REVEAL: Multimodal Vision–Language Alignment of Retinal Morphometry and Clinical Risks for Incident AD and Dementia Prediction

Seowung Leem[1] leem.s@ufl.edu
Lin Gu[2] rin.tani.e8@tohoku.ac.jp
Chenyu You[3,4] chenyu.you@stonybrook.edu
Kuang Gong[1] KGong@bme.ufl.edu
Ruogu Fang[*1] Ruogu.Fang@bme.ufl.edu
[1] *J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, United States*
[2] *Research Institute of Electrical Communication, Tohoku University, Japan*
[3] *Department of Applied Mathematics & Statistics, Stony Brook University, United States*
[4] *Department of Computer Science, Stony Brook University, United States*

Editors: Under Review for MIDL 2026

Abstract

The retina provides a unique, noninvasive window into Alzheimer's disease and dementia, capturing early structural changes through morphometric features, while systemic and lifestyle risk factors reflect well-established contributors to AD and dementia susceptibility long before clinical symptom onset. However, current retinal analysis frameworks typically model imaging and risk factors separately, preventing them from capturing the joint multimodal patterns that are critical for early risk prediction. Moreover, existing methods rarely incorporate mechanisms to organize or align patients with similar retinal and clinical characteristics, limiting their ability to learn coherent cross-modal associations. To address these limitations, we introduce REVEAL (**RE**tinal-risk **V**ision-language **E**arly **A**lzheimer's **L**earning) that aligns color fundus photographs with individualized disease-specific risk profiles for incident AD and dementia prediction on average 8 years before diagnosis (range: 1–11 years). Because real-world risk factors are structured questionnaire data, we first translate them into clinically interpretable narratives compatible with pretrained vision-language models (VLMs). We further propose a group-aware contrastive learning (GACL) strategy that clusters patients with similar retinal morphometry and risk factors as positive pairs, strengthening multimodal alignment. This unified representation-learning framework substantially outperforms state-of-the-art retinal imaging models paired with clinical text encoders, as well as general VLMs, demonstrating the value of jointly modeling retinal biomarkers and clinical risk factors. By providing a generalizable, noninvasive approach for early AD and dementia risk stratification, REVEAL has the potential to enable earlier interventions and improve preventive care at the population level.

Keywords: Retinal morphometry, risk factors, Alzheimer's disease and related dementia, Vision-language alignment, Contrastive learning

[*] Corresponding Author

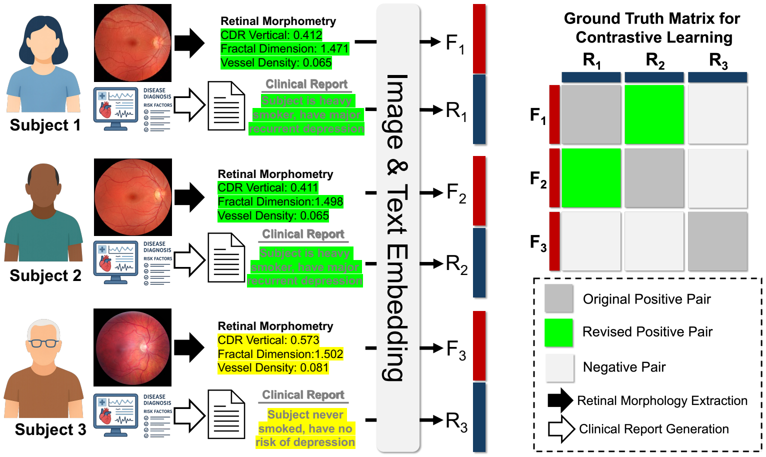

Figure 1: Schematic of clinical scenario and proposed method.

1. Introduction

Alzheimer's disease and dementia are progressive neurodegenerative diseases that manifest years before clinical symptom onset. Early identification of individuals at risk is critical for timely intervention and prevention. The retina offers a unique, noninvasive window into AD and dementia. Retinal morphometric features, referring to a set of quantitative measurements characterizing the size, shape, and structure of retinal components, have been shown to reflect early neurodegenerative changes and amyloid-β or tau deposition in the brain (Cheung et al., 2021; Koronyo et al., 2017; Ravichandran et al., 2025; Byun et al., 2021; Snyder et al., 2016). Parallel to retinal alterations, AD and dementia risk is strongly influenced by systemic and lifestyle factors (Leshner et al., 2017; Sprecher et al., 2017; Xiong et al., 2023; Hayden et al., 2024; Huszár et al., 2024; Livingston et al., 2024). While retinal morphometry captures early neurodegenerative signatures, risk factors provide complementary information on modifiable factors that contribute to disease susceptibility. This convergence suggests that jointly modeling retinal biomarkers and systemic risk factors could improve early AD and dementia prediction beyond what either modality can achieve alone.

Despite this potential, current approaches typically analyze retinal images and risk factors separately, limiting their ability to capture the complex multimodal relationships underlying preclinical AD and dementia. Conventional contrastive learning frameworks often fail to align patients who share both retinal and systemic risk characteristics, leading to overlooked clinical commonalities (Figure 1). Moreover, structured risk-factor data from questionnaires cannot be directly incorporated into standard vision-language models (VLMs), which are pretrained on natural language, creating a modality gap.

To address these challenges, we introduce **REVEAL** (**RE**tinal-risk **V**ision-language **E**arly **A**lzheimer's **L**earning), a novel VLM-based framework that integrates retinal morphometric features with individualized disease-specific risk profiles. Structured risk factors are first transformed into clinically meaningful narratives using large language models, enabling seamless multimodal representation learning. We further propose a **group-aware contrastive learning strategy** that leverages intra-modality similarity to identify clinically aligned individuals, capturing shared pathophysiological patterns across subjects. This

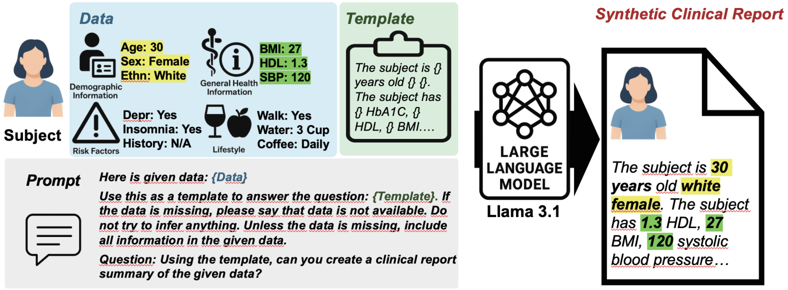
Data
Age: 30
Sex: Female
Ethn: White
Demographic Information
BMI: 27
HDL: 1.3
SBP: 120
General Health Information
Depr: Yes
Insomnia: Yes
History: N/A
Risk Factors
Walk: Yes
Water: 3 Cup
Coffee: Daily
Lifestyle
Subject
Template
The subject is {} years old {} {}. The subject has {} HbA1C, {} HDL, {} BMI....
Prompt
Here is given data: {Data}
Use this as a template to answer the question: {Template}. If the data is missing, please say that data is not available. Do not try to infer anything. Unless the data is missing, include all information in the given data.
Question: Using the template, can you create a clinical report summary of the given data?
LARGE LANGUAGE MODEL
Llama 3.1
Synthetic Clinical Report
The subject is 30 years old white female. The subject has 1.3 HDL, 27 BMI, 120 systolic blood pressure…

Figure 2: Schematic overview of how a synthetic clinical report is generated.

approach allows REVEAL to learn unified representations that more accurately reflect the interplay between retinal biomarkers and systemic risk factors, offering improved early AD and dementia risk stratification. Our work has the following contributions:

- We introduce **REVEAL**, the first framework to jointly model fundus images and individualized AD and dementia risk factors by translating structured questionnaires into clinically meaningful narratives compatible with pretrained VLMs.
- We propose a **group-aware contrastive learning strategy** that identifies subjects sharing similar retinal morphometry and risk profiles, enabling coherent and clinically aligned multimodal representation learning.
- REVEAL achieves state-of-the-art performance in predicting incident AD and dementia on average 8 years before clinical onset (AD: mean = 8.68 years, range = 2.38–11.58 years; dementia: mean = 8.49 years, range = 1.50–11.58 years) over retinal-only, clinical-text, and general VLM baselines, providing a generalizable, noninvasive approach for population-level early AD and dementia risk stratification.

2. Method

2.1. Overview of REVEAL Framework

The REVEAL framework was designed to operate in two stages. First, it aligned fundus images with individualized AD and dementia risk factors using a CLIP-style contrastive learning approach with our novel image-text pairing strategy. This enabled the model to learn multimodal relationships between colored fundus photography (CFP) and biological, phenotypic, and clinical markers of preclinical AD and dementia. Second, the learned joint representations were utilized in a downstream classifier to predict incident, preclinical AD and dementia (see Section 3.1 for details).

2.2. Constructing Clinical Report and Group-Aware Labels for Contrastive Learning

2.2.1. Synthetic Clinical Report Generation

Direct application of CLIP was not feasible because the risk factors are represented as structured, tabular variables rather than natural-language descriptions. However, alignment of fundus images with risk factors required a shared representation space that VLM can operate on. To bridge the modality gap between structured risk-factor variables and the natural-language input required by VLMs, we synthesized standardized clinical-style

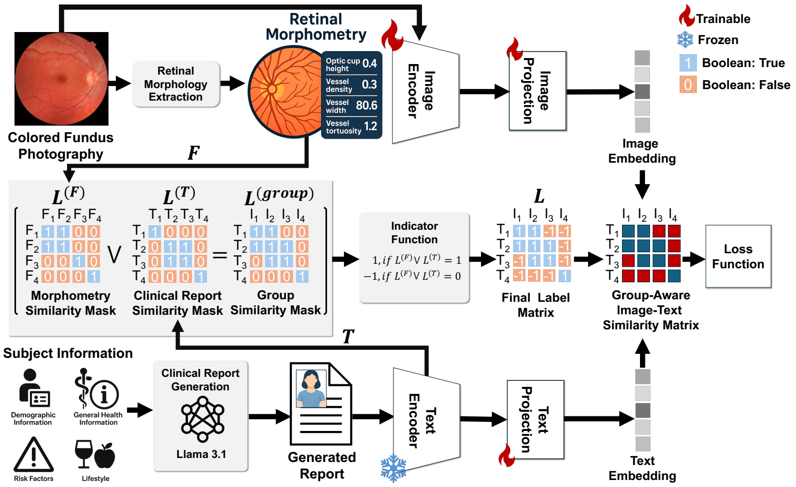

Figure 3: Schematic overview of how GACL is performed.

narratives from tabular health data (Figure 2). This transformation enabled the VLM to interpret the tabular risk factors in a linguistically contextualized form and facilitates multimodal alignment between fundus images and clinical attributes relevant to AD and dementia. Using the LLaMA-3.1 API as the text generation engine (Grattafiori et al., 2024), we converted each participant's risk factor profile into a synthetic clinical report. For each subject, the LLM was provided with (1) a template prompt, (2) the subject's structured risk factor values, and (3) explicit instructions for generating a concise medical summary. The template was adapted from the "Patient Information" section of the CARE clinical case report guideline (Gagnier et al., 2013), ensuring that the synthesized narratives follow established clinical documentation conventions. The input prompt was designed to map the tabular information 1:1 into a template to prevent potential variability (Appendix A). This process produced consistent, clinically meaningful text representations that enable seamless integration of structured health information into our multimodal predictive framework.

2.2.2. Group-aware Contrastive Learning Strategy

Conventional CLIP-style frameworks often fail in the medical domain (Radford et al., 2021). Prior studies showed that naive CLIP approaches struggle to capture the complex semantic relationships between images and disease-level information, highlighting the need for domain-specific strategies (Wang et al., 2022; Eslami et al., 2023). In our context, individuals sharing both retinal and systemic AD and dementia risk characteristics must be grouped during training, since conventional contrastive learning only treats image-text pairs from the same subject as positive matches. To mitigate these gaps and enable the model to capture shared pathophysiological patterns across different modalities, we designed a group-aware contrastive learning (GACL). To introduce explicit clinical grounding, our GACL leverages morphometric features extracted directly from CFP, rather than solely on latent representations from image encoders. This addressed limitations from prior works that attempted to improve the shortcomings of conventional CLIP by introducing the image-level or latent-level similarity (Du et al., 2024; Wu et al., 2024), which lacked explicit clinical grounding to find phenomenologically similar individuals with clinical relevance. The GACL was inspired by (Bulat et al., 2024).

As shown in Figure 3, $\mathbf{F} \in \mathbb{R}^{N \times K}$ and $\mathbf{T} \in \mathbb{R}^{N \times D}$ denote the z-normalized morphometric feature matrix and l2-normalized embeddings clinical report matrix for all N samples in a training batch, respectively. Here, K represents the number of morphometric features and D denotes the dimensionality of the text embedding space. To quantify the pairwise relationship within each modality, we computed the intra-morphometry similarity matrices $\mathbf{S}^{(\mathbf{F})} \in \mathbb{R}^{N \times N}$ for the fundus image and intra-clinical report similarity matrix $\mathbf{S}^{(\mathbf{T})} \in \mathbb{R}^{N \times N}$ for text.

$$\mathbf{S}^{(\mathbf{F})} = \mathbf{F} \cdot \mathbf{F}^{\top}, \quad \text{and} \quad \mathbf{S}^{(\mathbf{T})} = \mathbf{T} \cdot \mathbf{T}^{\top}. \tag{1}$$

Each entry in $\mathbf{S}^{(\mathbf{F})}$ characterizes how similar the retinal morphometric profiles of two subjects are, with a larger value indicating closer structural resemblance. Likewise, $\mathbf{S}^{(\mathbf{T})}$ captures the semantic similarity between the clinical report embeddings, demonstrating the degree to which two subjects share encoded risk factor profiles. To identify subjects with similar characteristics, we thresholded both similarity matrices using modality-specific thresholds $\boldsymbol{\tau_F}$ and $\boldsymbol{\tau_T}$, yielding binary similarity masks $\mathbf{L}^{(\mathbf{F})}$ and $\mathbf{L}^{(\mathbf{T})}$. In each mask, a value of **1 (Boolean True)** indicates a similar sample pair, while **0 (Boolean False)** indicates a dissimilar pair.

$$\mathbf{L}^{(\mathbf{F})} = \begin{cases} 1(True), & \text{if } \mathbf{S}^{(\mathbf{F})} > \boldsymbol{\tau_F} \\ 0(False), & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{L}^{(\mathbf{T})} = \begin{cases} 1(True), & \text{if } \mathbf{S}^{(\mathbf{T})} > \boldsymbol{\tau_T} \\ 0(False), & \text{otherwise} \end{cases} \tag{2}$$

To integrate information across modalities, we obtained a group similarity mask $\mathbf{L}^{(\mathbf{group})}$ by applying a logical OR operation between two modality-specific masks. Finally, the group similarity mask was mapped by an indicator function, resulting in a contrastive learning-compatible final label matrix $\mathbf{L}$, where entries of 1 were preserved, and 0s were converted to -1. This formulation preserved similarity relationships across modalities, ensuring that image-text alignment benefits from both structural consistency (from morphometry) and semantic consistency (from clinical reports). By reinforcing agreement between intra-modal similarity, image-text pairings were improved to maximize the learning efficiency between retinal morphometric features and risk factors.

$$\mathbf{L}^{(\text{group})} = \begin{cases} 1, & \text{if } \mathbf{L}^{(\mathbf{F})} \vee \mathbf{L}^{(\mathbf{T})} = 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{L} = \begin{cases} 1, & \text{if } \mathbf{L}^{(\text{group})} = 1, \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

2.3. Image-Text Alignment Learning with REVEAL

2.3.1. REVEAL Architecture

The REVEAL framework was built on a standard contrastive vision-language learning setup to capture joint patterns between fundus images and AD and dementia risk factors. As shown in Figure 3, we used RETFound (Zhou et al., 2023) as the image encoder and GatorTron (Yang et al., 2022) as the text encoder, adding only lightweight projection layers to align their feature dimensions. During each forward pass, a raw fundus image and its synthesized clinical report were encoded and projected into a shared latent space. Retinal morphometrics and clinical narratives were further integrated into the GACL procedure to construct a label matrix. Finally, a group-aware image-text similarity matrix was computed using image embedding, text embedding, and a label matrix. The trainable REVEAL components were denoted as "flame" in Figure 3. This design enabled REVEAL to leverage both retinal imaging priors from foundation models and semantic priors from clinically trained language models.

2.3.2. Contrastive learning

With GACL, the conventional contrastive objective was no longer applicable because it accommodated only a single positive pair per sample. Therefore, we adopted the loss from the prior work (Bulat et al., 2024) to support multiple clinically aligned pairs.

$$\mathcal{L} = -\frac{1}{N_{\text{img}} N_{\text{txt}}} \sum_{i=1}^{N_{\text{img}}} \sum_{j=1}^{N_{\text{txt}}} \log \left(\frac{1}{1 + \exp\big(l_{ij} \left(-s_{ij}/\tau + \beta\right)\big)} \right), \tag{4}$$

N_{img} and N_{txt} denote the number of images and texts in a training batch. The label term $l_{ij} \in \{+1, -1\}$ is the (i, j)-th entry of the final label matrix L, with $l_{ij} = 1$ indicating a similar (positive) image–text pair and $l_{ij} = -1$ indicating a dissimilar (negative) pair. The similarity value s_{ij} is computed as the cosine similarity between the corresponding i-th image and j-th text embeddings obtained from the REVEAL framework. The temperature parameter is fixed at $\tau = 0.07$. The bias term β is introduced to stabilize early training by reducing the initial loss, which is otherwise dominated by the large number of negative pairs. Including β, all hyperparameters (learning rate, eps, weight decay) and similarity thresholds ($\boldsymbol{\tau_F}$ and $\boldsymbol{\tau_T}$) were chosen using an Optuna, hyperparameter optimization framework (Akiba et al., 2019), which identifies the optimal configuration within user-defined search ranges (details in Appendix B).

2.4. Study Population and Data Preprocessing

2.4.1. Subject Selection

Table 1: Demographic characteristics of the UK Biobank participants across the training, validation, and test cohorts.

	Train (n=30,462)	Validation (n=3,384)	Test (n=5,396)
Gender: (male %)	45.10	45.41	45.10
Age: mean (s.d)	55.53 (8.24)	55.78 (8.12)	55.52 (8.17)
Ethnicity: (British %)	84.08	83.51	88.51

Color fundus photographs (CFPs) and AD and dementia-related risk factors were obtained from the UK Biobank (Sudlow et al., 2015). A total of 39,242 participants with high-quality CFPs were included and allocated into training (n=30,462), validation (n=3,384), and test (n=5,396) sets (Table 1, preprocessing details in Section 2.4.2). These splits each served a distinct role within the REVEAL framework. The training and validation sets were used solely in Stage 1 for representation alignment, with the validation set guiding hyperparameter tuning and similarity-threshold selection, while the test set was reserved for Stage 2 AD and dementia prediction.

All participants who later developed incident AD or dementia were assigned to the test set, and only participants free of both prevalent and incident disease were included in the training and validation sets. Incident diagnoses were identified using UK Biobank dementia fields (42018, 42020, 42022, 42024). Among individuals with high-quality CFPs, 86 developed incident AD (mean time to diagnosis: 8.68 years; range: 2.38–11.58) and 93 developed dementia of any subtype (mean: 8.49 years; range: 1.50–11.58).

To form the final evaluation cohort, control subjects without incident AD and dementia were sampled from the test pool to achieve an approximate 12% disease prevalence, consistent with estimates for adults aged $\geq$65 years (Xiaopeng et al., 2025), while maintaining age and gender matched distributions (AD controls = 1,077; dementia controls = 1,139). From this cohort, 931 subjects (862 controls, 69 AD) for AD prediction and 985 subjects (911 controls, 74 dementia) for dementia prediction were used to train SVM models with 5-fold cross-validation. The remaining subjects, 232 (215 controls, 17 AD) for AD prediction and 247 (228 controls, 19 dementia) for dementia prediction, were held out as an independent test set. Cohort characteristics for downstream prediction tasks are provided in Tables 6 and 7 (Appendix C), and the distribution of onset years for AD and dementia are shown in Figure 4 (Appendix D)

2.4.2. Risk Factor Compilation and Retinal Image Processing

A comprehensive set of demographic, behavioral, cognitive, and lifestyle variables was compiled as risk factors based on established epidemiological evidence (Leshner et al., 2017; Sprecher et al., 2017; Xiong et al., 2023; Hayden et al., 2024; Huszár et al., 2024; Livingston et al., 2024). The full list of these risk factors are provided in Appendix E. For the CFPs, image preprocessing and retinal morphometric feature extraction were carried out using the AutoMorph fundus morphology quantification pipeline (Zhou et al., 2022). A total of 136,994 CFPs were available from the initial UK Biobank assessment visit. AutoMorph first applied a convolutional neural network–based quality-control module that classified images as low, moderate, or good quality. Following automated quality filtering and subsequent manual review, 66,251 high-quality images from 39,242 participants were retained for analysis. From these curated images, AutoMorph produced a structured set of retinal morphometric features (K=17; full list provided in Appendix F). These structural features have been shown in prior research to exhibit measurable differences in both preclinical and clinical stages of AD and dementia (Frost et al., 2013; Sharafi et al., 2019; Valenti, 2011; Ong et al., 2014; Armstrong et al., 2021). To maintain consistent anatomical orientation across eyes, all right-eye images were horizontally flipped before feature extraction.

3. Experiments

3.1. Downstream Tasks

3.1.1. Incident AD and incident dementia prediction

We evaluated REVEAL on two prediction tasks: incident AD and incident dementia. For both tasks, we trained a multimodal SVM with an RBF kernel to perform binary classification, distinguishing individuals who later developed AD and dementia from those who remained cognitively normal. The SVM produced probabilistic outputs, providing likelihood estimates for being AD/dementia-positive versus control. Each subject was represented by a concatenated multimodal feature vector composed of L2-normalized CFP image embeddings and text embeddings extracted from the REVEAL encoders. Because not all participants had both left and right CFPs available, a single CFP embedding was randomly sampled from the available views for each subject. Class-weighted training was used to mitigate the imbalance between incident cases and controls. SVM hyperparameters (C and γ) were tuned using 5-fold cross-validation, and the best-performing model was subsequently evaluated on the independent hold-out test set. All reported results correspond to this final evaluation.

3.1.2. Comparison models

To evaluate REVEAL, we compared its performance with several strong fundus-based foundation models: RETFound (CFP) (Zhou et al., 2023), RET-CLIP (Du et al., 2024), and KeepFIT-CFP (Wu et al., 2024), as well as multimodal vision-language models, including PMC-CLIP (Lin et al., 2023) and BiomedCLIP (Zhang et al., 2025). Because RETFound provides only image embeddings, we paired it with GatorTron (Yang et al., 2022) to enable multimodal representation. In addition to these baselines, we trained a tabular SVM using clinical variables and CFP-derived morphometric features, applying most-frequent imputation for categorical variables and KNN imputation (k = 5) for continuous variables. All models followed the same training and testing protocol as the multimodal SVM. Each experiment was repeated 10 times with different random seeds, and we report the average performance across runs.

3.1.3. Threshold Evaluation of REVEAL Framework

In REVEAL, thresholds $\boldsymbol{\tau_F}$, and $\boldsymbol{\tau_T}$ from GACL determine which image-text pairs should be grouped to share information, to learn shared representations among phenomenologically similar samples. Thresholds that are too low introduce noise by aligning dissimilar pairs, whereas thresholds that are too high restrict the model's ability to capture meaningful cross-modal relationships. To assess their influence on predictive performance, we trained the model using varying threshold configurations. In each experiment, one threshold was fixed at the optimal value determined during optimization, while the other was varied systematically. Threshold candidates were chosen from the quartiles of the morphometric and text similarity distributions in the development set.

3.1.4. Evaluating Clinically Grounded Similarity in GACL

As previously noted in Section 2.2.2, prior works have attempted to remedy the shortcomings of conventional CLIP by incorporating image-level or latent-level similarity. To evaluate the contribution of clinically grounded similarity in GACL, we compared downstream prediction performance under two configurations: (1) GACL using morphometric features as the source of image-image similarity, and (2) GACL using similarity computed directly from the image embeddings produced by the image encoder. This comparison allowed us to isolate the benefit of explicit clinical grounding for identifying phenotypically similar subjects and enhancing downstream AD and dementia prediction.

3.2. Result

3.2.1. Group-aware contrastive learning improves the Incident AD and dementia Prediction

In the incident AD prediction task (Table 2), REVEAL achieved the best performance across nearly all evaluation metrics, including AUROC, balanced accuracy, F1-Score, and Matthew's Correlation Coefficient (MCC). Notably, the multimodal SVM trained on REVEAL embeddings substantially outperformed a baseline SVM trained directly on tabular risk factors and raw retinal morphometric features, demonstrating that vision-language embeddings effectively transform raw modalities into enriched representations. Incorporating

GACL further improved performance by aligning patients with similar retinal morphometry and risk profiles, enhancing overall predictive power. In the broader incident-dementia prediction task (Table 3), the SVM using REVEAL embeddings again outperformed baseline SVMs and other vision-language models. These results indicate that group-aware alignment strengthens multimodal representation learning, with the greatest impact observed in incident dementia, where retinal structural features closely correspond to disease-specific biomarkers. Importantly, all CFPs in embedding learning were collected from cognitively normal participants at baseline, emphasizing that REVEAL, combined with a multimodal SVM, can identify preclinical AD and dementia risk by leveraging the complementary information between retinal morphometry and systemic risk factors.

3.2.2. Impact of Thresholds on REVEAL Performance

The relative percentage differences in downstream prediction performance between the model trained with the optimal threshold and those trained under varying $\boldsymbol{\tau_F}$ and $\boldsymbol{\tau_T}$ in REVEAL are shown in Figure 5 of Appendix G. Compared to the performance metric from the REVEAL with the optimal threshold (gray horizontal line, where values below 0 indicate worse performance and values above 0 indicate improvement), other models trained with different image or text thresholds did not yield better performance in most cases for both AD and dementia. For AD, using highest $\boldsymbol{\tau_F}$ produced best accuracy, F1-score and MCC, but at the cost of a reduced AUROC. This highlights the importance of carefully calibrated thresholds, as multimodal associations are highly sensitive to pairing phenotypically similar pairs and avoiding weakly related alignments. Distinct trends were observed between image and text modalities. For images, higher thresholds demonstrated better performance, suggesting that lower thresholds introduce noise by forcing dissimilar samples to be similar. Conversely, for text embeddings, lower thresholds led to higher predictive performance, indicating that learning benefits when a broader range of semantically related texts are considered similar. However, the generalizability of these trends requires further validation in other domains and different datasets.

3.2.3. Impact of Clinical vs. Latent Similarity in GACL

The AD and dementia prediction results using morphometric features and the model's latent features in image-image similarity computation of GACL are shown in Table 4. For this experiment, the threshold for the image latent was determined as the third quartile of

Table 2: Performance of the incident Alzheimer's Disease prediction task. The best results for each modality are in bold text.

	AUROC	Balanced Accuracy	F1-Score	MCC
Baseline SVM	0.5721	0.5517	0.1290	0.0561
KeepFIT-CFP	0.5264	0.5411	0.1481	0.0455
BiomedCLIP	0.5941	0.5492	0.1468	0.0670
RETCLIP	0.5738	0.5633	0.1749	0.1050
PMC-CLIP	0.5281	0.5671	0.1953	0.1241
RETFound+GatorTron	0.6418	0.5835	0.1656	0.1286
Ours (no GACL)	0.5951	0.5622	0.2129	0.0763
Ours (with GACL)	**0.6439**	**0.6029**	**0.2270**	**0.1562**

Table 3: Performance of the incident dementia prediction task. The best results for each modality are in bold text.

	AUROC	Balanced Accuracy	F1-Score	MCC
Baseline SVM	0.5714	0.5431	0.1349	0.0595
KeepFIT-CFP	0.4903	0.4879	0.1113	-0.0157
BiomedCLIP	0.5430	0.5101	0.1091	0.0072
RETCLIP	0.6029	0.5397	0.1500	0.0751
PMC-CLIP	0.4605	0.5550	0.1725	0.0971
RETFound+GatorTron	0.6596	0.6121	0.2239	0.1532
Ours (no GACL)	0.5886	0.5768	0.1912	0.1003
Ours (with GACL)	**0.6782**	**0.6294**	**0.2599**	**0.1915**

the similarity distribution in the development set ($\boldsymbol{\tau_F}$=0.9974). In both the incident AD and dementia prediction cases, incorporating morphometric features consistently yielded superior performance. This indicates that clinically grounded morphometric similarity provides a more reliable and meaningful signal for identifying individuals who share similar retinal and systemic phenotypes, enabling richer and more discriminative representational learning.

Table 4: Performance of the incident AD and dementia prediction with different image-image similarity methods

	AUROC	Balanced Accuracy	F1-Score	MCC
AD				
Latent Feature	0.6110	0.5699	0.1920	0.1015
Morphometric Feature	**0.6439**	**0.6029**	**0.2270**	**0.1562**
Dementia				
Latent Feature	0.6550	0.6149	0.2286	0.1568
Morphometric Feature	**0.6782**	**0.6294**	**0.2599**	**0.1915**

4. Conclusion

In this paper, we present REVEAL, a multimodal VLM framework that improves embedding learning for incident AD and dementia prediction by explicitly aligning retinal morphometric features with individualized risk factors. Our group-aware contrastive learning strategy identifies clinically meaningful groups and patients with similar retinal and risk profiles, and enhances cross-modal representation learning. This alignment improves AD and dementia prediction diagnosed after an average of 8 years after baseline visit. These gains demonstrate that multimodal alignment reflects the strong correspondence between AD-specific risk factors and retinal structural features. Moreover, transforming structured clinical data into narrative form leverages the semantic richness of pretrained language models, further strengthening multimodal associations and boosting predictive performance. These results underscore the value of clinically contextualized representation learning in VLMs for early AD and dementia risk stratification.

Acknowledgments

This research has been conducted using data from UK Biobank, a major biomedical database under application ID 48388. This material is based upon work supported by the National Science Foundation under Grant No. (NSF 2123809).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, New York, NY, USA, July 2019. ACM.

Grayson W Armstrong, Leo A Kim, Filippos Vingopoulos, et al. Retinal imaging findings in carriers with PSEN1-associated early-onset familial alzheimer disease before onset of cognitive symptoms. *JAMA Ophthalmol.*, 139(1):49–56, January 2021.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182, 2024.

Min Soo Byun, Sung Wook Park, Jun Ho Lee, Dahyun Yi, So Yeon Jeon, Hyo Jung Choi, Haejung Joung, Un Hyung Ghim, Un Chul Park, Yu Kyeong Kim, Seong A Shin, Hyeong Gon Yu, Dong Young Lee, and KBASE Research Group. Association of retinal changes with alzheimer disease neuroimaging biomarkers in cognitively normal individuals. *JAMA Ophthalmol.*, 139(5):548–556, May 2021.

Carol Y. Cheung, Vincent Mok, Paul J. Foster, Emanuele Trucco, Christopher Chen, and Tien Yin Wong. Retinal imaging in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(9):983–994, September 2021. ISSN 0022-3050, 1468-330X. doi: 10.1136/jnnp-2020-325347. URL https://jnnp.bmj.com/content/92/9/983. Publisher: BMJ Publishing Group Ltd Section: General neurology.

Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. RET-CLIP: A Retinal Image Foundation Model Pre-trained with Clinical Diagnostic Reports, August 2024. URL http://arxiv.org/abs/2405.14137. arXiv:2405.14137 [cs].

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.88. URL https://aclanthology.org/2023.findings-eacl.88/.

S Frost, Y Kanagasingam, H Sohrabi, J Vignarajan, P Bourgeat, O Salvado, V Villemagne, C C Rowe, S Lance Macaulay, C Szoeke, K A Ellis, D Ames, C L Masters, S Rainey-Smith, R N Martins, and AIBL Research Group. Retinal vascular biomarkers for early detection and monitoring of alzheimer's disease. *Transl. Psychiatry*, 3(2):e233, February 2013.

Joel J Gagnier, Gunver Kienle, Douglas G Altman, David Moher, Harold Sox, David Riley, and CARE Group*. The CARE guidelines: Consensus-based clinical case reporting guideline development. *Glob. Adv. Health Med.*, 2(5):38–43, September 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

K.M. Hayden, M.M. Mielke, J.K. Evans, R. Neiberg, D. Molina-Henry, M. Culkin, S. Marcovina, K.C. Johnson, O.T. Carmichael, S.R. Rapp, B.C. Sachs, J. Ding, H. Shappell, L. Wagenknecht, J.A. Luchsinger, and M.A. Espeland. Association between Modifiable Risk Factors and Levels of Blood-Based Biomarkers of Alzheimer's and Related Dementias in the Look AHEAD Cohort. *JAR life*, 13:1–21, January 2024. ISSN 2534-773X. doi: 10.14283/jarlife.2024.1. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10775955/.

Zsolt Huszár, Alina Solomon, Marie Anne Engh, Vanda Koszovácz, Tamás Terebessy, Zsolt Molnár, Péter Hegyi, András Horváth, Francesca Mangialasche, Miia Kivipelto, and Gábor Csukly. Association of modifiable risk factors with progression to dementia in relation to amyloid and tau pathology. *Alzheimer's Research & Therapy*, 16:238, October 2024. ISSN 1758-9193. doi: 10.1186/s13195-024-01602-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11515263/.

Yosef Koronyo, David Biggs, Ernesto Barron, David S. Boyer, Joel A. Pearlman, William J. Au, Shawn J. Kile, Austin Blanco, Dieu-Trang Fuchs, Adeel Ashfaq, Sally Frautschy, Gregory M. Cole, Carol A. Miller, David R. Hinton, Steven R. Verdooner, Keith L. Black, and Maya Koronyo-Hamaoui. Retinal amyloid pathology and proof-of-concept imaging trial in Alzheimer's disease. *JCI Insight*, 2(16), August 2017. ISSN 0021-9738. doi: 10.1172/jci.insight.93621. URL https://insight.jci.org/articles/view/93621. Publisher: American Society for Clinical Investigation.

Alan I. Leshner, Story Landis, Clare Stroud, and Autumn Downey, editors. *Preventing Cognitive Decline and Dementia: A Way Forward*. National Academies Press, Washington, D.C., September 2017. ISBN 978-0-309-45959-4. doi: 10.17226/24782. URL https://www.nap.edu/catalog/24782.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents, March 2023. URL http://arxiv.org/abs/2303.07240. arXiv:2303.07240 [cs].

Gill Livingston, Jonathan Huntley, Kathy Y Liu, Sergi G Costafreda, Geir Selbæk, Suvarna Alladi, David Ames, Sube Banerjee, Alistair Burns, Carol Brayne, Nick C Fox, Cleusa P Ferri, Laura N Gitlin, Robert Howard, Helen C Kales, Mika Kivimäki, Eric B Larson, Noeline Nakasujja, Kenneth Rockwood, Quincy Samus, Kokoro Shirai, Archana Singh-Manoux, Lon S Schneider, Sebastian Walsh, Yao Yao, Andrew Sommerlad, and Naaheed Mukadam. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. *The Lancet*, 404(10452):572–628, August 2024. ISSN 0140-6736. doi: 10.1016/S0140-6736(24)01296-0. URL https://doi.org/10.1016/S0140-6736(24)01296-0. Publisher: Elsevier.

Yi-Ting Ong, Saima Hilal, Carol Yim-Lui Cheung, et al. Retinal vascular fractals and cognitive impairment. *Dement. Geriatr. Cogn. Dis. Extra*, 4(2):305–313, May 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Swetha Ravichandran, Peter J Snyder, Jessica Alber, Charles F Murchison, Lauren E Chaby, Andreas Jeromin, and Edmund Arthur. Association and multimodal model of retinal and blood-based biomarkers for detection of preclinical alzheimer's disease. *Alzheimers. Res. Ther.*, 17(1):19, January 2025.

Sayed Mehran Sharafi, Jean-Philippe Sylvestre, Claudia Chevrefils, Jean-Paul Soucy, Sylvain Beaulieu, Tharick A Pascoal, Jean Daniel Arbour, Marc-André Rhéaume, Alain Robillard, Céline Chayer, Pedro Rosa-Neto, Sulantha S Mathotaarachchi, Ziad S Nasreddine, Serge Gauthier, and Frédéric Lesage. Vascular retinal biomarkers improves the detection of the likely cerebral amyloid status from hyperspectral retinal images. *Alzheimers Dement. (N. Y.)*, 5(1):610–617, October 2019.

Peter J Snyder, Lenworth N Johnson, Yen Ying Lim, Cláudia Y Santos, Jessica Alber, Paul Maruff, and Brian Fernández. Nonvascular retinal imaging markers of preclinical alzheimer's disease. *Alzheimers Dement. (Amst.)*, 4(1):169–178, October 2016.

Kate E. Sprecher, Rebecca L. Koscik, Cynthia M. Carlsson, Henrik Zetterberg, Kaj Blennow, Ozioma C. Okonkwo, Mark A. Sager, Sanjay Asthana, Sterling C. Johnson, Ruth M. Benca, and Barbara B. Bendlin. Poor sleep is associated with CSF biomarkers of amyloid pathology in cognitively normal adults. *Neurology*, 89(5):445–453, August 2017. ISSN 0028-3878. doi: 10.1212/WNL.0000000000004171. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5539733/.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380465/.

Denise A Valenti. Alzheimer's disease and glaucoma: imaging the biomarkers of neurodegenerative disease. *Int. J. Alzheimers. Dis.*, 2010:793931, January 2011.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256/.

Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise, May 2024. URL http://arxiv.org/abs/2405.11793. arXiv:2405.11793 [cs].

Zhu Xiaopeng, Yu Jing, Lai Xia, Wang Xingsheng, Deng Juan, Long Yan, and Li Baoshan. Global burden of alzheimer's disease and other dementias in adults aged 65 years and older, 1991-2021: population-based study. *Front. Public Health*, 13:1585711, July 2025.

Jiayue Xiong, Rozina Bhimani, and Lisa Carney-Anderson. Review of Risk Factors Associated With Biomarkers for Alzheimer Disease. *The Journal of Neuroscience Nursing: Journal of the American Association of Neuroscience Nurses*, 55(3):103–109, June 2023. ISSN 1945-2810. doi: 10.1097/JNN.0000000000000705.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5(1):1–9, December 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00742-2. URL https://www.nature.com/articles/s41746-022-00742-2. Publisher: Nature Publishing Group.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, January 2025. URL http://arxiv.org/abs/2303.00915. arXiv:2303.00915 [cs].

Yukun Zhou, Siegfried K. Wagner, Mark A. Chia, An Zhao, Peter Woodward-Court, Moucheng Xu, Robbert Struyven, Daniel C. Alexander, and Pearse A. Keane. AutoMorph: Automated Retinal Vascular Morphology Quantification Via a Deep Learning Pipeline. *Translational Vision Science & Technology*, 11(7):12, July 2022. ISSN 2164-2591. doi: 10.1167/tvst.11.7.12. URL https://doi.org/10.1167/tvst.11.7.12.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, October 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x. URL https://www.nature.com/articles/s41586-023-06555-x. Publisher: Nature Publishing Group.

Appendix A. Full template for clinical report generation

Template: The subject is <age> years old <ethnic background> <sex>. The average total household of this subject is in between <economic status>. The subject has <HbA1C> HbA1C, <HDL> HDL, <BMI> BMI, <systolic blood pressure> systolic blood pressure,

<diastolic blood pressure> diastolic blood pressure. For lifestyle, the subject is in <employment status>. The subject is <smoking history>, has <depression>, has sleep deprivation <sleep deprivation>, and drinks alcohol <alcohol use>. The subject had his first cannabis at age <age of cannabis initiation> and used cannabis <cannabis use> times. The subject visits family <frequency of family visit>, and <number of leisure activity>. For physical activity, the subject walks <duration of walked 10+ minutes> minutes <number of days/week of walked 10+ minutes> days per week, exercises moderately <duration of moderate activity> minutes for <number of days/week of moderate activity> days a week, and exercises vigorously <duration of vigorous exercise> minutes for <number of days/week of vigorous activity> days a week. For diet, the subject has <cooked vegetable intake> tablespoons of cooked vegetables, <raw vegetable intake> tablespoons of raw vegetables, <fresh fruit intake> tablespoons of fresh fruit, and <dried fruit intake> dried fruit. In addition, the subject has oily fish <oily fish intake>, non-oily fish <non oily fish intake>, processed meat <processed meat intake>, poultry <poultry intake>, beef <beef intake>, lamb <lamb intake>, and pork <pork intake>. The subject has <bread intake> slices of bread per week, with <spread type>. The subject drinks <milk type>, <tea intake> cups of tea, <coffee intake> cups of coffee, <water intake> cups of water per day. The subject puts <salt added to food> in his diet. For cognitive function, the subject remembered <numeric memory> digits in the numeric memory test, scored <fluid intelligence> in a fluid intelligence test, completed trail #1 in <trail-making test A duration> deciseconds with <trail-making test A error counts> errors, and completed trail #2 in <trail-making test B duration> deciseconds with <trail-making test B error counts> errors.

When a risk factor was unavailable (e.g., age of cannabis initiation), the report stated: **No cannabis use was reported at that age** in the <age of cannabis initiation> section.

Appendix B. Implementation details and hyperparameter discovery

The dimension of the projection layer for both image and text encoders was fixed at 1024. The batch size was fixed at 16. The parameter search space and determined values for REVEAL are available in Table 5. The ranges for $\boldsymbol{\tau_F}$ and $\boldsymbol{\tau_T}$ were determined by the 3rd quartile to the 4th quartile range of retinal morphometric similarities and pseudo-clinical report similarity in 85% of the development set. Based on Optuna, learning rate was determined as 2.42e-4, eps was determined as 8.61e-7, weight decay was set to 0.0232, thresholds were determined as $\boldsymbol{\tau_F}$=0.9481 and $\boldsymbol{\tau_T}$=0.9808. When training without GACL, we used the standard InfoNCE loss.

Table 5: Hyperparameter search space and optimal values

Hyperparameter	**Range (min, max)**	**Optimal Value**
learning rate	1e-6, 5e-4	2.42e-4
eps	1e-9, 1e-6	8.61e-7
weight decay	1e-6, 1e-1	0.0232
$\boldsymbol{\tau_F}$	0.2853, 0.9949	0.9480
$\boldsymbol{\tau_T}$	0.9548, 0.9979	0.9808
$\boldsymbol{\beta}$	-5, 0	-0.6319

Appendix C. Demographic information of incident AD and dementia subjects and controls

Table 6: SVM train-test splits and demographic characteristics of subjects with incident Alzheimer's Disease (AD) and controls

	With incident AD (n=86)	**Without incident AD (n=1077)**
SVM$_{train}/SVM_{test}$	69/17	862/215
Gender: # male (%)	45 (52.33)	550 (51.07)
Age: mean (s.d)	64.23 (3.81)	64.31 (3.73)
Ethnicity: caucasian %	86.05	97.55

Table 7: SVM train-test splits and demographic characteristics of subjects with incident dementia and controls

	With incident dementia (n=93)	**Without incident dementia (n=1139)**
SVM$_{train}/SVM_{test}$	74/19	911/228
Gender: # male (%)	50 (53.76)	607 (53.29)
Age: mean (s.d)	64.54 (3.87)	64.24 (3.84)
Ethnicity: caucasian %	86.02	97.28

Appendix D. Distribution of disease onset of Alzheimer's Disease and Dementia

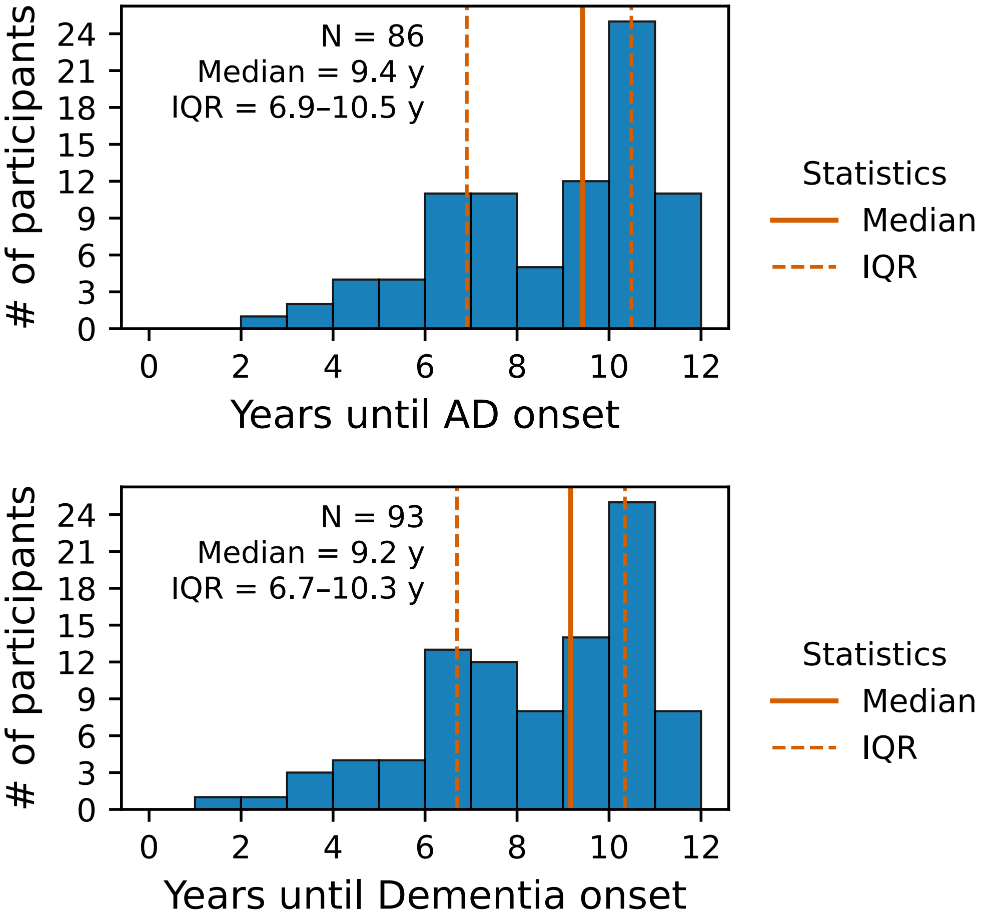

Figure 4: The years until onset of Alzheimer's Disease and dementia. IQR denotes interquartile range.

Appendix E. Full list of AD and dementia risk factors used in this study

- Demographic Information ($d = 5$): Age, sex, economic status, ethnic background, employment status
- General Health Information ($d = 11$): BMI, HbA1C, HDL, systolic/diastolic blood pressure, numeric memory, fluid intelligence, Trail-Making Test A/B duration and error counts
- Risk Factors ($d = 6$): Depression, sleep deprivation, alcohol use, smoking history, cannabis use, age of cannabis initiation
- Physical activity ($d = 6$): Number and Duration of days/week walked 10+ minutes, Number and Duration of days/week of moderate physical activity 10+ minutes, Number and Duration of days/week of vigorous physical activity 10+ minutes

- Social and leisure activities ($d = 2$): Frequency of friend&family visit, number of leisure activity
- Dietary habits ($d = 18$): cooked vegetable intake, raw vegetable intake, fresh fruit intake, dried fruit intake, oily fish intake, non-oily fish intake, processed meat intake, poultry intake, beef intake, lamb intake, pork intake, milk type, spread type, bread intake, salt added to food, tea intake, coffee intake, water intake

Appendix F. Full list of fundus-based retinal morphometry used in this study

- Optic nerve head features($k = 2$): Vertical and horizontal cup-to-disc ratios.
- Vascular features ($k = 15$): Fractal dimension, fractal density, distance tortuosity, squared curvature tortuosity, and tortuosity density for artery, vein, and both combined.

Appendix G. Impact of Thresholds on REVEAL Performance

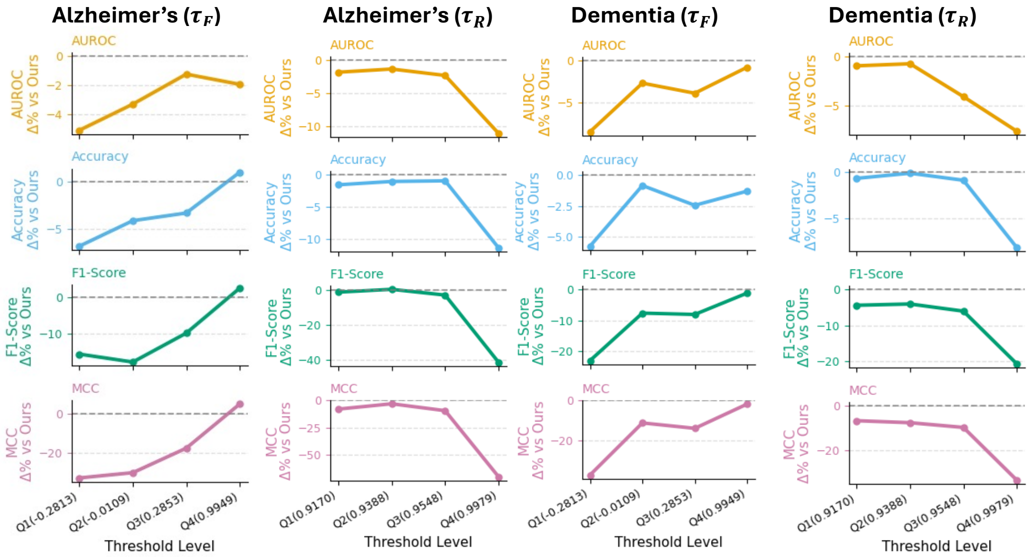

Figure 5: Effect (% difference) of varying thresholds on the incident AD and dementia prediction task.