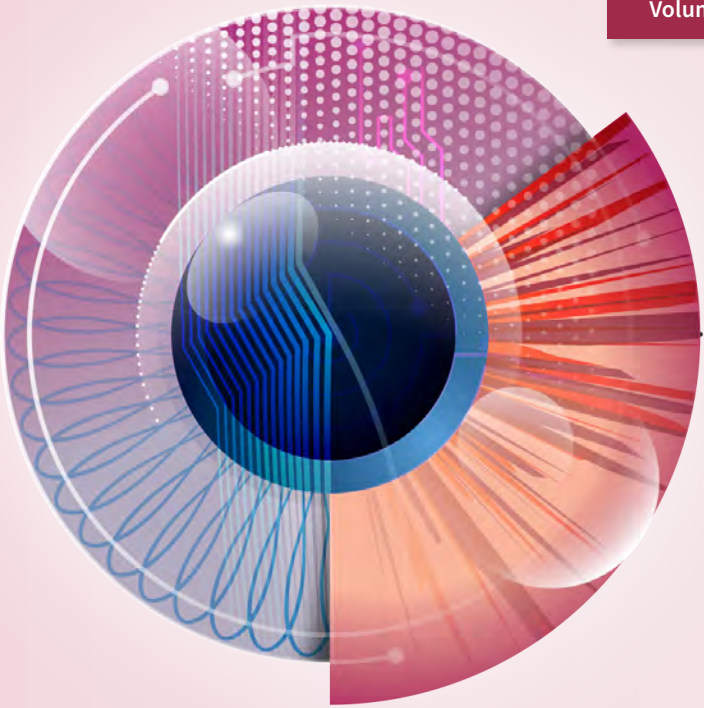
aivojournal.com
Volume 1 | Supplement



## Abstracts from SAIVO 2025 Inaugural Meeting
*Where AI Meets Vision*

# Artificial Intelligence in Vision & Ophthalmology

*Official Journal of the Society for Artificial Intelligence in Vision and Ophthalmology (SAIVO)*

ISSN
Print: 3051-2328 | Online: 3117-4035

# SAIVO

Society for AI in Vision and Ophthalmology

Established 2025

SAIVO - the Society for Artificial Intelligence in Vision and Ophthalmology - is the world's first professional society formally dedicated to AI and big data in eye care. SAIVO stands at the intersection of medicine, technology, and scientific innovation, with a shared commitment to improving patient outcomes through responsible, patient-centered AI.

WWW.SAIVO.ORG

# AIVO

Artificial Intelligence in Vision and Ophthalmology

While the rapid advance of imaging technologies in ophthalmology is making available a continually increasing number of data, the interpretation of such data is still very challenging and this hinders the advance in the understanding of ocular diseases and their treatment. Interdisciplinary approaches encompassing ophthalmology, physiology, mathematics, engineering, and computer science have shown great capabilities in data analysis and interpretation for advancing basic and applied clinical sciences. Artificial Intelligence in Vision and Ophthalmology (AIVO) was created with the aim of providing a forum for interdisciplinary approaches integrating mathematical and computational methods with experimental and clinical studies to address open problems in ophthalmology. AIVO welcomes articles that investigate questions related to the anatomy, physiology and function of the eye in health and disease.

For further information on AIVO's focus and scope as well as manuscript submissions:

**www.aivojournal.com**
**aivo@aivojournal.com**

## Copyright





**Publisher**
Kugler Publications
P.O. Box 20538
1001 NM Amsterdam
The Netherlands
info@kuglerpublications.com
www.kuglerpublications.com

**ISSN**
Online: 3117-4035
Print: 3051-2328

**Manuscript submissions**
Author guidelines and templates are available via the website, through which all manuscripts should be submitted. For inquiries please contact us via e-mail.

**Publication frequency**
AIVO uses the Continuous Article Publication (CAP) model. Articles are published online as soon as they are ready.

**Advertising inquiries**
AVIO offers online and in print sponsorship and advertising opportunities. Please contact Kugler Publications to for inquiries.

**Submit your article now:**

## Open access policy

AIVO is fully open access without requiring any publication fee from the authors. Publication fees will be introduced in 2026.

## SAIVO – Society for Artificial Intelligence in Vision and Ophthalmology

AIVO is the official journal of SAIVO. SAIVO is the first society in the world formally dedicated to artificial intelligence and big data in the fields of vision and ophthalmology, brings together experts in medicine, science, and technology to advance the safe and effective use of artificial intelligence in eye care. SAIVO supports innovation, education, and collaboration to improve diagnosis, treatment, and outcomes for patients around the world. SAIVO is committed to advancing the development, validation, and clinical integration of artificial intelligence technologies in eye care. More information on how to join: https://www.saivo.org/

SAIVO
Society for AI in Vision and Ophthalmology
Established 2025

## Disclaimers

# Table of Contents

# 4. Assessment of correctness, content omission and risk of harm in large language model responses to ophthalmology CME questions

Jacqueline Chen[1], Amanda Lu[2], Rohan Verma[3], Li Wang[4], Douglas Koch[4], Allison Chen[4]

*[1]Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, United States; [2]University of California Los Angeles, Jules Stein Eye Institute, Los Angeles, United States; [3]Mann Eye Institute, Houston, United States; [4]Baylor College of Medicine, Department of Ophthalmology, Houston, United States*

**Purpose/Background**

Large language models (LLMs) have demonstrated strong performance on multiple-choice (MC) ophthalmology examinations. However, the accuracy and quality of their free-text, or prose, responses remain largely unexplored. This study aims to evaluate the accuracy, completeness, and clinical reliability of prose outputs generated by two LLMs in response to ophthalmology continuing medical education (CME) questions.

**Methods**

A set of Basic and Clinical Science Course (BCSC) questions, along with corresponding MC answer options sourced from the American Academy of Ophthalmology (AAO) question bank, were used as prompts for two LLMs: OpenAI's GPT-4 and Google Vertex's Gemini Pro 1.5. Both the MC selections and the generated prose explanations were independently assessed by three board-certified ophthalmologists. Evaluations were based on a previously validated rubric, which scored responses for accuracy, evidence of correctness, completeness, presence of bias, and potential for clinical harm.

**Results**

ChatGPT-4 achieved a significantly higher multiple-choice accuracy rate compared to Gemini Pro 1.5 (82.5% [99/120] vs. 49.2% [59/120]; $p < 0.05$). Despite demonstrating a high frequency of correct reasoning in prose explanations (92% for ChatGPT-4 and 88% for Gemini Pro 1.5), both models also exhibited substantial limitations. Instances of incorrect reasoning were observed in 42% and 58% of responses, respectively; inappropriate content appeared in 29% and 36%; missing critical information was noted in 42% and 30%; and concerningly, 36% (ChatGPT-4) and 44% (Gemini Pro 1.5) of outputs included content with potential for physical or emotional harm.

**Conclusions**

While ChatGPT-4 demonstrated strong performance in multiple-choice accuracy, both LLMs exhibited notable deficiencies in their prose responses, including factual inaccuracies, omissions of critical content, and instances of potentially harmful

information. These findings underscore the necessity of expert oversight and rigorous validation before integrating LLM-generated content into patient-facing ophthalmologic applications.

## 5. Comparison of three generative artificial intelligence (AI) large language models on pediatric ophthalmology and strabismus board-style questions

Albert Yang[1], Melinda Chang[2]

*[1]USC Keck School of Medicine, Los Angeles, United States; [2]CHLA, Los Angeles, United States*

**Purpose/Background**

Large language models (LLMs) are increasingly used in medicine, yet the accuracy of medical information provided by LLMs, especially in subspecialties such as pediatric ophthalmology, remains unclear. The purpose of this study was to evaluate and compare the accuracy and reasoning quality of three LLMs—ChatGPT-4.5, OpenEvidence, and Claude Sonnet 4—in answering pediatric ophthalmology and strabismus board-style questions from the American Academy of Ophthalmology Basic and Clinical Science Course (BCSC) Self-Assessment Program (SAP).

**Methods**

Three LLMs were queried with 100 randomly selected multiple-choice questions from the BCSC SAP pediatric ophthalmology and strabismus bank. Images were excluded to maintain parity across models. Questions were categorized by phase of care (*e.g.*, Diagnosis, Workup, Management) and subspecialty topic (*e.g.*, strabismus and amblyopia, pediatric glaucoma, pediatric neuro-ophthalmology). Accuracy was measured as the percentage of correct answers. Reasoning quality was assessed by a board-certified pediatric ophthalmologist reviewer who determined whether the rationale behind each correct answer was accurate. Statistical comparisons between models and human performance were conducted using Fisher's exact test.

**Results**

ChatGPT-4.5 achieved the highest accuracy (88%), followed by OpenEvidence (83%) and Claude (78%). Only ChatGPT-4.5 performed significantly better than human respondents, who had an average of 70% accuracy ($p = 0.003$). Performance varied by topic. For example, ChatGPT-4.5 achieved 100% accuracy on questions related to pediatric retina, uveitis, and oculoplastics, while human performance was weakest in pediatric glaucoma and embryology (67% accuracy). Among correctly answered questions, OpenEvidence had the highest rate of accurate reasoning (95%), significantly outperforming Claude (79%, $p = 0.0034$) and marginally outperforming ChatGPT-4.5 (86%, $p = 0.0652$). Errors in reasoning, including hallucinated explanations, were more common in general-purpose models ChatGPT-4.5 and Claude

compared to OpenEvidence.

**Conclusions**
LLMs are capable of surpassing average human performance on pediatric ophthalmology board-style questions. ChatGPT-4.5 demonstrated the highest accuracy, whereas OpenEvidence offered the most consistently accurate reasoning. Given the potential for hallucinations, especially in general-purpose models, LLMs should be used cautiously as supplemental educational tools. Future work should expand model evaluation across other clinical domains and incorporate image-based assessment.

## 6. Deep Learning-Based Super-Resolution of Whole Eye MRI Provides 8x Higher Resolution With 90-Second Scan Times

Quan V. Hoang[1,2,3,4], Huanye Li[1,2], K. Bailey Freund[5], Lawrence A. Yannuzzi[5], Stanley Chang[4], Jack Grinband[6]

*[1]Dept. Ophthalmology, National University of Singapore, Singapore, Singapore; [2]Singapore Eye Research Institute, Singapore, Singapore; [3]Institute of Molecular and Cell Biology, ASTAR, Singapore, Singapore; [4]Dept of Ophthalmology, Columbia University, New York, United States; [5]Vitreous Retina Macula Consultants of New York (VRMNY), New York, United States; [6]Dept of Radiology, Columbia University, New York, United States*

**Background**
Conventional MRI approaches for ocular imaging are highly susceptible to motion artifacts due to involuntary eye movements. To mitigate this, rapid acquisition protocols are often used, but suffer from reduced spatial resolution.

**Purpose**
Enhance spatial resolution of T2-weighted MRIs acquired with short acquisition times using a deep learning-based super-resolution (SR) pipeline.

**Methods**
To develop a convolutional neural network (CNN) for super-resolution, we trained on T2-weighted brain MRIs from the Human Connectome Project (HCP, n=302 adults), and applied the CNN to an independent cohort of T2-weighted eye scans (n=25; ages=22-84). Eye MRIs were acquired using a Philips 3T MRI with an 8-channel phased-array head coil and a fat-suppressed, axial T2-weighted volumetric sequence with four 90-second scans (two coronal and two axial slice orientations, 45-degree rotation between each along the y-axis) at resolutions of 0.5x0.5mm in-plane and 2.0mm slice thickness. To generate the processing pipeline, the HCP brains were rotated to simulate the eye orientations and downsampled by a factor ranging from 2 to 9. Wiener deconvolution was applied to estimate the point-spread

function of blurring introduced during downsampling and registration. Multiple feed-forward deep CNN models, each corresponding to a different upsampling factor, were trained on brain MRIs and applied to the low-resolution eye images. Each model took four rotated low-resolution 3D volumes as input, mimicking the multi-directional eye MRI protocol and generated a high-resolution output. Trained to integrate spatial information across the four directions, the models generalized well to eye data despite being trained on brain images. To assess image quality of the SR image, we compared the spatial gradient across upsampling factors.

**Results**
The SR eye images demonstrated a significantly higher spatial gradient with upsampling factors from 2 to 7, plateauing at a factor of 8, corresponding to an effective resolution increase from 0.5x0.5x2.0mm/voxel to 0.5x0.5x0.25mm/voxel. No further improvement was observed for factors 8 to 9.

**Conclusion**
Our CNN-based image SR pipeline achieved up to 8 times higher voxel resolution for whole-eye MRI, enabling rapid acquisition while minimizing fixation loss. The resulting image quality supported accurate whole-eye length measurements, comparable to those from standard clinical instruments such as IOLMaster.

## 7. Artificial Intelligence -Based Analysis of Choroidal Biomarkers to Predict Treatment Response in Neovascular Age-Related Macular Degeneration

Mohamed Morsy, Nehal Nailesh Mehta, Amr L. Ali, Dirk-Uwe Bartsch, Lingyun Cheng, William Freeman

*Shiley Eye Institute , University of California , San Diego, San Diego, United States*

**Purpose**
To evaluate subfoveal choroidal thickness (CT) and choroidal volume (CV) as biomarkers of anti-VEGF treatment response in neovascular age-related macular degeneration (nAMD) using an AI-based OCT analysis platform, and to assess their behavior across four clinical response groups.

**Methods**
A total of 120 pairs ofOCT B-scans from eyes with nAMD were analyzed following intravitreal anti-VEGF treatment. Eyes were categorized into four response groups: improved, worsened, stable on treatment and stable off treatment. OCT B-scans were processed using the RetinAI software platform, which performed automated segmentation of retinal layers and delineation of the chorio-scleral junction. Subfoveal CT was manually measured using a digital caliper guided by AI-generated segmentation, while central 1 mm CV was computed directly by the software.

Additional OCT biomarkers,including central 1 mm retinal thickness, pigment epithelial detachment (PED) volumeand total fluid volume, were measured and correlated with CT and CV changes.

**Results**

Mean subfoveal CT decreased by 15.3% in the improved group ($p< 0.001$) and increased by 7.2% in the worsened group ($p< 0.001$), with negligible change in the stable-on and stable-off groups ($p= 0.597$ and$p= 0.546$). CV decreased by 11% in the improved group ($p< 0.001$) and increased by 3% in the worsened group ($p< 0.001$), with minimal change in the stable groups ($p= 0.328$ and$p= 0.293$). At baseline, the worsened group exhibited significantly higher CT and CV than the improved group ($p= 0.0028$ and$p= 0.0027$). Changes in retinal thickness, PED volume, and fluid volume paralleled CT and CV trends. The AI-based platform enabled consistent and accurate segmentation across all scans.

**Conclusion**

Subfoveal CT and CV correlate with anti-VEGF treatment response in nAMD, with higher baseline values associated with poor anatomical outcomes. This is the first study to evaluate both metrics across four distinct response groups using automated AI-based OCT segmentation, supporting their potential role as predictive biomarkers in personalized treatment planning.

## 8. Development and Validation of a Web-Based Platform for AI Based Periorbital Measurement in Low-Resource Settings

George R. Nahass[1], Sasha Hubschman[1], Benjamin Beltran[1], Bhavana Kolli[1], Caitlin Berek[1], James D. Edmonds[1], R.V. Paul Chan[1], Chad A. Purnell[1], James W. Larrick[1], Jacob van der Ende[2], Pete Setabutr[1], Darvin Yi[1], Ann Q. Tran[1]

*[1]University of Illinois, Chicago, United States; [2]Quina Care, Putumayo, Ecuador*

**Background**

Periorbital measurements such as margin reflex distances (MRD1/2), palpebral fissure height, and scleral show are critical in diagnosing and managing conditions like ptosis and disorders of the eyelid. We developed and evaluated *Glorbit*, a lightweight, browser-based application for automated periorbital distance measurement using artificial intelligence (AI), designed for deployment in low-resource clinical environments. The goal was to assess its usability, cross-platform functionality, and readiness for real-world field deployment.

**Methods**

The application integrates a DeepLabV3 segmentation model into a modular image processing pipeline with secure, site-specific Google Cloud storage. Glorbit

supports offline mode, local preprocessing, and cloud upload through Firebase-authenticated logins. The full workflow—metadata entry, facial image capture, segmentation, and upload—was tested. Post-session, participantscompleted a Likert-style usability survey.

**Results**
Glorbit successfully ran on all tested platforms, including laptops, tablets, and mobile phones across major browsers. A total of 15 volunteers were enrolled in this study where the app completed the full workflow without error on 100% of patients. The segmentation model succeeded on all images, and average session duration was 101.7 ± 17.5 seconds. Usability scores on a 5-point Likert scalewere uniformly high: intuitiveness and efficiency (5.0 ± 0.0), workflow clarity (4.8 ± 0.4), output confidence (4.9 ± 0.3), and clinical usability (4.9 ± 0.3).

**Conclusions**
Glorbit is a functional, cross-platform solution for standardized periorbital measurement in clinical and low-resource settings. By combining a local image processing with secure, modular data storage and offline compatibility, the tool enables scalable deployment and ethically governed datacollection. These features support broader efforts in AI-driven oculoplastics including future development of real-time triage tools and multimodal datasets for personalized ophthalmic care.

## 9. Evaluating ChatGPT-4's Role in Diagnosing and Grading Diabetic Retinopathy from Fundus Images

Ishan Bhanot[1], Nitin Rangu[1], David Seo[1], Vinay Shah[2]

*[1]University of Oklahoma College of Medicine, Oklahoma City, United States; [2]Retina Consultants of Oklahoma, Oklahoma City, United States*

**Purpose**
ChatGPT has emerged as one of the most popular and accessible AI models. Initially developed as a natural language model (NLM), ChatGPT-4 evolved to have multimodal capabilities, allowing for image inputs. Its potential as a clinical tool has been explored, particularly in diabetic retinopathy (DR), but little research exists on its ability to analyze fundus images and grade DR. This study is the first large-scale assessment of ChatGPT-4's performance in grading DR and macular edema (ME). Accurate grading could aid physicians, especially in resource-limited settings.

**Methods**
This retrospective, non-randomized study used the Indian Diabetic Retinopathy Image Dataset (IDRiD), containing 516 publicly available fundus images, with permission from its authors. Existing research indicates optimal Results occur when ChatGPT generates its own prompts from task descriptions. Therefore, we asked ChatGPT to create the best prompt for our goals. Each image was then assessed in

a separate chat session with memory off. ChatGPT-4 provided diagnoses for DR and ME and graded both conditions. Statistical analysis included confusion matrices, performance metrics, and ROC curves (AUC values). Grading accuracy was measured via exact match rates, mean absolute error (MAE), and quadratic kappa scores.

**Results**
For DR diagnosis, ChatGPT-4 achieved an accuracy of 79.7%, sensitivity 81.6%, specificity 75.6%, precision 87.4%, and F1 score 84.4%. Grading accuracy had an exact match rate of 47.7%, MAE 0.82, and kappa 0.51 (95% CI 0.46–0.59), indicating moderate agreement. For ME diagnosis, accuracy was 82.0%, sensitivity 84.1%, specificity 75.6%, precision 84.4%, and F1 score 84.2%. ME grading achieved a 76.9% exact match, MAE 0.33, and kappa 0.71 (95% CI 0.65–0.76), showing substantial agreement.

**Conclusion**
ChatGPT-4 shows promise for DR diagnosis but has limited grading accuracy compared with specialized AI systems such as AEYE-DS, which reports sensitivity of 92.6% and specificity of 95.3%. However, its accessibility and evolving capabilities suggest potential clinical value, particularly in underserved areas. Performance was notably stronger for ME diagnosis and grading, achieving substantial agreement. While ChatGPT-4 cannot replace ophthalmologists, it could assist in early detection of DR and ME, potentially reducing vision loss in resource-limited settings.

## 11. Assessment of retinal microvascular structure before and after anti-VEGF treatment for diabetic macular edema using an OCTA-based AI-Inferred fluorescein angiography system

Takao Hirano[1], Yoshiaki Chiku[1], Ken Hoshiyama[1], Hideaki Mizobe[2], Toba Shuhei[2], Toshinori Murata[1]

*[1]Shinshu University, Nagano, Japan; [2]Canon Inc, Tokyo, Japan*

**Purpose**
Optical coherence tomography angiography (OCTA) has become popular because it can visualize retinal vessels without the use of contrast dye or the risk of allergic reactions. However, a key limitation of OCTA is its reduced ability to detect microaneurysms (MAs) and vascular leakage compared to fluorescein angiography (FA). This study evaluated a new system that uses AI to create FA-like images from OCTA data to identify changes in microaneurysms (MAs) and leakage before and after anti-VEGF therapy for diabetic macular edema (DME).

**Methods**
We obtained 6×6 mm OCTA images (OCT-S1, Canon, Tokyo, Japan) and FA images from five eyes of five patients with DME before and after three intravitreal injections of faricimab. We assessed the number of microaneurysms (MAs) within a 3 mm central area of each FA (early phase), OCTA, and OCTA-based AI-inferred FA (early phase) image before and after anti-VEGF therapy. We also evaluated the macular leakage area within a 3-mm central area of each FA (late phase) and OCTA-based, AI-inferred FA (late phase) image.

**Results**
The number of MAs decreased significantly after treatment compared to baseline in the FA (early phase), OCTA, and OCTA-based AI-inferred FA (early phase) images (19.0 ± 10.8, 6.8 ± 3.1, 16.6 ± 9.2; 5.8 ± 5.3, 1.2 ± 1.8, 3.0 ± 3.5; $P < 0.05$, $P < 0.01$, $P < 0.05$). The total number of MAs identified by OCTA was only 32.2% (40/124) of those determined by FA; however, OCTA-based AI-inferred FA identified 86.3% (107/124) of the MAs. The macular leakage area decreased significantly from before to after treatment in FA (late phase): 2.86 ± 0.95 $mm^2$ to 0.35 ± 0.12 $mm^2$ ($P < 0.01$). A significant reduction was also observed with OCTA-based AI-inferred FA from before to after treatment (2.49 ± 1.09 $mm^2$, 0.27 ± 0.17 $mm^2$, $P < 0.05$).

**Conclusion**
OCTA-based AI-inferred FA more precisely illustrated the reduction in MAs and macular leakage area after anti-VEGF treatment compared to OCTA. AI-inferred FA can non-invasively show retinal circulatory dynamics and thus has potential to aid in DME management.

## 12. Validating a Deep Learning Algorithm to Identify Patients with Glaucoma using Systemic Electronic Health Records

John Xiang, Rohith Ravindranath, Sophia Wang

*Department of Ophthalmology, Byers Eye Institute, Stanford University, Stanford, United States*

**Purpose**
The goal of this study is to explore whether a GPS model trained on All of Us electronic health record (EHR) data can predict if patients seen in an independent academic eye center have high probability of glaucoma based on their systemic EHR data

**Methods**
We finetuned a pre-trained GPS model for use on a Stanford clinic cohort and evaluated on a held-out test set (N=4532). Model inputs included features from EHR such as demographics, systemic diagnoses, medications, laboratory Results, and physical examination measurements. Glaucoma status was determined from

diagnosis codes. The model architecture included two autoencoders to transform one-hot encoded medication and diagnosis data into dense feature representations fed into a convolutional neural network for prediction. During finetuning, we investigated the effect on performance of varying the dataset size and the number of trainable model layers.

For evaluation, we used standard model performance metrics including area under the receiver operating characteristic curve (AUROC) and positive predictive value (PPV) on the test set. Model calibration curves were computed comparing model-predicted glaucoma probability to actual rates of glaucoma diagnosis and clinical measures such as intraocular pressure (IOP), cup-to-disc ratio (CDR), and glaucoma treatment rates.

**Results**

The best fine-tuned model had an AUROC of 0.829 and PPV of 0.709 for detection of glaucoma diagnosis from systemic EHR features. Patients in the highest decile of predicted glaucoma probability (0.9-1) had the highest actual rates of glaucoma diagnosis (0.709) and treatment with medications, surgery, or laser (0.742), and on average, the highest maximum recorded IOP (23.83 mmHg) and CDR (0.612). Increasing the number of trainable layers increased model performance up to a threshold (15), and likewise with the amount of training data (80% comparable to 100%).

**Conclusion**

With transfer learning from a large pre-trained model, we were able to identify which patients at Stanford have high probability of glaucoma using only systemic EHR features, with performance comparable to testing on the original dataset. In addition, the Stanford GPS model was well-calibrated when evaluated against clinical measures such as IOP, CDR, and glaucoma treatment rates.

## 13. Implementation of a Prescreening Artificial Intelligence Algorithm for Geographic Atrophy Trial Enrollment

Roomasa Channa, Amitha Domalpally, Tom Saunders, Asha Jain, Jonathan Chang, Barbara Blodi, Kathleen Schildroth, Kimberly Stepien, Mihai Mititelu, Patricia Saab, Kevin Kurt

*University of Wisconsin, Madison, United States*

**Purpose/Background**

Accurate measurement of geographic atrophy (GA) area is a key enrollment criterion in clinical trials, yet no clinical tools provide rapid, reliable quantification. This gap contributes to high screen-failure rates (50–70%), creating logistical and financial burden. We prospectively evaluated the performance of a locally deployed,

validated artificial intelligence (AI) algorithm for GA area measurement with Fundus Autofluorescence (FAF) when used for prescreening at the clinic, compared with the standard Wisconsin Reading Center (WRC) screening process.

**Methods**
Inclusion criteria were GA area between 1.25 and 23 $mm^2$, unifocal or multifocal lesions, with subfoveal involvement in ≥25% of participants. Heidelberg Spectralis FAF images from a single retina clinic were analyzed by the AI algorithm, which generated GA area measurements and segmentation masks in a clinically readable report. Retina specialists reviewed the reports for accuracy and used the AI-based measurements to prescreen and exclude ineligible patients before trial referral. Regardless of prescreening outcome, all images were sent to the WRC for standard eligibility determination by certified graders. Agreement rates and reasons for discrepancy between AI and human assessments were recorded, and mean GA areas were compared.

**Results**
At clinic prescreening, of 107 eyes screened, AI identified 27 eyes as outside the target range; retina specialists agreed with AI segmentation in 73% of eyes. Discrepancies were most often due to misinterpretation of foveal involvement and omission of small GA lesions. At the WRC, certified graders determined 39 eyes were ineligible, agreeing with the AI report in 26 cases; the most common reason for inaccuracy was the presence of neovascular AMD. Mean GA area was 9.26 ± 8.83 $mm^2$ by AI and 9.03 ± 9.23 $mm^2$ by graders. Prescreening with AI at the clinic reduced the screen-failure rate from 36% to 24%.

**Conclusion**
A locally deployed AI algorithm integrated into the clinic workflow demonstrated good agreement with reading center grading for GA eligibility in a clinical trial. While not a replacement for human review, AI prescreening can reduce the number of ineligible patients referred for formal screening, improving efficiency and lowering trial screening burden.

## 14. Bringing Ophthalmic Clinical Trials into the AI Era

Maya Samman[1], Mark Barakat[2],[3], Ram Yalamanchili[4]

*[1]Arizona College of Osteopathic Medicine, Midwestern University, Phoenix, United States; [2]Retinal Macular Institute of Arizona, Phoenix, United States; [3]University of Arizona - Phoenix, Phoenix, United States; [4]Tilda Research, San Francisco, United States*

**Purpose/Background**
Ophthalmic clinical trials face increased complexity, administrative burden, and multi-dimensional logistical constraints. Artificial Intelligence (AI) teammates can automate repetitive administrative work, streamline data-centric workflows,

standardize protocol adherence, and reduce documentation errors. Their adoption has demonstrated operational impact: reducing query burden, improving data quality, accelerating startup/ investigator site file (ISF) documents, and maintaining inspection readiness. This poster presentation delivers updated Results across diverse ophthalmology clinical research sites, highlighting sustained and scalable improvements.

**Methods**
Query data was pulled from electronic data capture (EDC) and Investigator Site File systems for a before-and-after comparison of performance before onboarding AI teammates. The site was running the study for about 2 months prior to AI implementation.

**Results**
The early 2025 data showed a 90% reduction in queries per visit, from 3.8 to 0.4 (-90%), mean query resolution time decreased from about 14 days to 2 days, >10,000 startup/regulatory documents quality-control reviewed with 92% AI decision acceptance, and 102,000+ data points entered into electronic data capture (EDC). Over 2,500 research hours were spent on the platform.

**Conclusion**
AI is proving to be more than an exciting and novel technology, but it is also offering a durable infrastructure to modernize ophthalmic clinical trials. AI teammates are being seamlessly integrated in real-world ophthalmology trials to increase productivity, reduce coordinator burnout, and accelerate data turnaround times, all while impacting site staff satisfaction. These findings may inform best practices for integrating AI into clinical trial operations, enabling clinical research coordinators and principal investigators to improve efficiency and move the needle of patient care.

## 15. AI-Based Ensemble Model for Detecting Eye-Rubbing Behavior to Assess Keratoconus Risk in Down Syndrome Using Wearable Devices

Binh Duong Giap[1], Jefferson Lustre[1], Joshua Ong[1], Anitha Venugopal[2], Nambi Nallasamy[1]

*[1]University of Michigan, Ann Arbor, United States; [2]Aravind Eye Hospital, Tirunelveli, India*

**Purpose**
Keratoconus (KCN) occurs at a relatively high rate in individuals with Down syndrome (DS), and eye rubbing (ER) has been identified as a potential factor in its progression. This study aims to develop and validate an AI-powered ensemble system to detect ER behavior using sensor data from wearable devices.

**Methods**
Motion sensor data, including 3-axis rotation, acceleration, and quaternion, was collected from 20 healthy participants while performing five non-eye rubbing (NER) activities and eye rubbing (ER) behaviors. Each NER activity lasted 300 seconds, while ER behaviors lasted 30 seconds, all recorded at 100Hz using wearable devices on both wrists. By applying a sliding window of 250 timepoints to each timeseries, a total of 24,960 samples with 13 channels each was generated. An AI-based ensemble system was then developed, combining a 1D convolutional neural network with long short-term memory (1D CNN-LSTM) trained on the time domain, and a 3D CNN trained on the frequency domain (scalograms), to classify samples as either NER or ER, with the predictions from the two models combined using weighted soft voting. The 3D CNN had four convolutional blocks with normalization, pooling, and dropout, followed by global average pooling and dense layers, while the 1D CNN-LSTM combined two convolutional layers, two stacked LSTM layers with dropout and normalization, and global average pooling before the dense classifier.

**Results**
The ensemble was trained and validated on 80% of the dataset (16 subjects) using 5-fold cross-validation for 100 epochs, with a batch size of 16 and the Adam optimizer with an initial learning rate of 0.001. The remaining 20% of the dataset (4 subjects) was used for evaluation. The model demonstrated strong classification performance, achieving an average AUC of 97.39% (±2.53%) and F1-score of 75.16% (±13.50%) with 5-fold CV and an AUC of 98.77% and F1-score of 73.78% on the testing set.

**Conclusion**
The developed AI system successfully identified eye rubbing behaviors using 13-channel motion sensor data collected from wrist-based wearable devices. This method has the potential to facilitate large-scale evaluation of risk factors contributing to the onset and progression of KCN in individuals with DS.

# 16. Performance and Feasibility of an AI Model for RPE Loss Quantification in OCT Imaging of Geographic Atrophy

Amitha Domalpally[1,2], Robert Slater[2], Rushi Mankad[3], Madeline Pflasterer-Jennerjohn[2], Rachel Linderman[1,2], Roomasa Channa[2]

*[1]Wisconsin Reading Center, Madison, United States; [2]A-EYE Research Unit, UW, Madison, United States; [3]University of Wisconsin, Madison, United States*

**Background/Purpose**
Artificial intelligence (AI) models for Retinal pigment epithelial (RPE) loss detection in optical coherence tomography (OCT) imaging show promise, but their practical use as geographic atrophy (GA) area measurement tools depends on both accuracy and feasibility of integration. The Purpose of this study was to evaluate the implementation potential of an AI model for RPE loss quantification in OCT, using predefined thresholds for grader edits to assess feasibility as a semi-automated tool in a reading center workflow.

**Methods**
RPE loss was annotated by expert graders using 3D Slicer with an edge-detection method based on criteria for complete RPE and outer retina atrophy (cRORA). The AI model was trained on 319 OCT scans and tested on 643 OCT scans(243 eyes) from two independent GA clinical trials at the Wisconsin Reading Center. Graders rated AI outputs using a four-tier scale: Category 1 (excellent, no edits), Category 2 (good, minor edits), Category 3 (adequate, moderate edits), and Category 4 (poor, major corrections or unusable). Predefined thresholds for acceptable semi-automated implementation were: ≥70% in categories 1–2, ≤20% in category 3, and ≤10% in category 4.

**Results**
Mean RPE loss (GA) area was 7.2 ± 4.4 $mm^2$ for human graders and 7.2 ± 4.6 $mm^2$ for AI predictions. The average Dice coefficient across all scans was 0.88. Scoring distribution was 141 scans (35.3%) in Category 1, 167 (41.9%) in Category 2, 68 (17.0%) in Category 3, and 23 (5.8%) in Category 4. Dice scores by category were 0.96, 0.92, 0.83, and 0.31, respectively.

**Conclusion**
This AI model for RPE loss meets predefined usability criteria for semi-automated deployment in a Reading Center workflow. Applying objective edit-based thresholds offers a practical framework for assessing AI readiness for clinical trial integration, balancing accuracy with efficiency.

# 17. AI-Powered Surgical Phase Classification and Segmentation System for DMEK Surgical Analysis

Hamza Khan[1,2], Christian Reinhardt[1], Joshua Ong[1], Jefferson Lustre[1], Chanon Thanitcul[1], Keely Likosky[1], Binh D. Giap[1], Ossama Mahmoud[1,2], Bradford Tannen[1], Nambi Nallasamy[1,3]

*[1]Department of Ophthalmology and Visual Sciences, Kellogg Eye Center, University of Michigan, Ann Arbor, United States; [2]School of Medicine, Wayne State University, Detroit, United States; [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States*

**Purpose**

Descemet Membrane Endothelial Keratoplasty (DMEK) is an endothelial keratoplasty technique with a complex intraoperative workflow. Quantitative analysis of DMEK surgical performance requires accurate identification of surgical phases at the frame level and boundary localization at the phase level. We developed and validated a convolutional neural network–based system for automated DMEK phase classification using fully annotated surgical videos.

**Methods**

Seventy-seven DMEK surgery videos were annotated frame-by-frame into 15 non-overlapping surgical phases, with non-informative segments removed and visually similar steps merged. The dataset was divided into training (53 videos), validation (12 videos), and testing (12 videos) sets.

A convolutional neural network (DenseNet-169), pretrained on a surgical instrument image dataset, was adapted for 15-class classification by replacing its final layers with a 256-node fully connected layer and a softmax output layer. Input frames were resized to 240×135 pixels, normalized, and augmented during training with random rotations, translations, shears, zooms, and horizontal flips to improve generalizability. All videos were re-indexed to ensure consistent phase labeling prior to model ingestion. The network was trained using the RMSprop optimizer (learning rate $1\times10^{-5}$) with categorical cross-entropy loss, a batch size of 16, and early stopping based on validation loss. Testing was performed on held-out videos with no temporal overlap with training or validation sets. Metrics included per-class and micro-averaged accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC).

**Results**

The model achieved a micro-averaged accuracy of 0.93, precision of 0.72, recall of 0.72, F1-score of 0.71, and ROC-AUC of 0.84. High-performing classes corresponded to visually distinctive phases with stable instrument/tissue configurations, while lower scores were associated with phases exhibiting substantial inter-case variability or visual similarity to adjacent steps.

**Conclusions**
We demonstrate the feasibility of frame-level DMEK phase recognition using a 2D CNN trained on a fully annotated surgical video dataset. The approach generalizes across surgeons and cases without explicit temporal modeling and establishes a performance baseline for future multi-modal, temporally aware models. This system enables large-scale, objective analysis of DMEK surgical workflows, supporting skill assessment, targeted feedback, and quality improvement in corneal transplantation.

## 18. Bibliometric Analysis of Ophthalmology Artificial Intelligence Research and Venture Capital Investment: A 10-Year Comparative Study (2016-2025)

Victoria Gu[1,2], Jingli Guo[2], Quan Dong Nguyen[2]

*[1]Georgetown University Medical Center, Washington, United States; [2]Stanford University Byers Eye Institute, Palo Alto, United States*

**Purpose**
To quantify the disparity between research productivity and commercial translation in ophthalmology artificial intelligence through comparative analysis of bibliometrics trends and venture capital investment patterns.

**Methods**
A systematic bibliometrics analysis was performed using Web of Science and PubMed databases to identify all AI-related ophthalmology publications from 2016-2025. Venture capital investment data was extracted from PitchBook for companies with ophthalmic-relevant AI tools over the same timeframe. Primary outcomes included annual publication volume, citation metrics (h-index, mean citations per article), and commercial translation indicators (FDA approvals, funding rounds, patent applications). Statistical correlations were assessed using Pearson correlation coefficients.

**Results**
A total of 21,725 AI ophthalmology publications were identified, demonstrating exponential growth from 245 articles in 2016 to 5,297 in 2023. Mean citations per article declined from 34.2 (2016-2018) to 12.7 (2021-2023) despite increased publication volume. Only 5 autonomous AI diagnostic systems received FDA clearance during this period, yielding a publication-to-approval ratio of 4,345:1. Cumulative venture capital investment totaled $126 million across 23 ophthalmology AI companies, with 62% concentrated in diabetic retinopathy screening platforms. Investment peaks consistently lagged research publication surges by 24-36 months ($r=0.73$, $p<0.05$). Patent applications showed moderate correlation with publication volume ($r=0.58$, $p=0.02$) but weak correlation with subsequent FDA

approvals (r=0.31, p=0.24).

**Conclusions**
Despite growth in AI ophthalmology research output, commercial translation remains significantly constrained. The marked decline in citation impact concurrent with publication volume expansion suggests potential research fragmentation and an increased volume of lower impact literature. These findings underscore the need for enhanced industry-academic partnerships and interdisciplinary collaboration between clinicians and AI developers. Without improved integration, clinicians relying on academic literature may become disconnected from commercially viable AI innovations, potentially excluding physician expertise from the development process of tools that will ultimately impact patient care.

**Financial Disclosure**
None.

## 20. Machine learning-based prediction of adverse events after corneal transplantation using donor and tissue characteristics

Karthik Reddy[1], Susan Hurlbert[2], Shahzad Mian[3]

*[1]University of Michigan Medical School, Ann Arbor, United States; [2]Eversight, Ann Arbor, United States; [3]University of Michigan, Ann Arbor, United States*

**Purpose**
Eye banks in the United States have strict regulations set by the FDA and EBAA to ensure safe donor selection and tissue characteristics to limit adverse events. While the overall rate of adverse events is low at approximately 0.14-1.2%, risk stratification is crucial to guide proper clinical observation and management. No models have been developed to aid in identification of patients who may be at higher risk for adverse events based on known tissue and donor characteristics.

**Methods**
This retrospective, single-center study investigated adverse events at a large, regional Midwest eye bank (Eversight, Ann Arbor, MI). Donor data was collected from multiple sources, including existing medical records, direct interviews, and autopsy findings.
Data preparation was conducted in Alteryx Designer (Alteryx, Irvine, CA). Machine learning models were developed in Python (version 3.9.6) using the mljar-supervised package (version 1.1.18). Baseline, decision tree, XGBoost, neural network, random forest, and ensemble model (XGBoost weight = 2, random forest weight = 3) were constructed using a 75%/25% train/test split. We evaluated discrimination using the area under the receiver operating characteristic curve (AUC) and

probability accuracy with the Brier score. Ninety-five percent confidence intervals (95% CI) for metrics were obtained by bootstrap resampling (1,000 samples).

**Results**

There were 129,000 tissues distributed for surgical use. Of these tissues, 176 adverse events were reported (0.14%). The most frequently reported adverse event was primary graft failure (n = 83, 47.2%). Ensemble model had the best performance with the lowest log loss (0.00749) compared to the baseline (0.010308) and better than the best single model (Random Forest, 0.007921). The ensemble model had an accuracy [95% CI] of 0.988 [0.987–0.989], Brier score 0.0013 [0.001–0.002], ROC AUC 0.782 [0.689–0.872], sensitivity 0.477 [0.316–0.623], specificity 0.989 [0.988–0.990], PPV 0.055 [0.033–0.081], and NPV 0.999 [0.999–1.000].

**Conclusion**

Machine learning prediction of adverse events in corneal transplantation shows modest promise for risk stratification. Limitations were due to rarity of the outcome. This may be addressed through national data pooling efforts to increase predictive ability. Future iterations of this model may benefit from oversampling techniques or penalized learning Methods.

## 21. Generating Humphrey Visual Field Points Using Neural Network and Random Forest from Previous and Current Quantitative Spectral-Domain OCT

Justin Bennie[1], Ossama Mahmoud[1], Karim Dirani[1], Victor Tawansy[1,2], Jayen Bastani[1], Mark Juzych[1]

*[1]Kresge Eye Institute, Detroit, United States; [2]Wayne State University School of Medicine, Detroit, United States*

**Purpose**

Glaucoma patients often undergo routine testing to evaluate progression of the disease. While quantitative spectral-domain (SD) OCT and 24-2 Humphrey Visual Fields are the most popular tools to track glaucoma progression with objective findings, the burden of technician manpower, patient motivation, and time are limiting factors for successful administration of visual field testing. The purpose of this project is to investigate the efficacy of generating a visual field using a pair of a previous visual field and SD OCT along with a new SD OCT.

**Methods**

13,819 pairs of previous SD OCT/24-2 Visual Fields and new SD OCT/24-2 Visual Fields were used in a neural network and random forest model. 11,870 samples were used to train both algorithms and tested 1,319 samples at random. These samples were compiled among 4,681 unique patients through 16 years of visual field and SD OCT data at a multi-center academic institution in the Midwest which primarily serves an

urban population. Exclusion criteria included a signal strength of <6 for SD OCT and a multitude of parameters of reliability for visual fields.

**Results**
The neural network yielded a mean test R2 of .5642 (range by element .22-.72) and a mean test MAE of 3.49. The random forest model yielded similar Results, with a mean test R2 of .5809 (range by element .26-.76) and a mean test MAE of 3.39. These models both contained the same two coordinates which had the worst performance within the model, with a MAE of 10.7 and 9.5 in the random forest algorithm. These two coordinates superimpose physiological blind spots. When excluding blind spots, the average MAE was 3.1 and 3.2 for the random forest and neural network respectively. The median interval between the two pairs of tests was 707.0 days.

**Conclusion**
The Results of our study illustrate promising findings of projecting visual fields. Being the gold standard in glaucoma progression, projection of visual fields is a preliminary step in glaucoma progression screening. Further investigation is needed to show the external validity of the model.

## 23. Large Language Models for Use in Diabetic Retinopathy Screening

Amir Akhavanrezayat[1], S. Saeed Mohammadi[2], Sahana Aggarwal[2], Kavina Aggarwal[2], Grant Wiarda[2], Kayla Nguyen[2], Emmanuel Sarmiento[2], Quan Dong Nguyen[1], Manjot Gill[2]

*[1]Byers Eye Institute at Stanford, Palo Alto, United States; [2]Northwestern University, Chicago, United States*

**Purpose**
This study evaluates the potential of large language models (LLMs) as low-cost tools for DR screening. By simulating the FDA-approved IDx-DR protocol, we assessed whether a custom GPT model could accurately detect more-than-mild DR (mtmDR), with the goal of expanding screening accessibility in underserved populations.

**Methods**
We developed a custom GPT (powered by ChatGPT-4o) instructed to follow the LumineticsCore™ (IDx-DR) screening criteria for more-than-mild diabetic retinopathy (mtmDR), defined as an ETDRS level ≥35 and/or clinically significant diabetic macular edema (CSDME). Gemini 2.5 Pro was evaluated with the same criteria. Performance on fundus photography (FP) was assessed using two publicly available datasets: MESSIDOR-2 (n=106; 66 mtmDR, 40 no/mild NPDR) and EyePACS (n=99; 56 mtmDR, 43 non-mtmDR). To assess detection of diabetic macular edema (DME), a separate OCT dataset (n=48; 24 DME, 24 normal) was used to evaluate identification of intraretinal and/or subretinal fluid.

**Results**
On MESSIDOR-2 (n=106), the custom GPT achieved a sensitivity of 90.77%, specificity of 97.50%, PPV of 98.33%, and NPV of 86.67% for mtmDR detection. Gemini 2.5 Pro achieved a sensitivity of 80.30%, specificity of 97.50%, PPV of 98.15%, and NPV of 75.00%. On EyePACS (n=99), the custom GPT demonstrated a sensitivity of 94.64%, specificity of 86.05%, PPV of 89.83%, and NPV of 92.50%, while Gemini 2.5 Pro achieved a sensitivity of 89.29%, specificity of 88.37%, PPV of 90.91%, and NPV of 86.36%. For OCT-based DME detection (n=48), ChatGPT-4o achieved a sensitivity of 95.83%, specificity of 100%, and PPV of 100%, while Gemini 2.5 Pro achieved a sensitivity of 95.83%, specificity of 95.65%, PPV of 95.83%, and NPV of 95.65%.

**Conclusion**
ChatGPT-4o and Gemini 2.5 Pro demonstrated high performance in detecting mtmDR and DME across multiple publicly available datasets. These findings support the potential of LLMs as cost-effective and accessible tools for diabetic retinopathy screening. Further validation in larger, more diverse real-world datasets is warranted.

## 24. Towards Automated Segmentation of Laser Choroidal Neovascularization (LCNV) Murine Model Utilizing a Synthetic Dataset: A Proof-of-Concept Study

Anthony Dongchau[1], Timothy Stout[2], Yasir Sepah[3]

*[1]Texas A&M University, College Station, United States; [2]Cullen Eye Institute, Houston, United States; [3]Byers Eye Institute, Palo Alto, United States*

**Purpose**
Our work presents a proof-of-concept feasibility study towards automatic segmentation and quantification of lesions on fluorescein angiography murine LCNV images as a time-efficient alternative to manual annotations. The study aims to reduce reliance on scarce manual annotations by establishing a quantitative baseline for synthetic-data–driven deep-learning segmentation of leakage areas. For this Purpose, we trained a U-Net model enhanced with the Convolutional Block Attention Module (CBAM) using a custom synthetic dataset.

**Methods**
Synthetic LCNV-like images were generated using the OpenCV (cv2) Python library. Key features included four quadrant-distributed leakage areas, murine-specific vasculature, and variation in global brightness. Leakage geometries and vascular morphologies were parameterized to span expected phenotypic variability. Controlled blur and noise were introduced to test model robustness. The CBAM-enhanced U-Net backbone was selected to improve edge localization and

attention to salient leakage regions. A seven-image validation set was used for pilot assessment, consistent with the proof-of-concept scope.

**Results**
After 15 training epochs, the model achieved a training loss of 0.1316, validation loss of 0.1320, DICE score of 0.5218, and IoU of 0.3525, indicating low-to-moderate segmentation performance. Qualitatively, the model performed best on discrete, well-bounded leakage areas and struggled with confluent or amorphous leakage. These limitations likely arose from under-representation of high-intensity leakage patterns and variable vessel calibers in the synthetic generator. The Results establish a reproducible baseline and highlight the importance of addressing domain-shift between synthetic and real images.

**Conclusions**
These modest performance metrics demonstrate the feasibility of synthetic-data–driven segmentation on murine LCNV imaging, while identifying key dataset limitations that restrict generalization. Future improvements will focus on enriching leakage phenotypes, expanding vascular variability, and adding "distraction splotches" resembling fluorescein diffusion artifacts. Planned model extensions include lesion-area quantification with confidence intervals and ablation studies to isolate the impact of CBAM and dataset parameters. To close the synthetic-to-real gap, we will also evaluate domain adaptation strategies such as style transfer and fine-tuning with a small expert-annotated set, followed by validation in a larger real-image cohort. Overall, this study frames synthetic-data–driven segmentation as a viable and reproducible starting point for iterative improvement in LCNV analysis.

## 25. Deep Learning for Pixel-Level Mapping of Reticular Pseudodrusen on Near-Infrared Reflectance

Leon von der Emde[1], Souvick Mukherjee[1], Dylan Wu[1], Emily Vance[1], Marco Ji[1], Mehdi Emamverdi[1], Tharindu De Silva[1], Alisa T. Thavikulw[1], Jayashree Kalpathy-Cramer Jayashree Kalpathy-Cramer[2], Amitha Domalpally[3], Catherine A. Cukras[1,4], Emily Y. Chew Emily Y. Chew[1], Tiarnan D.L. Keenan[1]

*[1]National Eye Institute, National Institutes of Health, Bethesda, United States; [2]University of Colorado Anschutz Medical Campus, Aurora, United States; [3]Wisconsin Reading Center, University of Wisconsin-Madison, Wisconsin, United States; [4]Roche Pharmaceuticals, Basel, Switzerland*

**Purpose/Background**
Reticular pseudodrusen (RPD) are a key biomarker in age-related macular degeneration (AMD), yet they are often recorded dichotomously, limiting quantitative and spatial evaluation. We sought to build and validate a deep-learning system

that produces pixel-wise RPD segmentations on near-infrared reflectance (NIR) images to enable robust quantification.

**Methods**
We assembled 508 NIR images from 117 eyes (70 participants) spanning a wide range of AMD severities, with and without RPD. Reading-center multimodal grading for RPD presence and NIR contour annotations, followed by pixel-level NIR annotation of every individual RPD lesion from four retinal specialists, was performed. Data were partitioned 80:20 into training and test subsets. A DeepLabv3–ResNet-18 model ("ReticularNet") was trained to generate pixel-level RPD masks. Performance on the test set was benchmarked against retina specialists. Overlap accuracy was measured by Dice similarity coefficient (DSC). Agreement for lesion number, pixel area, and contour area was assessed using intraclass correlation coefficients (ICCs).

**Results**
ReticularNet achieved a mean DSC of 0.36 (SD 0.16), surpassing each specialist (range 0.03–0.23; $p \leq 0.02$ for all) and the pooled specialists ($p < 0.0001$). ICCs were 0.44 for lesion count, 0.56 for pixel area, and 0.61 for contour area, with no significant systematic under- or overestimation (all $p \geq 0.24$). These ICCs were numerically higher than those of each specialist (per-specialist ICC ranges: −0.08 to 0.23, −0.05 to 0.40, and −0.09 to 0.58, respectively), and for all three parameters ReticularNet exceeded the pooled specialists ($p \leq 0.02$). Most specialists showed significant underestimation.

**Conclusion**
ReticularNet provides automated, pixel-level quantification of RPD on NIR images and outperforms retina specialists on overlap and agreement metrics for both area and lesion number. Enabling scalable quantitative and spatial RPD assessment may deepen insight into RPD as a biomarker in AMD.

## 26. AI-Enabled Fundus Imaging for Screening of Retinal Diseases and Glaucoma in South India

Neelam Pawar[1], Nambi Nallasamy[2], Meenakshi Ravindran[1], Syed Mohideen[1], Devendra Maheshwari[1]

*[1]Aravind Eye Hospital, Tirunelveli, India; [2]Kellogg Eye Center, Ann Arbor, United States*

**Purpose**
To evaluate the effectiveness of AI-assisted fundus imaging (Remidio Fundus on Phone NM-10 with Medios AI) for screening retinal diseases and glaucoma in the free section of a rural community hospital serving low-income areas in South India.

**Methods**
A cross-sectional screening initiative was conducted in the free outpatient section of Aravind Eye Hospital, Tirunelveli, Tamil Nadu, India between September 2024 and

July 2025. This section caters exclusively to patients from socioeconomically disadvantaged rural Backgrounds. Trained technicians obtained non-mydriatic fundus photographs using the Remidio Fundus Non-Mydriatic on an iPhone device. Medios AI analyzed images for diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma suspects.
All AI-detected heat maps flagged as DR, glaucoma suspect, or AMD cases were reviewed by a general ophthalmologist, with referrals to retina or glaucoma specialists at the Paying section of base hospital for further evaluation.

**Results**
Of 8,000 patients screened, aged 42– 79 years, 426 (5.3%) required referral for specialist evaluation. The largest group was DR (n = 222; 52.1%), followed by glaucoma suspects (n = 112; 26.3%), AMD (n = 62; 14.6%), and other retinal/optic conditions (n = 30; 7.0%). Screening was integrated seamlessly into the hospital's workflow, ensuring the early detection of sight-threatening diseases at no additional cost to patients, who often lack access to specialty care.

**Conclusion**
AI-assisted fundus imaging is a scalable, cost-effective model for disease triage in free hospital settings. By embedding AI analysis into routine care at the base hospital's free section, this approach ensures equitable access to early detection of DR, glaucoma, and AMD for poor rural populations, thereby preventing avoidable blindness while maintaining efficiency in high-volume services.

## 27. Systematic Review of Large Language Model–Generated Summaries in Ophthalmology: Educational Applications, Benefits, and Risks

John Monroe, John Williams, Khristy Tapiero, Steven Williams

*Norton College of Medicine, Upstate Medical University, Syracuse, United States*

**Purpose/Background**
Large language models (LLMs) are increasingly used to simplify complex medical information and generate plain-language summaries (PLSs). In ophthalmology, LLMs have been applied to case reports, clinical notes, and patient education materials. These summaries may improve learning efficiency, support interdisciplinary communication, and aid patient understanding. However, concerns remain about accuracy, hallucinations, and limited transparency. This systematic review evaluates published studies on LLM-generated summaries in ophthalmology, emphasizing their educational and clinical applications.

**Methods**
A systematic PubMed search was conducted through August 2025 using the terms

"ophthalmology," "artificial intelligence," "large language model," and "plain-language summary." Eligible studies assessed AI-generated outputs in education, clinical communication, or patient-facing contexts. Data were extracted on study design, setting, application, outcomes (readability, comprehension, accuracy, learner or patient perceptions), and reported limitations. Due to heterogeneity in design and outcome measures, findings were synthesized narratively.

**Results**
Multiple studies examined LLM-generated summaries in ophthalmology. Across the literature, AI-generated text demonstrated greater readability than original material, with improvements on standard indices. Randomized trials indicated that learners reported better comprehension and reduced cognitive load when using PLSs. Accuracy ratings by ophthalmologists were generally favorable, though occasional errors were noted, underscoring the need for oversight. Educational uses included case-based learning, where summaries improved efficiency, and test-question generation. Clinical applications ranged from guideline synopses to communication tools for non-specialist providers and patients. While benefits were consistent across studies, challenges such as hallucinations, variability between platforms, and interpretability concerns were emphasized.

**Conclusion**
LLM-generated summaries show promise in enhancing comprehension, efficiency, and learner satisfaction in ophthalmology education, while also supporting interdisciplinary and patient communication. Nonetheless, limitations in accuracy and transparency highlight the need for structured human review before curricular or clinical adoption. Future work should establish standardized evaluation metrics, expand to subspecialty contexts, and assess long-term educational outcomes. Artificial intelligence appears most valuable as an adjunct to—not a replacement for—traditional educational Methods, with rigorous validation and oversight essential for safe implementation.

## 28. Physician Involvement in FDA-Approved AI Ophthalmology Devices: A Multidimensional Assessment of Roles and Agency

Jingli Guo[1,2], Victoria Gu[3,1]

*[1]Stanford Byers Eye Institute, Palo Alto, United States; [2]Xinhua Hospital, affiliated to Shanghai Jiaotong University, Shanghai, China; [3]Georgetown University Medical Center, Washington D.C., United States*

**Background**
The integration of artificial intelligence into clinical ophthalmology necessitates examination of how physicians are positioned along a spectrum: from active technical

collaborators to primarily end-users of predetermined systems. We analyzed FDA-approved AI ophthalmology devices using a multidimensional framework to characterize physician involvement and agency across the development lifecycle.

**Methods**

Comprehensive analysis of 10 FDA-approved AI devices for ophthalmology (2018-2025), examining regulatory documents, clinical validation studies, and peer-reviewed publications. We developed a six-dimensional scoring framework (1-5 scale each) assessing: (1) Technical Co-Development, (2) Clinical Leadership & Study Design, (3) Training Data & Reference Standards, (4) Workflow Integration & User Experience Design, (5) Real-World Implementation & Continuous Improvement, and (6) Physician Agency & Decision-Making Authority. Based on dimensional profiles, devices were categorized as Comprehensive Co-Developer, Active Collaborator, Engaged Consultant, Moderate Participant, or Limited Participant.

**Results**

Of 10 FDA-approved devices analyzed, 40% (n=4) demonstrated Comprehensive Co-Developer patterns with physicians serving as technical decision-makers across multiple dimensions (mean score: 4.6/5.0), including IDx-DR, EyeArt, AEYE-DS, and Aurora AEYE. 40% showed Engaged Consultant patterns (mean score: 2.75/5.0) with physicians primarily in validation and advisory roles. 20% exhibited Limited Participant patterns (mean score: 2.0/5.0) with minimal physician decision-making authority. Technical Co-Development and Physician Agency showed the greatest variation (scores 1-5), while Clinical Leadership was more consistent across devices (scores 2-5).

**Conclusions**

While all FDA-approved AI ophthalmology devices incorporated physician input, only 40% demonstrated comprehensive physician collaboration as technical co-developers with meaningful decision-making authority throughout development. Current development patterns risk creating a bifurcated field where physicians serve primarily as validators rather than collaborators, potentially limiting optimal clinical integration and professional agency. Future studies should examine physician involvement in investigational AI ophthalmology device, as well as examine evolving forms of physician collaboration across emerging technologies, subspecialties, and regulatory environments.

# 29. Keratoconus detection using an artificial neural network with OCT-based indices

Yan Li, Jiachi Hong, Travis Redd, Xubo Song, David Huang

*Oregon Health & Science University, Portland, United States*

**Purpose/Background**

Unrecognized keratoconus is the primary risk factor for post-LASIK ectasia, a serious complication of this common refractive procedure. While corneal topography can detect most cases, very early-stage keratoconus remains difficult to identify.

**Methods**

We investigated a feed-forward artificial neural network (ANN) to distinguish keratoconus from non-keratoconus eyes using four OCT-based features: epithelial pattern standard deviation (Epi PSD), coincident-thinning (CTN) index, ectasia index, and epithelial modulation (EM) index. Grid searches optimized the ANN architecture and hyperparameters. The final network layer consisted of a single neuron with sigmoid activation, outputting the probability of keratoconus. Model performance was assessed by repeated 5-fold cross-validation (70% training, 10% validation, 20% testing) on 131 keratoconic eyes (manifest, subclinical, or forme fruste) and 148 control eyes (normal or warpage).

**Results**

The ANN achieved an average balanced accuracy of 94 ± 3%. Precision and recall were 99 ± 2% and 91 ± 3%, respectively, with an F1 score of 0.95 ± 0.02 and an AUC of 0.95 ± 0.02. The network correctly identified 100% of manifest and subclinical keratoconus cases and 56 ± 17% of forme fruste keratoconus cases. Specificity was high, with 98 ± 5% of normal eyes and 100 ± 0% of warpage eyes correctly classified as non-keratoconus.

**Conclusion**

The OCT indices based ANN demonstrated strong diagnostic performance, accurately differentiating keratoconus from non-keratoconus eyes, including challenging subclinical cases, and may serve as a valuable tool for improving ectasia risk screening before refractive surgery.

# 30. Evaluation of Deep Learning Models for Detecting Artifacts from Corneal In Vivo Confocal Microscopy Images

Dena Ballouz[1,2], Binh Duong Giap[1], Roni Shtein[1], Nambi Nallasamy[1,3]

*[1]Kellogg Eye Center, University of Michigan, Ann Arbor, United States; [2]Bascom Palmer Eye Institute, University of Miami, Miami, United States; [3]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States*

**Purpose**

Artifacts in corneal *in vivo* confocal microscopy (IVCM) images caused by patient movement, blinking, or acquisition errors can reduce image quality and affect diagnosis. Detecting these artifacts is essential for accurate interpretation. This study evaluates the performance of state-of-the-art deep learning models in identifying artifacts in corneal IVCM images.

**Methods**

A dataset of 11,947 corneal IVCM images of size 384x384 from 12 individuals with non-microbial keratitis was established to develop and validate deep learning models for the detection of artifacts. Images were labeled into two classes, Artifact (AR) and Non-Artifact (NAR), by an experienced confocal reader and cornea specialist at University of Michigan's Kellogg Eye Center. The dataset was randomly split into training (7 subjects, 6,623 images), validation (2 subjects, 3,446 images), and testing (3 subjects, 1,878 images) subsets. Seven deep learning models, DenseNet-121, DenseNet-201, ResNet-152, ResNet-50, VGG-19, VGG-16, and MobileNetV2, pretrained on ImageNet, were trained on the training set and tuned on the validation set under two settings: freezing and unfreezing convolutional layers, resulting in a total of 14 models evaluated. Data augmentation was applied with batch size of 64 to enhance model generalizability. The Adam optimizer was used with an initial learning rate of 0.001. The best-performing model weights were then evaluated on the testing set.

**Results**

All 14 models achieved high performance on the validation set, with an average F1-score of88.38% 2.18 and an AUC of94.77% 1.88. VGG-based architectures with the unfrozen convolutional layers achieved the highest performance, with an F1-score of 91% and AUC of 97%. On the testing set, the models achieved an average F1-score of79.14% 3.53 and an AUC of89.07% 3.08. Among these models, DenseNet-201 with convolutional layers frozen yielded the best performance, with F1-score of 86% and AUC of 94%.

**Conclusion**

Deep learning models can effectively detect artifacts in IVCM images with VGG- and DenseNet-based architectures showing the highest performance. These findings demonstrate the potential of automated artifact detection to improve image

quality assessment and support more accurate and reliable IVCM-based analysis of keratitis.

## 31. AI-Enabled Objective Assessment of Eye Stability During Delayed Sequential Bilateral Cataract Surgery

Jefferson Lustre[1], Duong Binh Giap[1], Ossama Mahmoud[2], Nambi Nallasamy[1]

*[1]University of Michigan, Ann Arbor, United States; [2]Kresge Eye Institute, Detroit, United States*

**Purpose**

Patients commonly report greater awareness and discomfort during second eye surgeries despite identical anesthesia, though the impact on patient cooperation remains unclear. Here, we pilot a novel approach for objectively assessing intraoperative patient eye movement differences between first and second surgeries in delayed sequential bilateral cataract surgery (DSBCS) using a deep learning-based system for intraoperative cataract surgery analysis.

**Methods**

A deep learning-based system, CatSkill, originally developed to assess the skills of cataract surgeons, was used to compute the mean and standard deviation (STD) of three cataract surgery assessment metrics (CSAMs), including LCP1 (distance between limbus centroid and Purkinje image 1), LCFC (distance between limbus centroid and video frame center), and LFL (focus level within the limbus region), for each pair of videos from bilateral cataract surgeries. The CSAMs were then analyzed to compare differences between the first and second surgeries. A dataset of 380 pairs of cataract surgeries on 190 patients, each receiving equivalent dosages of midazolam, fentanyl, and ondansetron under the same surgeon across both surgeries was established for analysis across 11 surgical phases using the system with emphasis on the “No activity” phase, wherein the eye was not actively manipulated by surgical instruments, to approximate eye stability. The phases of the surgical videos were identified automatically using a deep learning system, CatStep.

**Results**

In the setting of identical sedation and pain management, no significant differences in eye stability were observed between first and second cataract surgeries during “No activity” phases ($p\_LCFC_{std}$ = 0.498, $p\_LCP1_{std}$= 0.132, $p\_LFL_{std}$= 0.183). In downstream analysis, among five machine learning models evaluated, a Naïve Bayes classifier achieved the highest performance, with an AUC of 0.80, in distinguishing first and second eyes using the CSAM metrics.

**Conclusion**

This study demonstrates a novel use of the CatSkill and CatStep systems to

objectively compare patient cooperation during first and second eye cataract surgeries. Under identical anesthesia, CSAM-based analysis shows similar eye stability during no activity phases. This approach is promising for large-scale studies to further quantify intraoperative eye movement and optimize surgical strategies.

## 32. Integrating Large Language Models into Clinical Decision Support for Complex Uveitis Management

Negin Yavari[1], Christopher Or[1], Gunay Uludag[1], Mohammadbagher Rajabi[1], Ishaan Iyer[1], Gunjan Awatramani[1], S. Saeed Mohammadi[2], Dalia Elfeky[1], Jia-Horung Hung[1], Ngoc Trong Tuong Than[1], Amir Akhavanrezayat[1], Isabel Sendino Teronio[3], Irmak Karaca[1], Osama Elaraby[1], Azadeh Mobasserian[1], Jingli Guo[1], Quan Dong Nguyen[1]

*[1]Byers Eye Institute, Stanford University, Palo Alto, CA, United States; [2]Department of Ophthalmology, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States; [3]Department of Ophthalmology, Complejo Asistencial Universitario de Leon, Spain*

**Purpose**
To explore the feasibility of integrating large language models (LLMs) into clinical decision support (CDS) systems for the management of complex uveitis.

**Methods**
We designed a proof-of-concept LLM-based CDS framework that ingests de-identified clinical notes, imaging reports, and laboratory results from uveitis encounters. A curated knowledge base derived from published guidelines and trial data was paired with a retrieval-augmented generation (RAG) pipeline. The model was prompted to output structured recommendations including phenotype, suspected etiologies, disease activity, suggested work-up, treatment plan, and safety alerts. A panel of fellowship-trained uveitis specialists independently reviewed each case and established a consensus "reference standard." Model outputs were compared against this standard to assess accuracy of phenotype and etiology classification, concordance of work-up and treatment recommendations, and the rate of unsafe or contraindicated suggestions.

**Results**
Across 50 standardized uveitis cases, LLM predictions demonstrated perfect agreement with reviewer consensus for phenotype, activity status, primary etiology, safety flags, and red-flag recognition (accuracy and $\kappa = 1.0$). Agreement was moderate for systemic immunomodulatory therapy (IMT) (accuracy 62%, $\kappa = 0.61$) and steroid regimen recommendations (accuracy 54%, $\kappa = 0.42$), but lower for biologics (accuracy 40%, $\kappa = 0.38$), local therapies (accuracy 10%, $\kappa = 0.09$), and contraindications (accuracy 24%, $\kappa = 0.24$). No instances of unsafe treatment recommendations were identified when applying strict reviewer-defined contraindications.

**Conclusion**
LLM can accurately recognize uveitis type, activity, etiology, and safety risks, showing perfect agreement with experts. Performance was weaker for treatment choices (IMT, biologics, local therapy). No unsafe recommendations were detected, but refinement is needed before clinical use.

## 33. Accurate Machine-Learning (ML) Ellipsoid Zone (EZ) Measurements of Volume and EZ Total Loss of Optical Coherence Tomography (OCT) scans in Non-exudative Macular Degeneration Eyes

Vlad Diaconita[1], Hanna Coleman[2], Jason Slakter[2]
*[1]Columbia University, New York, United States; [2]Voiant Clinical, Waltham, United States*

**Purpose**
To segment the EZ layer with accuracies approaching those of expert retinal graders using deep learning.

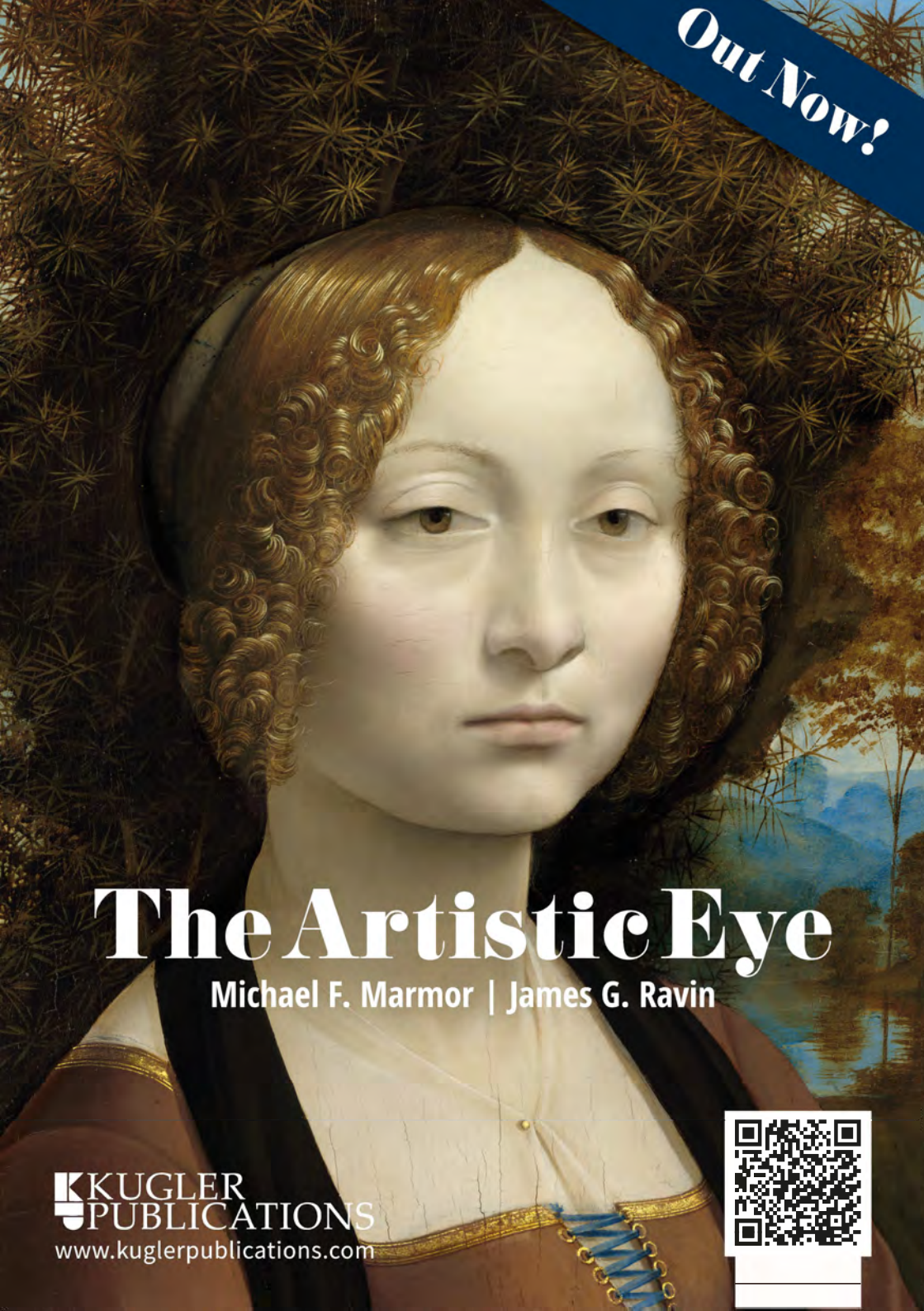
**Methods**
Two hundred sixteen (216) Optical Coherence Tomography (OCT) image volumes from 30 subjects were manually labeled by two independent retinal expert graders. They identified the ellipsoid zone (EZ), the retinal pigment epithelium (RPE), and Bruch's membrane (BM). Total EZ Loss zones were manually marked (Figure 1). To train our segmentation machine learning (ML) model, we use a modified version of the Sam2Unet architecture. Sam2Unet uses the well-studied SAM2 vision transformer as the backbone encoder for a more traditional U-Net segmentation framework. The effect is to capture hierarchical contextual features, making it powerful for delineating subtle or ambiguous boundaries like the thinning ellipsoid zone.
The training was run for 80 epochs using the AdamW optimizer with a "reduce on plateau" learning rate scheduler. We used 5-fold cross-validation stratified across subjects such that no subject's images were split across the train/test/validation sets. For each fold, we held out 20% of the training data as a validation set. We also augmented the data with a random combination of speckle noise, brightness/contrast/gamma adjustments, coarse dropouts, gaussian blur, and elastic deformations.

**Results**
The Pearson correlation of Total EZ Area Loss between manual expert graders and ML model was 0.87, with an EZ-RPE thickness volume correlation of 0.90. The Dice Score average across all cases was 0.82 +/- 0.07 (Figures 2 and 3). This reinforces the strength of the model at predicting EZ Area Loss and EZ-RPE thickness volumes in

OCT B-scans compared to independent expert graders.

**Discussion and Conclusion**

EZ-RPE thickness has been shown to correlate well with function and predict the rate of progression in geographic atrophy. Clinical correlation between EZ biomarkers and progression of AMD relies on accurate, reproducible and defensible data which can be assessed at multiple time points. We show that a highly performant ML model can accurately report EZ-RPE thickness volume and EZ Total Area Loss, allowing for deployment in clinical and research environments.


Out Now!
The Artistic Eye
Michael F. Marmor | James G. Ravin
KUGLER PUBLICATIONS
www.kuglerpublications.com

# The International Congress of Advanced Technologies and Treatments in Ophthalmology

**April 24-26, 2026 | Kaunas, Lithuania**

## OCULOMICS THROUGH OCULAR IMAGING

***Registration and Abstract submissions will open in September 2025***

ICATTO

ICATTO.COM

# OCULOMICS THROUGH OCULAR IMAGING

## Exploring the Eye as a Window to Systemic Health

*Announcing an exclusive expert meeting focused on Oculomics, an emerging field at the intersection of Ophthalmology, Imaging, and Systemic Disease Biomarkers*

**April 24-26, 2026 | Location: Kaunas, Lithuania | ICATTO.COM**

Join leading researchers, clinicians, imaging specialists, mathematicians, physicists, IT experts to discuss how advanced ocular imaging technologies — including OCT, OCT-A, and AI-assisted diagnostics — are transforming our understanding of systemic and neurodegenerative diseases through insights gained from the eye.

**Highlights:**

- Cutting-edge developments in ocular imaging and analytics
- AI- powered phenotyping
- Biomarker discovery – the role of tear fluid, retina, optic nerve
- Multidisciplinary perspectives on translating oculomic data into clinical practice
- Opportunities for collaboration and innovation

This meeting is designed for professionals in ophthalmology, neurology, biotechnology, imaging sciences, data analytics, and systems biology who are passionate about integrating ocular imaging into broader health diagnostics.

**Scientific Program Committee Chairs:**

Alon Harris ( USA)
Ingrida Januleviciene (Lithuania)

ICATTO

# Artificial Intelligence in Vision & Ophthalmology

While the rapid advance of imaging technologies in ophthalmology is making available a continually increasing number of data, the interpretation of such data is still very challenging and this hinders the advance in the understanding of ocular diseases and their treatment. Interdisciplinary approaches encompassing ophthalmology, physiology, mathematics, engineering, and computer science have shown great capabilities in data analysis and interpretation for advancing basic and applied clinical sciences. Artificial Intelligence in Vision and Ophthalmology (AIVO) was created with the aim of providing a forum for interdisciplinary approaches integrating mathematical and computational methods with experimental and clinical studies to address open problems in ophthalmology. AIVO welcomes articles that investigate questions related to the anatomy, physiology and function of the eye in health and disease.

SAIVO
Society for AI in Vision and Ophthalmology
Established 2025

*Official Journal of the Society for Artificial Intelligence in Vision and Ophthalmology (SAIVO)*

**www.aivojournal.com**
**Published by Kugler Publications**
**www.kuglerpublications.com**