AMERICAN ACADEMY OF OPHTHALMOLOGY®

Strong versus Weak Data Labeling for Artificial Intelligence Algorithms in the Measurement of Geographic Atrophy

Amitha Domalpally, MD, PhD,[1,2] *Robert Slater, PhD,*[1] *Rachel E. Linderman, PhD,*[1,2] *Rohit Balaji,*[2] *Jacob Bogost,*[1] *Rick Voland, PhD,*[2] *Jeong Pak, PhD,*[2] *Barbara A. Blodi, MD,*[1] *Roomasa Channa, MD,*[2] *Donald Fong, MD,*[3] *Emily Y. Chew, MD*[4]

Purpose: To gain an understanding of data labeling requirements to train deep learning models for measurement of geographic atrophy (GA) with fundus autofluorescence (FAF) images.

Design: Evaluation of artificial intelligence (AI) algorithms.

Subjects: The Age-Related Eye Disease Study 2 (AREDS2) images were used for training and cross-validation, and GA clinical trial images were used for testing.

Methods: Training data consisted of 2 sets of FAF images; 1 with area measurements only and no indication of GA location (Weakly labeled) and the second with GA segmentation masks (Strongly labeled).

Main Outcome Measures: Bland–Altman plots and scatter plots were used to compare GA area measurement between ground truth and AI measurements. The Dice coefficient was used to compare accuracy of segmentation of the Strong model.

Results: In the cross-validation AREDS2 data set (n = 601), the mean (standard deviation [SD]) area of GA measured by human grader, Weakly labeled AI model, and Strongly labeled AI model was 6.65 (6.3) mm^2, 6.83 (6.29) mm^2, and 6.58 (6.24) mm^2, respectively. The mean difference between ground truth and AI was 0.18 mm^2 (95% confidence interval, [CI], −7.57 to 7.92) for the Weakly labeled model and −0.07 mm^2 (95% CI, −1.61 to 1.47) for the Strongly labeled model. With GlaxoSmithKline testing data (n = 156), the mean (SD) GA area was 9.79 (5.6) mm^2, 8.82 (4.61) mm^2, and 9.55 (5.66) mm^2 for human grader, Strongly labeled AI model, and Weakly labeled AI model, respectively. The mean difference between ground truth and AI for the 2 models was −0.97 mm^2 (95% CI, −4.36 to 2.41) and −0.24 mm^2 (95% CI, −4.98 to 4.49), respectively. The Dice coefficient was 0.99 for intergrader agreement, 0.89 for the cross-validation data, and 0.92 for the testing data.

Conclusions: Deep learning models can achieve reasonable accuracy even with Weakly labeled data. Training methods that integrate large volumes of Weakly labeled images with small number of Strongly labeled images offer a promising solution to overcome the burden of cost and time for data labeling.

Financial Disclosures: Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100477 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

The development of artificial intelligence (AI) models for medical imaging typically involves a well-defined pathway involving identification of an application, image curation, development of AI architecture, training, validation, and deployment.[1,2] Although there is an abundance of literature with detailed information on model architecture and performance metrics, the crucial steps of image curation and the development of training data sets are often overlooked and underreported. It is well known that the parameters of the training data can significantly impact the accuracy and generalizability of the resulting model.[3,4]

Preparing imaging data for training AI models is a complex process that involves several steps.[5,6] Although there is significant emphasis on the quality and diversity of the training data, less attention is often paid to the critical role of data labeling and ground truth.[7] Ground truth refers to the data label that is linked to each image and serves as the reference standard for training the model. Choosing an appropriate label is crucial and depends on the specific task at hand. For example, binary classifiers require a presence/absence label, whereas segmentation models require annotations, where the pathology is outlined. Segmentation essentially assigns a label to each pixel of the image where the annotation can be classified as presence and lack of annotation as absence. In Figure 1, the image is segmented to identify

https://doi.org/10.1016/j.xops.2024.100477
ISSN 2666-9145/24

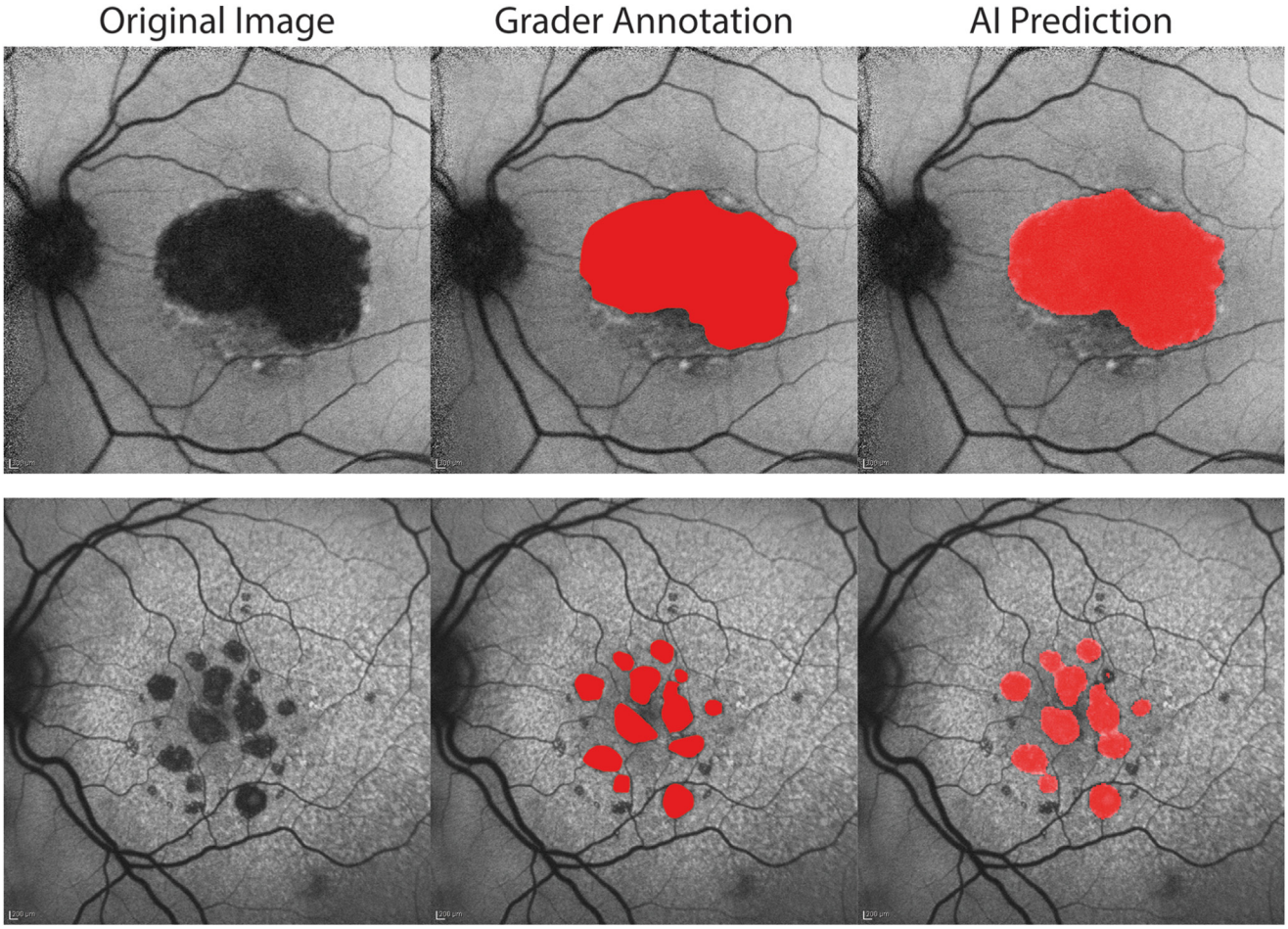

Figure 1. Autofluorescence images of eyes with geographic atrophy (GA), grader annotation, and artificial intelligence (AI) prediction shown for a unifocal (top row) and multifocal (bottom row) GA.

regions of geographic atrophy (GA). Every red pixel is linked to GA and lack of red pixels to non-GA area.

Semantic segmentation is a type of image annotation that involves outlining a region of interest or generating a mask that can be used for training. Semantic segmentation provides a precise means of identification and localization of pathology within an image. Because semantic segmentation is time consuming, researchers in other fields have used weaker forms of annotations known as instance segmentations such as bounding boxes, scribbles, or point annotations on the image to successfully train AI models.[8]

Enlargement of GA area with fundus autofluorescence (FAF) images is an important outcome for clinical trials and there is a need for rapid and reliable measurement tools.[9,10] Artificial intelligence models have been successfully developed using FAF imaging for automated assessment of GA area using deep learning.[11–14] Training these models requires a large number of FAF images with segmentation of GA by reading centers. There are no publicly available data sets that fulfill these criteria, restricting model development to those with access to large pharmaceutical trial data sets with reading-center segmentation of GA.[14,15] The main objective of this project was to investigate the labeling requirements for training AI models in measurement of GA area. To achieve this, a comparison was made between AI models trained using Weakly labeled data (FAF images without segmentation of GA area) and Strongly labeled data (FAF images with segmentation).

Methods

Training and Cross-Validation Data Set

Age-Related Eye Disease Study 2 (AREDS2) was a multicenter randomized clinical trial designed to study the effects of oral supplements on progression to advanced age-related macular degeneration (AMD).[16] The study was conducted under institutional review board approval at each site, and written informed consent was obtained from all study participants. The research was conducted according to the tenets of the Declaration of Helsinki and complied with the Health Insurance Portability and Accountability Act. Participants at high risk of developing late AMD due to either bilateral large drusen or late AMD in 1 eye and large drusen in the fellow eye were enrolled. Development of either central GA or neovascular AMD was the primary AREDS2 study outcome.

An autofluorescence ancillary study was initiated to obtain FAF images from a subset of participating clinics (36 of 90 sites) based on availability of imaging equipment.[17] Sites were permitted to join the ancillary study at any time after imaging equipment became available during the study period between the first AREDS2 visit and 5-year follow-up visit (2007−2013). Fundus autofluorescence images were obtained using the Heidelberg Retinal Angiograph (HRA) by certified photographers. A single image was acquired at 30 degrees centered on the macula, captured in high-speed mode (768 × 768 pixels) using the automated real-time mean function set at 14. Images were exported as tiff format to the Wisconsin Reading Center (formerly Fundus Photograph Reading Center) for evaluation by certified graders.

Image Evaluation

For this project, FAF images with GA were included from AREDS2 study visits at year 4, 5, and 6, because that was the time frame where most sites with FAF capabilities joined the ancillary study. There were 1501 FAF images corresponding to these visits. Eyes were chosen randomly from the visit years 4, 5, and 6 for segmentation. Hypoautofluorescence or GA was classified as well-defined, homogenously black areas with a minimum size of 250 microns in its widest diameter. Areas of hypoautofluorescence within the entire macula-centered FAF image were demarcated using Photoshop (Adobe Inc. v 24.4.1) with a red outline and filled in with the paint bucket tool. Images were deemed ungradable and excluded from this study if the border of GA merged with peripapillary atrophy and could not be distinguished, if the GA extended outside the field of the image, or if poor image quality prevented clear delineation of GA borders. In Heidelberg FAF images, the macula was assumed to be involved if the hypoautofluorescent patch merged with the darkness of the macula and there was no clear region demarcating the 2 OCT images, which provide a more accurate assessment of foveal involvement were not available.

Images were calibrated using the burnt in calibration scale. The pixels in red were converted to area measurements in mm^2. Areas were summed for eyes with multifocal GA to yield a single value. For visualization, Grader and Predicted areas were put into the red channel of a red-green-blue image, whereas the black and white FAF image was converted to red-green-blue. The 2 images were then superimposed together to produce visualized annotations. Graders used Photoshop to implement the mask, the predictions were added manually using the Python Pillow package.

External Validation (Testing)

Validation was performed using screening visit FAF images from a phase 2 study conducted by GlaxoSmithKline (GSK) between 2011 and 2016 (NCT NCT01342926).[18] This was a multicenter study conducted across 40 centers in United States and Canada, and the study concluded that the experimental drug did not slow the enlargement rate of GA compared with placebo. Inclusion criteria required well-demarcated GA with an area of 1.9 to 17 mm^2 measured on color fundus photographs of the study eye. For multifocal GA, at least 1 of the foci had to be $\geq$ 1.9 mm^2, and the total area of GA had to measure $\leq$ 17 mm^2. Fundus autofluorescence images were obtained as supplementary images using the same procedures as AREDS2 but were exported to the reading center in the Heidelberg proprietary e2e format. Geographic atrophy segmentation was performed using the same procedures as mentioned previously in Heidelberg software. Images with annotations were exported in tiff format for AI validation.

AI Model Development

The EfficientNet Architecture was selected to be able to rapidly try and then scale architectures to create models.[19] The terms "Weakly Labeled" or Weak and "Strongly Labeled" or Strong are used for each model.

Weakly Labeled Model

The Weak model was trained on an EffcientNet-B5 with an input size of 512 × 512 and a single output, using Imagenet pre-trained weights. The area of GA was used as a target, and a fivefold cross-validation, split on subject ID, was used to estimate the performance of the model. The model was trained using Mean Squared Error and the Adam optimizer, using an early stopping of 3 epochs on the validation set. Early stopping criteria of 3, 5, and 10 showed no difference in final performance. This resulted in an average number of training epochs being 15.

Strongly Labeled Model

The Strong model was a Feature Pyramid Network with an EffcientNet-B5 encoder and 2 class outputs for image segmentation.[20] Again, the input size was 512 × 512 pixels, but the target was now a segmented image of GA with dimension 512 × 512. The same method of fivefold cross-validation using subject ID to split images was used to estimate performance. Once a prediction was received, the Dice score was calculated with the target, and the area was calculated by counting the number of pixels identified as GA and multiplying that count by the known pixel area in mm^2. Dice score was defined as 2 × Common Elements / (# of Elements in Set A + # Elements in set B), with set A being the target pixels and set B being the predicted pixels. Thus, a Dice coefficient close to 1 indicated close agreement or overlap between predicted and target areas. The grader measured area was calculated in the same manner, by counting the pixels of the target segmentation. The Dice loss (1 − Dice coefficient) was optimized, and early stopping was implemented by monitoring the average Dice coefficient on the validation subset over 3 rounds. Strong models averaged 20 epochs of training.

Data Preparation

Some input data from Heidelberg had an "information" bar that contained nonessential image data. This was cropped out by making the image perfectly square. Image sizes varied between 868 × 768 and 1636 × 1536 pixels. Cropping off the last 100-pixel rows removed the label bar and made the pictures square to prevent distortion when resizing to 512 × 512.

Training was conducted on a single nVidia Quadro RTX 5000. Batch size was set to 4 for both models, which was the maximum size for the Strong model, which could be trained on the single GPU at a 512 × 512 resolution. Annotated images were used to generate segmentation mask targets by selecting the grader annotation color as the segmented class and all others as background. Data augmentation (including rotations, flips, and contrast limited adaptive histogram equalization) was tried but had minimal effect on the performance metrics for both models.[12,21,22]

Artificial Intelligence Model Performance Metrics

Geographic atrophy characteristics were outlined using summary statistics. The Weakly labeled model provided an area output in mm^2, whereas the Strongly labeled model provided both area in mm^2 and segmentation of GA. Area in the Strong model was calculated using a sum of the GA labeled pixels and then

Table 1. Characteristics of GA in Internal Cross-Validation and External-Validation Images

	Training/Tuning (Cross-Validation)	Testing (External Validation)
Clinical trial	AREDS2	GSK BAM114341
Number of images (participants)	601 (271)	156 (100)
Camera	Heidelberg Spectralis	Heidelberg Spectralis
GA inclusion criteria	Not applicable	GA area from color photographs 1.9–17 mm^2
Mean (SD) area of GA mm^2	6.65 (6.30)	9.79 (5.60)
Subfoveal GA (%)	48%	63%
Multifocal GA (%)	18%	21%

AREDS2 = Age-Related Eye Disease Study 2; GA = geographic atrophy; GSK = GlaxoSmithKline; SD = standard deviation.

multiplying by the known pixel size. Model performance was measured using mean difference between AI vs. human grader measurement. Metrics were generated for each cross-validation data set, and summaries are presented. Similar metrics were also generated for the external validation (testing) data set. Intergrader agreement was also measured as mean difference and 95% confidence interval (CI). Scatter plots and Bland–Altman plots were used to compare the Weak and Strong models. Geographic atrophy segmentation between the Strong model and human grader was compared using Dice coefficient. A Dice coefficient closer to 1 indicates excellent agreement in spatial overlap of segmented pixels between AI and grader.

Results

The AREDS2 training data set included 601 FAF images from 362 eyes (271 participants) distributed into 5 cross-validation data sets (120 each). Of the 271 participants, both eyes were included in 94 (35%) and 1 eye in 177 (65%). Of the 362 eyes (601 images) included 9% with 3 visits, 31% with 2 visits, and 55% with 1 visit only. Of the eyes that met inclusion criteria, 0.9% (13/1501 images) were excluded due to ungradable image quality.

The mean area of GA was 6.65 (standard deviation [SD], 6.30; range, 0.1–36.3) mm^2. The testing (external validation) data set consisted of 156 images (156 eyes, 100 participants) with a mean area of 9.79 (SD, 5.60; range, 0.4–24.3) mm^2. Additional characteristics of the data sets are presented in Table 1.

Most published AI models for GA area measurement have been trained and tested on clinical trial data. These data sets are curated with inclusion/exclusion criteria and usually have an area range of 2.5 to 17.5 mm^2. The AREDS2 data set has a wide range of GA areas because no such criteria were applied specifically for GA. As seen in Table 1, the range of GA was 0.1 to 36.3 mm^2. A subset of AREDS2 data (n = 383) fitting clinical trial inclusion criteria (GA area 2.5–17.5 mm^2) was analyzed to compare performance against published models (Table S2, available at www.ophthalmologyscience.org).

Weakly Labeled Model Results

The Weakly labeled model provided an area output only, whereas the Strongly labeled model provided a segmentation mask from which area was derived. Comparison of the area measurements generated by the Weak and Strong models with the ground truth (grader measurements) for the cross-validation data set is shown in Table 3 and for the external validation data in Table 4. The mean difference between ground truth and AI in AREDS2 cross-validation data set is larger in Weak compared with Strong, for both the AREDS2 cross-validation (0.18 vs. −0.07 mm^2) and GSK testing data (−0.97 vs. −0.24 mm^2).

Figure 2 shows scatter plots, and Figure 3 shows Bland–Altman Plots for comparison of area measurements in both the AREDS2 cross-validation and GSK validation data sets. As seen in the pattern on the Bland–Altman plots

Table 3. Comparison of Performance Metrics between Weakly Labeled and Strongly Labeled AI Models in the Cross-Validation Data Set (AREDS2)

	Weakly Labeled Model (Trained on Area of GA Only: Images without Area of GA Outlined) n = 601	Strongly Labeled Model (Trained on Images with Area of GA Outlined) n = 601	Intergrader Agreement n = 47 (April 4, 2023)
Area with human graders (mm^2), mean (SD)	6.65 (6.30)	6.65 (6.30)	4.91 (4.95)
Area with AI (mm^2), mean (SD)	6.83 (6.29)	6.58 (6.24)	NA
Difference in area between AI and human measurement (mm^2), mean (95% CI)	0.18 (−7.57 to 7.92)	−0.07 (−1.61 to 1.47)	0.36 (−1.03 to 1.75)
R (correlation coefficient)	0.803	0.992	0.990
Dice coefficient	— output is numeric area only	0.885	

AI = artificial intelligence; AREDS2 = Age-Related Eye Disease Study 2; CI = confidence interval; GA = geographic atrophy; SD = standard deviation. Intergrader agreement is also shown for comparison.

Table 4. Comparison of Performance Metrics between Weakly Labeled and Strongly Labeled AI Models in the External Validation Data Set (GSK).

	Weakly Labeled Model (Trained on Area of GA Only) n = 156	Strongly Labeled Model (Trained on Segmentation of GA) n = 156
Area with human graders (mm^2), mean (SD)	9.79 (5.60)	9.79 (5.60)
Area with AI (mm^2), mean (SD)	8.82 (4.61)	9.55 (5.66)
Difference in area between human and AI measurement (mm^2), mean (95% CI)	−0.97 (−4.36 to 2.41)	−0.24 (−4.98 to 4.49)
R (correlation coefficient)	0.926	0.908
Dice coefficient	— Output is numeric area only	0.918

AI = artificial intelligence; CI = confidence interval; GA = geographic atrophy; GSK = GlaxoSmithKline; SD = standard deviation.

for the Weak models, there is a tendency to overcall smaller areas and undercall larger areas in both data sets.

The mean difference between grader and AI was similar to the full AREDS2 data set, with a mean difference of 0.27 (95% CI, −8.3 to 8.84) with the Weak model.

Strongly Labeled Model Results

In image segmentation tasks, the Dice coefficient is commonly used to evaluate how well the segmented regions match the ground truth regions. As such, the Dice coefficient can only be assessed with segmentations in the Strong model and is not available for the Weak model. The Dice coefficient for the Strong model was 0.88 with the AREDS2 data and improved to 0.92 for the GSK data.

The mean difference between grader and AI was −0.11 (95% CI, −1.63 to 1.41) with the Strong model. Although the Dice coefficient was 0.89 for the entire AREDS2 data, it changed to 0.92 when restricted to the clinical trial inclusion cohort. The distribution of the Dice coefficient with the area of GA for the AREDS2 data is shown in Figure 4. The Dice coefficient shows significant decrease for lesions < 2.5 mm^2.

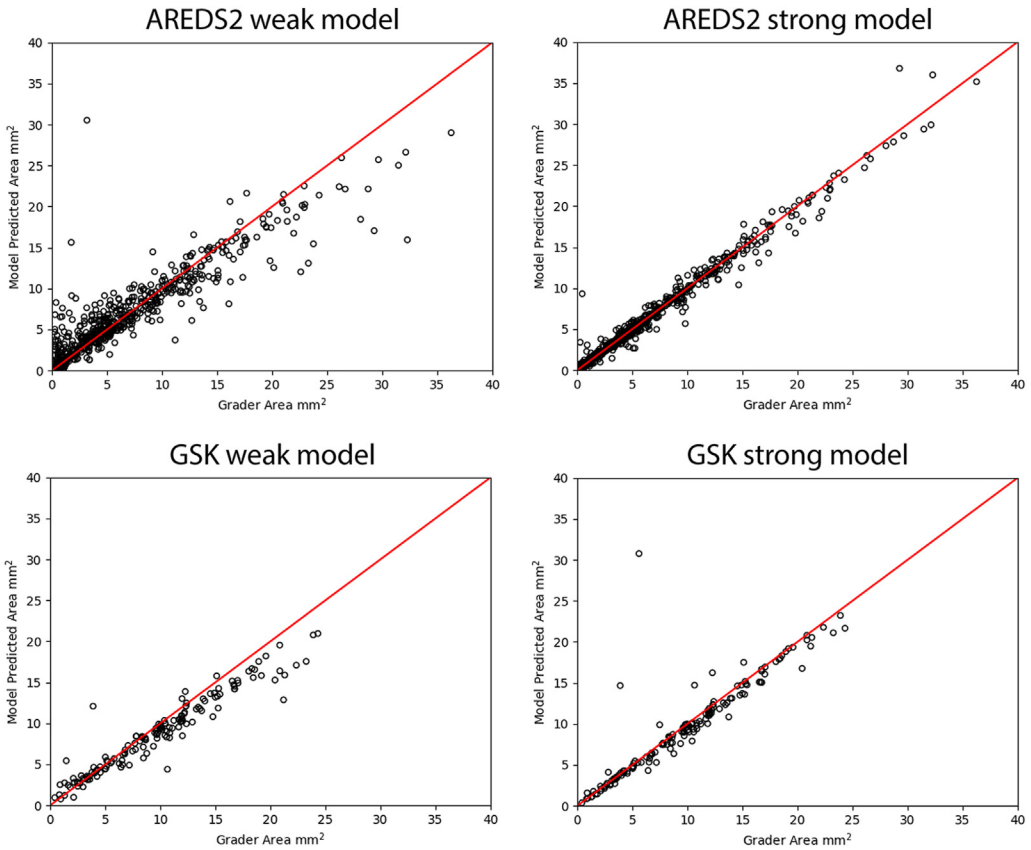

Figure 2. Scatter plots comparing measurement of area of geographic atrophy using artificial intelligence and human graders with the Age-Related Eye Disease Study 2 (AREDS2) weak model (top left), AREDS2 strong model (top right), GlaxoSmithKline (GSK) weak model (bottom left), and GSK strong model (bottom right).

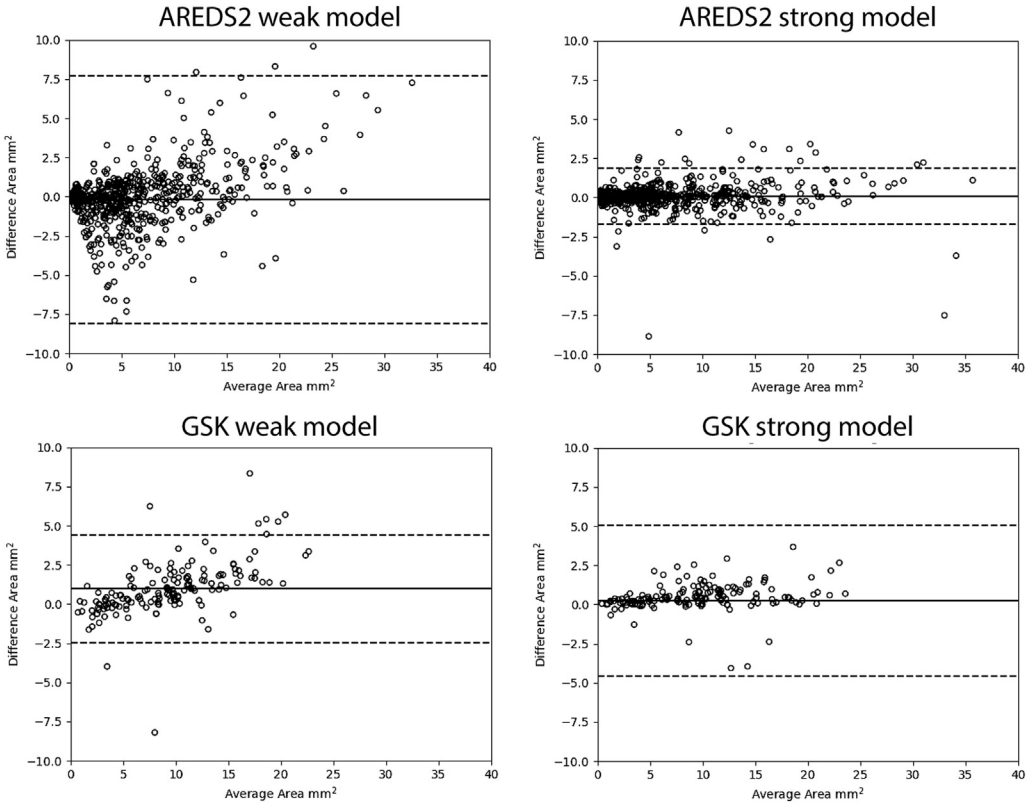

Figure 3. Bland Altman plots comparing measurement of area for geographic atrophy using artificial intelligence and human graders with the Age-Related Eye Disease Study 2 (AREDS2) weak model (top left), AREDS2 strong model (top right), GlaxoSmithKline (GSK) weak model (bottom left), and GSK right model (bottom right).

Discussion

Two AI algorithms were trained for measurement of GA area, a Weakly labeled model using images with measurement of GA and no indication of GA location and a Strongly labeled model using images with GA outlined on the image (segmentation masks). Both models demonstrated promising performance during cross-validation, showing good results based on the mean difference between AI and human measurements and on the Pearson correlation coefficient. In the AREDS2 cross-validation set (n = 601), the mean difference of the Weak model was 0.18 mm^2 (95% CI, −7.57 to 7.92; r = 0.80) compared with −0.07 mm^2 (95% CI, −1.61 to 1.47; r = 0.99) with the Strong model. However, the Strong model outperformed the Weak model, displaying higher and more consistent performance metrics. Although the mean difference is comparable, the scattering of the prediction points and wide confidence limits of the Weak model, as seen in Figures 2 and 3, show the instability of model prediction. The superior performance of the Strong model persisted even when tested on an external data set (−0.24 mm^2; 95% CI, −4.98 to 4.49, r = 0.91), whereas the Weak model had a larger mean difference (−0.97 mm^2;−4.36 to 2.41; r = 0.92). This is not unexpected, considering that the Strong model was trained on images with GA segmentations available, giving it an advantage. On the other hand, the Weak model had no such information and relied solely on numeric labels with area measurements. The Weak model faced the challenge of not only identifying areas of atrophy but also distinguishing normal anatomic structures like the

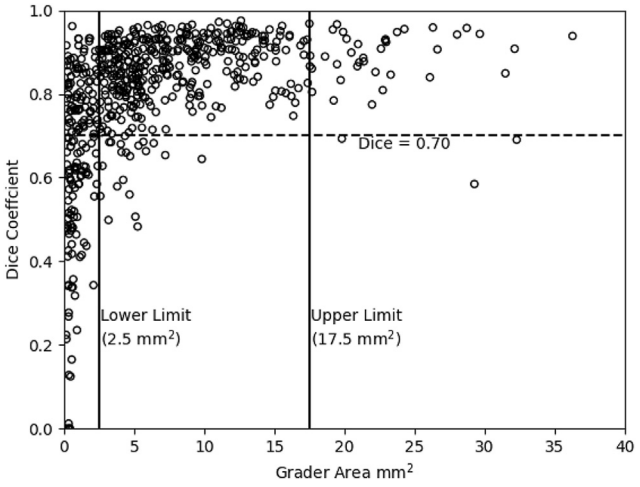

Figure 4. Distribution of Dice coefficient with area of geographic atrophy (GA). The Dice coefficient is lower when GA area is < 2.5 mm^2. Clinical trial enrollment is usually limited to the range of GA 2.5−17.5 mm^2.

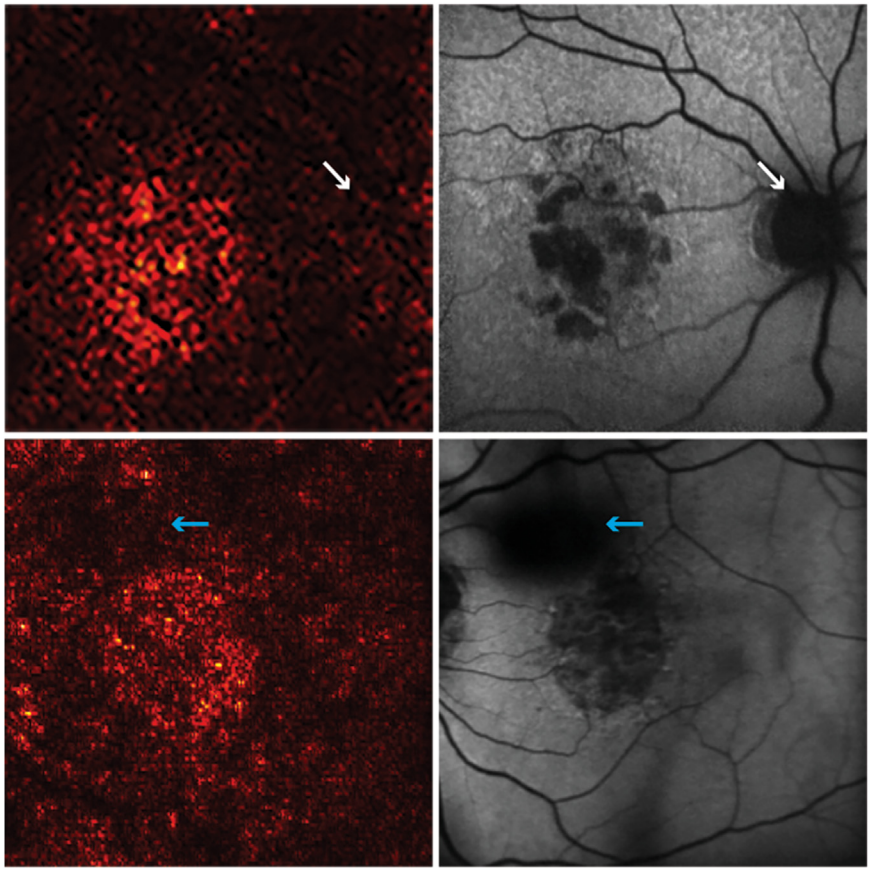

Figure 5. Saliency maps of the weak model predictions showing the pixels measured as geographic atrophy (GA) in hot colors. Shown in the top row, the optic nerve (white arrow), which is also hypoautofluorescent like GA, was not identified. In the bottom row, vitreous floaters (blue arrow) were also not identified as GA.

optic nerve and vessels, which also present as hypoautofluorescent regions, similar to GA.

We generated saliency maps to understand the Weak model predictions and identify the pixels used for measuring GA area as shown in Figure 5. These saliency maps were made by looking at the max of the input gradients during backpropagation and produce a single channel image. Apart from identifying normal anatomy, the Weak model accurately excluded shadows caused by vitreous floaters and did not include these in area measurements. This insight into the prediction of the Weak model demonstrates its capacity to autonomously learn and train on the appearance of GA, showcasing its ability to discern imaging features without the need for explicit annotations.

Although this project used 2 extremes of data labeling providing minimal labels to train 1 model and all information available to train the other, a hybrid of the 2 can be helpful to reduce the burden of data labeling and sample size. Unlike classification labels, which require presence/absence of disease, segmentation models require lesion area to be annotated in calibrated images. Generating such segmentations on many images is both time intensive and expensive. Therefore, it is encouraging that an AI model can be trained using images without segmentation and requires exploration of other segmentation efficient methods of training such as semisupervised learning or hybrid architectures.

Unlike classification models that use sensitivity/specificity to assess performance metrics, segmentation models rely on the Dice coefficient. The Dice coefficient ranges from 0 to 1, with 1 indicating excellent correlation of segmented pixels between the human grader and AI. The Dice coefficient in the AREDS2 data was 0.88 and for the GSK data, was 0.92. In comparison, the Dice coefficient with previously published AI models ranged from 0.89 to 0.98.[12,23] The Dice coefficient relies on the intersection of pixels between the 2 segmentations being compared and the total number of pixels in the ground truth and predictions. When the segmentation area is small, false positives and false negatives have a more substantial effect on the Dice coefficient compared with larger areas. Smaller areas have a smaller number of pixels to match on, and even if a few mismatches lower the Dice coefficient. This is clearly seen in the plot in Figure 4 where the Dice coefficient is lower for lesions < 2.5 mm^2. When the AREDS2 data, which ranges from 0.1 to 36.3 mm^2, were restricted to clinical trial cohort (2.5−17.5 mm^2), the Dice coefficient increased from 0.88 to 0.92. Figure 6 shows examples where

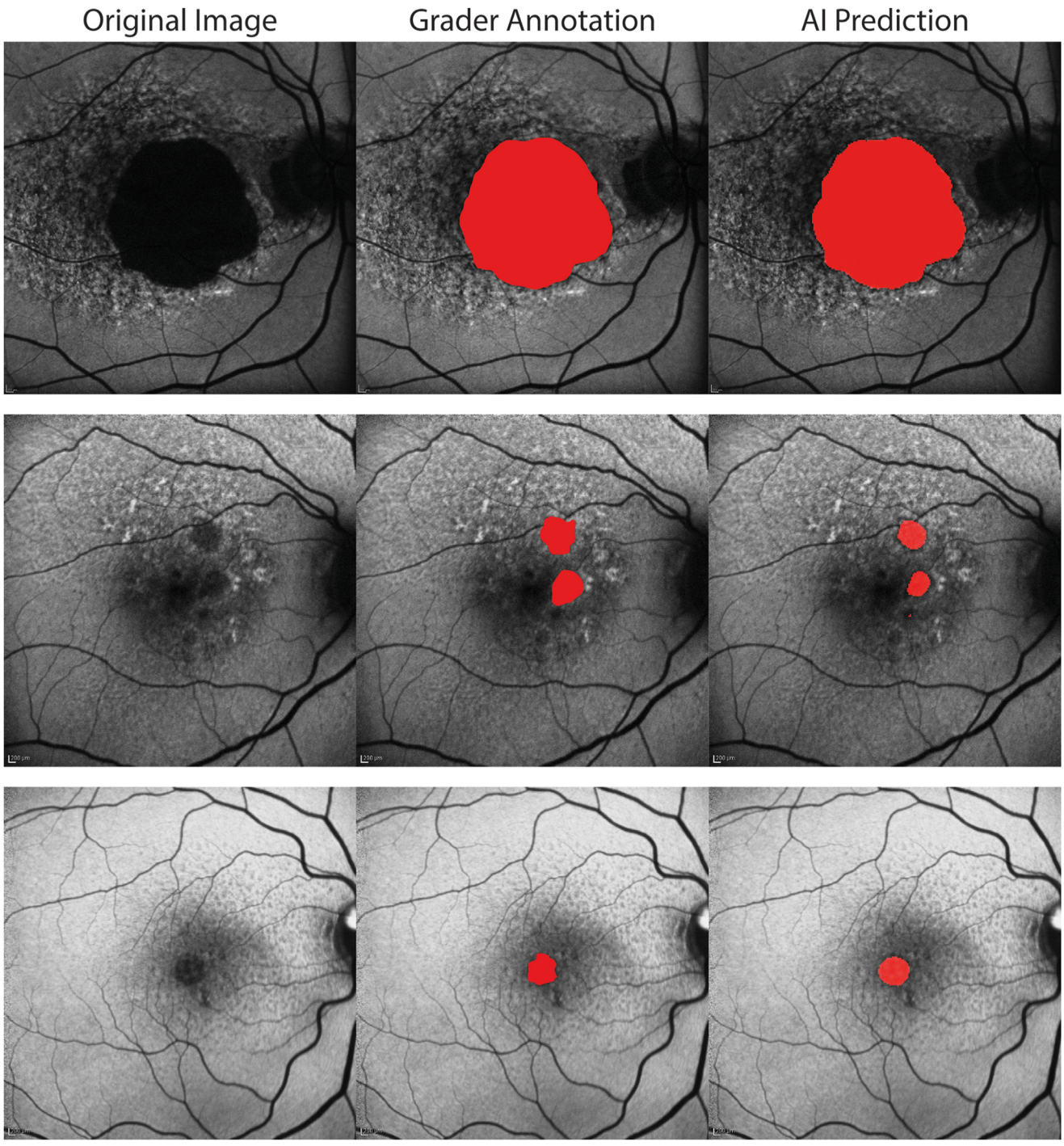

Figure 6. Dice coefficient is a commonly used metric to identify the degree of overlap between grader and artificial intelligence (AI) segmentation masks. The index ranges from 0-1 with 0 being no overlap and 1 being exact match. While all graders and AI segmentation appear similar visually with minimal differences across all 3 examples, the Dice coefficient is 0.98 for the top row, 0.77 for the middle row, and 0.47 for the bottom row. The Dice coefficient tends to penalize small areas more so than larger areas.

the GA area looks similar visually for the ground truth and AI prediction, but the Dice coefficient varies significantly depending on size of GA.

Artificial intelligence models for segmenting and measuring GA have been published using multiple modalities such as color fundus photography, OCT, and FAF.[12,15,23–27] These studies use GA interventional trial data for training purposes, as the images are readily segmented for training purposes. However, GA trials have specific area requirements, with most trials using an inclusion range of 2.5–17.5 mm^2.[28] In addition, trial-specific requirements exclude eyes with peripapillary atrophy or foveal

involvement, which are challenging images to annotate for graders. The AREDS2 study included eyes with intermediate AMD in 1 or both eyes and as such did not have an area cutoff, including both prevalent and incident GA. This is seen from the mean (SD) area of GA in the training cohort at 6.65 (6.30) mm^2, which is smaller than that seen in GA trials at 7.3−9.0 mm^2.[29,30] In addition, about 30% of training data were < 2.5 mm^2, and 7% were > 17.5 mm^2, indicating the diversity of the data set. One of the challenges with real-world implementation of AI models is degradation of model performance, primarily due to selective nature of the training data. Models trained on selected clinical trial data may not perform well in the real world. In contrast, the model in this project was trained on nearly real-world representative images with diverse presentations of GA from AREDS2 data set and tested on a clinical trial selective data. Despite this, a reduction on performance metrics was seen with the mean difference between AI and ground truth increasing from −0.07 (95% CI, −1.61 to 1.47) with AREDS2 to −0.24 (95% CI, −4.98 to 4.49) with GSK, indicating both an increase in mean difference and widening of CIs. These changes highlight the inherent challenges of generalization to external data sets and suggest that some level of accuracy loss should be expected with AI model testing. The findings underscore the importance of conducting repeated external validation to assess model performance on different data sets. By doing so, we can better understand the model's robustness and limitations in real-world scenarios, thereby enhancing its reliability and applicability in clinical settings.

Some of the common issues with measurement of GA using FAF images includes presence of peripapillary atrophy, identification of foveal involvement, variability in hypoautofluorescence, and image quality. The model performance varied in each of these situations with accurate segmentation in some and errors in others. Figures S7 to S9 (available at www.ophthalmologyscience.org) depict cases with segmentations in these scenarios. The minimum diameter of GA was 250 microns (0.05 mm^2). It is interesting that the AI model learns the minimum size from the segmentations despite no specific filters and does not annotate those lesions that fail to meet the threshold, as shown in Figure 1 (bottom image).

This project is an exploration of the ability of AI to get a better understanding of the labeling needs for training algorithms, using GA areas as the use case. The strength of the paper is in a large training data set with a diverse range of GA phenotypes from multiple clinics, meticulous reading center measurements, and use of external data set for testing to ensure model performance. Limitations of the study include lack of proprietary imaging formats to use semi-automated quantification such as Region Finder.[31] Multimodal imaging, including OCT or infrared imaging, was also not available in the training data.

This is the era of therapy for GA, with many clinical trials underway. There is an urgent need for clinical monitoring of GA lesion size, and automated measurement of GA using deep learning is the pathway forward. The findings of this study shed light on the labeling requirements of images, which is an essential step toward training robust AI models. To strike a balance between model performance and labeling resources, a hybrid approach seems promising, which can capitalize on the availability of weak labels to guide the training process while benefiting from strong labels for fine-tuning and refinement. As the field of deep learning continues to advance, further research into innovative labeling techniques and data augmentation approaches may open new avenues for more efficient and reliable AI models. Ultimately, this study serves as a stepping stone toward harnessing the full potential of AI technologies in advancing the management and treatment of GA.

Acknowledgments

This publication is based on research using data from GSK that has been made available through CSDR secured access. GSK has not contributed to or approved, and is not in any way responsible for, the contents of this publication. The authors thank both GSK and CSDR for providing us data and access.

Footnotes and Disclosures

Originally received: September 11, 2023.
Final revision: November 15, 2023.
Accepted: January 19, 2024.
Available online 26 January 2024. Manuscript no. XOPS-D-23-00223R1.

[1] A-EYE Research Unit, Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, Wisconsin.

[2] Wisconsin Reading Center, Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, Wisconsin.

[3] Annexon Biosciences, Brisbane, California.

[4] Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, Maryland.

Emily Y. Chew, MD, the Editor-in-Chief of this journal, was recused from the peer-review process of this article and had no access to information regarding its peer-review.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosure(s):

A.D.: Support − Annexon Biosciences.

D.F.: Employee − Annexon Biosciences.

The other authors have no proprietary or commercial interest in any materials discussed in this article.

The AREDS2 study was supported by the intramural program funds and contracts from the National Eye Institute/National Institutes of Health (NEI/NIH), Department of Health and Human Services, Bethesda, MD. Contract No. HHS-N-260-2005-00007-C. ADB Contract No. N01-EY-5-0007. This work was supported by partial unrestricted funds provided to the A-EYE Research Unit, University of Wisconsin by Annexon Biosciences. This work was also supported in part by an unrestricted grant from Research to Prevent Blindness, Inc. to the UW Madison Department of Ophthalmology and Visual Sciences.

HUMAN SUBJECTS: Human subjects were included in this study. The study was conducted under institutional review board approval at each site and written informed consent was obtained from all study participants. The research adhered to the tenets of the Declaration of Helsinki and complied with the Health Insurance Portability and Accountability Act.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Domalpally, Slater

Data collection: Domalpally, Slater, Linderman, Balaji, Bogost, Pak

Analysis and interpretation: Domalpally, Slater, Voland

Obtained funding: Domalpally

Overall responsibility: Domalpally, Slater, Linderman, Blodi, Channa, Fong, Chew

Abbreviations and Acronyms:
AI = artificial intelligence; **AMD** = age-related macular degeneration; **AREDS2** = Age-Related eye Disease Study 2; **CI** = confidence interval; **GA** = geographic atrophy; **GSK** = GlaxoSmithKline; **FAF** = fundus autofluorescence; **SD** = standard deviation.

Keywords:
Artificial intelligence, Geographic atrophy, Dry AMD, Data labeling.

Correspondence:
Amitha Domalpally, MD, PhD, 301 S Westfield Rd, Suite 200, Madison, WI 53717. E-mail: domalpally@wisc.edu.

References

1. Dow ER, Keenan TDL, Lad EM, et al. From data to deployment: the collaborative community on ophthalmic imaging roadmap for artificial intelligence in age-related macular degeneration. *Ophthalmology*. 2022;129:e43−e59.
2. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations. *Prog Retin Eye Res*. 2019;72:100759. https://doi.org/10.1016/j.preteyeres.2019.04.003.
3. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264−1272. https://doi.org/10.1016/j.ophtha.2018.01.034.
4. Nielsen KB, Lautrup ML, Andersen JKH, et al. Deep learning-based algorithms in screening of diabetic retinopathy: a systematic review of diagnostic performance. *Ophthalmol Retina*. 2019;3:294−304. https://doi.org/10.1016/j.oret.2018.10.014.
5. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295:4−15. https://doi.org/10.1148/radiol.2020192224.
6. Harvey H, Glocker B. A standardized approach for preparing imaging data for machine learning tasks in radiology. In: Ranschaert ER, Morozov S, Algra PR, eds. *Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks*. Springer International Publishing; 2019:61−72.
7. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3:e51−e66. https://doi.org/10.1016/S2589-7500(20)30240-5.
8. Sharma R, Saqib M, Lin CT, Blumenstein M. A survey on object instance segmentation. *SN Comput Sci*. 2022;3:499. https://doi.org/10.1007/s42979-022-01407-3.
9. Schaal KB, Rosenfeld PJ, Gregori G, et al. Anatomic clinical trial endpoints for nonexudative age-related macular degeneration. *Ophthalmology*. 2016;123:1060−1079. https://doi.org/10.1016/j.ophtha.2016.01.034.
10. Biarnés M. Deep learning in geographic atrophy: the best is yet to come. *Lancet Digit Health*. Oct 2021;3:e617−e618. https://doi.org/10.1016/s2589-7500(21)00204-1.
11. Arslan J, Samarasinghe G, Benke KK, et al. Artificial intelligence algorithms for analysis of geographic atrophy: a review and evaluation. *Transl Vis Sci Technol*. 2020;9:57. https://doi.org/10.1167/tvst.9.2.57.
12. Arslan J, Samarasinghe G, Sowmya A, et al. Deep learning applied to automated segmentation of geographic atrophy in fundus autofluorescence images. *Transl Vis Sci Technol*. 2021;10:2. https://doi.org/10.1167/tvst.10.8.2.
13. Miere A, Capuano V, Kessler A, et al. Deep learning-based classification of retinal atrophy using fundus autofluorescence imaging. *Comput Biol Med*. 2021;130:104198. https://doi.org/10.1016/j.compbiomed.2020.104198.
14. Yang Q, Anegondi N, Steffen V, Rabe C, Ferrara D, Gao SS. Multi-modal geographic atrophy lesion growth rate prediction using deep learning. *Invest Ophthalmol Vis Sci*. 2021;62:235−235.
15. Anegondi N, Gao SS, Steffen V, et al. Deep learning to predict geographic atrophy area and growth rate from multimodal imaging. *Ophthalmol Retina*. 2023;7:243−252. https://doi.org/10.1016/j.oret.2022.08.018.
16. Chew EY, Clemons T, SanGiovanni JP, et al. The Age-related Eye Disease Study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). *Ophthalmology*. 2012;119:2282−2289. https://doi.org/10.1016/j.ophtha.2012.05.027.
17. Domalpally A, Danis R, Agron E, et al. Evaluation of geographic atrophy from color photographs and fundus autofluorescence images: Age-Related Eye Disease Study 2 report number 11. *Ophthalmology*. 2016;123:2401−2407. https://doi.org/10.1016/j.ophtha.2016.06.025.
18. Rosenfeld PJ, Berger B, Reichel E, et al. A randomized phase 2 study of an anti-amyloid β monoclonal antibody in geographic atrophy secondary to age-related macular degeneration. *Ophthalmol Retina*. 2018;2:1028−1040. https://doi.org/10.1016/j.oret.2018.03.001.
19. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. *Proc Mach Learn Res*. 2019;97:6105−6114.
20. Lin TY, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. *CVPR*. 2017:936−944.
21. Reza AM. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for real-time image enhancement. *J VLSI Signal Process Syst Signal Image Video Technol*. 2004;38:35−44. https://doi.org/10.1023/B:VLSI.0000028532.53893.82.
22. Alwakid G, Gouda W, Humayun M. Deep learning-based prediction of diabetic retinopathy using CLAHE and ESRGAN for enhancement. *Healthcare (Basel)*. 2023;11. https://doi.org/10.3390/healthcare11060863.
23. Spaide T, Jiang J, Patil J, et al. Geographic atrophy segmentation using multimodal deep learning. *Transl Vis Sci Technol*. 2023;12:10. https://doi.org/10.1167/tvst.12.7.10.
24. Chu Z, Wang L, Zhou X, et al. Automatic geographic atrophy segmentation using optical attenuation in OCT scans with deep learning. *Biomed Opt Express*. 2022;13:1328−1343. https://doi.org/10.1364/boe.449314.
25. Keenan TD, Dharssi S, Peng Y, et al. A deep learning approach for automated detection of geographic atrophy from

color fundus photographs. *Ophthalmology*. 2019;126:1533−1540. https://doi.org/10.1016/j.ophtha.2019.06.005.

26. Liefers B, Colijn JM, González-Gonzalo C, et al. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. *Ophthalmology*. 2020;127:1086−1096. https://doi.org/10.1016/j.ophtha.2020.02.009.
27. Zhang G, Fu DJ, Liefers B, et al. Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study. *Lancet Digit Health*. 2021;3:e665−e675. https://doi.org/10.1016/S2589-7500(21)00134-5.
28. Sivaprasad S, Chandra S, Kwon J, et al. Perspectives from clinical trials: is geographic atrophy one disease? *Eye (Lond)*. 2023;37:402−407. https://doi.org/10.1038/s41433-022-02115-1.
29. Jaffe GJ, Westby K, Csaky KG, et al. C5 Inhibitor avacincaptad pegol for geographic atrophy due to age-related macular degeneration: a randomized pivotal phase 2/3 trial. *Ophthalmology*. 2021;128:576−586. https://doi.org/10.1016/j.ophtha.2020.08.027.
30. Liao DS, Grossi FV, El Mehdi D, et al. Complement C3 inhibitor pegcetacoplan for geographic atrophy secondary to age-related macular degeneration: a randomized phase 2 trial. *Ophthalmology*. 2020;127:186−195. https://doi.org/10.1016/j.ophtha.2019.07.011.
31. Schmitz-Valckenberg S, Brinkmann CK, Alten F, et al. Semiautomated image processing methods for identification and quantification of geographic atrophy in age-related macular degeneration. *Invest Ophthalmol Vis Sci*. 2011;52:7640−7646. https://doi.org/10.1167/iovs.11-7457.