Archives of
Ophthalmological Research

Original Article

DOI: 10.51271/AOR-0056

Multimodal retinal imaging-based assessment of large language models for the differential diagnosis of retinal vascular occlusion

Memduh Kurt, Ayna Sariyeva Ismayilov*, Mahmut Oğuz Ulusoy

Department of Ophthalmology, University of Health Sciences Bursa Yüksek İhtisas Training and Research Hospital, Bursa, Turkiye

Cite this article: Kurt M, Sariyeva Ismayilov A, Ulusoy MO. Multimodal retinal imaging-based assessment of large language models for the differential diagnosis of retinal vascular occlusion. *Arch Ophthalmol Res.* 2026;3(3):52-56. doi:10.51271/AOR-0056

Received: 29/06/2026 ◆ **Accepted:** 21/09/2026 ◆ **Published:** 24/09/2026

ABSTRACT

Aims: To compare the diagnostic and treatment recommendation performance of three large language models (LLMs)-ChatGPT-4o, Gemini, and Microsoft Copilot-in the differential diagnosis of retinal vascular occlusions using multimodal retinal imaging.

Methods: This observational study included 75 patients, comprising 50 retinal vein occlusion (RVO; 25 branch and 25 central) and 25 retinal artery occlusion (RAO; 6 branch and 19 central) cases. Each LLM independently evaluated three imaging datasets: (1) color fundus photographs, (2) fundus photographs combined with optical coherence tomography (OCT), and (3) fundus photographs combined with OCT and fundus fluorescein angiography (FFA). A standardized prompt was used for all analyses. Diagnostic accuracy and treatment recommendation accuracy were compared among the models using Cochran's Q and post-hoc McNemar tests.

Results: Significant differences in diagnostic accuracy were observed among the three models under all imaging conditions (all $p<0.001$). Using fundus photographs alone, diagnostic accuracy was 21.3% for ChatGPT-4o, 12.0% for Gemini, and 1.3% for Copilot. After adding OCT images, ChatGPT-4o and Gemini each achieved an accuracy of 21.3%, whereas Copilot remained at 1.3%. With multimodal imaging (fundus photographs+OCT+FFA), diagnostic accuracy increased to 28.0% for ChatGPT-4o and 37.3% for Gemini, while Copilot again remained unchanged at 1.3%. Pairwise comparisons demonstrated no significant difference between ChatGPT-4o and Gemini across imaging conditions, whereas both significantly outperformed Copilot. For treatment recommendations, Gemini achieved the highest accuracy (57.3%), followed by ChatGPT-4o (26.7%) and Copilot (10.7%) ($p<0.001$). Gemini performed significantly better than both ChatGPT-4o and Copilot, and ChatGPT-4o also outperformed Copilot.

Conclusion: Although multimodal retinal imaging improved the diagnostic performance of current LLMs, their overall diagnostic accuracy remained inadequate for independent clinical use in retinal vascular occlusions. Gemini demonstrated the greatest benefit from multimodal image integration and provided the most accurate treatment recommendations. These findings suggest that current LLMs should be regarded as supportive clinical tools rather than autonomous diagnostic systems, while future multimodal Artificial Intelligence models trained on ophthalmology-specific datasets may achieve substantially improved performance.

Keywords: Retinal vascular occlusion, retinal vein occlusion, retinal artery occlusion, large language models, multimodal imaging, Artificial Intelligence

INTRODUCTION

In recent years, Artificial Intelligence (AI) technologies have become an integral component of daily life and an essential tool for addressing complex problems. This technological transformation has created new opportunities in medicine and health sciences. In the field of ophthalmology, in particular, the ability of AI algorithms to analyze data derived from imaging modalities and detect pathological findings offers considerable potential for the development of clinical decision-support systems.

According to the International Diabetes Federation (IDF), the global prevalence of diabetes mellitus is expected to reach 700 million by 2045.[1] Therefore, image-based autonomous diagnostic systems in the field of ophthalmology have predominantly focused on diabetic retinopathy. A Google DeepMind deep learning algorithm trained on diabetic retinopathy datasets demonstrated high sensitivity, specificity, and accuracy in detecting diabetic retinopathy using color fundus photographs.[2] Considering the anticipated decline in the number of physicians per patient, the use of autonomous

***Corresponding Author:** Ayna Sariyeva Ismayilov, sariyevaayna@hotmail.com

systems is expected to become increasingly widespread. Moreover, several studies have suggested that AI-assisted screening may be more cost-effective than conventional screening performed by ophthalmologists.[3]

In addition, algorithms capable of classifying age-related macular degeneration (AMD) according to the AREDS severity scale are currently available.[4] These studies indicate that AI can achieve success rates approaching 99% through image-based deep learning. However, without requiring task-specific deep learning training, readily accessible large language models (LLMs), such as Chat Generative Pre-trained Transformer-4 (ChatGPT-4o; OpenAI), Google Gemini (Alphabet Inc.), and Microsoft Copilot (Microsoft Corporation), are increasingly used by both patients and physicians in diagnostic and therapeutic processes in healthcare. Today, these systems play a critical role not only in the detection of various fundus pathologies but also in the screening and early diagnosis of systemic diseases through retinal assessment.[5] Researchers in Singapore[6,7] and South Korea[8,9] have developed deep learning systems (DLSs) implemented in public healthcare settings to detect referable eye diseases, including glaucoma, AMD, and diabetic retinopathy, with high accuracy.

Although the increasing use of LLMs in healthcare presents both advantages and limitations for physicians and patients, these tools may offer substantial potential when applied with appropriate clinical awareness and consideration of potential biases. While AI has demonstrated competence in information delivery, its reliability and diagnostic accuracy in medical practice remain to be fully established through comprehensive academic research. Therefore, this study aims to assess the performance and clinical reliability of current LLM platforms in the interpretation of imaging findings associated with retinal vascular diseases.

METHODS

Ethics

The study protocol was formally approved by the University of Health Sciences Bursa Yüksek İhtisas Training and Research Hospital Medical Sciences Ethics Committee (Date: 22.04.2026, Decision No: 2024-TBEK 2026/04-10) in accordance with the Declaration of Helsinki.

Study Design

A total of 75 patients were included in the study, comprising 50 patients diagnosed with retinal vein occlusion (RVO; 25 central and 25 branch retinal vein occlusions) and 25 patients diagnosed with retinal artery occlusion (RAO; 19 central and 6 branch retinal artery occlusions).

The mean duration of symptoms before hospital presentation was 4.02±2.64 days (range, 0-8 days) in the RAO group and 15.04±3.14 days (range, 1-21 days) in the RVO group. Fundus photography, OCT, and FFA were performed on the day of presentation, and these examinations represented the patients' initial imaging assessments. Patients with chronic or diagnostically ambiguous retinal vascular occlusions were excluded. All patients were evaluated and managed by experienced retina specialists independently of the LLM-generated responses. The LLMs were subsequently provided with the same initial imaging data used by the retina specialists for diagnosis.

The diagnostic performances of large language models (LLMs) were evaluated using three different datasets: (1) fundus photographs obtained at a 50-degree angle, including the optic disc, macula, and major vascular arcades; (2) a combination of fundus photographs and optical coherence tomography (OCT) images; and (3) a triple-modality combination consisting of fundus photographs, OCT images, and fundus fluorescein angiography (FFA) images. The analyses were performed independently by each model using ChatGPT-4o, Gemini, and Copilot software.

The fundus photographs (Topcon TRC-50DX, Japan), OCT images (RTVue XR Avanti, Optovue, Inc., Fremont, CA, USA), and FFA images (Topcon TRC-50DX, Japan) used in the study were retrospectively obtained from the archive of the Depatment of Retina of Bursa Yüksek İhtisas Training and Research Hospital. A multimodal approach was adopted to ensure that the responses generated by the AI models were based directly on visual data. The data input process was carried out between December 27, 2025, and January 4, 2026. Each model was queried using a standardized prompt: "I will present you with an image; can you diagnose the disease?"

The models' abilities to interpret pathological findings in the images and their diagnostic accuracy rates were recorded. To minimize potential memory bias and prevent the models from learning from previous queries, each session was terminated after every analysis, and a new chat was initiated. Fundus photographs, OCT images, and FFA combinations were queried separately in different chat windows, and the accuracy of the findings and diagnoses was recorded prospectively. As examinations from 75 patients were presented to each AI application in three different combinations, a total of 675 queries were analyzed within the scope of the study.

Statistical Analysis

All data analyses were performed using IBM SPSS Statistics version 23.0. The normality of continuous variables was assessed using the Kolmogorov-Smirnov test. Continuous variables that did not show a normal distribution were presented as median minimum-maximum values, while categorical variables were expressed as numbers and percentages.

The responses of the AI models regarding diagnosis and treatment recommendations were coded as "correct" or "incorrect." Since the same 75 cases were evaluated by all models, Cochran's Q test was used to compare the accuracy rates of the three models (Figure 1). When a significant difference was detected in the overall comparison, pairwise post-hoc comparisons were performed using McNemar's test.

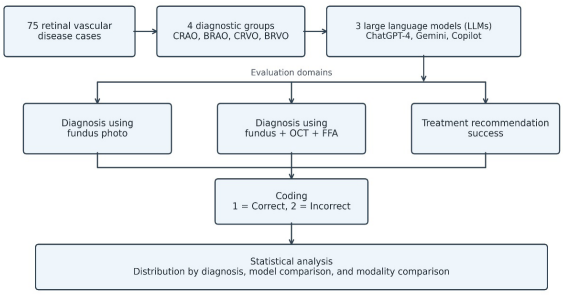

Figure 1. Study flow diagram

CRAO: Central retinal artery occlusion, BRAO: Branch retinal artery occlusion, CRVO: Central retinal vein occlusion, BRVO: Branch retinal vein occlusion, ChatGPT: Chat Generative Pre-trained Transformer, OCT: Optical coherence tomography, FFA: Fundus fluorescein angiography

RESULTS

A total of 50 patients with retinal vein occlusion, including 25 with branch retinal vein occlusion (BRVO) and 25 with central retinal vein occlusion (CRVO), and 25 patients with retinal artery occlusion, including 6 with branch retinal artery occlusion (BRAO) and 19 with central retinal artery occlusion (CRAO), were included in the evaluation. As the data of 75 patients were presented to each AI application in three different combinations, a total of 675 queries/prompts were performed.

In the RAO group, there were 7 female (28%) and 18 male (72%), whereas the RVO group included 23 female (46%) and 27 male (54%) (p=0.105). The mean age was 62.52±11.95 years in the RAO group (range: 40-86) and 64.62±11.52 years in the RVO group (range: 43-75). No statistically significant difference was observed between the groups in terms of age (p=0.465) (Table 1).

Table 1. Distribution of age and gender according to diagnostic groups

Diagnostic group	n	Age, median (min-max)	Female n (%)	Male n (%)
CRAO	19	64 (19-76)	4 (21.1)	15 (78.9)
BRAO	6	63.5 (55-76)	3 (50.0)	3 (50.0)
CRVO	25	69 (51-92)	11 (44.0)	14 (56.0)
BRVO	25	57 (40-85)	12 (48.0)	13 (52.0)
Total	75	65 (19-92)	30 (40.0)	45 (60.0)

Age data are presented as median (minimum-maximum), and gender data are presented as n (%). Min: Minimum, Max: Maximum, CRAO: Central retinal artery occlusion, BRAO: Branch retinal artery occlusion, CRVO: Central retinal vein occlusion, BRVO: Branch retinal vein occlusion

When fundus photographs alone were evaluated, a significant difference was observed among the three models (Cochran's Q, p<0.001). ChatGPT-4o demonstrated the highest diagnostic accuracy, followed by Gemini, whereas Copilot showed the lowest performance. Pairwise comparisons revealed no significant difference between ChatGPT-4o and Gemini, while both models significantly outperformed Copilot (Table 2).

The addition of OCT did not improve the diagnostic accuracy of ChatGPT-4o or Gemini compared with fundus photography alone. Nevertheless, a significant difference remained among the three models, with ChatGPT-4o and Gemini performing significantly better than Copilot. There was no significant difference between ChatGPT-4o and Gemini (Table 2).

With the addition of FFA to fundus photography and OCT, diagnostic performance increased for both ChatGPT-4o and Gemini, with Gemini achieving the highest overall diagnostic accuracy. However, the difference between ChatGPT-4o and Gemini remained statistically nonsignificant. Both models continued to significantly outperform Copilot, whose diagnostic accuracy remained unchanged across all imaging conditions (Table 2).

For treatment recommendations, a significant difference was observed among the three models (p<0.001). Gemini demonstrated the highest accuracy and significantly outperformed both ChatGPT-4o and Copilot. ChatGPT-4o also performed significantly better than Copilot (Table 2; Figure 2).

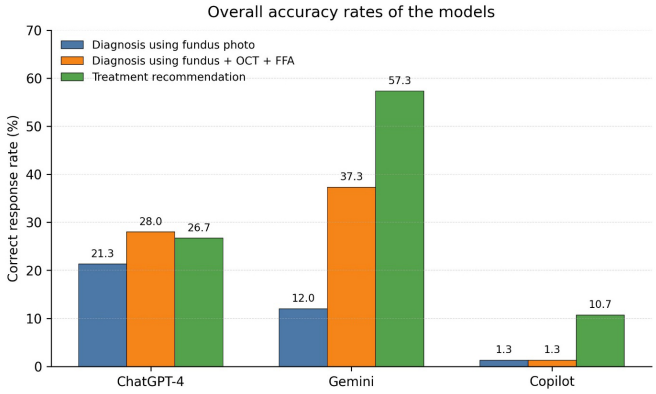

Figure 2. Overall accuracy rates of the models

OCT: Optical coherence tomography, FFA: Fundus fluorescein angiography, ChatGPT: Chat Generative Pre-trained Transformer

Detailed diagnostic and treatment recommendation accuracy rates and pairwise comparisons are presented in Table 2.

DISCUSSION

CRAO is estimated to occur at an incidence of 1-2 cases per 100,000 individuals per year.[10] The importance of early diagnosis is not limited to the intended visual prognosis; it is also critical in terms of the patient's morbidity and mortality risk. The interval between symptom onset and presentation to

Table 2. Overall model performance according to assessment domains

Assessment domain	Model	Correct n (%)	Overall p (Cochran's Q)	ChatGPT-4o vs Gemini	ChatGPT-4o vs Copilot	Gemini vs Copilot
Diagnosis using fundus photographs	ChatGPT-4o	16 (21.3)	<0.001	0.143	<0.001	0.021
	Gemini	9 (12.0)				
	Copilot	1 (1.3)				
Diagnosis using fundus photographs+OCT	ChatGPT-4o	16 (21.3)	<0.001	1.000	<0.001	<0.001
	Gemini	16 (21.3)				
	Copilot	1 (1.3)				
Diagnosis using fundus photographs+OCT+FFA	ChatGPT-4o	21 (28.0)	<0.001	0.248	<0.001	<0.001
	Gemini	28 (37.3)				
	Copilot	1 (1.3)				
Treatment recommendation	ChatGPT-4o	20 (26.7)	<0.001	<0.001	0.013	<0.001
	Gemini	43 (57.3)				
	Copilot	8 (10.7)				

Data are presented as the number and percentage of correct responses. Overall comparisons were performed using Cochran's Q test. Pairwise pos- hoc comparisons were performed using McNemar's test. ChatGPT: Chat Generative Pre-trained Transformer, OCT: Optical coherence tomography, FFA: Fundus fluorescein angiography

an ophthalmologist is of crucial importance. Unfortunately, in some clinical scenarios, this interval may extend to several days. With the increasing use of AI-based fundus image screening tools, this delay is expected to decrease.

RVO is the second most common retinal vascular disease leading to visual impairment and blindness after diabetic retinopathy. In 2015, the global prevalence of RVO, BRVO, and CRVO among individuals aged 30-89 years was reported as 0.77%, 0.64%, and 0.13%, respectively, corresponding to approximately 28.06 million, 23.38 million, and 4.67 million affected individuals worldwide.[11] Early diagnosis and timely treatment contribute positively to final visual acuity outcomes.

Previous studies have also demonstrated that RVO is associated with cardiovascular risk factors such as arterial hypertension, family history of stroke, and atrial fibrillation.[12,13] Therefore, early diagnosis of retinal vascular diseases may provide important benefits in terms of reducing morbidity and mortality. Through early detection, underlying systemic conditions such as hypertension, diabetes mellitus, or occult thrombophilia/hypercoagulability in younger patients may be identified at an earlier stage, thereby potentially reducing the risk of systemic thromboembolic events.

Limitations

This study comparatively demonstrated the clinical capabilities of current LLMs in the diagnosis and treatment planning of retinal vascular diseases. Our findings reflect the sensitivity of AI models to multimodal data input, as well as their current limitations within clinical decision-support mechanisms.

An important distinction should be made between our findings and those reported by Gulshan et al.,[2] who developed and validated a deep convolutional neural network specifically designed for the detection of diabetic retinopathy and diabetic macular edema from retinal fundus photographs. Their algorithm was trained on a large, task-specific dataset of 128,175 retinal images and achieved very high diagnostic performance, with area under the receiver operating characteristic curve values of approximately 0.99 in two independent validation datasets. In contrast, the models evaluated in our study were general-purpose LLMs that were not specifically developed or fine-tuned for the diagnosis of retinal vascular occlusions. Therefore, the substantially lower diagnostic performance observed in our study should not be interpreted as evidence that AI is inherently inadequate for ophthalmic image interpretation. Rather, the difference likely reflects the fundamentally different design and training objectives of these systems. Task-specific deep-learning algorithms are optimized for a predefined ophthalmic classification task using large, carefully annotated image datasets, whereas general-purpose LLMs are designed to perform a broad range of language and multimodal tasks and may lack disease-specific optimization. Furthermore, Gulshan et al.[2] evaluated a well-defined diabetic retinopathy detection task using standardized fundus photographs, whereas our study addressed the more challenging differential diagnosis of retinal vascular occlusions, including CRAO, CRVO, and BRVO, and further examined whether incorporating complementary OCT and FFA information could improve performance. Thus, our findings highlight both the potential and the current limitations of general-purpose LLMs in ophthalmic image interpretation and suggest that their performance should be interpreted differently from that of task-specific deep-learning systems.

Overall, all three models demonstrated limited diagnostic performance when interpreting retinal vascular occlusions from fundus photographs alone, with ChatGPT-4o showing relatively better performance than Gemini and Copilot. The addition of OCT did not substantially improve ChatGPT's overall diagnostic performance, whereas Gemini showed a more favorable ability to recognize CRAO and cherry-red spot findings when OCT information was incorporated. The most notable improvement was observed with the combined presentation of fundus photography, OCT, and FFA, particularly for Gemini, suggesting that access to complementary multimodal imaging information may enhance the interpretation of retinal vascular abnormalities by some general-purpose LLMs. In contrast, Copilot showed consistently poor performance across all imaging configurations, indicating that the addition of multimodal information alone was insufficient to improve its image-based diagnostic capability. Treatment recommendations followed a similar pattern, with Gemini demonstrating greater consistency with current clinical management principles than ChatGPT-4o and Copilot. Interestingly, Copilot occasionally provided appropriate treatment recommendations despite poor diagnostic performance, largely because it could recognize findings such as cystoid macular edema on OCT.

From a clinical perspective, our findings suggest that general-purpose LLMs currently have limited utility as standalone diagnostic tools for retinal vascular occlusions. Although the integration of multimodal imaging, particularly the combination of fundus photography, OCT, and FFA, improved diagnostic performance in some models, the maximum accuracy achieved remained insufficient to support independent clinical decision-making. Nevertheless, the observed improvement with multimodal information suggests that LLMs may have a potential role as supportive tools by integrating complementary imaging findings and assisting clinicians in generating differential diagnoses or management considerations. Importantly, such applications should be considered adjunctive rather than substitutive, and all diagnostic and treatment decisions should remain under the supervision of an ophthalmologist. To our knowledge, this is the first study to evaluate general-purpose LLMs in the multimodal interpretation of retinal vascular occlusions, including CRAO, CRVO, and BRVO, while also assessing their treatment recommendations. By evaluating different combinations of fundus photography, OCT, and FFA, our study provides preliminary evidence regarding both the potential and current limitations of general-purpose LLMs in this specific ophthalmic context.

CONCLUSION

To the best of our knowledge, this is the first study in the literature to compare the performance of current LLMs in the diagnosis and treatment recommendations of retinal vascular diseases using multimodal data. Our findings demonstrated that, despite the limitations of AI in single-image analysis, the integration of clinical imaging data, including fundus photography, OCT, and FFA, through a multimodal approach can significantly improve diagnostic performance. Despite their current limitations, these models hold potential as "clinical assistants," and further training with more specific

datasets may pave the way for a new era in the management of retinal diseases.

ETHICAL DECLARATIONS

Ethics Committee Approval

This study was approved by the University of Health Sciences Bursa Yüksek İhtisas Training and Research Hospital Medical Sciences Ethics Committee (Date: 22.04.2026, Decision No: 2024-TBEK 2026/04-10).

Informed Consent

Written informed consent was obtained from all individual participants prior to their inclusion in the study. Participants were fully informed about the study's aims, procedures, potential risks and benefits, and their rights-including the right to withdraw at any time without consequence. All participants voluntarily signed a written informed consent form.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: MK, ASI, MOU; Design: MK, ASI, MOU; Control: MK, ASI, MOU; Resources: MK, ASI, MOU; Materials: MK, ASI, MOU; Data Collection and/or Processing: MK, ASI, MOU; Analysis and/or Interpretation: MK, ASI, MOU; Literature Review: MK, ASI, MOU; Writing the Article: MK, ASI, MOU; Critical Review: MK, ASI, MOU.

REFERENCES

1. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res Clin Pract.* 2019;157:107843. doi:10.1016/j.diabres.2019.107843
2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216
3. Huang XM, Yang BF, Zheng WL, et al. Cost-effectiveness of Artificial Intelligence screening for diabetic retinopathy in rural China. *BMC Health Serv Res.* 2022;22(1):260. doi:10.1186/s12913-022-07655-6
4. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology.* 2019;126(4):565-575. doi:10.1016/j.ophtha.2018.11.015
5. Li Z, Wang L, Wu X, et al. Artificial Intelligence in ophthalmology: the path to the real-world clinic. *Cell Rep Med.* 2023;4(7):101095. doi:10.1016/j.xcrm.2023.101095
6. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152
7. Wong DCS, Kiew G, Jeon S, Ting D. Singapore eye lesions analyzer (SELENA): the deep learning system for retinal diseases. In: Grzybowski A, editor. Artificial Intelligence in Ophthalmology. Cham: Springer; 2021:177-185. doi:10.1007/978-3-030-78601-4_13
8. Xie Y, Nguyen QD, Hamzah H, et al. Artificial Intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health.* 2020;2(5):240-249. doi:10.1016/S2589-7500(20)30060-1
9. Dismuke C. Progress in examining cost-effectiveness of AI in diabetic retinopathy screening. *Lancet Digit Health.* 2020;2(5):e212-e213. doi:10.1016/S2589-7500(20)30077-7
10. Leavitt JA, Larson TA, Hodge DO, Gullerud RE. The incidence of central retinal artery occlusion in Olmsted County, Minnesota. *Am J Ophthalmol.* 2011;152(5):820-823. doi:10.1016/j.ajo.2011.05.005
11. Song P, Xu Y, Zha M, Zhang Y, Rudan I. Global epidemiology of retinal vein occlusion: a systematic review and meta-analysis of prevalence, incidence, and risk factors. *J Glob Health.* 2019;9(1):010427. doi:10.7189/jogh.09.010427
12. Ponto KA, Elbaz H, Peto T, et al. Prevalence and risk factors of retinal vein occlusion: the Gutenberg health study. *J Thromb Haemost.* 2015;13(7):1254-1263. doi:10.1111/jth.12982
13. Frederiksen KH, Stokholm L, Frederiksen PH, et al. Cardiovascular morbidity and all-cause mortality in patients with retinal vein occlusion: a Danish nationwide cohort study. *Br J Ophthalmol.* 2023;107(9):1324-1330. doi:10.1136/bjophthalmol-2022-321225