Causal Inference with Multi-Modal Foundation Models: A Case Study of Anti-VEGF Injections in Diabetic Macular Edema

Siqi Sun[1], Cindy X. Cai[2,3], Ruochong Fan[1], Saiyu You[1], Diep Tran[2], P. Kumar Rao[4], Marc A. Suchard[5], Yixin Wang[6], Linying Zhang[1]

[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] 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

1. Background

Estimating causal effects from observational health data remains a significant challenge due to confounding, especially when data are captured in diverse formats such as medical images, clinical notes, and structured tabular features.[1-5] Traditional causal inference methods are predominantly designed for structured tabular data, limiting their ability to adjust for information embedded in unstructured modalities.

Recent advances in foundation models have demonstrated strong performance and generalizability in processing unstructured data such as images and text.[6-8] However, to the best of our knowledge, no existing methods have successfully integrated multi-modal foundation models into causal inference frameworks for estimating treatment effects.

In this work, we propose a novel multi-modal causal inference pipeline that leverages foundation models to adjust for confounding present in both structured EHR data and medical images. We apply this pipeline to a real-world comparative effectiveness study in ophthalmology, evaluating the vision-improving effects of intravitreal anti-vascular endothelial growth factor (VEGF) therapies in patients with diabetic macular edema (DME) using a large clinical dataset comprising 132,108 macular optical coherence tomography (OCT) images and over 8,000 EHR features.

2. Methods

2.1 Data source and study objective

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.

We adopted an active-comparator new-user cohort design to compare the effectiveness of aflibercept and bevacizumab in improving visual acuity (VA) 1-year after treatment among patients with DME. The study included adults (≥18 years) with a diagnosis of DME who initiated treatment with one of two anti-VEGF agents—aflibercept or bevacizumab —between January 1, 2018, and December 31, 2024. Ranibizumab was excluded due to an insufficient number of exposed patients. For each patient, the earliest anti-VEGF drug exposure date was designated as the index date.

2.2 Outcome definition

Visual acuity (VA) measurements were extracted from EHR flowsheets at two timepoints for each eye. Baseline VA was defined as the most recent measurement within one year prior to the treatment initiation date. Post-treatment VA was defined as the measurement closest to 12 months (within a window of 10 to 24 months) following treatment initiation. All Snellen scores were converted to logMAR values using the formula:

$$logMAR = -log_{10}(Snellen),$$

where a lower logMAR value indicates better visual acuity. The change in logMAR was defined as

$$\Delta logMAR \; = \; logMAR_{post} \; - \; logMAR_{baseline},$$

which was calculated separately for each eye. The outcome was defined as follows:

- Vision improvement (Y = 1) if $\Delta logMAR$ ≤ -0.2 in either eye.
- No improvement (Y = 0) if $\Delta logMAR$ > –0.2 in both eyes.

Analyses were further stratified by baseline VA, the worse VA group was defined as either eye had $Snellen_{baseline}$ ≤ 20/50, or equivalently, $logMAR_{baseline} \geq 0.4$.

2.3 Feature extraction

Structured EHR features*:* We extracted demographics, diagnoses, medications, procedures, measurements, and observations from the 1-year pre-treatment window, which leads to approximately 8,000 features.

OCT images*:* The OCT image immediately prior to the treatment start date was selected for each patient. Patients without pre-treatment OCT scans (<1%) were excluded. Images came from two major devices—Spectralis and Cirrus—in approximately equal proportions. Each Spectralis volume scan included 25 OCT images per eye per imaging session, whereas each Cirrus volume scan included 128 OCT images per eye per session. A total of 132,108 OCT scans were included in this study.

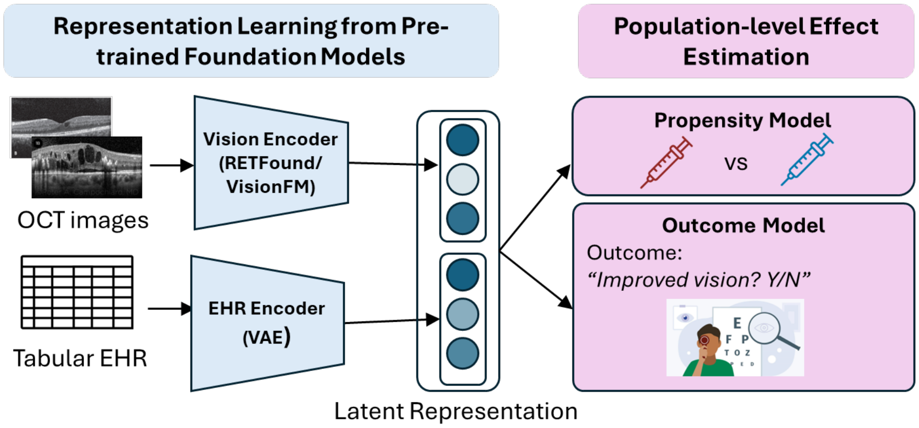

Figure 1. Multi-modal causal inference framework for population-level effect estimation.

2.4 Low-dimensional embedding from generative models

Due to the high dimensionality of both EHR and OCT imaging data, we used modality-specific encoders to learn low-dimensional embeddings. For structured EHR data, given the lack of well-validated pre-trained foundation models, we trained a variational autoencoder (VAE)[9] on our dataset (excluding demographics) and used its encoder to extract latent representations. We performed a grid search over latent dimensions (16 to 2048) and selected 1024 dimensions based on the best AUROC for outcome prediction.

We leveraged two existing pre-trained foundation models in ophthalmology to extract latent embedding for OCT images. RETFound is a self-supervised transformer model trained on over 1.6 million OCT images to learn retina-specific visual embeddings. RETFound embedding has 1024 dimensions.[7] VisionFM is a vision foundation model pretrained on 3.4 million ophthalmic images across eight modalities, including OCT and its embedding has 3072 dimensions.[8]

When estimating treatment effects across both modalities, we concatenated the EHR embedding with the OCT embedding, along with demographics, and used them as inputs for treatment effect estimation.

2.5 Doubly robust multi-modal ATE estimator

We used the Augmented Inverse Probability Weighting (AIPW) estimator to compute the ATE.[13,14]

Propensity score model We estimated the treatment assignment probability (i.e., the propensity score) using L1-regularized logistic regression:

$$\hat{e}(Z_i) = P(T_i = 1 \mid Z_i) = \frac{1}{1 + \exp(-\beta^T Z_i)},$$

where $\hat{e}(Z_i)$ is the estimated propensity score, T_i is the treatment indicator (1 = aflibercept, 0 = bevacizumab), and Z_i is the embedding.

Outcome model We trained separate L1-regularized logistic regression models to estimate the outcome under treatment and control:

$$\hat{Y}(1, Z_i) = P(Y = 1 | A = 1, Z_i) = expit(\alpha_1^T Z_i),$$

$$\hat{Y}(0, Z_i) = P(Y = 1 | A = 0, Z_i) = expit(\alpha_0^T Z_i),$$

where $\hat{Y}(t, Z_i)$ is the estimated potential outcome for treatment t using the embedding Z.

AIPW estimator The Augmented Inverse Probability Weighting (AIPW) estimator is given by:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{Y}(1, Z_i) - \hat{Y}(0, Z_i) + \frac{T_i\left(Y_i - \hat{Y}(1, Z_i)\right)}{\hat{e}(Z_i)} - \frac{(1 - T_i)\left(Y_i - \hat{Y}(0, Z_i)\right)}{1 - \hat{e}(Z_i)}\right],$$

where $\hat{\tau}$ is the ATE and Y_i is the observed outcome. The AIPW estimator combines the outcome model and inverse-probability-weighted components for double robustness.

Variance estimation To estimate the variance of the ATE, we used the efficient influence function approach. We computed the influence function:

$$\phi_i = \hat{Y}(1, Z_i) - \hat{Y}(0, Z_i) + \frac{T_i\left(Y_i - \hat{Y}(1, Z_i)\right)}{\hat{e}(Z_i)} - \frac{(1 - T_i)\left(Y_i - \hat{Y}(0, Z_i)\right)}{1 - \hat{e}(Z_i)}.$$

The variance of the ATE estimator $\hat{\tau}$ was then estimated by: $\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}(\phi_i - \hat{\tau})^2$. The 95% confidence intervals were constructed as $\hat{\tau} \pm 1.96 \times \sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}$.

2.5 Experimental design

We compared treatment effect estimates across different adjustment strategies:

1. Unadjusted
2. EHR only (EHR-VAE)
3. image only (RETFound)

4. Image only (VisionFM)
5. EHR + images (EHR-VAE + RETFound)
6. EHR + images (EHR-VAE + VisionFM)

We compared the ATE estimates and 95% CI from each model and evaluated: 1) The variation in ATE estimates with different adjustment strategies and 2) the sensitivity of ATE to the choice of foundation model for imaging (RETFound vs. VisionFM).

3. Results

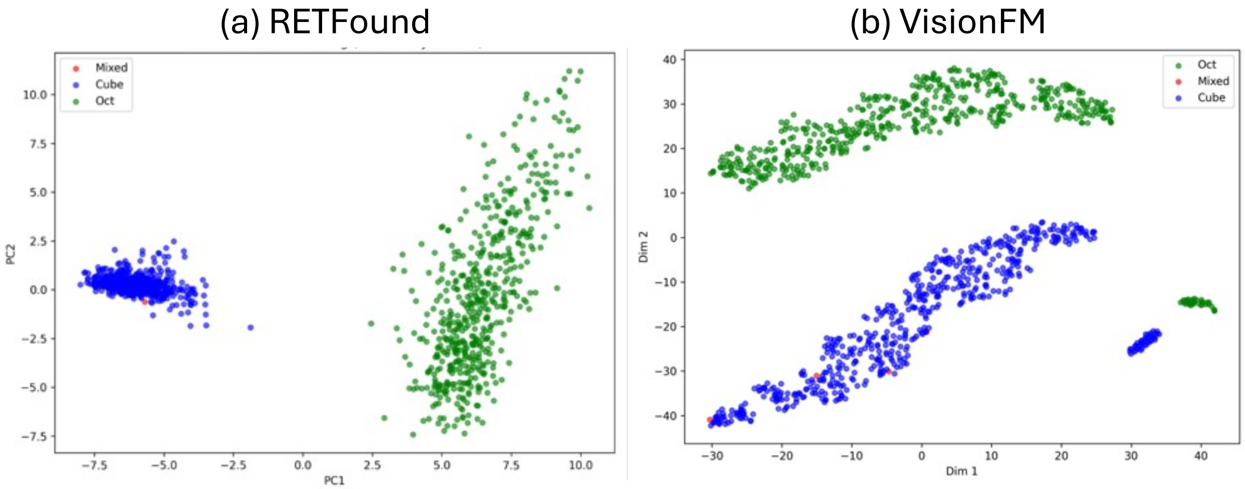

Figure 2. t-SNE visualization of latent embeddings generated by foundation models: (a) RETFound and (b) VisionFM. Each point represents a patient, and colors indicate the OCT imaging device. Clear separation by device suggests that both models capture device-specific features.

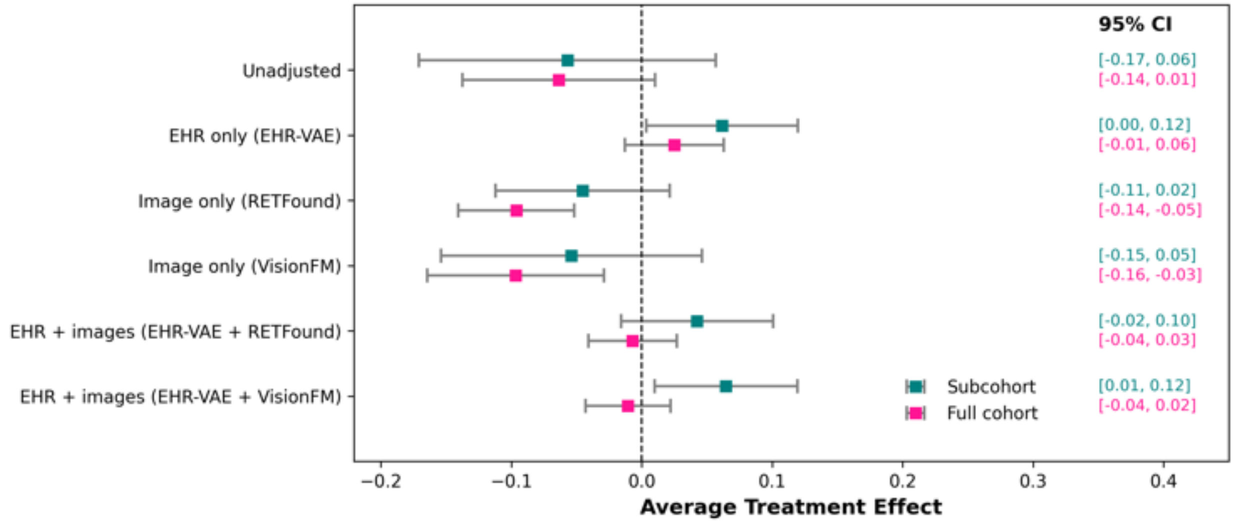

Figure 3. Average treatment effect estimation across adjustment strategies. The full cohort includes all patients in the study population and the sub-cohort includes patients with worse baseline VA. A positive ATE indicates that aflibercept is better at improving vision than bevacizumab.

There were 766 patients with DME, among whom 538 were new aflibercept users and 228 were new bevacizumab users. Of these, 336 (43.9%) patients had baseline visual acuity (VA) worse than 20/50.

Due to the lack of ground truth ATE, we qualitatively evaluated our estimates by comparing to relevant RCTs. The DRCR Retina Network Protocol T trial demonstrated that aflibercept yields greater mean VA gains than bevacizumab at one year in patients with baseline VA of 20/50 or worse, whereas in patients with better baseline VA (20/32 to 20/40), the mean VA improvement is similar between the two agents.[11] Across all adjustment strategies, the multi-modal approach was most consistent with these RCT findings: in the full cohort, the multi-modal estimate was not statistically significantly different from zero, indicating comparable effectiveness between the treatments, while in the worse VA subgroup, the multi-modal analysis found that aflibercept was associated with a higher probability of vision improvement compared to bevacizumab.

Our results showed that (1) the choice of modalities used for confounding adjustment can substantially influence treatment effect estimates, with multi-modal adjustment yielding estimates more consistent with those reported in the RCT; and (2) the two foundation models produced similar estimates, suggesting a degree of robustness in the learned imaging embeddings.

4. Conclusion

Our findings highlight the importance of incorporating multi-modal data for confounding adjustment, which produced treatment effect estimates consistent with established RCT evidence. These results also suggest that foundation models can robustly learn imaging features that contribute to reliable effect estimation in real-world settings.

References

[1] Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A. Review of multimodal machine learning approaches in healthcare. Information Fusion. 2025 Feb 1;114:102690.

[2] Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nature medicine. 2022 Sep;28(9):1773-84.

[3] Soenksen LR, Ma Y, Zeng C, Boussioux L, Villalobos Carballo K, Na L, Wiberg HM, Li ML, Fuentes I, Bertsimas D. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ digital medicine. 2022 Sep 20;5(1):149.

[4] Sun Z, Lin M, Zhu Q, Xie Q, Wang F, Lu Z, Peng Y. A scoping review on multimodal deep learning in biomedical images and texts. Journal of biomedical informatics. 2023 Oct 1;146:104482.

[5] Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, Lee AY, Kawasaki R, van Wijngaarden P, Grzybowski A, He M. Large language models and their impact in ophthalmology. The Lancet Digital Health. 2023 Dec 1;5(12):e917-24.

[6] Deshpande S, Wang K, Sreenivas D, Li Z, Kuleshov V. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. Advances in Neural Information Processing Systems. 2022 Dec 6;35:10931-44.

[7] Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, Liu T, Xu M, Lozano MG, Woodward-Court P, Kihara Y. A foundation model for generalizable disease detection from retinal images. Nature. 2023 Oct 5;622(7981):156-63.

[8] Qiu J, Wu J, Wei H, Shi P, Zhang M, Sun Y, Li L, Liu H, Liu H, Hou S, Zhao Y. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence.

NEJM AI. 2024 Nov 27;1(12):AIoa2300221.

[9] Kingma DP, Welling M. Auto-Encoding Variational Bayes arXiv preprint arXiv:1312.6114; 2013 Dec 20.

[10] Stewart MW. Anti-VEGF therapy for diabetic macular edema. Current diabetes reports. 2014 Aug;14:1-0.

[11] Wells JA, Glassman AR, Ayala AR, Jampol LM, Bressler NM, Bressler SB, Brucker AJ, Ferris FL, Hampton GR, Jhaveri C, Melia M. Aflibercept, bevacizumab, or ranibizumab for diabetic macular edema: two-year results from a comparative effectiveness randomized clinical trial. Ophthalmology. 2016 Jun 1;123(6):1351-9.

[12] Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, Kong T. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832. 2021 Nov 15.

[13] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005 Dec;61(4):962-73.

[14] Klaassen S, Teichert-Kluge J, Bach P, Chernozhukov V, Spindler M, Vijaykumar S. Doublemldeep: Estimation of causal effects with multimodal data. arXiv preprint arXiv:2402.01785. 2024 Feb 1.

[15] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. Proceedings of the IEEE. 2021 Feb 26;109(5):612-34.