AMERICAN ACADEMY OF OPHTHALMOLOGY®

Commentary

Machine versus Human Vertical Cup-Disc-Ratio Grading: Artificial Intelligence Screening for Glaucoma

Sophia Y. Wang, MD - *Palo Alto, California*

Glaucoma remains a leading cause of irreversible blindness globally,[1] with over 3 quarters of individuals in the United States unaware of their condition,[2] resulting in late presentation for eye care after irreversible vision loss has occurred. Despite this public health burden, national committees such as the US Preventive Services Task Force have not recommended population-wide screening.[3] One barrier stems from concerns over screening performance and the unsustainable burden of false-positive referrals. In addition, ophthalmic screening often requires specialized equipment for fundus photography and specialized personnel to take these photographs and interpret them. Thus, screening a general population may be difficult to scale.

Accurate machine learning (ML) grading of fundus photos for glaucoma could alleviate one potential barrier to glaucoma screening. In this issue, Taylor et al[4] (p. 1447) report on the development of an open-access ML model to grade vertical cup-disc-ratio (VCDR) from fundus photos (ML-VCDR), validated in a large cohort of 6304 participants in the European Prospective Investigation into Cancer (EPIC)-Norfolk Eye Study. The study's clinical gold standard for glaucoma at the patient level was established via direct participant examinations by consultant ophthalmologists, with additional diagnostic testing as indicated. The primary result is striking: the ML-VCDR model outperformed human graders of VCDR in predicting glaucoma status. In the left eyes, ML-VCDR achieved an area under the receiver operating characteristic curve of 88%, compared with human graders' 79% area under the receiver operating characteristic curve.

These results are intriguing because both humans and the model were tasked with identifying the VCDR; given that, why should 1 VCDR be more predictive of glaucoma status than the other? Obviously, the humans and the model disagree on VCDR rating; one naturally wonders which ratings more closely reflect the anatomical truth. Grading VCDR suffers from high interobserver variability[5] and may be prone to fatigue and other subjective effects. Human graders in this cohort estimated only 2 eyes to have a VCDR >0.8, whereas the ML model estimated 87 eyes above this threshold. Perhaps the ML model could be capturing a "truer" physical anatomy, or it could be standardizing an otherwise noisy metric in such a way that it becomes mathematically more predictive of specialist-diagnosed glaucoma than human-graded VCDR.

A strength of the study was the EPIC-Norfolk population, which, although not strictly a population-based screening cohort, is nonetheless a more realistic screening population than the highly curated case-control populations in which many previous glaucoma artificial intelligence diagnostic models have been evaluated. Evaluating any screening algorithm requires understanding how disease prevalence dictates clinical performance. For a given algorithm's sensitivity and specificity, the higher the disease prevalence, the higher the positive predictive value (PPV); conversely, the lower the disease prevalence, the lower the PPV. Although many prior ML models for glaucoma screening have reported excellent diagnostic performance,[6] when they are validated on highly curated case-control data sets that unrealistically balance healthy cases with glaucoma cases close to 50:50, the PPV of such algorithms would be inflated compared with the expected performance in a realistic population, where glaucoma prevalence is much lower. The result for deployment of such algorithms in a typical low-prevalence setting would be an unsustainable deluge of false positives because of poor PPV.

In this study, the overall participant-level prevalence of glaucoma suspect or glaucoma was 11.0%. The authors noted that at 95% specificity, the ML-VCDR model's right-eye PPV is a respectable 50.7%, although sensitivity drops to 50.5%. Although a 50% sensitivity may seem modest, it would capture a substantial portion of what would otherwise be completely undiagnosed glaucoma. However, several key limitations remain. The EPIC-Norfolk cohort is 99.4% White. The glaucoma phenotype in this population may not reflect challenges common to other populations, such as those with a high prevalence of comorbid high myopia. Thus, would-be users of ophthalmic artificial intelligence should always validate a model in their local population to ensure that its performance is generalizable to the particular ways disease presents in their own cohort. External validation across geographically and ethnically diverse populations can help to ensure acceptable model performance when deployed in a variety of settings.

Footnotes and Disclosures

Disclosure(s):

The author has completed and submitted the ICMJE disclosure form.

The author made the following disclosures:

© 2026 American Academy of Ophthalmology, Inc
Published by Elsevier Inc.
All rights are reserved, including those for text and data mining, AI training, and similar technologies.

http://doi.org/10.1016/j.ophtha.2026.09.008
ISSN 0161-6420/26

S.Y.W.: Grants or contracts from any entity − NIH NEI (K23EY03263501 and R01EY036893), Alcon Young Investigator Award.

Correspondence:
Sophia Ying Wang, MD, Stanford University School of Medicine. Stanford University School of Medicine, Department of Ophthalmology, Stanford Byers Eye Institute, 2370 Watson Ct, Palo Alto, CA 94303
E-mail: sywang@stanford.edu.

References

1. Wang Z, Xue CC, Li Y, et al. Global glaucoma prevalence: burden and projection to 2060. *Am J Ophthalmol*. 2026;283:324−335.
2. Shaikh Y, Yu F, Coleman AL. Burden of undetected and untreated glaucoma in the United States. *Am J Ophthalmol*. 2014;158:1121−1129.e1.
3. Chou R, Selph S, Blazina I, et al. Screening for glaucoma in adults: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA*. 2022;327:1998−2012.
4. Taylor TRP, Khasentino J, Luben RN, et al. A comparison of machine learning and human graders for glaucoma diagnosis from fundus images for population screening. *Ophthalmology*. 2026;133:1447−1457.
5. Abrams LS, Scott IU, Spaeth GL, et al. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. *Ophthalmology*. 1994;101:1662−1667.
6. Wu J-H, Nishida T, Weinreb RN, Lin J-W. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. *Am J Ophthalmol*. 2022;237:1−12.