AMERICAN ACADEMY OF OPHTHALMOLOGY®

Comparison of Deep Learning and Clinician Performance for Detecting Referable Glaucoma from Fundus Photographs in a Safety Net Population

Van Nguyen, MD,[1] Sreenidhi Iyengar,[2] Haroon Rasheed, MD,[1] Galo Apolo,[1] Zhiwei Li,[2] Aniket Kumar,[2] Hong Nguyen,[2] Austin Bohner, MD,[1] Kyle Bolo, MD,[1] Rahul Dhodapkar, MD,[1] Jiun Do, MD, PhD,[3] Andrew T. Duong, MD,[1] Jeffrey Gluckstein, MD,[1] Kendra Hong, MD,[1] Lucas L. Humayun,[4] Alanna James, MD,[1] Junhui Lee, MD,[1] Kent Nguyen, OD,[1] Brandon J. Wong, MD,[1,3] Jose-Luis Ambite, PhD,[2] Carl Kesselman, PhD,[2] Lauren P. Daskivich, MD,[1,5] Michael Pazzani, PhD,[2] Benjamin Y. Xu, MD, PhD[1]

Purpose: Develop and test a deep learning (DL) algorithm for detecting referable glaucoma.

Design: Retrospective cohort study.

Participants: A total of 6116 patients from the Los Angeles County (LAC) Department of Health Services (DHS) were included.

Methods: Fundus photographs and patient-level labels of referable glaucoma (cup-to-disc ratio ≥0.6) provided by 21 certified optometrists. A DL algorithm based on the Visual Geometry Group-19 architecture was trained using patient-level labels generalized to images from both eyes. Area under the receiver operating curve (AUROC), sensitivity, and specificity were calculated to assess algorithm performance using an independent test set that was also graded by 13 clinicians with 0 to 10 years of experience. Algorithm performance was tested using reference labels provided by either LAC DHS optometrists or an expert panel of 3 glaucoma specialists.

Main Outcome Measures: Area under the receiver operating curve, sensitivity, and specificity.

Results: The DL algorithm was trained using 12 998 images from 5616 patients (2086 referable glaucoma, 3530 nonglaucoma). In this data set, the mean age was 56.8 ± 10.5 years with 54.8% women, 68.2% Latinos, 8.9% Blacks, 6.0% Asians, and 2.7% Whites. One thousand images from 500 patients (250 referable glaucoma, 250 nonglaucoma) with similar demographics ($P \geq 0.57$) were used to test the algorithm. Algorithm performance matched or exceeded that of all independent clinician graders in detecting patient-level referable glaucoma based on LAC DHS optometrist (AUROC = 0.92) or expert panel (AUROC = 0.93) reference labels. Clinician grader sensitivity (range, 0.33–0.99) and specificity (range, 0.68–0.98) ranged widely and did not correlate with years of experience ($P \geq 0.49$). Algorithm performance (AUROC = 0.93) also matched or exceeded the sensitivity (range, 0.78–1.00) and specificity (range, 0.32–0.87) of 6 certified LAC DHS optometrists in the subsets of the test data set they graded.

Conclusions: A DL algorithm for detecting referable glaucoma trained using patient-level data provided by certified LAC DHS optometrists approximates or exceeds performance by ophthalmologists and optometrists, who exhibit variable sensitivity and specificity unrelated to experience level. Implementation of this algorithm in screening workflows could help reallocate resources and provide more reproducible and timely glaucoma care.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100751*

Supplemental material available at www.ophthalmologyscience.org.

Glaucoma is the leading cause of irreversible blindness worldwide, with prevalence growing from 64.3 million in 2013 to 111.8 million in 2040.[1,2] In the United States, glaucoma is projected to affect 7.3 million people by 2050 with the majority being racial minorities.[2] The rising burden of glaucoma in the United States is exacerbated by a critical shortage of eye care providers; the total supply of ophthalmologists is projected to decrease by 12%, although demand for eye care services is projected to increase by 24% by 2035.[3] Underserved racial minorities and individuals living in nonmetro areas who already experience difficulty accessing care will likely be disproportionately affected, thereby exacerbating ongoing disparities in glaucoma care.[4] For example, Blacks and Hispanics in the United States have significantly lower utilization of eye care services and higher risk of

https://doi.org/10.1016/j.xops.2025.100751
ISSN 2666-9145/25

glaucoma-related blindness and need for glaucoma surgery compared with non-Hispanic Whites.[4–11] Therefore, there is an urgent need to develop and implement novel interventions that address the impending eye care crisis by ensuring timely and equitable detection of at-risk individuals.

The Los Angeles County Department of Health Services (LAC DHS), the second largest municipal health system in the United States, has operated a teleretinal screening program for patients with diabetes since 2013.[12] Although the program primarily focuses on detecting diabetic retinopathy, it also screens for other ocular conditions, including cataracts and referable glaucoma. The referable glaucoma component of the program has been effective: between 2016 and 2018, 817 patients were referred for glaucoma evaluations, 534 (65.4%) patients successfully completed in-person evaluations, and 131 (24.5%) patients were diagnosed with glaucoma by LAC DHS clinicians.[13] Despite its success, the program is hindered by key workflow limitations. Reliance on certified optometrists to manually grade fundus photographs contributes to referral delays and takes time away from direct patient care. Manual grading by >20 LAC DHS optometrists also potentially introduces intergrader variability in disease detection.[14] Therefore, it is critical to consider alternative approaches for standardizing and streamlining referrals to ensure reproducibility and equity of care.

Artificial intelligence (AI), specifically deep learning (DL), is an emerging technology in health care that could enhance the reproducibility and efficiency of glaucoma screening, thereby enabling earlier detection and intervention. In this study, we develop a DL algorithm for detecting referable glaucoma from optic nerve photographs of patients in the LAC DHS teleretinal screening program. We also perform a rigorous validation of the algorithm by comparing its performance to a panel of 13 clinicians, including 4 fellowship-trained glaucoma specialists. This type of algorithm, once rigorously validated against the standard-of-care human grading, could be implemented to address the critical need for reproducible and scalable solutions in glaucoma screening, especially among vulnerable, safety net populations.

Methods

This study was approved by the University of Southern California Institutional Review Board. The study adhered to the tenets of the Declaration of Helsinki and complied with the Health Insurance Portability and Accountability Act. Informed consent was not obtained because this was a retrospective study utilizing deidentified patient data. All patient information used in the study was anonymized, ensuring that no identifiable personal information was accessed or used. As a result, the study posed minimal risk to participants, and the need for informed consent was waived by the Institutional Review Board in accordance with applicable ethical guidelines.

Data Source

The LAC DHS health system administers a primary care-based teleretinal screening program across 17 hospital- and community-based sites throughout LAC.[12] The program serves around 1750 diabetics per month. The LAC DHS patients participating in the program receive dilation and fundus photography by trained staff (medical assistants or licensed vocational nurses) using the Topcon NW400 and NW8 (Topcon Corporation) and Canon CR-2 AF Digital (Canon U.S.A. Inc) cameras. These photographs are evaluated primarily for diabetic retinopathy and secondarily for other ocular conditions such as referable glaucoma, defined as cup-to-disc ratio (CDR) ≥0.6, by 21 certified LAC DHS optometrists. Disease diagnoses, including for referable glaucoma, are recorded on the patient level. All patients ≥18 years of age with ≥1 fundus photograph taken between January 4, 2016 and December 2, 2022 were eligible for analysis.

A segmentation-free approach to detecting referable glaucoma was selected given: (1) generally superior diagnostic performance compared with segmentation-reliant approaches; and (2) lack of access to CDR and segmentation data in the LAC DHS data set.[15] Fundus photographs centered on the optic nerve from all patients diagnosed with referable glaucoma and a comparable number of patients diagnosed as nonglaucoma were retrieved for purposes of the AI algorithm development. All photographs underwent manual review. The photographs of low quality (e.g., blurry, underexposed or overexposed, or media opacities partially obscuring the optic nerve) were included to ensure the generalizability of algorithms to clinical screening environments. However, photographs were excluded if they could not be graded for glaucoma (e.g., media opacities totally obscuring the optic nerve, so out of focus that the optic nerve could not be delineated, or if the optic nerve was not in the field of view).

Fundus photographs were cropped and centered around the optic nerve head for analysis in a 2-step process that was programmed in Python. First, the program cropped each raw fundus image to the image region by removing any black or extraneous regions. Then, the program scanned the image using a sliding window approach that attempted to match the cropped image to the pattern of an optic disc. Once a potential match was found, the section of the image was saved as the final cropped image. If the program failed to locate or confirm an optic disc after multiple attempts, the entire uncropped image was saved. All images were manually reviewed to ensure cropping and centration were effective. Images where the optic disc was present but difficult to visualize due to occlusion or exposure issues were retained in the data set to represent clinical scenarios. Images without an optic disc were excluded. Images were resized to 224 by 224 pixels to reduce hardware demands during training. Images were preprocessed by normalizing red, green, blue channels and augmented through random rotation, translation, and perturbations to balance and contrast.

Algorithm Development and Validation

The LAC DHS data set was divided into development (80%) and test (20%) data sets. The development data set was further split into training (75%) and validation (25%) data sets. Some patients with multiple teleretinal screening visits were represented multiple times in the training and validation data sets, although reference labels by LAC DHS optometrists were unique for each visit. The test data set was used to derive a sample of 1000 test images from the latest observation of 500 patients (2 images per patient) with no overlap of patients with the training or validation data sets.

Twenty-one certified LAC DHS optometrists analyzed the photographs of both eyes; if ≥1 eye was referable for glaucoma, it was generalized to both eyes to create patient-level labels of referable glaucoma. These patient-level labels were used to train DL algorithms for detecting referable glaucoma at the eye level. A convolutional neural network was developed based on the Visual

Geometry Group-19 (VGG-19) architecture using the training and validation data sets labeled in this manner. The VGG-19 architecture was chosen due to its efficiency with image-based data while providing similar performance to other architectures, including InceptionV3, MobileNetV3, EfficientNetV2, and ResNet50V2. The average pooling layer was replaced by an adaptive pooling layer where bin size is proportional to input image size, enabling the convolutional neural network to be applied to input images of arbitrary sizes. Softmax regression was used to calculate the multinomial probability of the 2 classes with a cross-entropy loss used during training. All layers of the convolutional neural network were fine-tuned using backpropagation; optimization was performed using stochastic gradient descent with warm restarts. The code for data preprocessing and model training is linked: https://github.com/informatics-isi-edu/eye-ai-exec/blob/main/notebooks/VGG19/VGG19_Diagnosis_Train.ipynb. The hyperparameters for training are available in Table S1 (available at www.ophthalmologyscience.org).

The DL algorithm was tested using the 1000-image test data set, which was also graded by 13 clinicians (1 optometrist [K.N.], 7 general ophthalmologists [A.B., R.D., A.T.D., K.H., A.J., J.L., H.R.] 1 neuro-ophthalmologist [J.G.], and 4 glaucoma specialists [J.D., V.N., B.J.W., B.Y.X.]) with 0 to 15 full years of clinical experience as medical school or optometry school graduation. Before grading, each clinician was provided with a reference data set comprised of 20 images per CDR between 0.2 and 0.9 in 0.1-unit increments to establish a baseline degree of standardization among human graders. As one objective of the study was to assess the effect of clinician experience, the size of the sample data set was limited to avoid strongly biasing clinicians with less experience.

Three sets of reference labels (2 patient-level sets and 1 eye-level set) of the independent test set were used to assess algorithm performance. One patient-level set of reference labels was provided by the 21 certified LAC DHS optometrists who originally graded the photos in the test data set. Two sets of reference labels, 1 patient level and 1 eye level, were provided by 3 fellowship-trained glaucoma specialists (V.N., B.J.W., and B.Y.X.) who were among the 13 clinician graders. The majority diagnosis of referable glaucoma or nonglaucoma by the 3 glaucoma specialists (at least 2 of 3) determined the eye-level reference label for each individual image. These eye-level reference labels were combined to generate patient-level reference labels; a patient was positive for referable glaucoma if $\geq$1 eye was labeled as such.

Data Analysis

Demographic characteristics between the training and validation and test sets were compared using a 2-tailed Student *t* test or a chi-square test. The study cohort was stratified by glaucoma status based on LAC DHS optometrist labels to compare demographic and clinical characteristics. Continuous measures were summarized by means and standard deviations, and categorical measures were summarized by proportions and percentages. The area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), calibration curve, accuracy, precision, sensitivity, and specificity were calculated to assess algorithm performance compared with the sensitivity and specificity of individual clinician graders using all 3 sets of reference labels. Optimal algorithm probability thresholds for referable glaucoma detection were determined through Youden's index, which maximizes the sum of sensitivity and specificity. A subanalysis comparing LAC DHS optometrist and algorithm performance was performed for 6 LAC DHS optometrists who graded $\geq$70 test set images using the expert panel reference labels. A linear regression was performed to assess the association between grader sensitivity or specificity and years of clinical experience. Statistical tests were considered statistically significant when the P value < 0.05. Statistical analyses were performed using Python's SciPy statistics library.

Results

A total of 13 098 images were retrieved, and 12 998 images were included in the analysis after excluding 100 cropped images (0.76%) without visible optic nerves. The training data set had 8996 images from 4212 patients, the validation data set had 3002 images from 1404 patients, and the test data set had 1000 images from 500 patients. The 5616 patients (2086 referable glaucoma, 3530 nonglaucoma) in the training and validation data sets had a mean age of 56.8 $\pm$ 10.5 years, and there were 54.8% (N = 3091) women, 68.2% (N = 3826) Latino, 8.9% (N = 501) Black, 6.0% (N = 338) Asian, 2.7% (N = 153) White, and 14.2% (N = 798) Other or not specified race. The 500 patients (250 referable glaucoma, 250 nonglaucoma) in the test data set had a mean age of 57.3 $\pm$ 10.3 years, and there were 52.4%

Table 2. Baseline Demographics Stratified by Training and Validation or Test Data Set

Parameter	Training and Validation	Test	*P* Value
Age	56.8 ± 10.5	57.3 ± 10.3	0.30
Sex			0.57
Female	55.0% (N = 3091)	52.4% (N = 262)	
Male	42.8% (N = 2401)	44.8% (N = 224)	
Race			0.78
Latino	68.1% (N = 3826)	69.2% (N = 346)	
Black	8.9% (N = 501)	8.6% (N = 43)	
White	2.7% (N = 153)	2.6% (N = 13)	
Asian	6.0% (N = 338)	5.2% (N = 26)	
Other or not specified	14.2% (N = 798)	15.0% (N = 75)	
Glaucoma status			<0.001
Referable	37.1% (N = 2086)	50.0% (N = 250)	
Nonreferable	62.9% (N = 3530)	50.0% (N = 250)	

Statistical significance tested by 2-tailed Student *t* test or chi-square test.

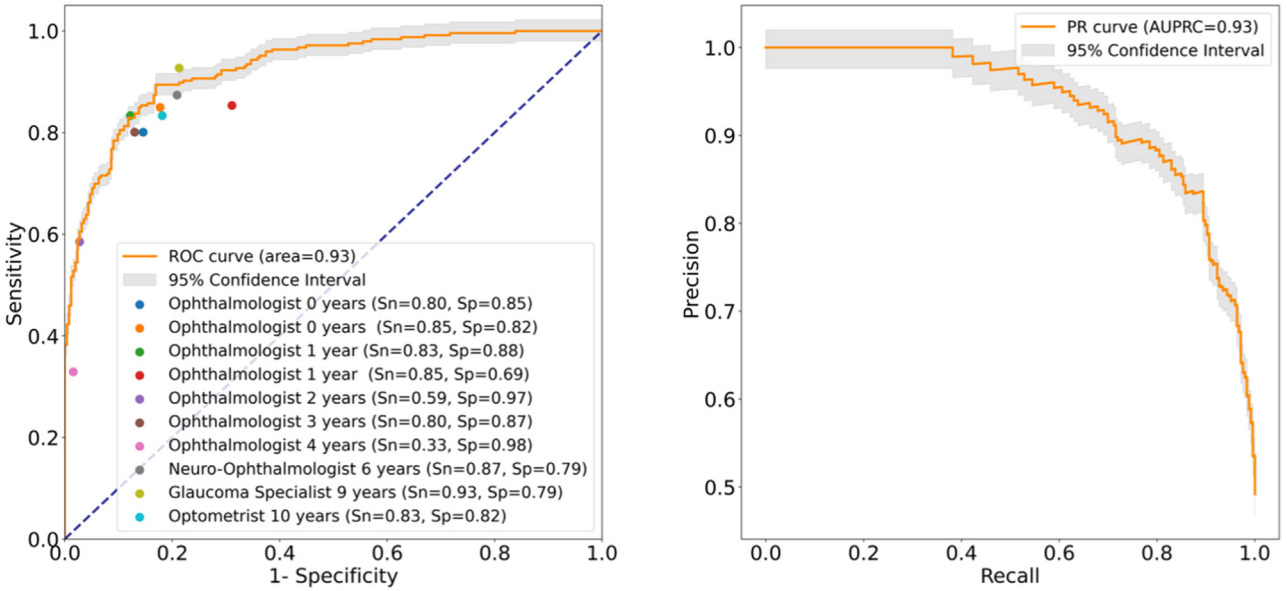

Figure 1. Patient-level algorithm and independent clinician performance with full years of experience (left) and precision–recall curve (right) when using patient-level expert panel reference labels. AUPRC = area under the precision–recall curve; PR = precision–recall; ROC = receiver operating characteristic; Sn = sensitivity; Sp = specificity.

(N = 262) women, 69.2% (N = 346) Latino, 8.6% (N = 43) Black, 5.2% (N = 26) Asian, 2.6% (N = 13) White, and 15.0% (N = 75) Other or not specified race. There was no difference in age ($P = 0.295$), race ($P = 0.781$), or sex ($P = 0.569$) between patients in the training and validation and test data sets (Table 2).

Algorithm performance for detecting referable glaucoma on the patient level based on expert panel labels of the test data set had an AUROC of 0.93 (95% confidence interval [CI], 0.91–0.95) and AUPRC of 0.93 (95% CI, 0.908–0.952), with optimal sensitivity of 0.89 and specificity of 0.83 (probability threshold = 0.50) (Fig 1). The calibration curve for this model is shown in Figure S2 (available at www.ophthalmologyscience.org). Individual graders had a sensitivity ranging from 0.33 to 0.99 and a specificity ranging from 0.68 to 0.98, including a sensitivity of 0.98 and specificity of 0.79 by a fourth glaucoma specialist (Fig 1). There was no association between full years of clinical experience and grader sensitivity ($P = 0.491$) or specificity ($P = 0.56$) (Fig 3).

Algorithm performance for detecting referable glaucoma on the patient level based on LAC DHS optometrist labels of the test data set had an AUROC of 0.92 (95% CI, 0.90–0.94) and AUPRC of 0.92 (95% CI, 0.899–0.948), with optimal sensitivity of 0.86 and specificity of 0.83 (probability threshold = 0.50) (Fig 4). Individual graders, including a fourth glaucoma specialist, had a sensitivity ranging from 0.32 to 0.91 and a specificity ranging from 0.61 to 0.98 (Fig 4).

Algorithm performance on the eye level based on expert panel labels of the test data set had an AUROC of 0.92 (95% CI, 0.90–0.93) and AUPRC of 0.88 (95% CI, 0.857–0.908), with optimal sensitivity of 0.88 and specificity of 0.82 (probability threshold = 0.49) (Fig S5, available at www.ophthalmologyscience.org). Individual graders had a sensitivity ranging from 0.28 to 0.99 and a specificity ranging from 0.74 to 0.99, including a sensitivity of 0.90 and specificity of 0.82 by a fourth glaucoma specialist (Fig S5). A summary of the classification metrics can be found in Table 3.

In the subanalysis of the 6 most frequent LAC DHS optometrist graders (N = 70–150 images), the DL algorithm (AUROC of 0.93) approximated or exceeded optometrist sensitivity (range, 0.78–1.0) and specificity (range, 0.32–0.87) for all 6 graders (Fig 6).

Discussion

In this study, we developed a DL algorithm for detecting referable glaucoma from fundus photographs of LAC DHS teleretinal screening patients that matched or exceeded performance by clinicians with a range of clinical expertise. The algorithm, trained on patient-level labels provided by 21 certified LAC DHS optometrists, demonstrated robust performance across 3 sets of reference labels. In addition, LAC DHS optometrists and independent ophthalmologists exhibited wide ranges of sensitivity and specificity that raise concerns about variability associated with human grading of fundus photographs. Our findings highlight potential benefits of adopting AI-based strategies to improve the reproducibility, timeliness, and scalability of glaucoma care, which could facilitate earlier glaucoma detection and intervention.

Although several DL algorithms for detecting referable or manifest glaucoma from fundus photographs have previously been reported, none have been as rigorously validated against the standard-of-care human grading as in the

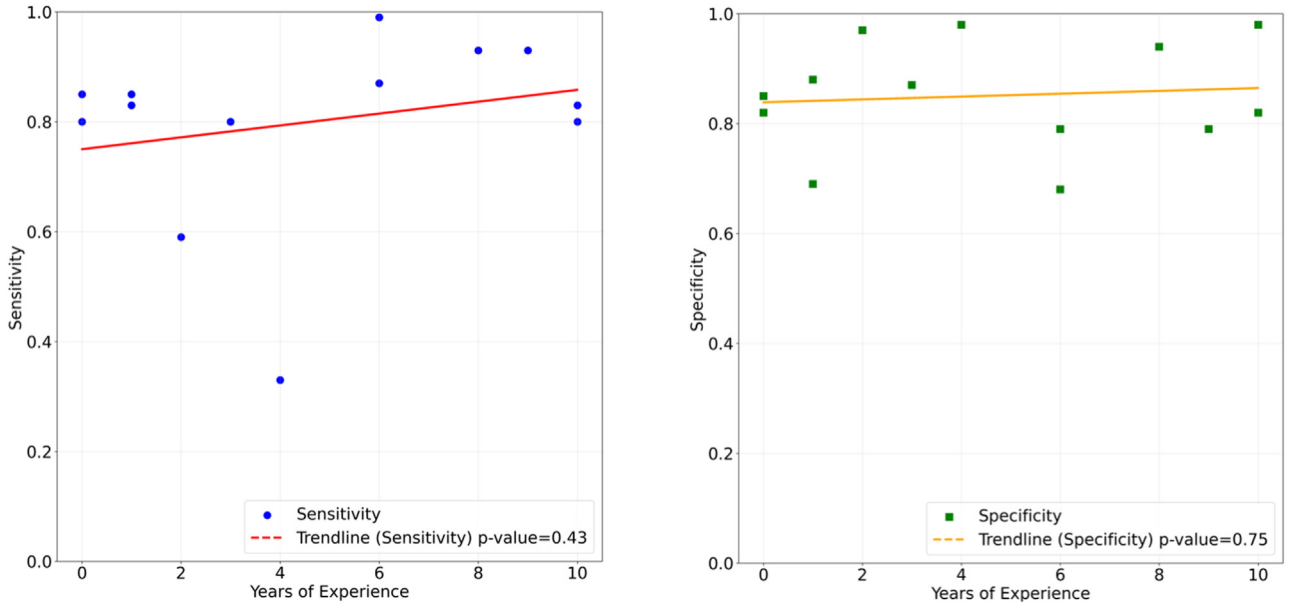

Figure 3. Correlation between sensitivity (left) or specificity (right) in detecting referable glaucoma and full years of clinical experience among independent clinician graders.

current study.[16–19] Our algorithm's performance (AUROC >0.9 and AUPRC >0.9) falls within the general range of performance demonstrated by these previous algorithms.[15–20] However, it is difficult to evaluate algorithm performance based solely on comparisons with previous algorithms due to interstudy differences in disease definitions, study populations, and AI methodology. Therefore, we focused on producing a higher level of evidence to instill confidence in LAC DHS clinicians, patients, and health care administrators, especially given our plan to implement the algorithm in a clinical teleretinal screening environment. In a rigorous comparison with human graders, our algorithm demonstrated excellent performance, matching or exceeding the sensitivity and specificity of 13 clinicians with a range of clinical experience. In a separate subanalysis, the algorithm also matched or outperformed 6 certified LAC DHS optometrists. This robust performance compared with current standard-of-care human grading provides evidence

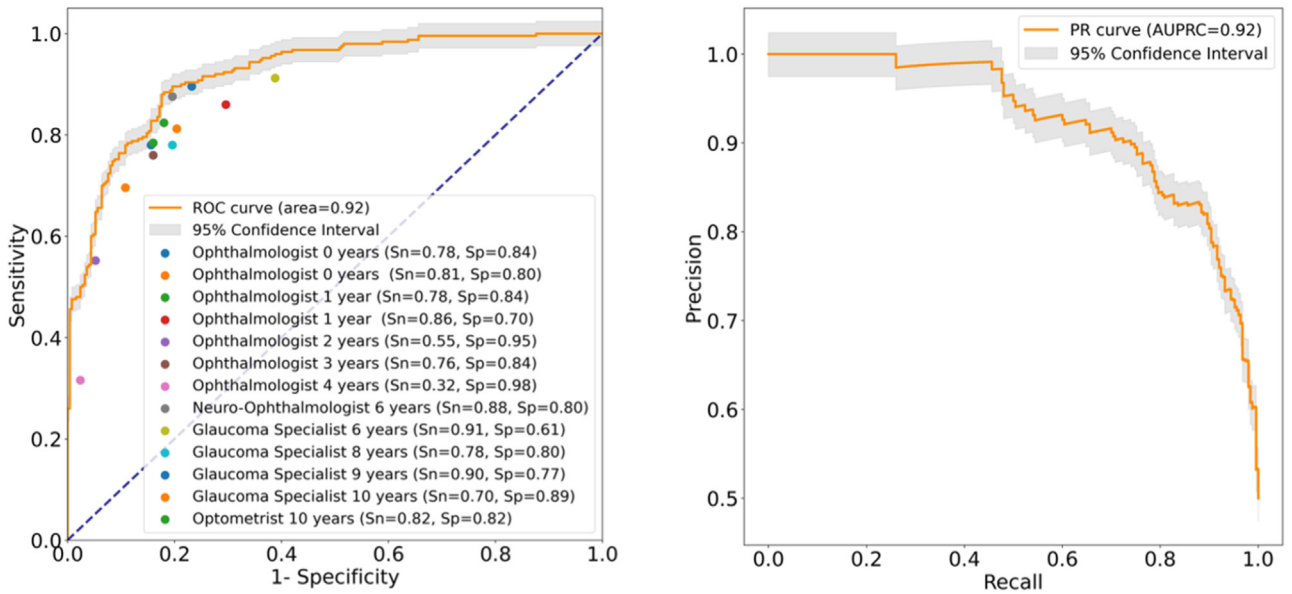

Figure 4. Patient-level algorithm and independent clinician performance with full years of experience (left) and precision–recall curve (right) when using patient-level LAC DHS optometrist reference labels. AUPRC = area under the precision–recall curve; DHS = Department of Health Services; LAC = Los Angeles County; PR = precision–recall; ROC = receiver operating characteristic; Sn = sensitivity; Sp = specificity.

Table 3. Summary of Classification Metrics at a Classification Threshold of 0.50

Parameter	Patient-Level Expert Panel Labels	Patient-Level LAC DHS Optometrist Labels	Eye-Level Expert Panel Labels
AUROC	0.930	0.920	0.920
AUPRC	0.930	0.920	0.880
Accuracy	0.858	0.846	0.838
Precision	0.835	0.831	0.754
Sensitivity (recall)	0.886	0.868	0.853
Specificity	0.831	0.824	0.829
F1 score	0.860	0.849	0.800

AUPRC = area under the precision–recall curve; AUROC = area under the receiver operating characteristic curve; DHS = Department of Health Services; LAC = Los Angeles County.

supporting algorithm integration into existing LAC DHS teleretinal screening workflows to improve the timeliness of referable glaucoma detection and reallocate optometrist time for direct eye care.

We tested our DL algorithm using 3 different sets of reference labels to assess the robustness of its performance. It is somewhat unsurprising that the algorithm matched or outperformed independent human graders when test labels were provided by the same LAC DHS optometrists who provided the training labels. However, it is interesting that the algorithm matched or outperformed independent human graders even when using test labels provided by an expert panel of 3 glaucoma specialists. The robust performance observed across test labels may partially stem from the diversity of training labels by 21 LAC DHS optometrists, which is likely advantageous when automating a task that is inherently variable on the individual grader level.[21] It may also partially stem from using reference labels provided by LAC DHS optometrists rather than specially trained study graders. Using training labels obtained in a clinical

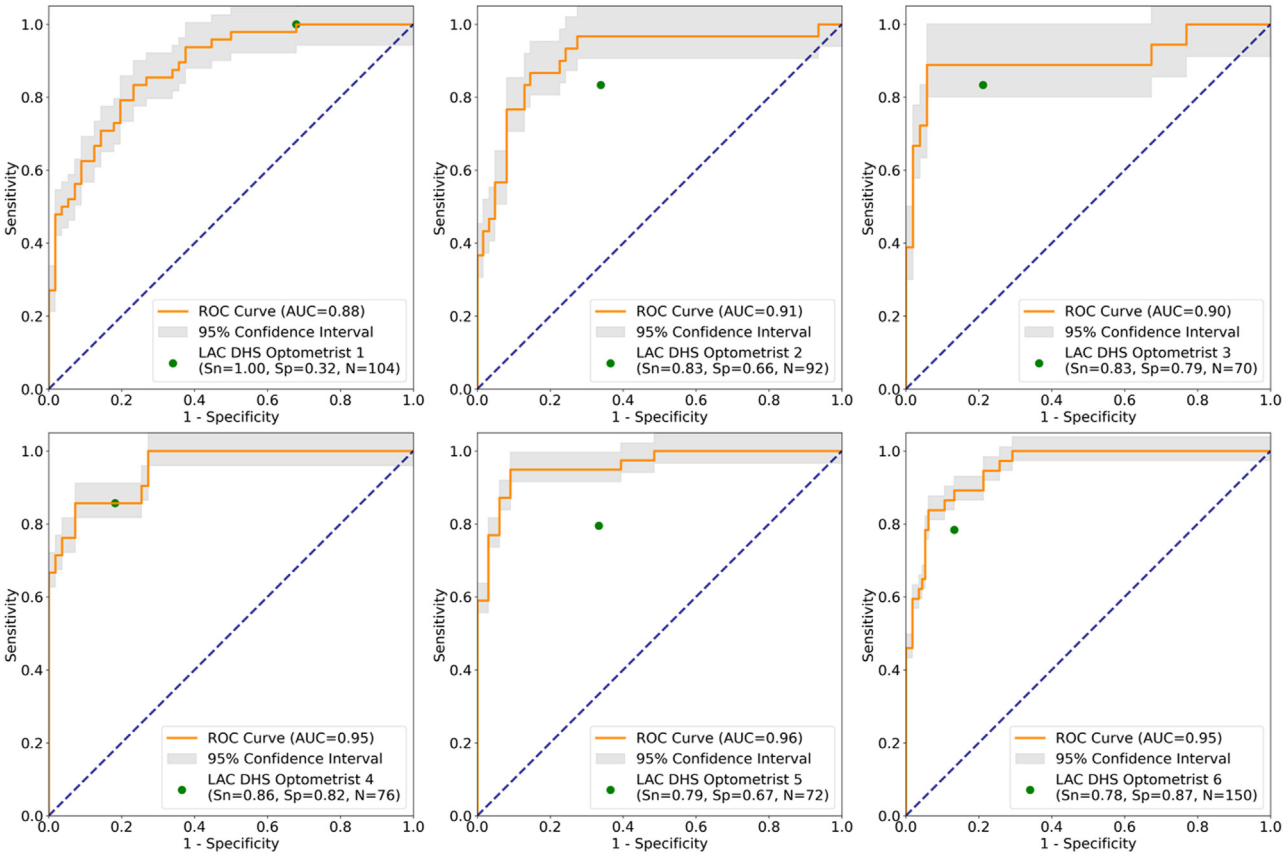

Figure 6. Subanalysis comparing the performance of the patient-level algorithm with that of 6 certified LAC DHS optometrists in subsets of the test data set when using patient-level expert panel reference labels. AUC = area under the curve; DHS = Department of Health Services; LAC = Los Angeles County; ROC = receiver operating characteristic; Sn = sensitivity; Sp = specificity.

environment as opposed to those obtained in a study environment could help minimize the Hawthorne effect, by which individuals modify their behaviors in response to being observed or scrutinized, thereby making the labels more applicable in clinical settings.[22] In contrast, graders of the test data set had a higher likelihood of being affected by the Hawthorne effect; performance observed among individual clinicians likely represents their best efforts. It is also interesting that our algorithm demonstrated robust and consistent performance across all 3 test sets despite being trained using patient-level labels generalized to images from both eyes. This labeling strategy was necessary due to LAC DHS optometrists diagnosing referable glaucoma on the patient level rather than eye level. Fortunately, any noise in the training labels resulting from intereye differences (e.g., 1 eye referable and the other not) did not seem to affect algorithm performance, which matched or exceeded both patient- and eye-level performance by the human graders.

The high degree of variability among clinicians in referable glaucoma detection regardless of experience level presents a significant barrier for teleglaucoma screening programs. Our finding is consistent with previous studies that reported high variability among optometrists and ophthalmologists in grading CDR or detecting manifest glaucoma from fundus photographs.[14,23] This highlights an important issue associated with human grading in teleretinal screening workflows: systematic biases by graders can lead to large-scale over- or underdetection of disease, making it difficult to standardize disease detection and limiting the scalability of teleglaucoma screening overall. This variability was also not correlated with experience level, which suggests that it may be an intrinsic property of graders that is not easily modifiable, even with extensive training. We are currently investigating the interhuman variability of CDR grading in a separate study. In contrast to human graders, AI algorithms can be trained using collective labels provided by many graders, potentially mitigating systematic biases associated with a small number of undercallers (high specificity) or overcallers (high sensitivity). Furthermore, AI algorithms also provide consistent and reproducible image analysis, and sensitivity and specificity can be tailored to suit the specific needs and capacities of individual health care systems. In addition, recalibration techniques may be used to ensure that predicted probabilities remain well-calibrated when applied to new patient populations. Therefore, the relatively unbiased, reproducible, and adaptable nature of certain AI algorithms may make them better suited for large-scale, high-throughput teleglaucoma screening.

Our study has some limitations. First, our training data reflects the unique demographics of the communities served by LAC DHS, which may limit algorithm generalizability in other populations.[2,6] This concern is mitigated by our primary intention to implement the algorithm locally in the LAC DHS teleretinal screening program. However, if the algorithm is implemented more widely in the future, it may benefit from retuning using data from local populations. Second, the test set was deliberately balanced with equal proportions of referable glaucoma and nonreferable glaucoma cases. Although this approach ensures that performance metrics such as sensitivity and specificity are not disproportionately influenced by class imbalance, it may lead to higher precision than would be observed in clinical settings where referable glaucoma is less prevalent. Third, the utility of glaucoma screening in the general population remains unclear, which calls into question the role of algorithms for detecting referable glaucoma.[24] However, LAC DHS serves a high-risk population that is predominantly Hispanic, which may explain why glaucoma referrals at a CDR cutoff of 0.6 are high yield, though the optimal cutoff may vary between different populations; around a quarter of LAC DHS teleretinal patients detected with referable glaucoma were diagnosed with manifest glaucoma after in-office evaluation.[13,25–27] Finally, our algorithm only evaluates single fundus photographs, which is simplistic compared with the comprehensive glaucoma evaluation.[13] However, it is important to point out that we plan to implement this algorithm in resource-constrained screening environments, where the cost of expensive diagnostic tests is prohibitive and the effectiveness of fundus photography alone has been demonstrated. Nevertheless, it is important to consider future opportunities to incorporate accessible factors, such as age and race, that could improve the predictive accuracy of glaucoma referrals and minimize the burden of false positives on the LAC DHS health system.[28]

In conclusion, the performance of our DL algorithm for detecting referable glaucoma matched or exceeded LAC DHS optometrists and independent clinicians with a range of clinical experience. The implementation of validated AI algorithms that approximate expert-level performance into existing clinical workflows could enhance the timeliness and quality of care while also conserving clinician time for direct patient care, which is a valuable commodity in resource-constrained health care systems providing care to underserved, safety net populations.[29–31] Artificial intelligence can also provide more reproducible and adaptable diagnostic capabilities, ensuring that more patients have consistent access to a higher standard of care.[32] However, further work is needed to address technical, ethical, and legal questions surrounding AI for glaucoma care before wide-spread implementation.[33–35]

Footnotes and Disclosures

Originally received: October 3, 2024.
Final revision: February 12, 2025.
Accepted: February 18, 2025.
Available online: February 25, 2025. Manuscript no. XOPS-D-24-00408R1.

[1] Department of Ophthalmology, Keck School of Medicine, Roski Eye Institute, University of Southern California, Los Angeles, California.

[2] Department of Computer Science, Information Sciences Institute, University of Southern California, Los Angeles, California.

[3] Department of Ophthalmology, Los Angeles General Medical Center, Los Angeles, California.

[4] Keck School of Medicine, University of Southern California, Los Angeles, California.

[5] Los Angeles County Department of Health Services, Los Angeles, California.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

B.Y.X.: Support − NIH NEI R01 EY035677, NIH NEI K23 EY032985; Grants or contracts − Ocular Therapeutix; Consultant − AbbVie, Alcon, Ocular Therapeutix; Receipt of equipment, materials, drugs, medical writing, gifts or other services − ArcScan, Heidelberg Engineering.

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

Supported by grants R01 EY035677 and K23 EY032985 from the National Eye Institute, National Institutes of Health, Bethesda, Maryland; a DHS-USC Safety Net Innovation Award from the Southern California Clinical and Translational Science Institute; an AI4Health Award from the University of Southern California; an unrestricted grant to the Department of Ophthalmology from Research to Prevent Blindness, New York, New York; and The Carlson Family Foundation. The research reported in this publication was also supported by the National Eye Institute of the National Institutes of Health under Award Number P30EY029220. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Support for Open Access publication was provided by the University of Southern California, Southern California Clinical and Translational Science Institute.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the University of Southern California Institutional Review Board. The study adhered to the tenets of the Declaration of Helsinki and complied with the Health Insurance Portability and Accountability Act. Informed consent was not obtained because this was a retrospective study utilizing deidentified patient data. All patient information used in the study was anonymized, ensuring that no identifiable personal information was accessed or used. As a result, the study posed minimal risk to participants, and the need for informed consent was waived by the Institutional Review Board in accordance with applicable ethical guidelines.

No animal subjects were used in this study.

Author Contributions:

Conception and design: V. Nguyen, Ambite, Kesselman, Daskivich, Pazzani, Xu

Data collection: V. Nguyen, Iyengar, Rasheed, Apolo, Li, Kumar, H. Nguyen, Bohner, Bolo, Dhodapkar, Do, Duong, Gluckstein, Hong, Humayun, James, Lee, K. Nguyen, Wong, Ambite, Kesselman, Daskivich, Pazzani, Xu

Analysis and interpretation: V. Nguyen, Iyengar, Rasheed, Apolo, Li, Kumar, H. Nguyen, Bohner, Bolo, Dhodapkar, Do, Duong, Gluckstein, Hong, Humayun, James, Lee, K. Nguyen, Wong, Ambite, Kesselman, Daskivich, Pazzani, Xu

Obtained funding: Xu

Overall responsibility: V. Nguyen, Ambite, Kesselman, Daskivich, Pazzani, Xu

Abbreviations and Acronyms:
AI = artificial intelligence; **AUPRC** = area under the precision−recall curve; **AUROC** = area under the receiver operating curve; **CDR** = cup-to-disc ratio; **CI** = confidence interval; **DHS** = Department of Health Services; **DL** = deep learning; **LAC** = Los Angeles County; **VGG** = Visual Geometry Group.

Keywords:
Artificial intelligence, Deep learning, Glaucoma, Screening, Telemedicine.

Correspondence:
Benjamin Y. Xu, MD, PhD, 1450 San Pablo Street, 4th Floor, Suite 4700, Los Angeles, CA 90033. E-mail: benjamin.xu@med.usc.edu.

References

1. GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the right to sight: an analysis for the Global Burden of Disease Study. *Lancet Glob Health*. 2021;9:e144−e160.
2. Vajaranant TS, Wu S, Torres M, Varma R. The changing face of primary open-angle glaucoma in the United States: demographic and geographic changes from 2011 to 2050. *Am J Ophthalmol*. 2012;154:303−314.e3.
3. Berkowitz ST, Finn AP, Parikh R, et al. Ophthalmology workforce projections in the United States, 2020 to 2035. *Ophthalmology*. 2024;131:133−139.
4. Davuluru SS, Jess AT, Kim JSB, et al. Identifying, understanding, and addressing disparities in glaucoma care in the United States. *Transl Vis Sci Technol*. 2023;12:18.
5. Tielsch JM, Sommer A, Katz J, et al. Racial variations in the prevalence of primary open-angle glaucoma. The baltimore eye survey. *JAMA*. 1991;266:369−374.
6. Varma R, Wang D, Wu C, et al. Four-year incidence of open-angle glaucoma and ocular hypertension: the Los Angeles Latino Eye Study. *Am J Ophthalmol*. 2012;154:315−325.e1.
7. Barquet-Pizá V, Siegfried CJ. Understanding racial disparities of glaucoma. *Curr Opin Ophthalmol*. 2024;35:97−103.
8. Yoo K, Apolo G, Zhou S, et al. Rates and patterns of diagnostic conversion from anatomical narrow angle to primary angle-closure glaucoma in the United States. *Ophthalmol Glaucoma*. 2023;6:169−176.
9. Shean R, Yu N, Guntipally S, et al. Advances and challenges in wearable glaucoma diagnostics and therapeutics. *Bioengineering (Basel)*. 2024;11:138.
10. Halawa OA, Kolli A, Oh G, et al. Racial and socioeconomic differences in eye care utilization among medicare beneficiaries with glaucoma. *Ophthalmology*. 2022;129:397−405.
11. Apolo G, Bohner A, Pardeshi A, et al. Racial and sociodemographic disparities in the detection of narrow angles before detection of primary angle-closure glaucoma in the United States. *Ophthalmol Glaucoma*. 2022;5:388−395.
12. Daskivich LP, Vasquez C, Martinez Jr C, et al. Implementation and evaluation of a large-scale teleretinal diabetic retinopathy screening program in the Los Angeles County Department of Health Services. *JAMA Intern Med*. 2017;177:642−649.
13. Yuen J, Xu B, Song BJ, et al. Effectiveness of glaucoma screening and factors associated with follow-up adherence

among glaucoma suspects in a safety-net teleretinal screening program. *Ophthalmol Glaucoma*. 2023;6:247–254.
14. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215–221.
15. Chaurasia AK, Greatbatch CJ, Hewitt AW. Diagnostic accuracy of artificial intelligence in glaucoma screening and clinical practice. *J Glaucoma*. 2022;31:285–299.
16. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125:1199–1206.
17. Medeiros FA, Jammal AA, Mariottoni EB. Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning. *Ophthalmology*. 2021;128:383–392.
18. Li F, Yan L, Wang Y, et al. Deep learning-based automated detection of glaucomatous optic neuropathy on color fundus photographs. *Graefes Arch Clin Exp Ophthalmol*. 2020;258:851–867.
19. Al-Aswad LA, Kapoor R, Chu CK, et al. Evaluation of a deep learning system for identifying glaucomatous optic neuropathy based on color fundus photographs. *J Glaucoma*. 2019;28:1029–1034.
20. Murtagh P, Greene G, O'Brien C. Current applications of machine learning in the screening and diagnosis of glaucoma: a systematic review and meta-analysis. *Int J Ophthalmol*. 2020;13:149–162.
21. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
22. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol*. 2014;67:267–277.
23. Harper R, Reeves B, Smith G. Observer variability in optic disc assessment: implications for glaucoma shared care. *Ophthalmic Physiol Opt*. 2000;20:265–273.
24. United States Preventive Services Taskforce. Recommendation: primary open-angle glaucoma: screening. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/primary-open-angle-glaucoma-screening; 2022. Accessed July 17, 2024.
25. Gordon MO, Beiser JA, Brandt JD, et al. The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120:714–720.
26. Mitchell P, Smith W, Attebo K, Healey PR. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye study. *Ophthalmology*. 1996;103:1661–1669.
27. Klein BE, Klein R, Sponsel WE, et al. Prevalence of glaucoma. The Beaver Dam Eye study. *Ophthalmology*. 1992;99:1499–1504.
28. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol*. 2020;9:42.
29. Xu BY, Chiang M, Chaudhary S, et al. Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images. *Am J Ophthalmol*. 2019;208:273–280.
30. Xu BY, Chiang M, Pardeshi AA, et al. Deep neural network for scleral spur detection in anterior segment OCT images: the Chinese American Eye study. *Transl Vis Sci Technol*. 2020;9:18.
31. Bolo K, Apolo Aroca G, Pardeshi AA, et al. Automated expert-level scleral spur detection and quantitative biometric analysis on the ANTERION anterior segment OCT system. *Br J Ophthalmol*. 2024;108:702–709.
32. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
33. Al-Aswad LA, Ramachandran R, Schuman JS, et al. Artificial intelligence for glaucoma: creating and implementing artificial intelligence for disease detection and progression. *Ophthalmol Glaucoma*. 2022;5:e16–e25.
34. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology*. 2022;129:e14–e32.
35. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit Med*. 2023;6:170.