BEHRT Architecture for Longitudinal EHR Modeling

Weihan Li

College of Computing

(A) Introduction — Motivation & Objective

Introduction

Longitudinal Electronic Health Records (EHRs) contain rich temporal patterns, but many traditional models ignore visit order and long-range dependencies. Li et al. (2020) proposed BEHRT, a Transformer-based architecture that treats diagnoses as tokens, visits as sentences, and a patient's history as a document, demonstrating strong performance on future-disease prediction.

Objective

To reproduce the BEHRT architecture on a different EHR dataset (MIMIC-III) and evaluate its ability to predict diagnosis codes that will occur in the next 12 months of a patient's history. The study further assesses predictive performance and model interpretability at both cohort and patient levels.

Real-world Motivation

Predicting diagnoses in the next 12 months has practical relevance in clinical risk stratification, enabling earlier identification of emerging chronic conditions and supporting proactive care planning. Such models may further benefit real-world healthcare operations—including resource allocation, patient management workflows, and optimization of care pathways—potentially reducing avoidable downstream costs.

(B) Methods — BEHRT Architecture & Data Processing

Dataset & Population

- Dataset: MIMIC-III (ICD-9 diagnosis codes from hospital admissions).
- Adult patients with ≥3 recorded visits.
- At least 12 months of follow-up after the index visit to support next-12-month prediction.
- This ensures sufficient longitudinal history, consistent with the cohort-selection principles used in BEHRT.

Baseline

- Bag-of-Codes + Logistic Regression: All past diagnoses aggregated into a single multi-hot vector per patient (no temporal structure).
- One-vs-rest logistic regression for multi-label prediction.
- This non-sequential baseline highlights the added value of modeling visit-level temporal patterns with BEHRT.

Model & Training

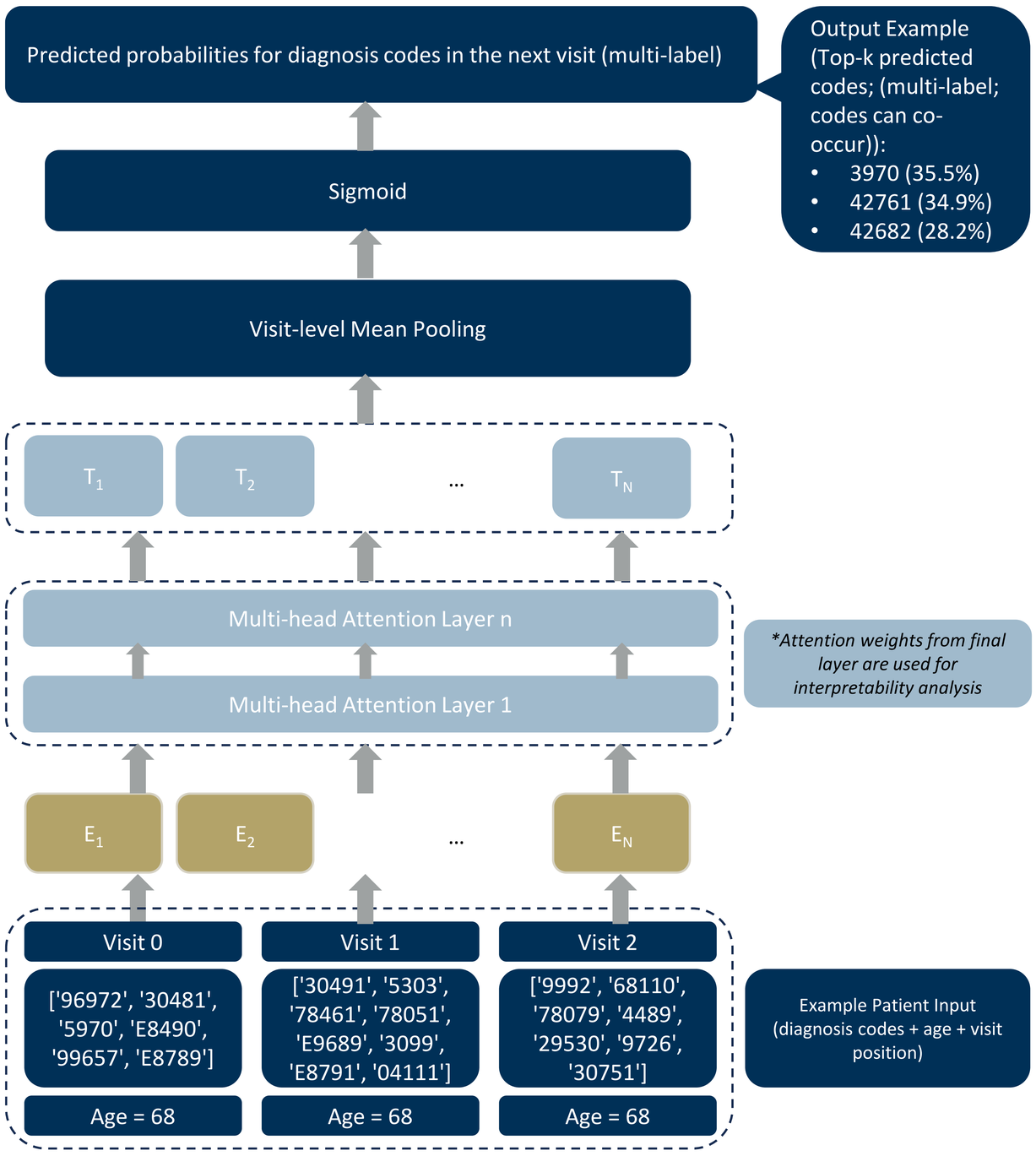
- **Model**: 6-layer Transformer encoder (hidden size 288, 12 attention heads, intermediate size 512, dropout 0.1), following the BEHRT design with reduced depth for computational efficiency.
- Visit-level mean pooling.
- Sigmoid output layer for multi-label diagnosis prediction.
- **Loss**: Binary cross-entropy.
- **Split**: 70% / 10% / 20% patient-level train / validation / test.

Input Representation

Each diagnosis token is represented by the sum of four embeddings:

- **Diagnosis embedding** – ICD-9 code identity
- **Position embedding** – token position in sequence
- **Age embedding** – patient age at visit
- **Visit-segment embedding** – alternating visit indicator

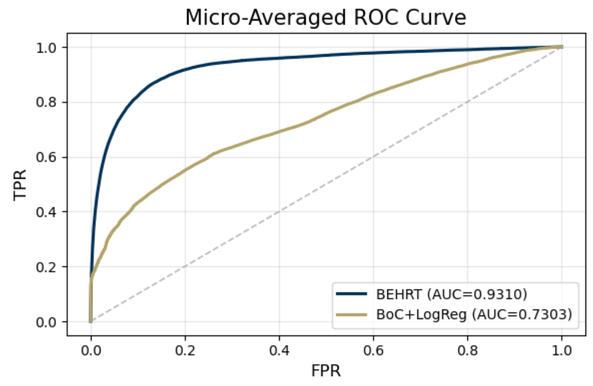
These summed embeddings form a longitudinal patient sequence following the BEHRT architecture.
(In Panel A, we illustrate this as E_token = E_diag + E_age + E_pos + E_seg.)

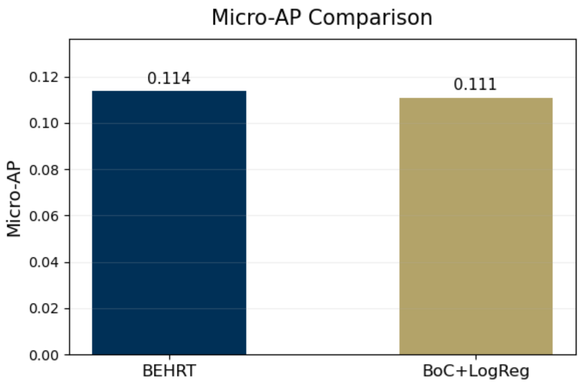


Architecture adapted from BEHRT (Li et al., 2020).

(C) Results — Cohort-level Predictive Performance

Micro-averaged evaluation on MIMIC-III



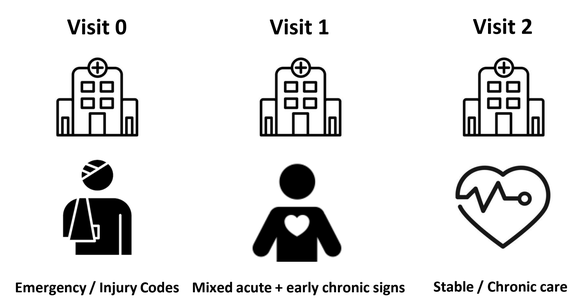


Model	Micro-AUC	Micro-AP
BEHRT	0.9310	0.114
BoC+LR	0.7303	0.111

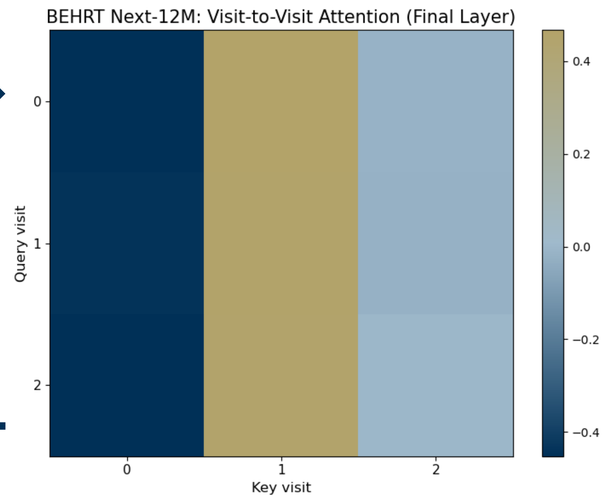
- BEHRT achieves significantly higher Micro-AUC and Micro-AP than a BoC + Logistic Regression baseline.
- Sequential modeling of longitudinal EHR histories provides a clear performance advantage.

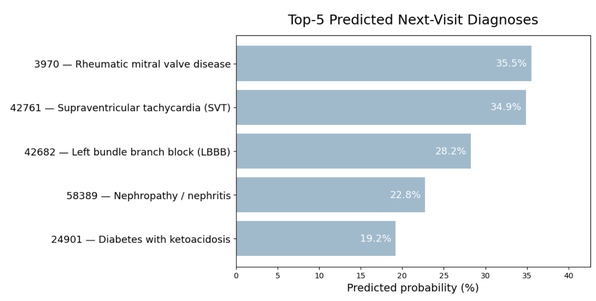
(D) Results — Patient-level Interpretation & Attention Patterns

Patient Age: 68
Prediction Window: Next 12 months



Visit-Level Attention Patterns





- Visits 1–2 receive the highest attention weights across all query visits.
- These visits contribute most to BEHRT's internal patient representation.
- Model predictions reflect a shift from acute injury to chronic cardiovascular monitoring.

(E) Conclusion — Summary & Future Directions

Conclusion

- We successfully reproduced the BEHRT architecture on the hospital-based MIMIC-III dataset and adapted it to a next-12-month multi-label diagnosis prediction task.
- BEHRT consistently outperforms a Bag-of-Codes + Logistic Regression baseline in AUROC and micro-average AP, demonstrating the advantage of modeling longitudinal visit sequences rather than flattened histories.

Interpretability & Clinical Relevance

- Patient-level attention analysis shows that BEHRT assigns the highest attention weights to follow-up visits containing early cardiovascular symptoms and monitoring patterns, rather than the initial injury visit.
- The top predicted diagnoses align with these attended visits, suggesting that BEHRT captures clinically meaningful trajectories that may support earlier risk stratification, pending further validation.

Future Work

- Incorporate additional EHR modalities (medications, labs, procedures) into the BEHRT framework.
- Evaluate transferability of BEHRT representations across different EHR datasets and clinical settings.
- Explore complementary interpretability approaches (e.g., gradient-based or counterfactual explanations) beyond attention weights.

Key Takeaways

1. Performance
BEHRT improves Micro-AUC by +0.20 over Bag-of-Codes + Logistic Regression.

2. Interpretability
Attention consistently focuses on visits with emerging chronic risk indicators.

3. Clinical relevance
Predicted diagnoses align with high-attention visits, suggesting potential for proactive risk identification.

4. Extendability
Architecture easily generalizes to multi-modal EHR data (labs, meds, procedures).