

Multimodal prediction of visual improvement in diabetic macular edema using real-world electronic health records and optical coherence tomography images

Siqi Sun¹, Cindy X. Cai^{2,3}, Ruochong Fan¹, Saiyu You¹, Diep Tran², P. Kumar Rao⁴, Marc A. Suchard⁵, Yixin Wang⁶, Cecilia S. Lee⁴, Aaron Y. Lee⁴, Linying Zhang^{1,7}

1. Institute for Informatics, Data Science and Biostatistics, Washington University in St. Louis

2. Wilmer Eye Institute, Johns Hopkins School of Medicine

3. Department of Biomedical Informatics and Data Science, Johns Hopkins School of Medicine

4. John F. Hardesty MD, Department of Ophthalmology and Visual Sciences, Washington University in St. Louis

5. Department of Biostatistics, University of California, Los Angeles

6. Department of Statistics, University of Michigan, Ann Arbor

7. Department of Medicine, Washington University in St. Louis

Background

Clinical decision-making is inherently multimodal. Clinicians must synthesize diverse sources of information, including demographic and clinical variables such as age, sex, comorbidities, medication history, laboratory test results that are typically stored as structured data in electronic health record (EHR) databases, as well as medical imaging data such as X-rays, magnetic resonance imaging (MRI), and pathology images. Multimodal learning methods aim to integrate these complementary data modalities and therefore have the potential to improve prognostic accuracy and better support clinical decision-making¹⁻⁵.

Multimodal learning is particularly promising in ophthalmology, where clinical decision-making heavily relies on imaging, such as optical coherence tomography (OCT)^{6,7}. In this study, we focus on diabetic macular edema (DME), a leading cause of vision impairment among working-age adults worldwide⁸. OCT captures key morphological features of DME. Although anti-vascular endothelial growth factor (anti-VEGF) therapy is effective for treating DME, treatment response varies substantially across patients, making early prediction of visual improvement important for personalized risk stratification and treatment planning^{9,10}.

Existing prognostic studies in DME have primarily relied on clinical and demographic variables, which do not capture retinal structural features visible on OCT¹⁰⁻¹². In this study, we develop a multimodal survival prediction framework integrating real-world OCT and EHR data to predict time to visual improvement in patients with DME. We systematically evaluate OCT representations derived from multiple ophthalmic foundation models and quantify their incremental prognostic value beyond EHR variables. Our study provides insights into the potential of integrating real-world multimodal data to improve prognostic modeling in ophthalmology.

Methods

Data sources and cohort definition

Electronic health records (EHRs) and OCT images were from Washington University/BJC HealthCare. The EHR database is structured in the OMOP Common Data Model (CDM) v5.4. OCT images were extracted from the clinical ophthalmic imaging database, accessed through Continuum. The study cohort was defined with the following inclusion criteria (Figure 1b): (1) study period between January 1, 2018 and December 31, 2024; (2) diagnosis of diabetic macular edema (DME) during the study period; (3) receipt of intravitreal anti-VEGF injection (aflibercept or bevacizumab); and (4) availability of injection laterality

(OD/OS/OU) to identify the treated eye, with OU entries duplicated and split into OD and OS to create eye-level observations. The index date was defined the date of the first anti-VEGF injection. The outcome was defined as the days from the first injection to the first post-treatment VA measurement demonstrating a ≥ 2 -line improvement in VA relative to baseline within 730 days of follow-up.

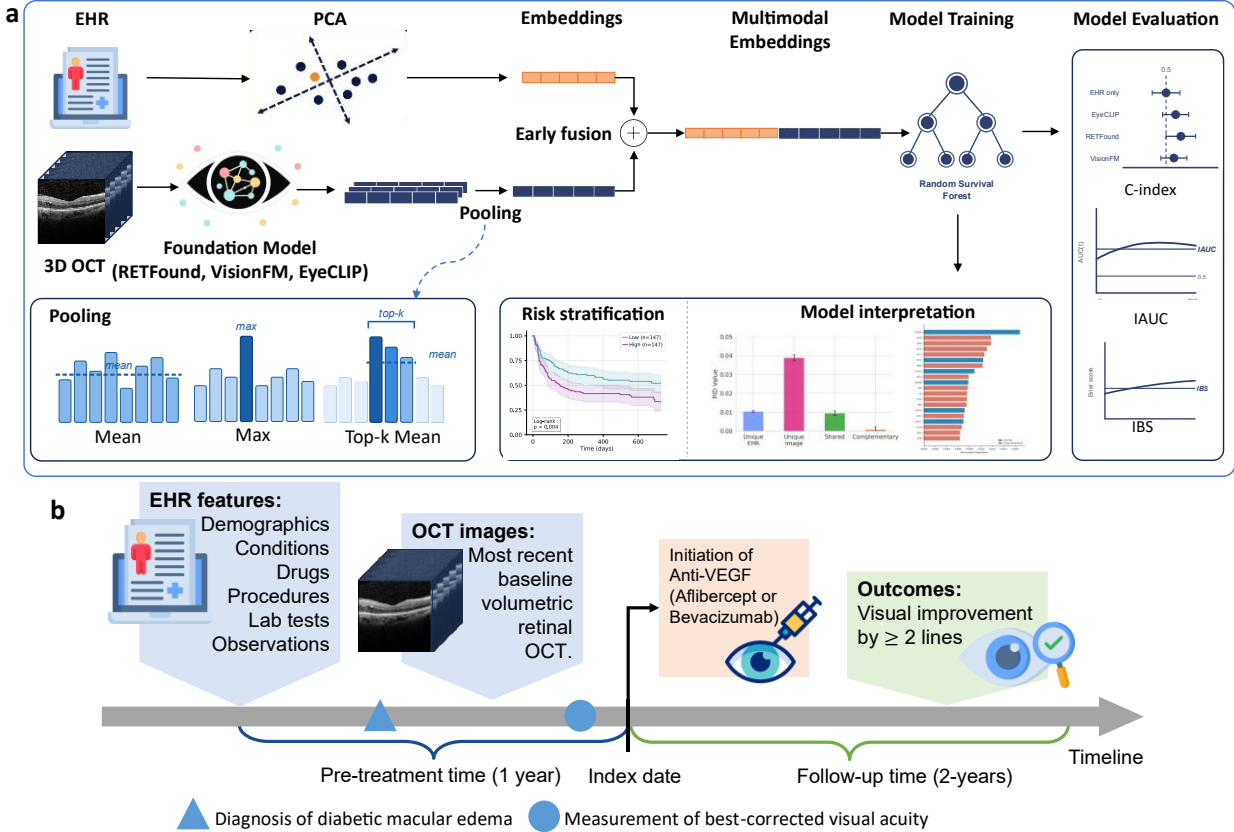


Figure 1. Study overview. (a) Multimodal modeling framework. (b) Study design and timeline.

OCT embeddings and pooling strategies

Each B-scan is resized to 224×224 pixels and embedded using pretrained ophthalmic foundation models: RETFound (1024-dimensional)¹³, VisionFM (3072-dimensional)¹⁴, and EyeCLIP (512-dimensional)¹⁵. Embeddings are computed per B-scan and then aggregated using mean pooling, max pooling, or top-k mean pooling to obtain eye-level OCT embedding.

EHR embeddings with principal component analysis

After excluding identifiers, outcome variables, treatment indicators, and other non-predictor fields, principal component analysis (PCA) was then applied to the remaining EHR features. The minimum number of principal components (PC) explaining 95% of the cumulative variance in the training data was retained, which included 445 PCs, and used as the EHR input to the survival model.

Model development and evaluation

All comparisons used a patient-level train/test split (80% train, 20% test) to prevent leakage between eyes of the same patient. We used a random survival forest (RSF) for all experiments. The evaluation metrics included Harrell's C-index, integrated time-dependent AUC (IAUC), and integrated Brier score (IBS).

Bootstrap resampling of the test set ($B=1000$) was used to estimate 95% confidence intervals for all evaluation metrics.

Multimodal data fusion strategies

For fusion sensitivity, we compared early fusion (*EF*) versus late fusion (*LF*). In early fusion, we concatenated EHR features $x_i \in R^{d_x}$ and OCT features $z_i \in R^{d_z}$ into a single input $U_i = [X_i; Z_i]$. In late fusion, we trained two separate RSF models to produce modality-specific risk scores $r_i^X = f_X(X_i)$ and $r_i^Z = f_Z(Z_i)$, computed validation C-indices c_X^{val} and c_Z^{val} , converted them to nonnegative weights $w_X = \max(C_X^{\text{val}} - 0.5, 0)$ and $w_Z = \max(C_Z^{\text{val}} - 0.5, 0)$, and formed the fused test risk score

$$r_i^{\text{LF}} = \begin{cases} \frac{w_X r_i^X + w_Z r_i^Z}{w_X + w_Z}, & w_X + w_Z > 0, \\ \frac{r_i^X + r_i^Z}{2}, & w_X + w_Z = 0. \end{cases}$$

We then evaluated the fused model using test-set Harrell’s C-index C_{LF} .

Partial information decomposition (PID)

PID provides a principled framework for quantifying how multiple input modalities jointly contribute to predicting an outcome by decomposing total information into shared (redundant), unique, and complementary (synergistic) components. Let X denote the EHR embedding from PCA, Z denote the eye-level OCT embedding, and $Y \in \{0,1\}$ denote the event indicator. We first trained three probabilistic multilayer perceptron (MLP) classifiers to estimate

$$\widehat{p}_X = p(y = 1 | X), \widehat{p}_Z = p(y = 1 | Z), \widehat{p}_{XZ} = p(y = 1 | X, Z),$$

each optimized with binary cross-entropy loss. We then trained a Q -estimator (a two-branch MLP) to learn a constrained joint distribution $q(X, Z, Y)$ subject to marginal constraints implied by $\widehat{p}_X$ and $\widehat{p}_Z$ using Sinkhorn–Knopp updates, and the model parameters were optimized to minimize the mutual information objective

$$\mathcal{L}_Q = \sum_{x,z,y} q(x, z, y) \log_2 \frac{q(x, z, y)}{q(x, z) q(y)}.$$

Results

The study cohort included 973 patients with DME initiated with anti-VEGF therapy, contributing 1,450 eyes (196,402 OCT B-scans) that initiated anti-VEGF therapy during 2018-2024. At the eye level, the mean baseline best-corrected visual acuity (BCVA) was 0.50 (SD, 0.40) logMAR. Overall, 718 of 1,450 eyes (49.5%) achieved a visual acuity improvement of at least two lines within 2 years of treatment initiation.

Our results showed that (1) across experiments, performance was largely consistent across OCT pooling and modality fusion strategies, with no meaningful differences observed. Therefore, we present results from the early fusion with mean pooling configuration as the primary analysis. The EHR-only model showed minimal prognostic discrimination. Incorporating OCT improved performance, although the magnitude of improvement varied across ophthalmic foundation models. RETFound yielded the largest prognostic gains; and (2) the PID patterns indicate that the incremental prognostic value of OCT is primarily driven by modality-specific information, and that the magnitude of this contribution is highly dependent

on the choice of ophthalmic foundation model, with RETFound providing the most informative OCT representations.

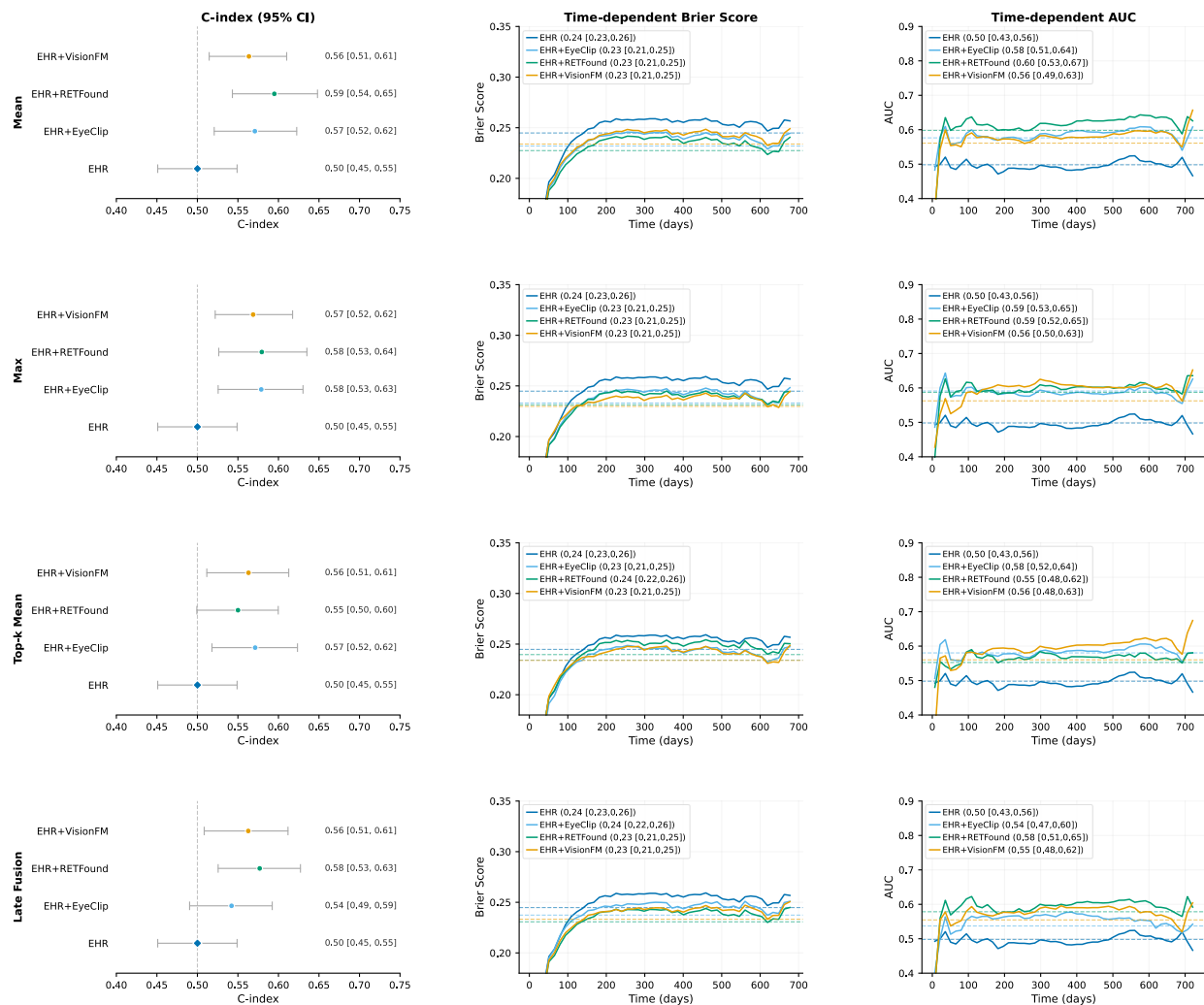


Figure 2. Comparative performance of EHR-only and multimodal survival models across OCT pooling and modality fusion strategies.

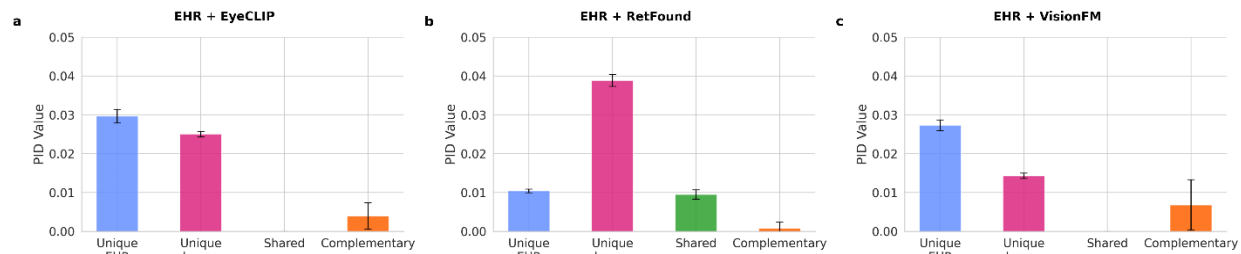


Figure 3. Partial information decomposition of prognostic information across modalities.

Conclusion

Our findings indicate that OCT provides complementary prognostic value beyond structured clinical data, but gains are modest and depend strongly on representation choice. Our results highlight both the promise of multimodal modeling for personalized prognosis and the need for rigorous, context-specific evaluation of foundation models in real-world clinical settings.

References

1. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nature medicine* **28**, 1773–1784 (2022).
2. Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I. & Lungren, M. P. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports* **10**, 22147 (2020).
3. Zhang, C. *et al.* M3care: Learning with missing modalities in multimodal healthcare data. in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* 2418–2428 (2022).
4. Kline, A. *et al.* Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine* **5**, 171 (2022).
5. Soenksen, L. R. *et al.* Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine* **5**, 149 (2022).
6. Wang, Z. *et al.* Artificial intelligence and deep learning in ophthalmology. *Artificial intelligence in medicine* 1519–1552 (2022).
7. Swanson, E. A. & Fujimoto, J. G. The ecosystem that powered the translation of OCT from fundamental research to clinical and commercial impact. *Biomedical optics express* **8**, 1638–1664 (2017).
8. Yau, J. W. *et al.* Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care* **35**, 556–564 (2012).
9. Nguyen, Q. D. *et al.* Ranibizumab for diabetic macular edema: results from 2 phase III randomized trials: RISE and RIDE. *Ophthalmology* **119**, 789–801 (2012).
10. Wells, J. A. *et al.* Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema. *The New England journal of medicine* **372**, 1193–1203 (2015).
11. Bressler, S. B. *et al.* Factors associated with changes in visual acuity and central subfield thickness at 1 year after treatment for diabetic macular edema with ranibizumab. *Archives of ophthalmology* **130**, 1153–1161 (2012).
12. Kuo, B. L. *et al.* Long-term treatment patterns for diabetic macular edema: up to 6-year follow-up in the IRIS® Registry. *Ophthalmology Retina* **8**, 1074–1082 (2024).
13. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
14. Qiu, J. *et al.* Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, A10a2300221 (2024).
15. Shi, D. *et al.* A multimodal visual–language foundation model for computational ophthalmology. *npj digital medicine* **8**, 381 (2025).