224-C0061
ARVO 2023

Data Labeling for Artificial Intelligence (AI) Algorithms for Measurement of Geographic Atrophy

Rohit Balaji[1], Robert Slater[2], Jacob Bogost[1], Gelique Ayala[1], Jeong W. Pak[1], Rick Voland[1], Barbara A. Blodi[1], Roomasa Channa[2], Donald Fong[3], Amitha Domalpally[1,2]

[1]WISCONSIN READING CENTER
[2]A-EYE Unit
[3]ANNEXON BIOSCIENCES
Department of Ophthalmology and Visual Sciences
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH
Research to Prevent Blindness

Background and Purpose

- AI models have impressive ability to segment geographic atrophy (GA) from fundus autofluorescence (FAF) images[1]
- Training AI models for accurate segmentation requires laborious pixel-level annotation of a large training dataset of FAF images
- We sought to understand the training requirements for AI algorithms to accurately segment and measure GA from FAF images

Methods

- Heidelberg FAF images from the Age-Related Eye Disease Study 2 were utilized[2]
- Training dataset: 512 FAF images AREDS2
- Testing dataset: 140 FAF images AREDS2

- GA was segmented on FAF images using planimetry and areas measured in mm^2 by trained and certified human graders
- Two models were used (Figure 1):
  - A **STRONG LABEL MODEL** trained using images and annotations with GA areas annotated
  - A **WEAK LABEL MODEL** trained with images and numerical area measurements of GA. No annotations were used

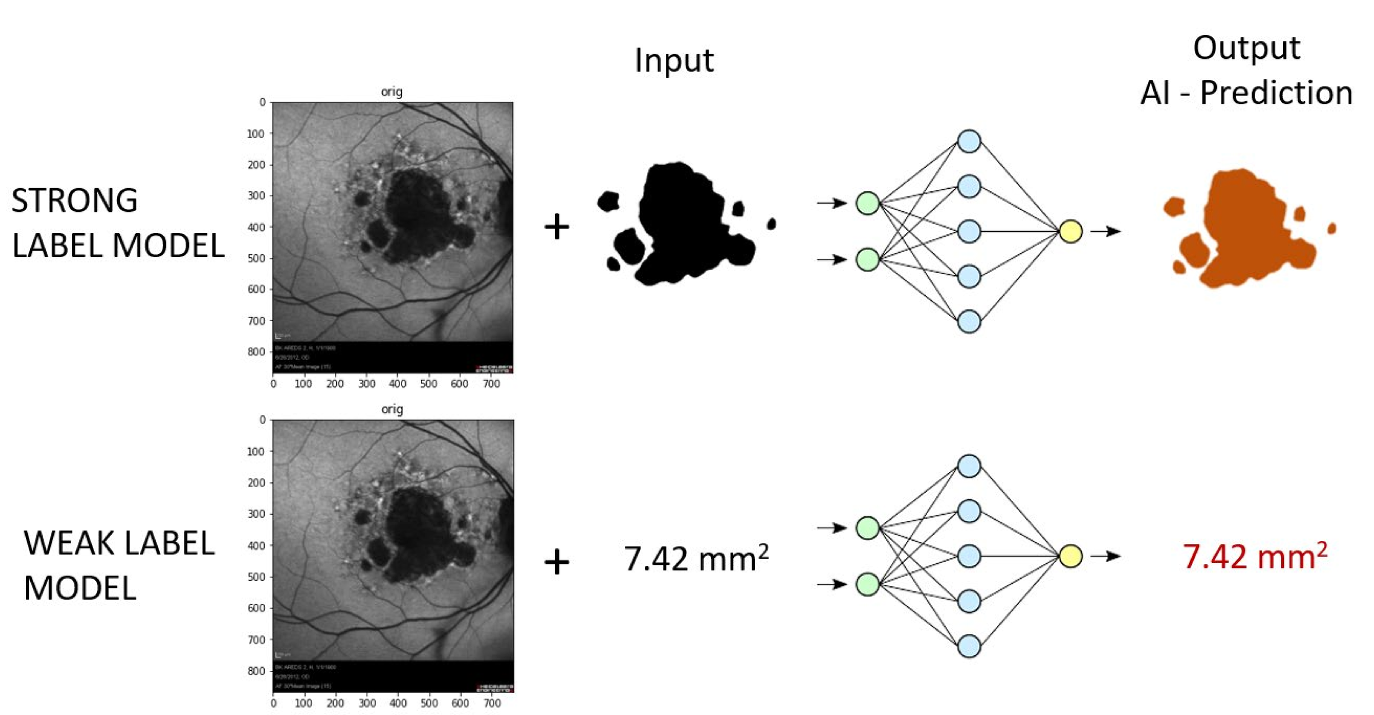

Figure 1. *Model development for GA area measurement*

Disclosures

Commercial interest disclosures: NONE for RB, RS, JB, GA, JP, RV, BB, RC, AD
DF is an employee of Annexon Biosciences
Partial unrestricted funds provided to the University of Wisconsin by Annexon Biosciences

Table 1. Strong vs Weak label model performance metrics

	Strong Label Model (image + annotation)		Weak Label Model (image + measurement)		Intergrader agreement
	All images	Clinical trial subset (area >2.5-17.5 mm^2)	All images	Clinical trial subset (area >2.5-17.5 mm^2)	
Image Number	140	89	140	89	47
Mean area (SD) (mm^2)	6.02 (5.31)	7.50 (3.58)	6.15 (5.17)	7.52 (3.57)	4.91 (4.95)
Mean difference in area Grader-AI (mm^2) (95% CI)	0.05 (-1.18, 1.28)	-0.03 (-1.25, 1.19)	0.18 (-2.86, 3.21)	-0.01 (-2.36, 2.35)	0.36 (-1.03, 1.75)
Intra-Class Correlation	0.993	0.986	0.958	0.948	0.988
Dice Coefficient	0.76	0.84	---	---	0.995

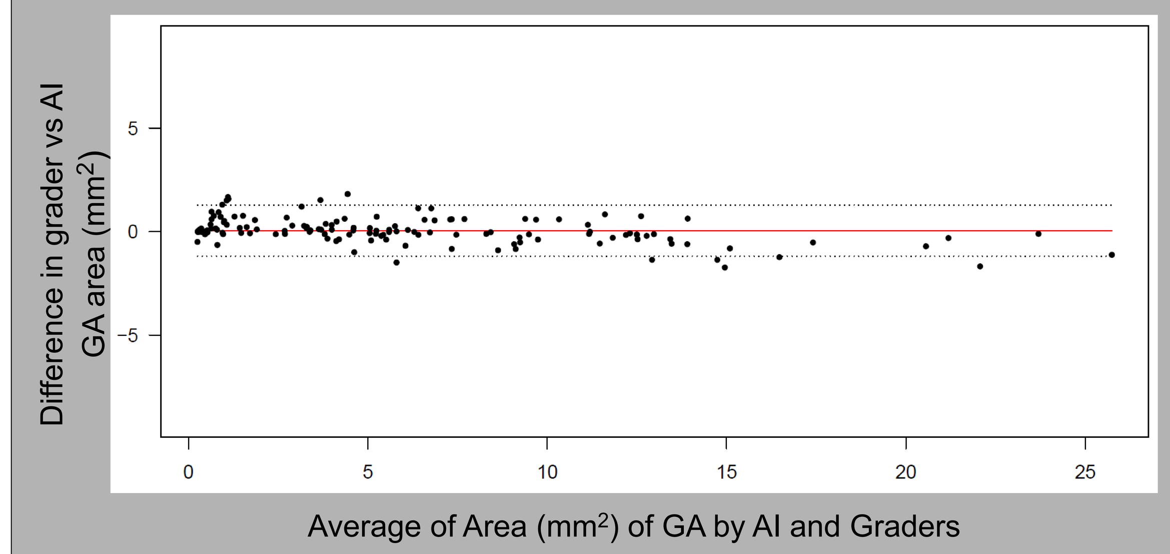

Figure 2. *Bland-Altman plot for all strong label data*

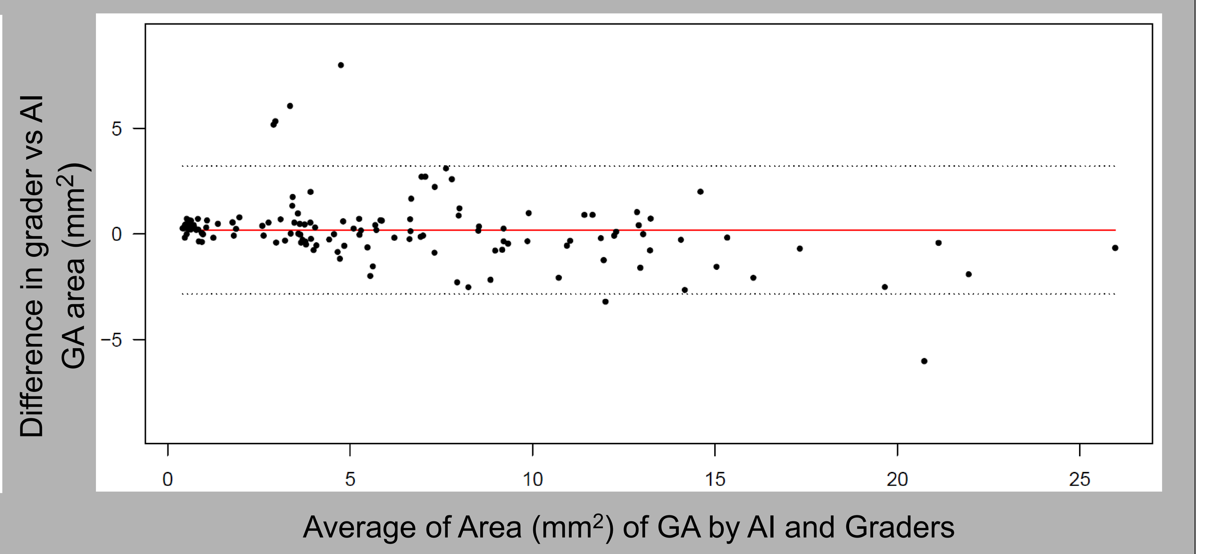

Figure 3. *Bland-Altman plot for all weak label data*

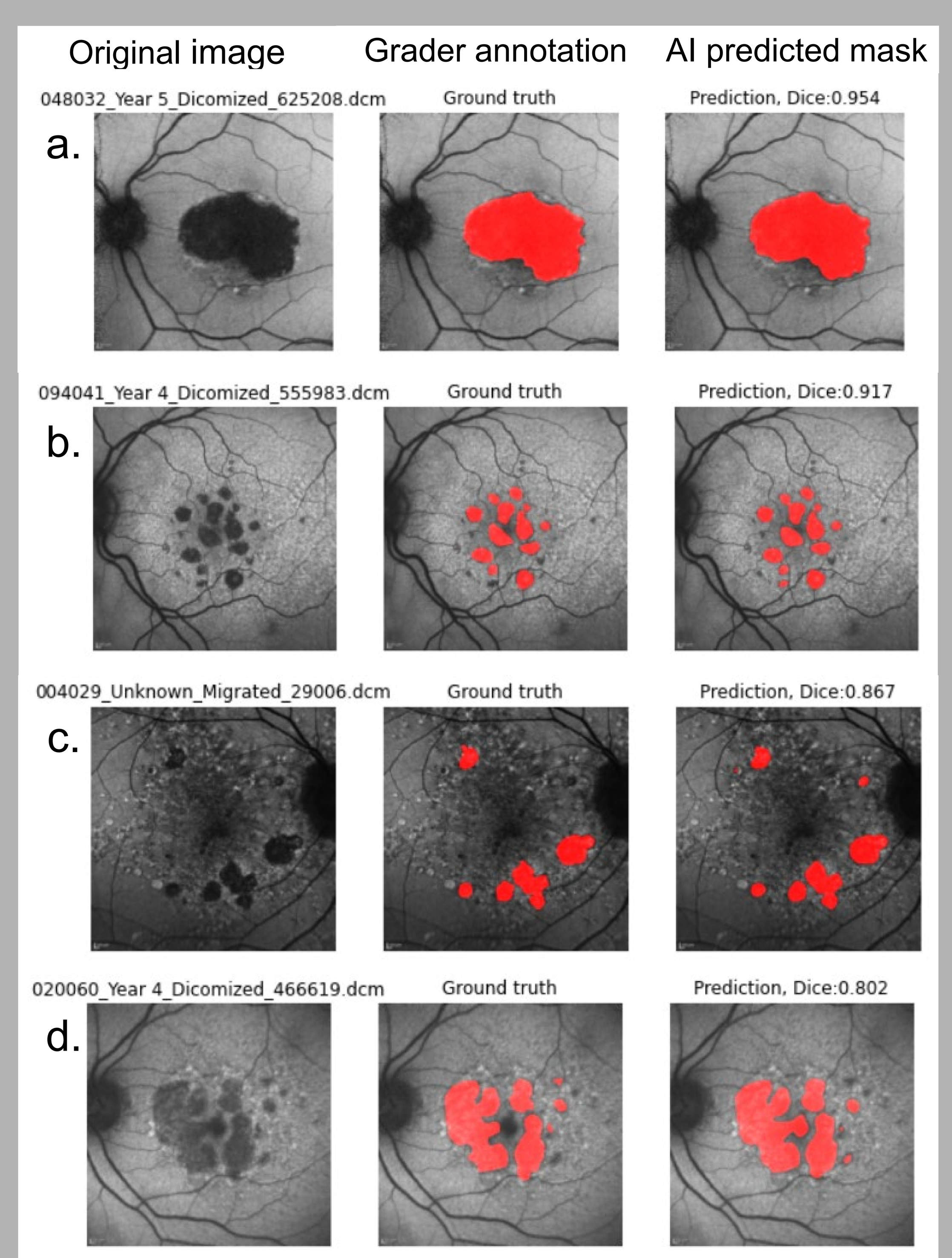

Figure 4. Examples of strong label model development*. The original image (left) is annotated by the grader (center) to train the AI model. The AI model then produces a prediction with a mask (right).* ***a.*** *Example of a unifocal lesion.* ***b.*** *Example of multifocal lesions.* ***c.*** *Example of lesions with complex background autofluorescence.* ***d.*** *Example of a complex annotation where foveal center was erroneously annotated by AI.*

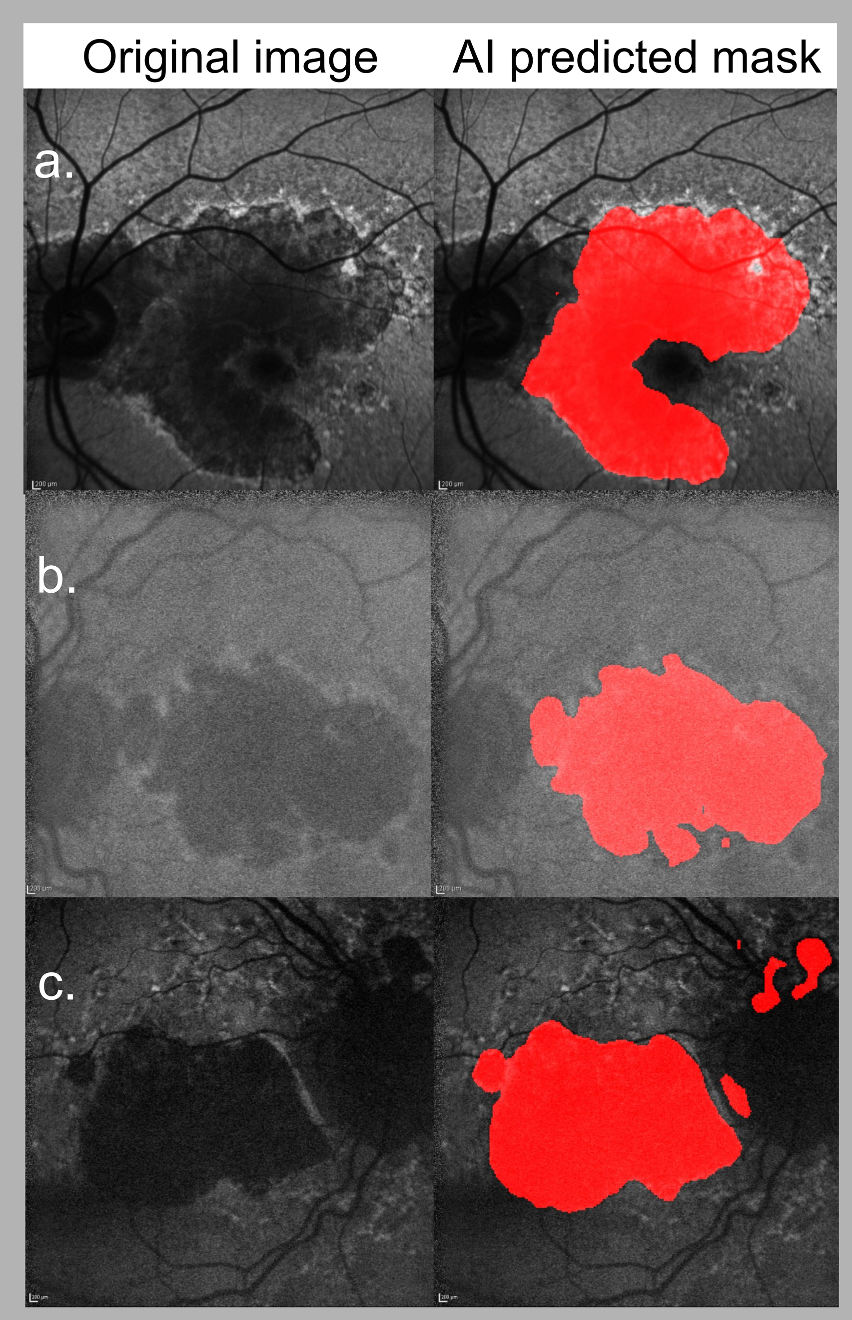

Figure 5. Examples of complex GA annotations using the strong label model*. The original image is on the left and the AI predicted mask is on the right.* ***a.*** *Complex lesion with GA and peripapillary atrophy (PPA) merged. The AI correctly omitted annotating the (PPA).* ***b.*** *Example of a poor quality FAF image. The AI has again annotated only the appropriate GA areas.* ***c.*** *Example of complex FAF image in which the AI erroneously picked up some PPA.*

Results

- GA characteristics in the dataset (512 eyes) included
  - Subfoveal GA (51%)
  - Junctional zone pattern (24%)
  - Background autofluorescence (64%)
  - Multifocal (26%)
- Comparison of model parameters when trained with strong labels and weak labels is shown in Table 1

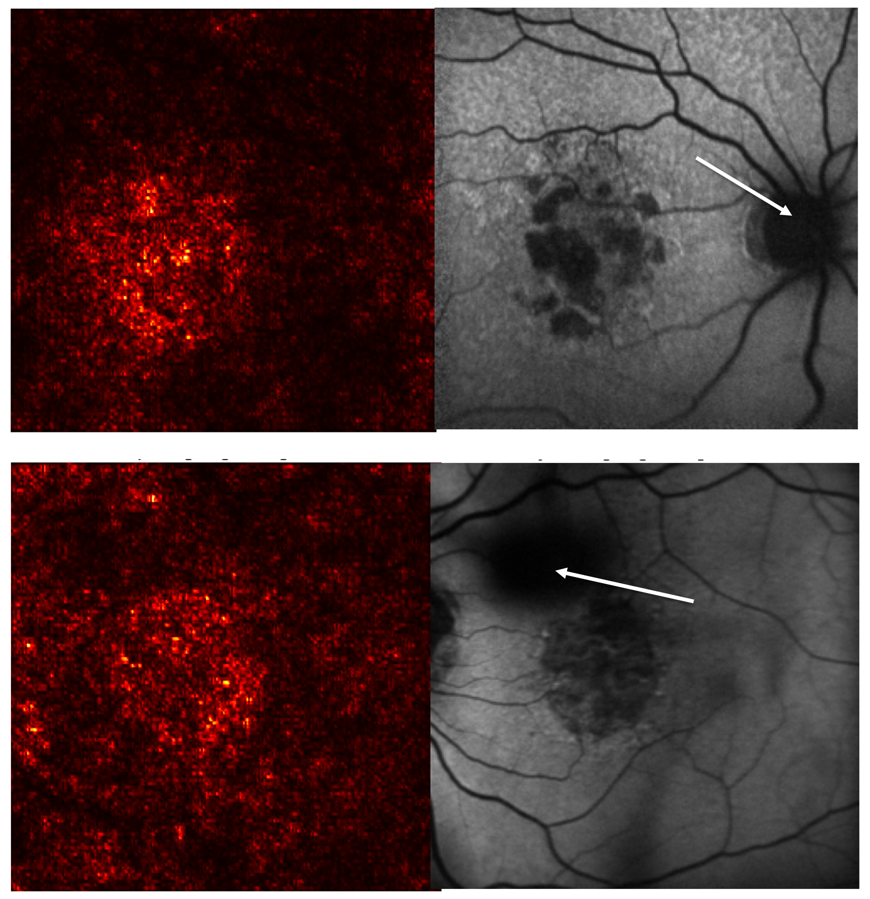

Figure 6. Saliency maps were used to understand regions of AI-predicted GA using the weak label model: *There are no annotations produced by the weak label model. Non-GA features including the optic nerve (arrow above) and vitreous floaters (arrow below) are omitted in the AI prediction.*

Conclusions

- AI models demonstrate good accuracy for identifying and measuring areas of GA on FAF images
- The weak label model provides a high level of accuracy, but the strong label model is more accurate
- Laborious human grader annotations may not be necessary to train AI models to segment GA
- A model that combines features of both the weak and strong label models may be more beneficial

References

1. Arslan, J., Samarasinghe, G., Benke, K. K., Sowmya, A., Wu, Z., Guymer, R. H., & Baird, P. N. (2020). Artificial Intelligence Algorithms for Analysis of Geographic Atrophy: A Review and Evaluation. *Translational vision science & technology*, *9*(2), 57. https://doi.org/10.1167/tvst.9.2.57

2. Domalpally, A., Danis, R., Agrón, E., Blodi, B., Clemons, T., Chew, E., & Age-Related Eye Disease Study 2 Research Group (2016). Evaluation of Geographic Atrophy from Color Photographs and Fundus Autofluorescence Images: Age-Related Eye Disease Study 2 Report Number 11. *Ophthalmology*, *123*(11), 2401–2407. https://doi.org/10.1016/j.ophtha.2016.06.025