Trust Your Neighbors: Multimodal Patient Retrieval for TBI Prognosis Supplementary Material

Pranav Manjunath, Brian Lerner, and Timothy W. Dunn

A. Patient Text format

Figure S1 illustrates the textual format used to represent each patient to the language model.

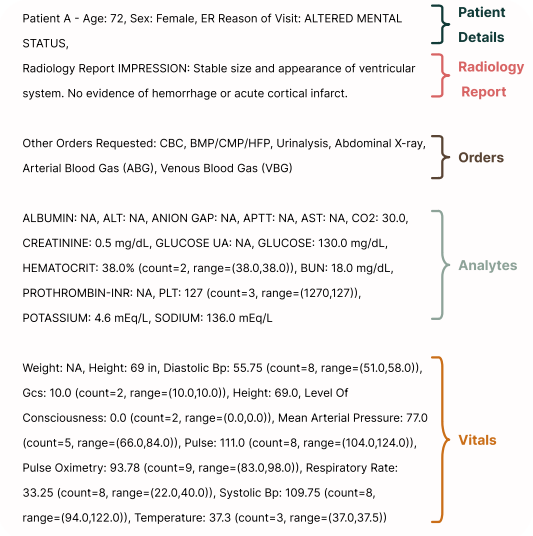

Fig. S1. Patient Data in the textual summary format for SLM

B. Experimental Design

All multimodal encoder training experiments were conducted using three NVIDIA A6000 GPUs with a batch size of 150 for training and validation. We systematically tuned hyperparameters, selecting an optimal temperature parameter of 0.3 from [0.1, 0.5] and a learning rate of 0.005 from [0.0005, 0.1]. Each multimodal architecture was trained for 500 epochs, and the model from the epoch with the lowest validation loss chosen as the final model. We compared CTNet image encoder with SOTA 3D image encoders — CTViT [1] and 3D-ResNet [2]. To evaluate how well each encoder captures radiology report semantics, we computed the normalized Hamming distance between each test patient's binary report labels and those of its retrieved neighbors, averaging across neighbors and test patients. Hamming distance was used due to the binary nature of the labels.

For the classifiers, RPEC was initially trained for 50 epochs, and the optimal model checkpoint was selected based on the lowest validation loss. Subsequently, the training and validation datasets were combined, and the model was retrained from scratch up to the selected optimal epoch before final evaluation on the test set. For patient search, approximate nearest neighbors from FAISS outperformed exact nearest neighbors in both retrieval and classification tasks, so we report only FAISS results. We compare our approach to baseline Nearest Neighborhood classifier and SOTA GraphSAGE [3]. For GraphSAGE, edges were assigned between patient embeddings with a cosine similarity exceeding a threshold that is defined by the mean pairwise cosine similarity of all patient embeddings in the dataset. For testing, we retrain the final model on combined training and validation data. All code, pretrained models, results will be on Github upon acceptance.

1) SLM Experimental Design: Both models were downloaded from HuggingFace and integrated into our local inference pipeline using the HuggingFace transformers library in Python, running on NVIDIA A6000 GPUs. Each model was configured with an maximum output length of 1,028 tokens and the temperature was set to 0 to ensure deterministic outputs from the language model during evaluation.

2) Generalizability Experiment Design: To evaluate the generalizability of our approach, we conducted two types of experiments: (i) Cross-institutional and (ii) Temporal. For the institutional experiment, the model was trained and validated on data from one hospital institution ($n = 5266$) and tested on data from another institution ($n = 4054$; 32.8% Admit). For the temporal experiment, the model was trained and validated on patient data from 2015–2018 ($n = 7276$) and tested on patient data from 2019 ($n = 2044$; Admit: 26.8%).

C. Ablation Study

Our ablation studies () highlight the impact of attention, classification mechanism, and radiology report representations (Report: R-Input vs. Label: L-Input) on ED disposition. We demonstrate that MMnet is less effective than the attention-based approach (MMANet) when utilizing CTNet as the image encoder, which differs from other model variants in its

This work is supported by an R01 from the NIH under grant number 101NS123275.

Pranav Manjunath is with the Department of Biomedical Engineering at Duke University, Durham, NC, USA (e-mail:pranav.manjunath@duke.edu)

Brian Lerner is with the Department of Electrical and Computer Engineering at Duke University, Durham, NC, USA (e-mail:brian.lerner@duke.edu)

Timothy W. Dunn is with the Department of Biomedical Engineering at Duke University, Durham, NC, USA (e-mail:timothy.dunn@duke.edu)

performance. RPEC, on average, outperforms GraphSAGE and Nearest Neighbors across architectures and encoders. Notably, for the RAPID-TBI encoder, using full radiology report text improves classification over report labels, a trend observed across most models. MMANet demonstrates improved performance when compared to the Meta-Transformer encoder.

TABLE S1
ABLATION STUDIES - CLASSIFICATION AUROC. R-MODEL: RADIOLOGY REPORT INPUT, L-MODEL: RADIOLOGY REPORT LABELS INPUT.

Encoder Configurations		R-Model			L-Model
Multimodal Architecture	**Image Encoder**	Nearest Neighbors	Graph SAGE	RPEC	RPEC
Meta-Transfomer	CTViT	0.780	0.831	0.799	0.786
Meta-Transformer	3D ResNet	0.729	0.675	0.758	0.774
Meta-Transformer	CTNet	0.727	0.732	0.771	0.835
MMNet	CTViT	0.756	0.794	0.805	0.804
MMNet	3D ResNet	0.764	0.801	0.803	0.795
MMNet	CTNet	0.773	0.800	0.798	0.811
MMANet	CTViT	0.751	0.788	0.799	0.771
MMANet	3D ResNet	0.760	0.786	0.797	0.791
MMANet	**CTNet**	0.832	0.835	**0.851**	0.800

D. Stratified Results

TABLE S2
TBI RESULTS ACROSS PATIENT SUBGROUPS FOR RPEC CLASSIFIER. PPV IS THE PREDICTIVE POSITIVE VALUE

Subgroup	Category	Count	F1	AUROC	PPV
TBI Severity	Mild	1316	0.614	**0.820**	0.468
	Moderate	61	0.882	0.774	0.911
	Severe	34	0.915	0.593	0.900
Sex	Female	1279	0.652	**0.858**	0.576
	Male	1049	0.657	0.841	0.513
Age	18-35	469	0.339	0.816	0.208
	36-64	823	0.607	**0.858**	0.469
	65+	971	0.694	0.802	0.597

E. Generalizability Results

Supplementary Figure S2 presents performance from the institutional and temporal experiments for the RPEC classifier, while Supplementary Tables S3 and S4 report results for the SLM classifiers.

1) RPEC: Across both institutional and temporal generalizability experiments (Fig. S2), the RPEC classifier achieved the highest performance, with AUROC values of 0.727 (95% CI: 0.715–0.738) and 0.764 (95% CI: 0.749–0.779), respectively. These results exceeded those of GraphSAGE (0.718 and 0.757) and the nearest neighbor baseline (0.705 and 0.738). Although the model achieved a higher AUROC in the temporal generalizability experiment (0.764 vs. 0.727), its F1 score at the optimal threshold was lower (0.554 ± 0.021 vs. 0.586 ± 0.013, 95% CI). This pattern suggests that while the model preserved strong overall discriminative ability over time, its probability calibration and threshold alignment deteriorated, potentially due to temporal shifts in class prevalence or feature distributions. In contrast, the institutional experiment demonstrated slightly lower discrimination but a more stable precision–recall balance at the decision threshold. F1 scores are additionally reported to facilitate comparison with SLM classifiers.

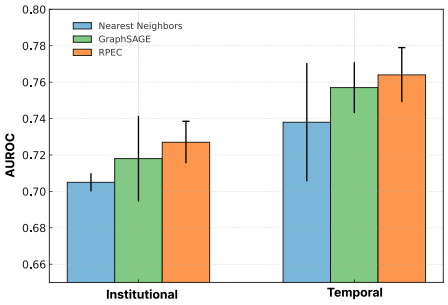

Fig. S2. RAPID-TBI Generalizability Experiments across Institutional and Temporal Shifts

2) SLM: Tables S3 and S4 summarize the performance of the SLM classifiers across varying numbers of retrieved neighbors (k). Consistent with trends observed for RPEC, the SLMs achieve stronger performance in the institutional generalizability setting compared to the temporal one when evaluated using F1 score. Among the SLM variants, retrieval with $k = 5$ yielded the best results, with Qwen and Phi attaining F1 scores of 0.627 and 0.622 in the institutional experiment, and 0.592 and 0.575 in the temporal experiment, respectively.

TABLE S3
INSTITUTIONAL GENERALIZABILITY OF SLM CLASSIFIER

k	Qwen		Phi	
	F1 Score	Sensitivity	F1 Score	Sensitivity
0	0.437	0.384	0.574	0.611
1	0.566	0.527	0.624	0.764
3	0.589	0.616	0.617	0.852
5	0.627	0.750	0.622	0.863
7	0.620	0.756	0.605	0.914

TABLE S4
TEMPORAL GENERALIZABILITY OF SLM CLASSIFIER

k	Qwen		Phi	
	F1 Score	Sensitivity	F1 Score	Sensitivity
0	0.352	0.299	0.577	0.719
1	0.548	0.529	0.558	0.750
3	0.535	0.535	0.543	0.819
5	0.592	0.734	0.575	0.912
7	0.590	0.759	0.563	0.936

F. Optimal Number of Retrieved Patients

To identify the optimal number of retrieved patients, we examined how AUROC and F1 varied with different k values (Fig. S3). We hypothesize that this reflects a trade-off between informativeness and noise: smaller k yields highly similar and informative neighbors, while larger k introduces greater heterogeneity that dilutes the signal. The thresholds differ across metrics because F1 is more sensitive to precision–recall trade-offs, whereas AUROC is more robust and declines only when neighborhoods become excessively broad.

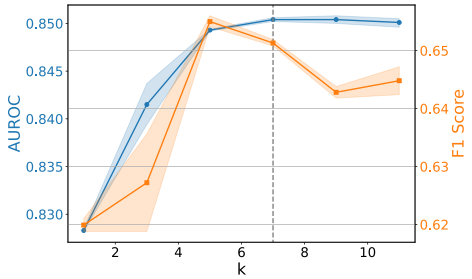

Fig. S3. RPEC classification results at different numbers of retrieved patients. K=7 is chosen as our overall model

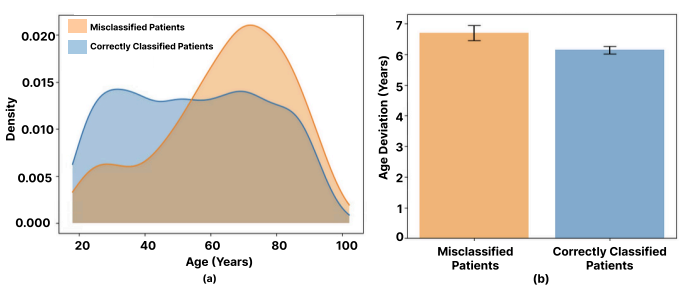

Fig. S4. Discrepancy analysis of RAPID-TBI predictions. (a) Kernel density estimate (KDE) plots showing the age distribution of misclassified versus correctly classified patients. (b) Average age deviation between each test patient and their top-7 retrieved neighbors, stratified by misclassified and correctly classified cases.

G. Discrepancy Analysis

We analyzed the misclassified test cases to better understand model errors (Supplementary Fig. S4a). The KDE plot shows that these patients were typically older and had greater age deviations (Supplementary Fig. S4b) from their retrieved neighbors. Despite misclassification, the model had a high mean nDCG of 0.811, suggesting that the retrieved neighbors remained aligned with the predicted outcome. Representative false-negative and false-positive examples further illustrate characteristic failure patterns (Fig. S5, Fig. S6).

H. Clinical Implementation

Fig. S7 illustrates a potential clinical implementation of RAPID-TBI to assist at bedside. When a patient arrives, encounter-level data—including demographics, vitals, labs, notes, and imaging—are entered into the EHR. A clinician can then select the "Similar Patients" option, prompting the EHR to invoke the RAPID-TBI service. The patient's multimodal data are encoded into an embedding and compared against a secure, hospital-hosted vector database containing curated, de-identified historical encounters. The system retrieves the top- k most similar patients and generates an emergency-department disposition prediction, which is returned to the EHR alongside the retrieved examples. At the bedside, clinicians can review these predictions and similar cases directly within their workflow to inform decision-making. Under the hood, RAPID-TBI integrates imaging data from PACS (DICOM format) with EHR-derived laboratories, vitals, and clinical documentation represented through FHIR resources—Patient, Encounter,

	Test Patient	Retrieved Patient 1	Retrieved Patient 2	Retrieved Patient 3
Age (Years)	28.1	30.0	35.8	31.0
Sex	Female	Male	Male	Male
GCS	14	15	15	15
CT Findings	Bilateral basal ganlia hyperdensity	Right scalp contusions, small focal hematoma posterior to the right ear	No Acute Intracranial Abnormalities	No Acute Intracranial Abnormalities
Pulse	108.4	83	93	101.8
BP	118/84	136/72	150/84	133/84
PLT (10^9)	261	334	337	230
Sodium	137	136	139	132
Prothrombin INR	NA	1.0	1.3	NA
APTT	NA	20.0	33.1	NA
Actual ED Disposition	Admission	Discharge	Admission	Discharge

Fig. S5. Discrepancies between a test patient (False Negative: ground-truth admitted, predicted discharged) and its top-3 retrieved neighbor

	Test Patient	Retrieved Patient 1	Retrieved Patient 2	Retrieved Patient 3
Age (Years)	70.6	73.5	79.2	62.06
Sex	Female	Female	Female	Female
GCS	15	15	13	14
CT Findings	No Acute Intracranial Abnormalities	No Acute Intracranial Abnormalities	No Acute Intracranial Abnormalities	No Acute Intracranial Abnormalities
Pulse	75.6	84	79.2	85.6
BP	158/83	168/75	NA	122/82
PLT (10^9)	407	234	220	290
Sodium	135	139	131	141
Prothrombin INR	1.1	1.0	NA	NA
APTT	36.7	30.5	NA	NA
Actual ED Disposition	Discharge	Discharge	Admission	Admission

Fig. S6. Discrepancies between a test patient (False Positive: ground-truth discharged, predicted admitted) and its top-3 retrieved neighbor

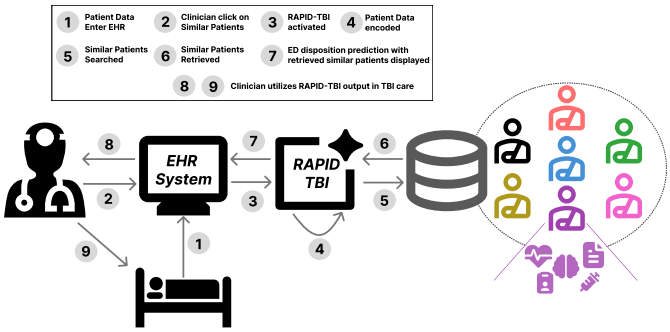

Fig. S7. Potential Clinical Implementation Workflow of RAPID-TBI in TBI Care

Observation, Condition, ImagingStudy, and DiagnosticReport. All embeddings are stored and queried in a de-identified form, ensuring retrieval without PHI exposure, and results can be surfaced as FHIR GuidanceResponse or RiskAssessment objects to maintain seamless interoperability.

REFERENCES

[1] I. E. Hamamci, S. Er, F. Almas *et al.*, "Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography," Oct. 2024, arXiv:2403.17834 [cs].

[2] K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," Aug. 2017, arXiv:1708.07632 [cs]. [Online]. Available: http://arxiv.org/abs/1708.07632

[3] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," Sep. 2018, arXiv:1706.02216 [cs].