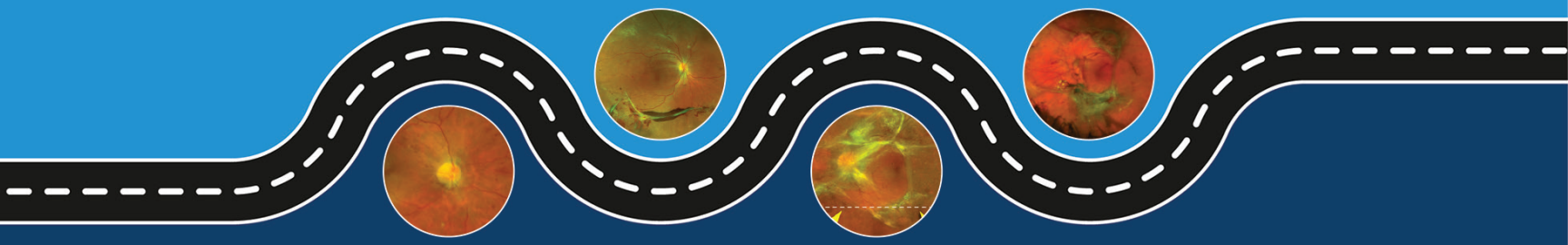
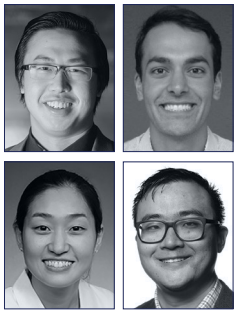


THE PATH TO VALIDATING AI SCREENING SYSTEMS

A close look at seven systems for diabetic retinopathy screening reveals significant performance differences.

BY RANDY Y. LU; ANAND RAJESH; CECILIA S. LEE, MD, MS; AND AARON Y. LEE, MD, MSCI



Artificial intelligence (AI) screening algorithms are a promising solution to the growing global diabetic retinopathy (DR) screening burden. Many AI algorithms have been shown to perform at or above the level of human experts on DR classification tasks when evaluated on their internal datasets.[1-4] However, these algorithms may underperform on larger, external validation datasets due to a lack of generalizability, overfitting, or underspecificity.[5-7]

Discrepancies between internal and external validation performance can be concerning, given that many of these algorithms are already commercially available; two algorithms (IDx-DR [Digital Diagnostics] and EyeArt [Eyenuk]) have FDA approval.[3,8]

Here, we share our findings after validating seven commercially available DR screening algorithms on a large-scale dataset collected from two Veterans Affairs (VA) hospitals.

HEAD-TO-HEAD VALIDATION

Our multicenter, noninterventional, head-to-head device validation study included seven commercially available AI-based DR screening algorithms from five participating companies.[9] The validation dataset consisted of 311,604 fundus photographs from 23,724 veterans from the Seattle VA Puget Sound Health Care System (HCS) and the Atlanta VA HCS. In addition, a randomly sampled subset of 7,379 images were regraded using double-masked arbitration.

Non-referable DR was defined as no DR (International Clinical Diabetic Retinopathy Severity Scale [ICDR] of 0), and referable DR was defined as the presence of any DR (ICDR 1-4) by the VA standard.[9]

The results showed substantial differences in overall performance between the algorithms. Using the original VA teleretinal grades as the reference standard, algorithm sensitivity ranged from 50.98% to 85.90%, specificity from 60.42% to 83.69%, negative predictive value from 82.72% to 93.69%, and positive predictive value from 36.46% to 50.80%. Overall, the algorithms achieved higher negative predictive values using the Atlanta data set (90.71% to 98.05%) compared with the Seattle data set (77.57% to 90.66%). In contrast, the positive predictive values ranged from 24.80% to 39.07% in the Atlanta data set, which was lower than the Seattle data set (42.04% to 62.92%).[9]

When the arbitrated grades from the 7,379 regraded images were used as the new reference standard, most algorithms performed worse in terms of both sensitivity and specificity compared with the VA teleretinal graders. While the VA teleretinal graders achieved an overall

AT A GLANCE

- Discrepancies between internal and external validation performance of diabetic retinopathy screening algorithms can be concerning.
- In the author's validation study of seven commercially available diabetic retinopathy screening algorithms, sensitivity ranged from 50.98% to 85.90% and specificity ranged from 60.42% to 83.69%.
- The goal of further study is to ensure that automated screening algorithms can maintain adequate performance standards regardless of variables such as race, image quality, and coexisting disease.

sensitivity of 82.22% and specificity of 84.36%, only one algorithm approached a similar level of performance in both sensitivity (80.47%) and specificity (81.28%). Two algorithms achieved higher sensitivities than the VA teleretinal graders (92.71% and 92.71%) but were also statistically less specific. Only one algorithm achieved higher specificity (90.00%), at the cost of a statistically lower sensitivity than the VA teleretinal graders.[9]

REFERRABLE THRESHOLDS

When we performed a sensitivity analysis for different thresholds of disease severity, we found that, although most algorithms had higher sensitivities when the threshold was raised to moderate DR or worse, none of the algorithms were better than human graders in identifying referable disease when analyzed by DR severity. When the referrable threshold was raised to severe DR or worse, the sensitivity of one algorithm only reached 74.42%. Thus, regional- and site-specific differences in the thresholds for referrable DR may be an important factor that can affect downstream model performance.[9]

OTHER VALIDATION WORKS

In a separate validation study, Tufail et al assessed the performance of three DR screening algorithms using 102,856 images from 20,258 patients.[6] The authors found that two algorithms achieved acceptable sensitivity for referable retinopathy (85% and 94%); however, both algorithms had low specificity, contributing to false positive rates of 47.7% and 80%, respectively. The third algorithm also classified all episodes as diseased or ungradable, resulting in a 100% sensitivity rate but also a 100% false positive rate.

Meanwhile, smaller studies validating individual DR screening algorithms have reported stronger results, with one study reporting a sensitivity of 100% and specificity of 82% when validated on 2,680 patients undergoing DR screening in Valencia, Spain.[10]

Another DR screening algorithm that was validated on 4,504 fundus images from five urban centers in Zambia showed clinically acceptable performance in detecting referable DR with sensitivity and specificity of 92.25% and 89.04%, respectively.[11] Still, single-algorithm studies are difficult to interpret, and multi-algorithm studies are better for direct head-to-head comparison of different algorithms.

CLINICAL IMPLICATIONS

The performance of many AI-based DR screening algorithms may differ significantly when being evaluated using large-scale external validation datasets. The discrepancy in performance highlights the issue of algorithm generalizability, or how well a model performs for all subsets of unseen data. Generalizability concerns often arise when the training and validation datasets are sufficiently different, which can be attributed to variations in image collection protocols, image quality, device manufacturers, or demographic factors.

Our study demonstrated site-specific differences in algorithm performance, and we hypothesized that differences in imaging protocols, disease prevalence, and patient demographics may be notable contributing factors.[9]

As more automated screening algorithms are introduced, they should be validated on datasets that are representative of the population in which they are deployed. While our study was strengthened by the head-to-head comparison and a large real-world dataset, the VA population may not reflect the general population.

Furthermore, a meta-analysis found that many AI-based DR screening algorithms often used the same datasets for training and external validation.[12-16] While these datasets are typically graded by trained ophthalmologists, many exclude ungradable images, are limited in size, lack extensive demographic information, and may not capture the full underlying distribution of disease.[5,12,17,18] This highlights the importance of conducting additional validation in diverse populations, as well as the need for prospective, interventional trials after clinical integration and regulatory approval. The goal is to ensure that automated screening algorithms can maintain adequate performance standards regardless of variables such as race, image quality, and coexisting disease.

HURDLES TO OVERCOME

Automated DR screening systems have shown potential in helping to alleviate the DR screening burden. However, many challenges still exist, and we must exercise caution when interpreting the performance of an algorithm that is trained and validated solely on internally curated datasets.

Large-scale external validation studies, although challenging to conduct, serve as the best indicators for an algorithm's true performance. Future work should aim to develop automated AI-based screening algorithms that are flexible, efficient, and able to demonstrate robust performance on the populations in which they are deployed. ■

CAN'T-MISS DISCLAIMER

Although automated diabetic retinopathy (DR) screening systems can greatly expand access, they do not replace routine eye examinations. Current commercial DR screening systems are approved only to diagnose referrable DR using specific devices and protocols. Sole reliance on automated screening systems may miss additional important features, such as undiagnosed glaucoma, macular degeneration, retinal detachments, or choroidal melanomas. DR screening systems should supplement traditional eye examinations to expand screening access, while also upholding a high standard of care.

Financial Support: NIH/NEI K23EY029246 (A.Y.L); Latham Vision Innovation Award, the C. Dan and Irene Hunter Endowed Professorship, and an unrestricted grant from Research to Prevent Blindness.

1. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351-357.
2. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211-2223.
3. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
4. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7):962-969.
5. Wu JH, Liu TYA, Hsu WT, Ho JHC, Lee CC. Performance and limitation of machine learning algorithms for diabetic retinopathy screening: meta-analysis. *J Med Internet Res*. 2021;23(7):e23863.
6. Tufail A, Rudisill C, Egan C, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology*. 2017;124(3):343-351.
7. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. *JAMA Netw Open*. 2018;1(5):e182665.
8. Eyenuk announces FDA clearance for EyeArt autonomous AI system for diabetic retinopathy screening [press release]. August 5, 2020. Accessed June 13, 2022. www.eyenuk.com/us-en/articles/diabetic-retinopathy/eyenuk-announces-eyeart-fda-clearance
9. Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care*. 2021;44(5):1168-1175.
10. Shah A, Clarida W, Amelon R, et al. Validation of automated screening for referable diabetic retinopathy with an autonomous diagnostic artificial intelligence system in a Spanish population. *J Diabetes Sci Technol*. 2021;15(3):655-663.
11. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1(1):e35-e44.
12. Nagpal D, Panda SN, Malarvel M, Pattanaik PA, Zubair Khan M. A review of diabetic retinopathy: datasets, approaches, evaluation metrics and future trends [Preprint published online June 22, 2021]. *J King Saud University - Computer and Information Sciences*.
13. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal Stereol*. 2014;33(3):231.
14. Kamble VV, Kokate RD. Automated diabetic retinopathy detection using radial basis function. *Procedia Comput Sci*. 2020;167:799-808.
15. Porwal P, Pachade S, Kamble R, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a database for diabetic retinopathy screening research. *Brown Univ Dig Addict Theory Appl*. 2018;3(3):25.
16. Diabetic retinopathy detection. Accessed June 21, 2022. www.kaggle.com/c/diabetic-retinopathy-detection
17. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 2020;34(3):451-460.
18. Lakshminarayanan V, Kheradfallah H, Sarkar A, Jothi Balaji J. Automated detection and diagnosis of diabetic retinopathy: a comprehensive survey. *J Imaging Sci Technol*. 2021;7(9):165.

AARON Y. LEE, MD, MSCI
- Associate Professor, Department of Ophthalmology, C. Dan and Irene Hunter Endowed Professorship, Roger and Angie Karalis Johnson Retina Center, University of Washington, Seattle
- leeay@uw.edu
- Financial disclosure: Research Grants/Support (Carl Zeiss Meditec, Novartis, Regeneron, Santen, US FDA); Consultant (Genentech/Roche, Johnson & Johnson)

CECILIA S. LEE, MD, MS
- Associate Professor, Klorfine Family Endowed Chair, Director, Clinical Research, Department of Ophthalmology, Roger and Angie Karalis Johnson Retina Center, University of Washington, Seattle
- Financial disclosure: None

RANDY Y. LU
- MD candidate, Roger and Angie Karalis Johnson Retina Center, University of Washington, Seattle
- Financial disclosure: None

ANAND RAJESH
- MD candidate, Roger and Angie Karalis Johnson Retina Center, University of Washington, Seattle
- Financial disclosure: None