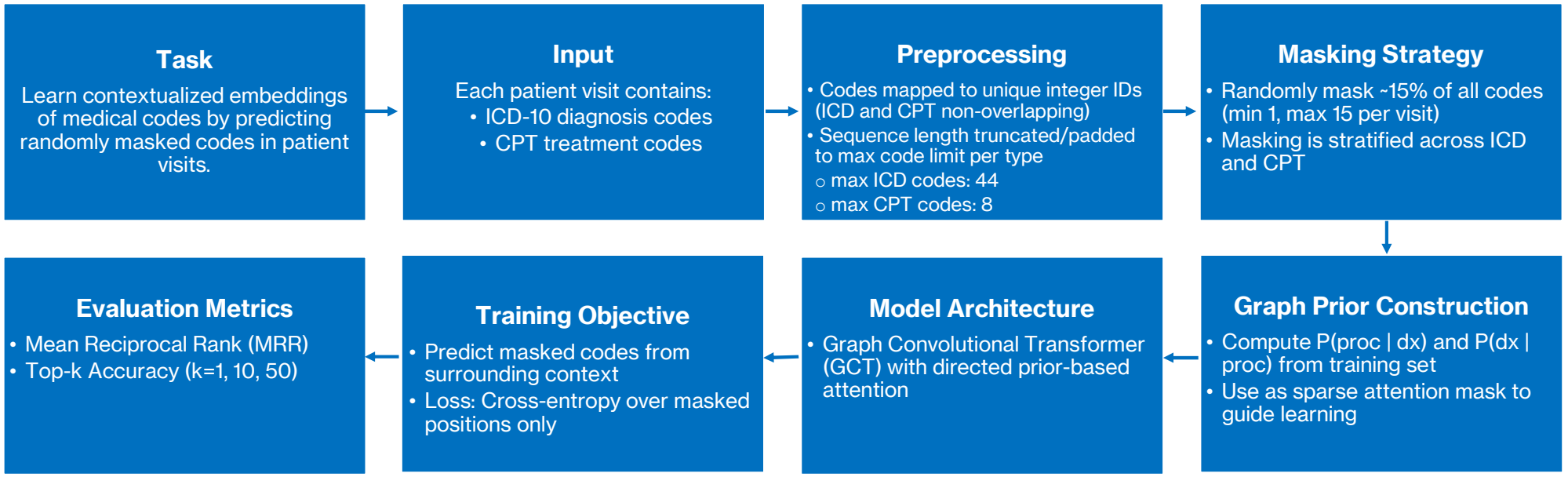
Graph Convolutional Networks (GCN)



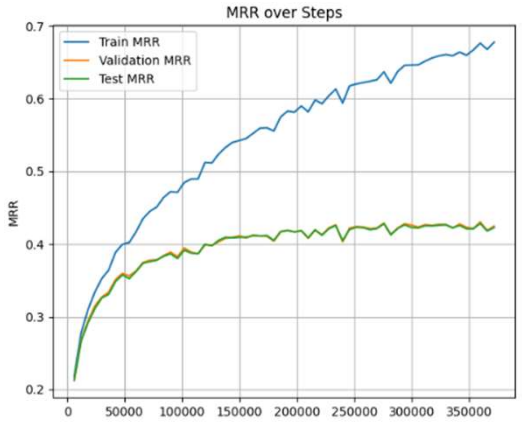
Transformer Architecture

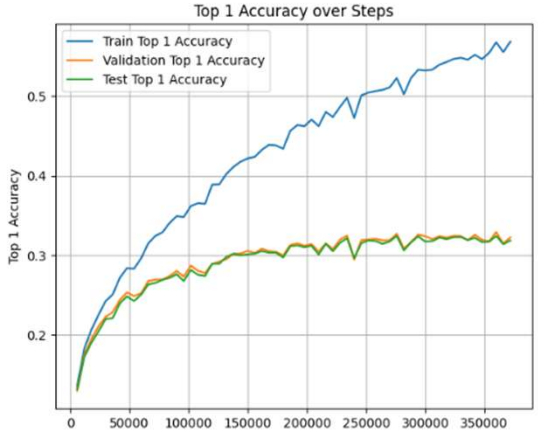


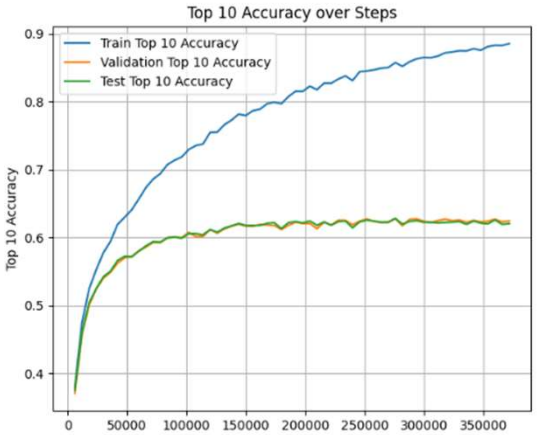
MASKED CODE MODELING (MCM)

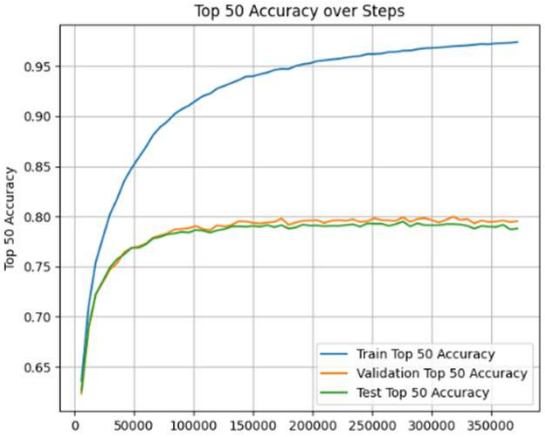


MCM PERFORMANCE





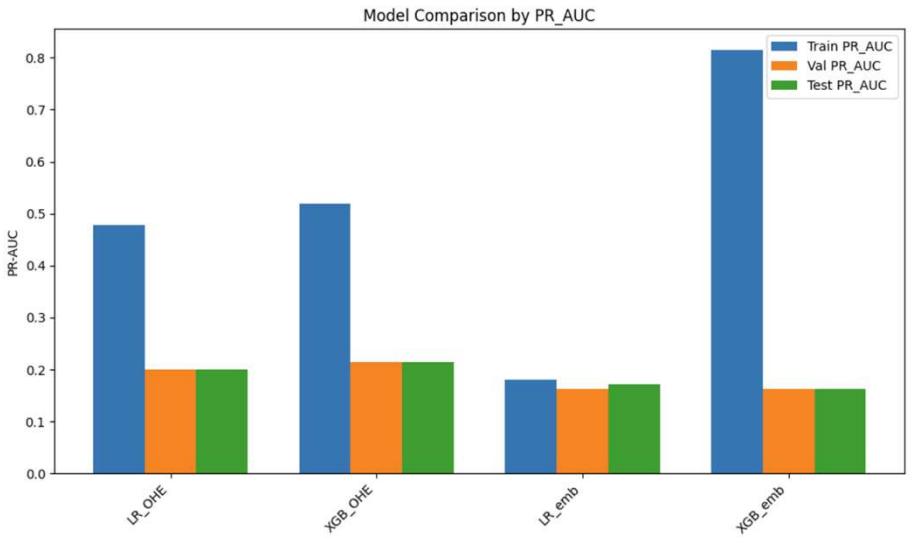




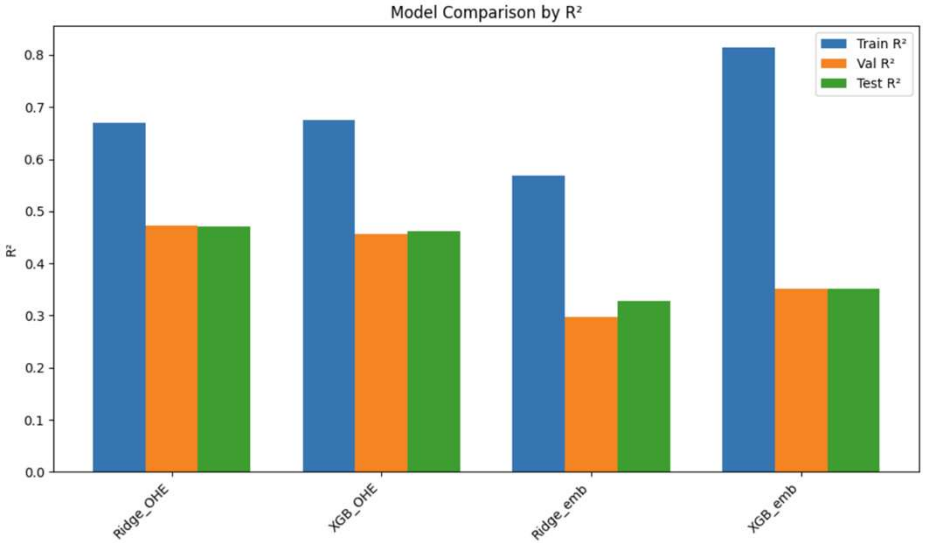
- **Training Summary**
  - Total Iterations: 375,000
  - Checkpointing: Model weights and metrics saved every 6,000 iterations
- **Evaluation Metrics**
  - Mean Reciprocal Rank (MRR)
  - Top-K Accuracy: Top-1, Top-10, Top-50
- **Final Model (Checkpoint @ 276,000)**
  - Mean Reciprocal Rank (MRR): 0.43
  - Top-1 Accuracy: 32.70%
  - Top-10 Accuracy: 62.76%
  - Top-50 Accuracy: 79.93%
- **Baseline / Random Chance**
  - Top-1 Accuracy ≈ 1/13,000 ≈ 0.0077%

MCM VALIDATION

Readmission



Length of Stay (LOS)



LIMITATIONS OF MCM-ONLY PRETRAINING

- **Task-Specific Embeddings**

Embeddings are tightly coupled to the code reconstruction objective, limiting transferability to clinical prediction tasks

- **Shallow Clinical Semantics**

Lack of visit-level supervision restricts the ability to capture patient context and trajectory

- **Weak Downstream Alignment**

Performance on predictive tasks like readmission and LOS indicates limited utility of learned representations

- **No Outcome Guidance**

Absence of supervised signals during training means embeddings are not optimized for clinical relevance

MULTI-TASK MODELING (HERMÈS)

Heterogeneous-task
Embedding
Representations with
Multi-loss
Equilibration via
Structured attention

- **Solution: Multi-Task Learning**
- **Multi-Task Setup**
  1. **Masked Code Prediction** – Predict randomly masked medical codes
  2. **Readmission** – Binary classification
  3. **Mortality** – Binary classification
  4. **Length of Stay (LOS)** – Regression (in days)
  5. **ESI Level** – Multi-class classification (6 levels)
  6. **Specialty Prediction** – Multi-label classification (e.g., cardiology, oncology)

KEY CHALLENGES

- **The Multi-Task Challenge**
  - Different tasks produce losses on very different scales
  - Need dynamic balancing to ensure all tasks learn effectively
- **Equal Weight**
  - Assigns the same fixed weight to every task's loss.
  - Simple and requires no extra parameters.
  - **Limitation:** Ignores loss scale differences – high-magnitude losses dominate, leading to imbalance.
- **Homoscedastic Uncertainty Weighting (HUW)**
  - *Kendall et al., "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," CVPR 2018*
  - Introduce one learnable log-variance $\log\sigma_i$ per task

$$\mathcal{L} = \sum_{i\in\text{reg}} \left[\frac{1}{2\,\sigma_i^2}\,\mathcal{L}_i^{\text{MSE}} + \frac{1}{2}\log(\sigma_i^2) \right] + \sum_{j\in\text{cls}} \left[\frac{1}{\sigma_j^2}\,\mathcal{L}_j^{\text{CE}} + \log(\sigma_j) \right]$$

  - L_i^{MSE} is the mean-squared error for regression.
  - L_i^{CE} is the cross-entropy for classification.

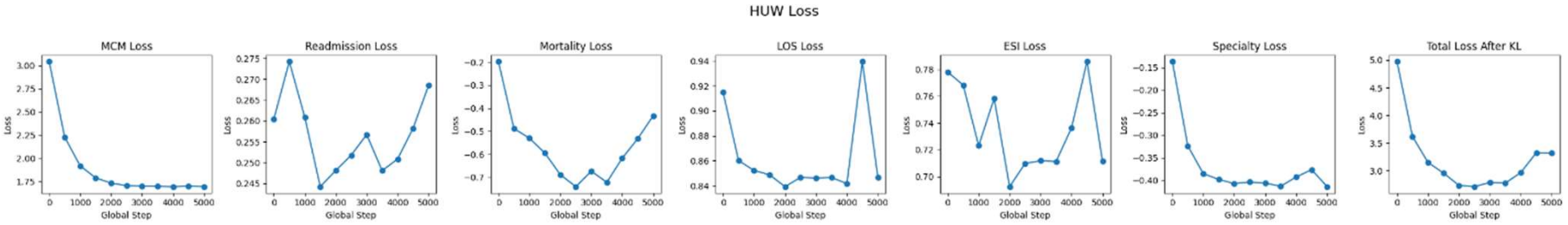
  - Weight each task's loss L_i by $\frac{1}{\sigma_i^2}$
  - Tasks with **higher uncertainty** get **down-weighted** automatically.
  - **Limitation** : Can still face gradient conflicts and risk loss explosion when σ becomes very small.

SOLUTION

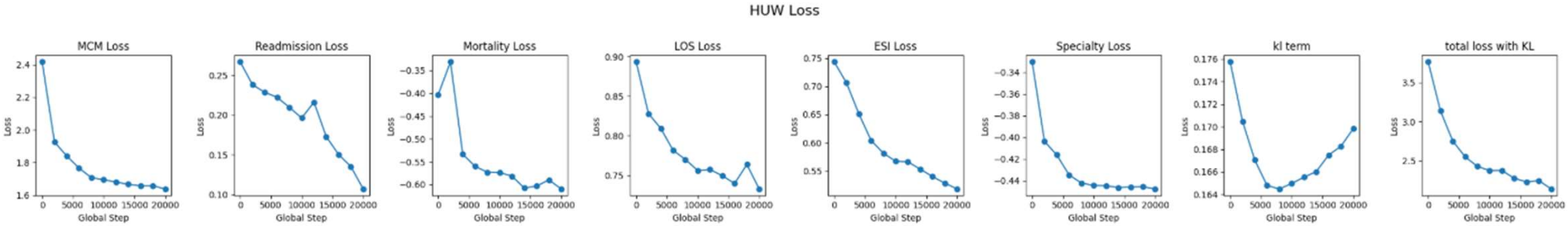
- **Gradient Normalization (GradNorm)**
  - *Chen et al., "GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," ICML 2018*
  - Introduce trainable per task weights w_i applied to each L_i.
  - Compute per-task gradient norms $g_i = \|\nabla_\theta(\omega_i L_i)\|$ w.r.t. shared parameters.
  - Define a target norm $\widetilde{g_i} = \bar{G}\left(\frac{L_i}{\bar{L}}\right)^\alpha$ where $\bar{G}$ = avg. of all g_i, $\bar{L}$ = avg. of all L_i, and α tunes rigidity.
  - Minimize $\Sigma_i |g_i - \widetilde{g_i}|$ to equalize training rate across tasks.
  - Dynamically adjusts w_i so all tasks contribute equally during optimization.
- **HUW + GradNorm**
  - HUW handles **static** scale differences via uncertainty-based weighting.
  - GradNorm corrects **dynamic** imbalance by aligning gradient norms.
  - Together, they mitigate gradient conflicts and improve stability in multi-task learning.
- **KL Divergence Regularization**
  - Encourage consistency in attention distribution across layers.
  - Add a regularization term to the total loss. : $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{multi-task}} + \lambda_{\text{KL}} \cdot \mathcal{D}_{\text{KL}}(A_i \| A_{i+1})$
  - The coefficient λ controls the regularization strength, balancing fit to data with smoothness in attention pattern.

GRADIENT CONFLICT

HUW only

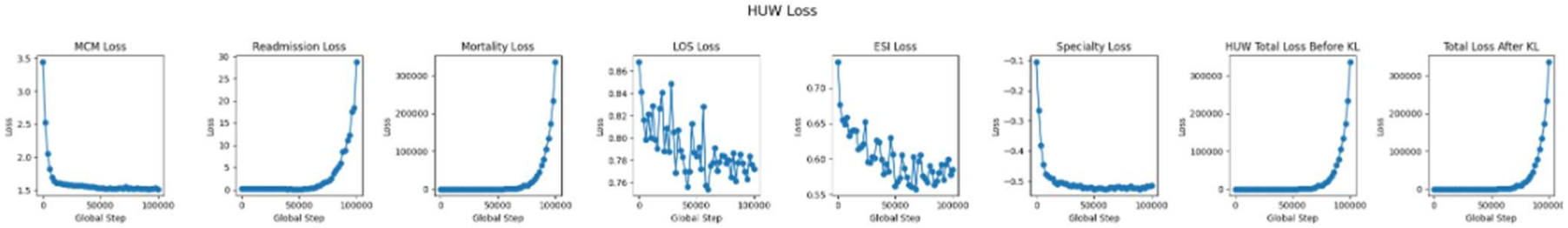


HUW + GradNorm + KL

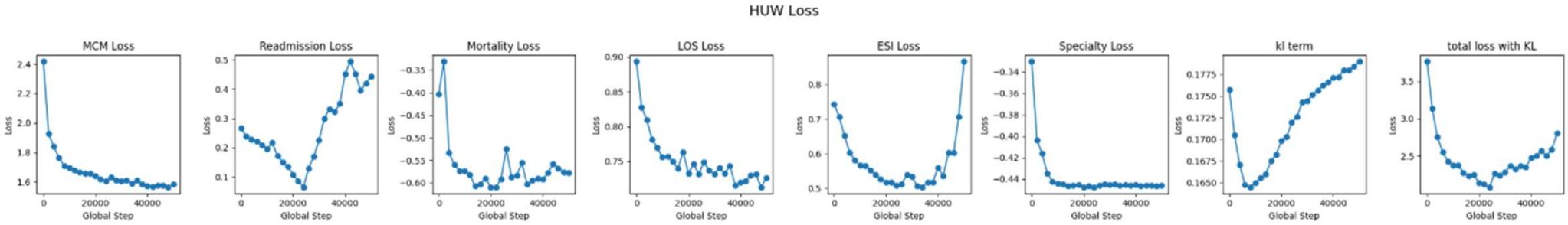


LOSS EXPLOSION

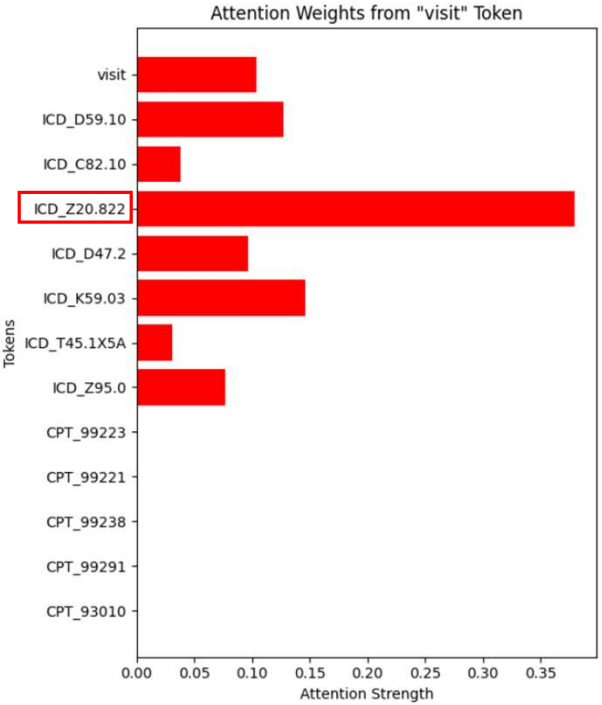
HUW only

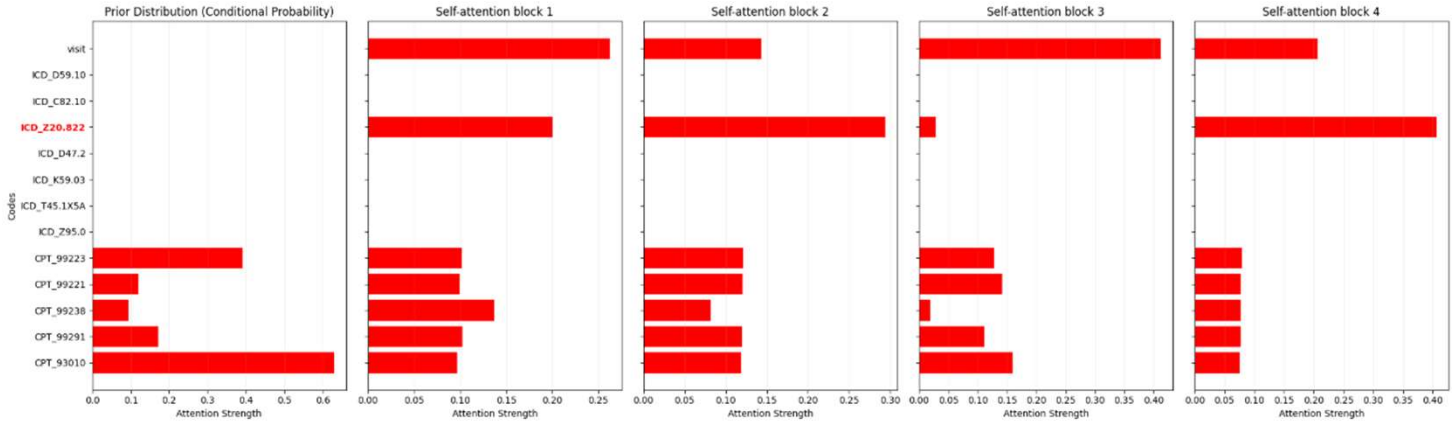


HUW + GradNorm + KL



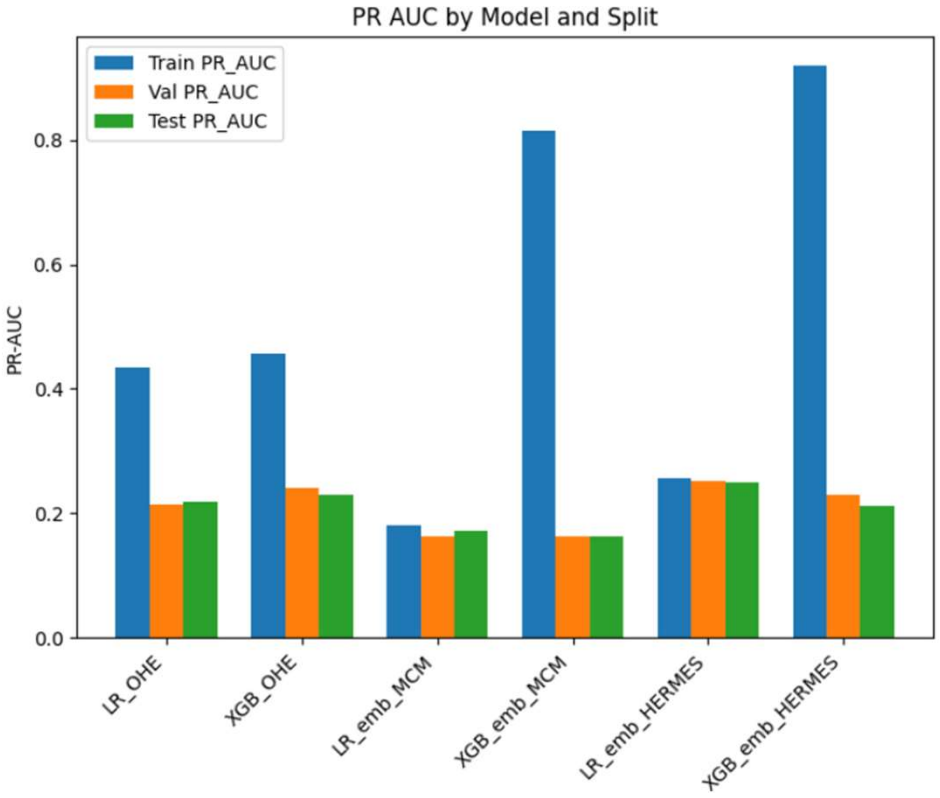
GRAPH RECONSTRUCTION WITH ATTENTIONS



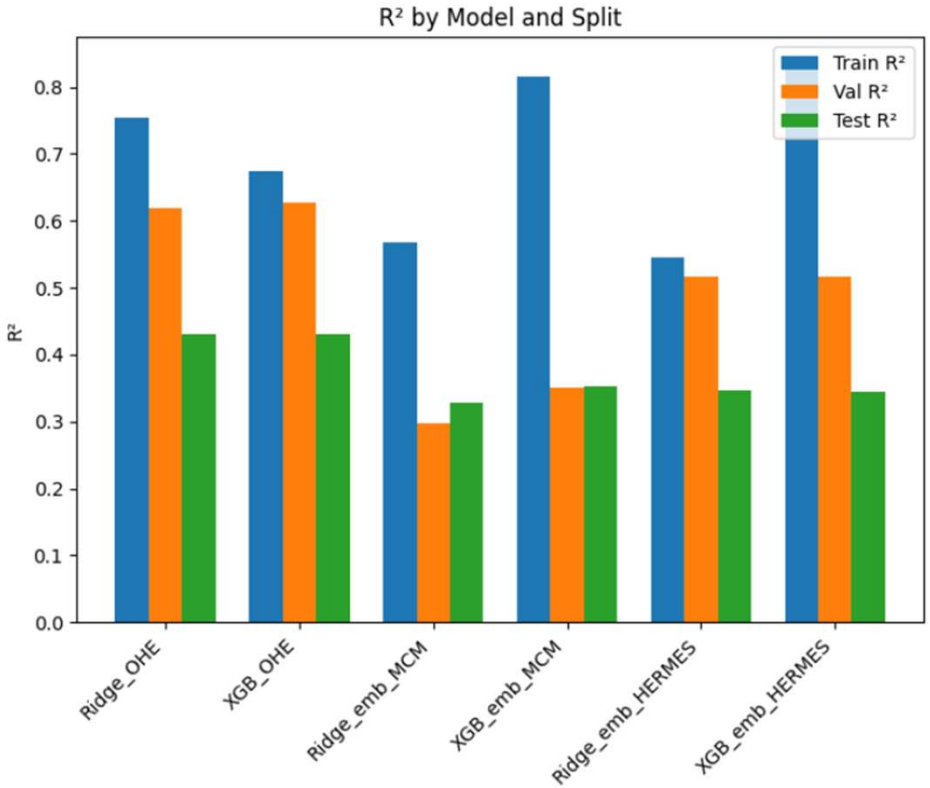


HERMÈS VALIDATION

Readmission

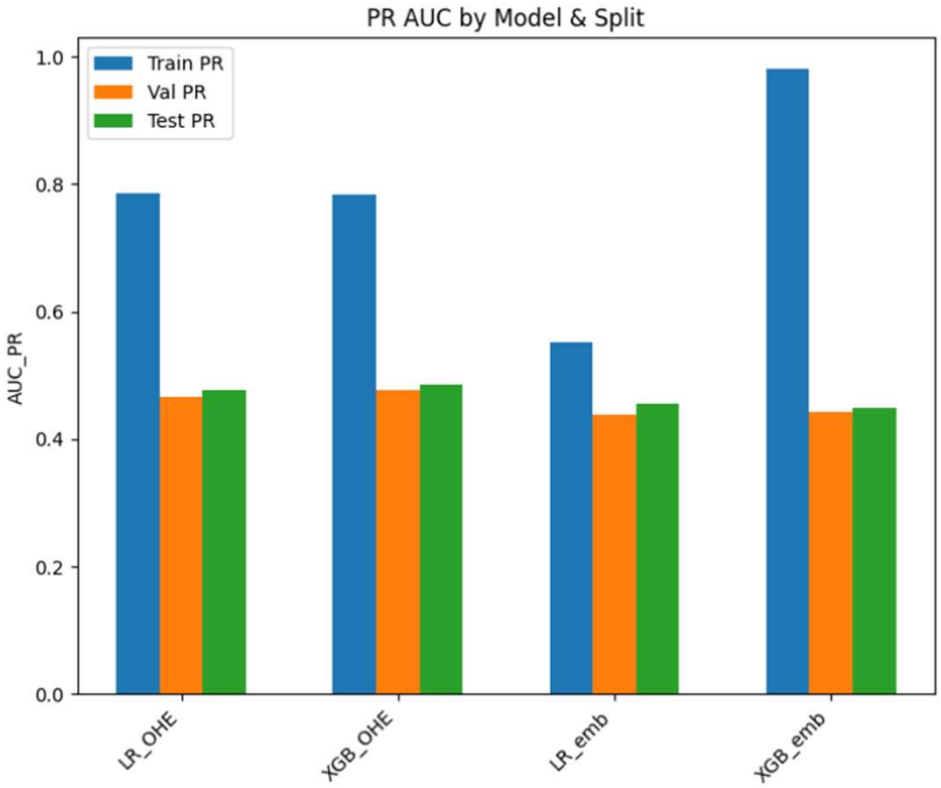


LOS

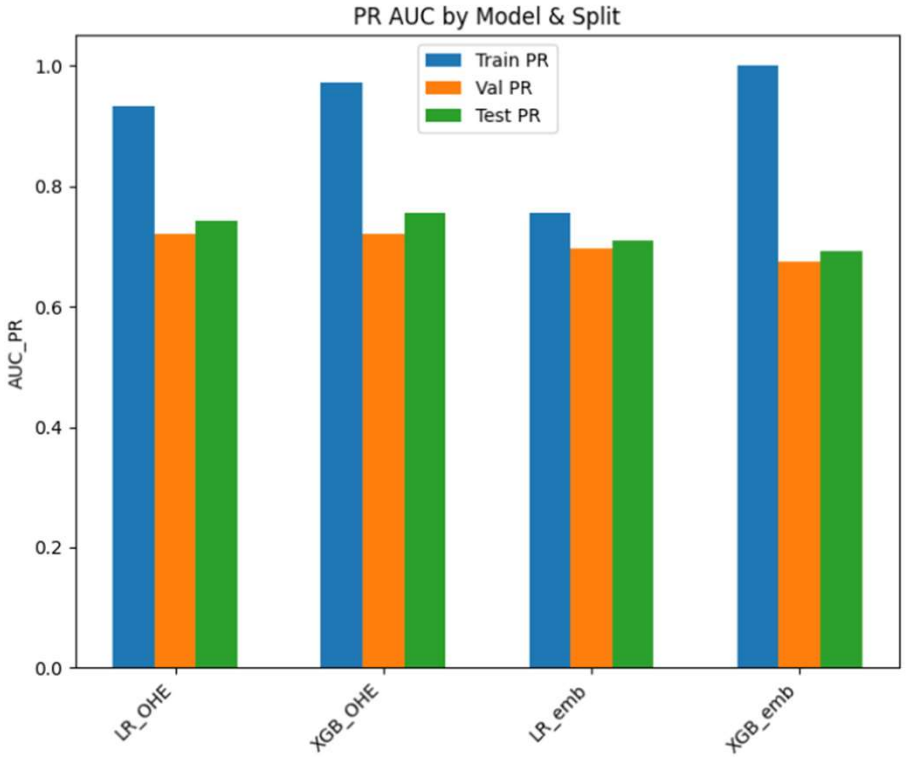


HERMÈS VALIDATION

ESI

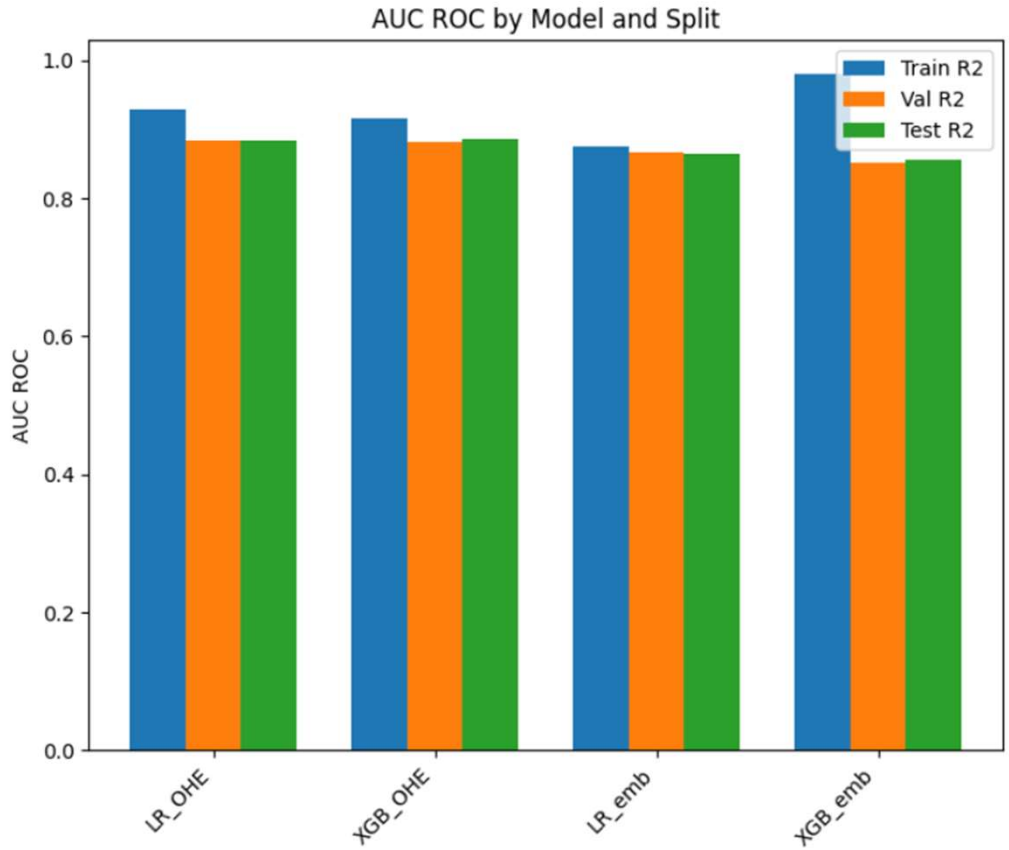


Mortality



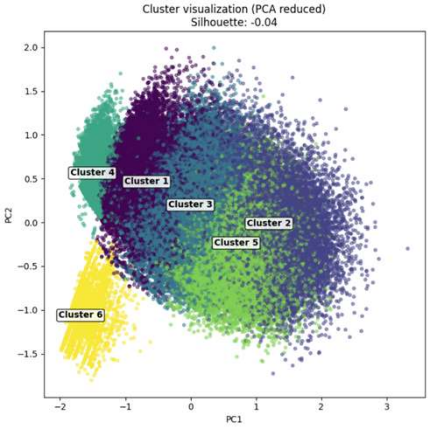
HERMÈS VALIDATION

Disposition

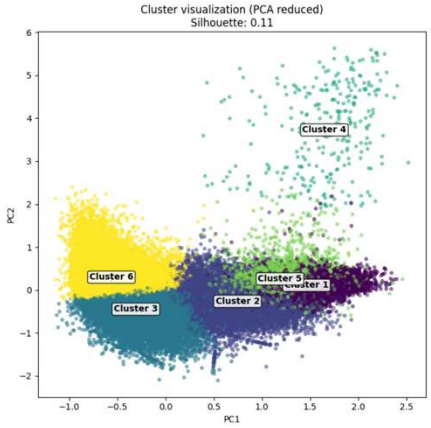


CLUSTER ANALYSIS

One-hot: >13k sparse features



512-d embeddings



Cluster summary: 512-d embeddings

- Cluster 1
  - Number of Visits: 9,458
  - Summary: Labor & delivery service line – complicated pregnancies, fetal monitoring, vaginal/C-section deliveries, and postpartum care.
- Cluster 2
  - Number of Visits: 10,501
  - Summary: Ortho & spine heavy (arthroplasty, lumbar stenosis/disc disease) with pre-op evaluations and some GI access care; matches elective surgical population.
- Cluster 3
  - Number of Visits: 31,225
  - Summary: Acute surgical and cardiology mix – gallbladder/appendix stone disease, renal colic, coronary workups; frequent ED/obs throughput.
- Cluster 4
  - Number of Visits: 280
  - Summary: Small, high-acuity NICU cohort – prematurity, RDS, jaundice – frequent resuscitation, invasive lines, and neonatal intensive care.
- Cluster 5
  - Number of Visits: 8,898
  - Summary: Well-baby nursery with frequent circumcisions and routine labs; notable hypoglycemia/macrosomia cases and a subset requiring NICU-level care.
- Cluster 6
  - Number of Visits: 41,140
  - Summary: Broad medical inpatient mix with chronic lung disease, ulcers, malnutrition, pleural effusions/malignancy; frequent central lines, endoscopy, and TEE.

ADVANTAGES OF USING EMBEDDINGS

1. **Foundation Model**
2. **Transferability**
3. **Efficiency vs. performance**
4. **Better generalization**

NEXT STEPS

1. Operationalize embeddings
2. Expand the graph
3. Broaden data coverage
4. Revise the task set

THANK YOU FOR YOUR ATTENTION

Jon Sege
jsege@wphospital.org
linkedin.com/in/jonsege

Vincent Pan, CFA
vpan@wphospital.org
linkedin.com/in/vincent-pan-cfa