JAMA Ophthalmology | Original Investigation

Detecting Glaucoma in the Ocular Hypertension Study Using Deep Learning

Rui Fan, PhD; Christopher Bowd, PhD; Mark Christopher, PhD; Nicole Brye; James A. Proudfoot, MS; Jasmin Rezapour, MD; Akram Belghith, PhD; Michael H. Goldbaum, MD; Benton Chuter, MS; Christopher A. Girkin, MD; Massimo A. Fazio, PhD; Jeffrey M. Liebmann, MD; Robert N. Weinreb, MD; Mae O. Gordon, PhD; Michael A. Kass, MD; David Kriegman, PhD; Linda M. Zangwill, PhD

IMPORTANCE Automated deep learning (DL) analyses of fundus photographs potentially can reduce the cost and improve the efficiency of reading center assessment of end points in clinical trials.

OBJECTIVE To investigate the diagnostic accuracy of DL algorithms trained on fundus photographs from the Ocular Hypertension Treatment Study (OHTS) to detect primary open-angle glaucoma (POAG).

DESIGN, SETTING, AND PARTICIPANTS In this diagnostic study, 1636 OHTS participants from 22 sites with a mean (range) follow-up of 10.7 (0-14.3) years. A total of 66 715 photographs from 3272 eyes were used to train and test a ResNet-50 model to detect the OHTS Endpoint Committee POAG determination based on optic disc (287 eyes, 3502 photographs) and/or visual field (198 eyes, 2300 visual fields) changes. Three independent test sets were used to evaluate the generalizability of the model.

MAIN OUTCOMES AND MEASURES Areas under the receiver operating characteristic curve (AUROC) and sensitivities at fixed specificities were calculated to compare model performance. Evaluation of false-positive rates was used to determine whether the DL model detected POAG before the OHTS Endpoint Committee POAG determination.

RESULTS A total of 1147 participants were included in the training set (661 [57.6%] female; mean age, 57.2 years; 95% CI, 56.6-57.8), 167 in the validation set (97 [58.1%] female; mean age, 57.1 years; 95% CI, 55.6-58.7), and 322 in the test set (173 [53.7%] female; mean age, 57.2 years; 95% CI, 56.1-58.2). The DL model achieved an AUROC of 0.88 (95% CI, 0.82-0.92) for the OHTS Endpoint Committee determination of optic disc or VF changes. For the OHTS end points based on optic disc changes or visual field changes, AUROCs were 0.91 (95% CI, 0.88-0.94) and 0.86 (95% CI, 0.76-0.93), respectively. False-positive rates (at 90% specificity) were higher in photographs of eyes that later developed POAG by disc or visual field (27.5% [56 of 204]) compared with eyes that did not develop POAG (11.4% [50 of 440]) during follow-up. The diagnostic accuracy of the DL model developed on the optic disc end point applied to 3 independent data sets was lower, with AUROCs ranging from 0.74 (95% CI, 0.70-0.77) to 0.79 (95% CI, 0.78-0.81).

CONCLUSIONS AND RELEVANCE The model's high diagnostic accuracy using OHTS photographs suggests that DL has the potential to standardize and automate POAG determination for clinical trials and management. In addition, the higher false-positive rate in early photographs of eyes that later developed POAG suggests that DL models detected POAG in some eyes earlier than the OHTS Endpoint Committee, reflecting the OHTS design that emphasized a high specificity for POAG determination by requiring a clinically significant change from baseline.

JAMA Ophthalmol. doi:10.1001/jamaophthalmol.2022.0244
Published online March 17, 2022.

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Linda M. Zangwill, PhD, Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093-0946 (lzangwill@health.ucsd.edu).

In clinical trials assessing the efficacy of disease treatments, such as medical or surgical intervention, the most important consideration arguably is the primary end point. The primary end point serves as an indicator of clinical trial success as it measures efficacy of the assessed treatment.

The Ocular Hypertension Treatment Study (OHTS)[1,2] began as a randomized clinical trial designed to determine the safety and efficacy of topical ocular hypotensive medication in delaying or preventing the onset of primary open-angle glaucoma (POAG) in eyes with ocular hypertension. In OHTS, the primary end point was development of POAG in one or both eyes, defined as reproducible, clinically significant glaucomatous optic disc changes or reproducible glaucomatous visual field (VF) defects.[3] Assessment of optic disc and VF changes were determined by masked readers at the independent Optic Disc Reading Center (ODRC) and Visual Field Reading Center (VFRC). The final POAG determination was decided by a 3-member end point committee of glaucoma experts who reviewed the photographs and VFs to determine whether observed changes were due to POAG or another disease (eg, ischemic optic neuropathy).

The use of 2 reading centers and a masked end point committee is a demanding, laborious, and complicated task. In addition, as agreement among the OHTS Endpoint Committee members' assessment was required, several consensus grading sessions were necessary before a final end point was determined. The 3 committee members reached unanimity in 61% of the end points in the first round of independent reviews, 32.2% in the second round (reviews were completed independently to resolve disagreement), and 6.8% in a final round (a consensus conference telephone call).[4] This end point determination process was used to ensure high specificity, which was appropriate for the OHTS but not necessarily needed for all clinical trials.

Recent improvements in machine learning methods have resulted in automated glaucoma detection methods that could be useful for automating end point determination in clinical trials.[5] Specifically, machine learning end points have the potential to reduce the need for manual assessment, thereby improving the reproducibility of the end point determinations. For instance, deep learning (DL) approaches, including deep convolutional neural networks, have been used to detect glaucoma and estimate structural and VF defects from fundus photographs.[6-13] Besides the increased consistency and potential cost savings, an additional benefit of using these methods is that they provide a probability of disease output that may be used to achieve a target sensitivity and specificity by varying classification cutoffs. An advantage of using the OHTS data set for DL model development to detect POAG is that it is a large multicenter study that included a variety of cameras, technicians, and study participants and the OHTS Endpoint Committee definition of glaucoma.

The current report assessed automated diagnosis of POAG by DL algorithms trained and tested on OHTS fundus stereophotographs. Generalizability was assessed in independent samples from the OHTS and 3 external test sets. We hypothesized that DL models trained on OHTS photographs would successfully classify eyes as having POAG or as healthy in independent test sets at an acceptable level, suggesting that automated classification can supplant the need for multitiered expert assessment of optic disc images in clinical trials. We also compared results from models trained on the OHTS ODRC and VFRC POAG assessment to describe the relative effectiveness of each stage of OHTS photograph and VF classification to assess the accuracy for detecting conversion to POAG in eyes with ocular hypertension.

Key Points

Question Can deep learning analysis of fundus photographs be used to automate the determination of primary open-angle glaucoma (POAG) end points in clinical trials?

Finding In this diagnostic study of 1636 participants in the Ocular Hypertension Treatment Study (OHTS), deep learning models trained and tested on 66 715 photographs achieved high diagnostic accuracy for detecting POAG determined by the OHTS Endpoint Committee using medical information, fundus photographs, and visual fields.

Meaning The high diagnostic accuracy of deep learning models using OHTS photographs suggests that it has the potential to automate end point determination and improve the efficiency of POAG clinical trials.

Methods

Data Collection

The OHTS[1,2] was initiated in 1994 as the first large randomized clinical trial to document the safety and efficacy of topical ocular hypotensive medication in preventing or delaying the onset of VF and/or optic nerve damage in participants with ocular hypertension. Details of the study methods have been reported previously.[1,2] At study entry, written informed consent was obtained from each participant. For this analysis, institutional review board approval was not needed because only deidentified data were used. The current report followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

The OHTS recruited 1636 participants with ocular hypertension with elevated intraocular pressure from 22 sites. Each participant was seen twice a year for Humphrey 30-2 VF testing and once a year for stereoscopic optic nerve head (ONH) photography. The demographic and clinical characteristics included age, self-reported race and ethnicity, self-reported sex, intraocular pressure, central corneal thickness, and refractive status. At study entry, all participants were required to have normal-appearing ONHs based on review of stereoscopic optic disc photographs and VFs within normal limits, as determined by the ODRC and VFRC, respectively. After each visit, the ODRC compared the baseline test with the follow-up test to determine if there was evidence of glaucomatous changes. Specifically, if 2 consecutive sets of photographs demonstrated change from the baseline, the case was reviewed by the 3 masked glaucoma specialist OHTS Endpoint Committee members. Similarly, if the VFRC determined that 3 consecutive sets of VFs were abnormal, then the case was reviewed by

the OHTS Endpoint Committee. Each OHTS Endpoint Committee member independently reviewed the participant's medical history and compared baseline with follow-up photographs to determine whether the changes were clinically significant and attributable to POAG and whether changes in VFs were due to POAG. The advantages of using an end point committee in the OHTS have recently been reported.[4] In brief, using an end point committee had a significant effect on the accuracy of POAG incidence rate, with 16.3% of study participants reaching an unadjudicated all-cause end point but only 9.5% of participants developing a POAG OHTS Endpoint Committee–adjudicated end point. As treatment is unlikely to affect a study participant without POAG, removal of these unadjudicated, all-cause end points led to a more accurate estimate of the efficacy of treatment; treatment reduced the POAG-adjudicated end point by 56% (relative risk, 0.44; 95% CI, 0.31-0.61), while it reduced all-cause end points by 33% (relative risk, 0.67; 95% CI, 0.54-0.84).[4]

For this report, we used OHTS photographs collected during the OHTS randomized clinical phase 1 trial from 1994 to 2002 and the longitudinal follow-up OHTS phase 2 trial from 2002 to 2009 to determine whether DL algorithms can accurately classify eyes based on optic disc changes and VF changes identified by the ODRC, VFRC, and OHTS Endpoint Committees as the ground truth. All photographs, regardless of quality, were included. Specifically, we trained 5 DL algorithms to identify each of the following 5 outcomes:

- OHTS Endpoint Committee determination
  - Model 1: Optic disc changes attributable to POAG by Endpoint Committee
  - Model 2: VF changes attributable to POAG by Endpoint Committee
  - Model 3: Optic disc or VF changes attributable to POAG by Endpoint Committee
- Reading center determination
  - Model 4: Optic disc changes attributable to POAG by ODRC
  - Model 5: VF changes attributable to POAG by VFRC

Photographs taken on or after the initial classification of POAG by the Endpoint Committee determinations were included as POAG for DL models 1, 2, and 3. For the reading center determinations, photographs taken on the visit determined by the ODRC or VFRC as POAG were considered POAG for models 4 and 5, respectively. In contrast to the ground truth used in the DL models for the OHTS Endpoint Committee determinations, POAG was not inferred on photographs taken after the initial reading center determination of change unless the eye was considered as POAG by the OHTS Endpoint Committee.

DL Models, Training, and Selection

Details of data set preparation, including data augmentation, are described in the eMethods and eFigure 1 in the Supplement. In our experiments, a ResNet-50[14] model pretrained on the ImageNet database (Stanford Vision Lab)[15] was fine-tuned for POAG detection. As illustrated in eFigure 2 in the Supplement, we modified the fully connected layer so that it could output 2 scalars indicating the probability of healthy and POAG classes, with respect to the model classifications.

The OHTS data set was divided into training, validation, and test sets, with 85%, 5%, and 10% of participants, respectively, included in each set, so that all images from a single participant were included in the same partition. Each of the 5 models used the exact same training, validation, and test sets.

The DL model training was carried out on 2 NVIDIA Geforce RTX 2080 Super GPUs, each with 8-GB GDDR6 memory. Because the OHTS data set is imbalanced with most eyes not developing POAG (1299 of 1636 [79.4%]), we implemented additional class weights into the loss function (eMethods in the Supplement).

Performance Evaluation

The trained DL models were evaluated on the OHTS test set as well as 3 additional independent test data sets of optic disc photographs labeled as glaucoma or healthy: (1) ACRIMA,[16] (2) Large-scale Attention-Based Glaucoma (LAG),[17] and (3) Diagnostic Innovations in Glaucoma Study (DIGS) and African Descent and Glaucoma Evaluation Study (ADAGES).[18] The ACRIMA public data set provides cropped fundus photographs. We applied the same cropping strategy to the LAG and DIGS/ADAGES data sets that we applied to the OHTS photographs (eMethods and eFigure 1 in the Supplement). Because these data sets are only used to test the generalizability of our DL model, we did not apply flipping and rotation operations to the external independent data sets.

Performance in distinguishing between healthy eyes and eyes with glaucoma was evaluated using sensitivity, specificity, precision, and area under the receiver operating characteristic curve (AUROC). To evaluate model accuracy for detecting early glaucoma, we conducted a subset analysis in eyes with a VF mean deviation (MD) better than −6 dB. The AUROC scores of different models were statistically compared using a clustered bootstrap approach to address multiple images from the same eyes.[19] To help evaluate clinical utility, sensitivity at 4 fixed levels of specificity (80%, 85%, 90%, and 95%) was evaluated. We also calculated precision (also known as positive predictive value) at fixed specificities, which is particularly informative along with recall (also known as sensitivity) in imbalanced data sets.[20] Furthermore, Grad-CAM++,[21] a common network explanation and visualization technique, was used to help understand the model decision-making process. Compared with saliency and occlusion map-based DL model visualization methods, Grad-CAM++ has several advantages including that it (1) can be used for any DL model, (2) does not require model retraining to produce saliency maps, (3) uses softmax scores as weights to remove the dependence on unstable gradients, and (4) removes irrelevant noise to create a saliency map.

Statistical Analysis

Baseline patient-level characteristics are presented as means with 95% CIs for continuous variables and counts with percentages for categorical variables. The statistical significance of comparisons between patient-level characteristics across training, validation, and test data sets was determined by χ^2 tests for categorical variables. For eye-level characteristics, mean and confidence interval estimates were derived from linear mixed-effects models, with a random intercept to

Table 1. Characteristics of the Ocular Hypertension Treatment Study Training, Validation, and Test Sets

Characteristic	No. (%)			*P* value
	Training set	Validation set	Test set	
Participants, No.	1147	167	322	NA
Eyes, No.	2294	334	644	NA
Eye visits, No.	26 313	3807	7220	NA
Mean age (95% CI), y	57.2 (56.6 to 57.8)	57.1 (55.6 to 58.7)	57.2 (56.1 to 58.2)	>.99
Sex				
Female	661 (57.6)	97 (58.1)	173 (53.7)	.44
Male	486 (42.4)	70 (41.9)	149 (46.3)	
Self-reported race				
European descent	871 (75.9)	120 (71.9)	238 (73.9)	.43
African descent	276 (24.1)	47 (28.1)	84 (26.1)	
Mean baseline visual field MD (95% CI), dB	−0.02 (−0.10 to 0.06)	0.02 (−0.20 to 0.23)	−0.12 (−0.27 to 0.04)	.50
Baseline photograph-based vertical cup-disc ratio, mean (95% CI)	0.39 (0.38 to 0.40)	0.36 (0.33 to 0.40)	0.39 (0.37 to 0.41)	.28
Did not develop glaucoma[a]				
Participants, No.	955	138	267	NA
Eyes, No.	2046	298	569	NA
Eye visits, No.	22 796	3272	6208	NA
Mean visual field MD (95% CI), dB	−0.18 (−0.25 to −0.12)	−0.09 (−0.25 to 0.07)	−0.16 (−0.28 to −0.05)	.53
Developed a POAG end point by visual field or photograph				
Participants, No.	192	29	55	NA
Eyes, No.	248	36	75	NA
Eye visits, No.	3517	535	1012	NA

Abbreviations: MD, mean deviation; NA, not applicable; POAG, primary open-angle glaucoma.

[a] The number of eyes includes all eyes for the did not develop glaucoma group plus fellow eyes from a subset of participants in the developed a POAG end point by visual field or photograph group in which one eye developed a POAG end point and the fellow eye did not.

account for within-participant correlations. All *P* values were 2-tailed, and significance was set at $P < .05$. All analyses were performed using R version 3.6.3 (The R Foundation).

Results

A total of 1147 participants were included in the training set (661 [57.6%] female; mean age, 57.2 years; 95% CI, 56.6-57.8), 167 in the validation set (97 [58.1%] female; mean age, 57.1 years; 95% CI, 55.6-58.7), and 322 in the test set (173 [53.7%] female; mean age, 57.2 years; 95% CI, 56.1-58.2) (**Table 1**). The fundus photograph-based DL models detected conversion to POAG with good accuracy. Specifically, the best diagnostic accuracy of the DL model was achieved for the OHTS Endpoint Committee POAG attribution based on optic disc changes (model 1; AUROC, 0.91; 95% CI, 0.88-0.94) followed by either optic disc or VF changes (model 3; AUROC, 0.88; 95% CI, 0.82-0.92) and VF only change (model 2; AUROC, 0.86; 95% CI, 0.76-0.93) (**Table 2**). The AUROCs of the ODRC and VFRC POAG attribution by optic disc photographs and by VFs were 0.89 (95% CI, 0.85-0.92) and 0.83 (95% CI, 0.76-0.88), respectively. The diagnostic accuracy of detecting early POAG (VF MD of −6 dB or greater) also was generally high for the OHTS Endpoint Committee and the ODRC, with AUROCs ranging from 0.83 (95% CI, 0.70-0.91) to 0.90 (95% CI, 0.87-0.94). Model performance was lower for the VFRC POAG determination of early glaucoma (AUROC, 0.80; 95% CI, 0.72-0.86).

To determine whether photograph quality was associated with model performance, we used an objective DL algorithm to assign a quality metric in a subset of OHTS fundus photographs.[22] Including only the highest-quality images (approximately 73% of the test eyes contributing at least 1 photograph to the analysis) in a post hoc analysis, the algorithm increased the model accuracy (AUROC) from 0.86 (95% CI, 0.76-0.93) to 0.90 (95% CI, 0.87-0.93) for model 2 (OHTS Endpoint Committee determination of VF changes) and from 0.83 (95% CI, 0.76-0.88) to 0.87 (95% CI, 0.82-0.91) for model 5 (VFRC determination). No improvement was found for models 1 (OHTS Endpoint Committee determination of optic disc changes), 3 (OHTS Endpoint Committee determination of optic disc or VF changes), or 4 (ODRC determination).

We also investigated the model's accuracy if fewer visits were included in the test set, as is often observed in other studies. Using 3 randomly chosen visits per eye instead of using all OHTS photographs resulted in similar findings to those obtained using all photographs, with AUROCs ranging from 0.83 (95% CI, 0.75-0.89) for model 5 to 0.91 (95% CI, 0.85-0.95) for model 1 (eTable 2 in the Supplement).

Given that the OHTS data set is imbalanced with a much larger proportion of eyes without POAG than with POAG, it is more likely that the model will identify an eye without POAG as having POAG (ie, false-positives) than classifying the smaller number of eyes with POAG as not having POAG (false-negatives). For this reason, we also reported sensitivity (recall) and precision (positive predictive value) at fixed specificities (Table 2). The sensitivity decreased with increasing specificity, while precision values increased with increasing specificity.

The false-positive rates at 90% specificity were higher in photographs of eyes with ocular hypertension acquired be-

Table 2. Diagnostic Accuracy of DL Model Performance in Identifying Primary Open-Angle Glaucoma by the Ocular Hypertensive Treatment Study (OHTS) Endpoint Committee and Optic Disc and Visual Field Reading Centers

POAG detection modality	POAG, No.			AUROC (95% CI)		Specificity						
	Participants	Eyes	Visits	All eyes	Eyes with early glaucoma (VF MD ≥−6 dB)	80%		85%		90%		95% Sensitivity	Precision
						Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision		
OHTS Endpoint Committee													
Optic disc and/or visual field changes	52	71	352	0.88 (0.82-0.92)	0.86 (0.79-0.91)	0.81	0.21	0.77	0.26	0.69	0.32	0.56	0.43
Optic disc changes	41	56	262	0.91 (0.88-0.94)	0.90 (0.87-0.94)	0.86	0.17	0.81	0.21	0.73	0.26	0.56	0.35
Visual field changes	35	41	195	0.86 (0.76-0.93)	0.83 (0.70-0.91)	0.80	0.14	0.76	0.17	0.69	0.22	0.55	0.31
Reading center													
Optic Disc Reading Center	60	77	318	0.89 (0.85-0.92)	0.87 (0.83-0.91)	0.83	0.20	0.77	0.23	0.71	0.30	0.56	0.40
Visual Field Reading Center	61	78	242	0.83 (0.76-0.88)	0.80 (0.72-0.86)	0.69	0.15	0.64	0.18	0.58	0.22	0.48	0.32

Abbreviations: AUROC, area under the receiver operating characteristic curve; MD, mean deviation; POAG, primary open-angle glaucoma; VF, visual field.

fore they reached a POAG end point compared with the false-positive rate of eyes with ocular hypertension that did not develop POAG (OHTS Endpoint Committee for optic dis or VF changes [model 3]: 27.5% [56 of 204] vs 11.4% [50 of 440]; OHTS Endpoint Committee for optic disc changes [model 1]: 24.7% [47 of 190] vs 8.4% [38 of 454]; OHTS Endpoint Committee for VF changes [model 2]: 21.2% [46 of 217] vs 5.4% [23 of 427]). The **Figure** illustrates the increasing probability of observing a false-positive over the course of the OHTS phase 1 and 2 trials in eyes that eventually developed POAG compared with eyes that did not. The mean time between the occurrence of the first false-positive result and the development of POAG ranged from 4.5 years (95% CI, 3.4-5.6) for the optic disc photograph POAG end point (model 1) to 5.2 years (95% CI, 4.1-6.3) for the optic disc or VF POAG end point (model 3) (eTable 1 in the Supplement).

Table 3 shows the diagnostic accuracy of the DL models trained based on the OHTS optic disc end points on the 3 independent clinical data sets, which was lower compared with the OHTS test set (DIGS/ADAGES: AUROC, 0.74; 95% CI, 0.69-0.79; ACRIMA: AUROC, 0.74; 95% CI, 0.70-0.77; LAG: AUROC, 0.79; 95% CI, 0.78-0.81).

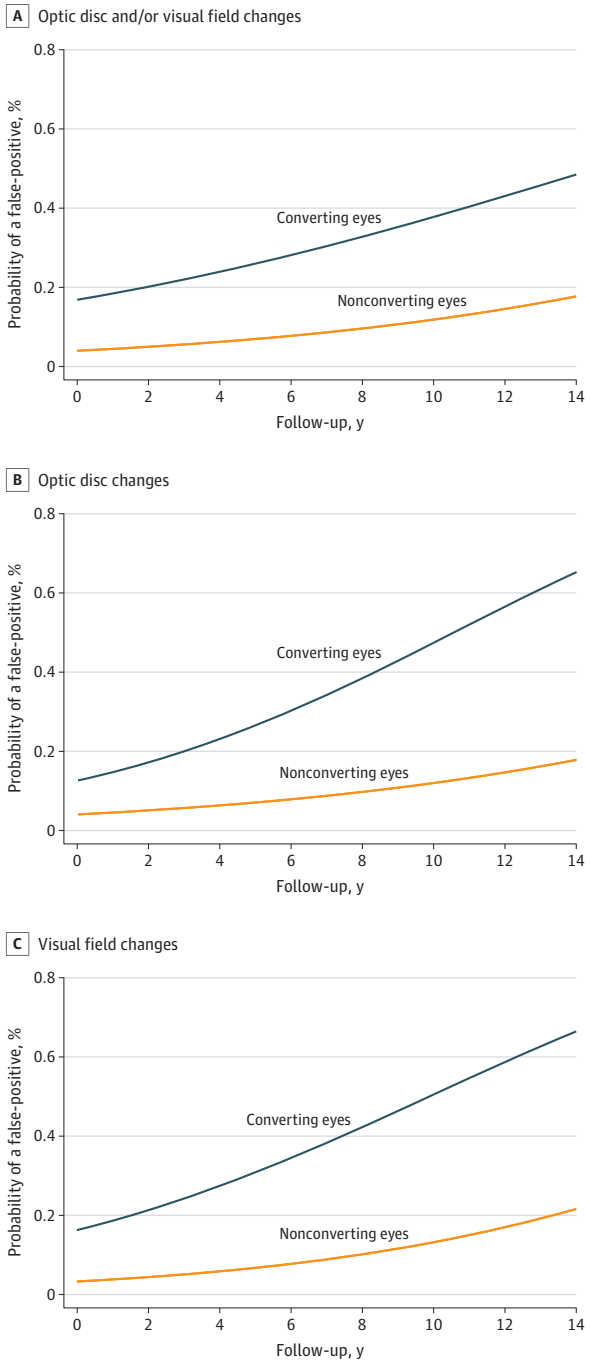
We used Grad-CAM++[21] to determine which regions of the photographs were saliently important for the DL models' decision-making (eFigure 3 in the Supplement). The Grad-CAM++ visualization results suggest that the region within the ONH had the most impact on model decisions. The neuroretinal rim areas are identified as most important, and the periphery contributed comparatively little to clear model decisions for both healthy eyes and eyes with POAG in correct and incorrect classifications. Borderline results were those in which *p* ranged from 0.3 to 0.7 and seem to be less focused on the ONH regions.

Discussion

These results suggest that the DL models using fundus photographs only can provide good accuracy for the determination of glaucomatous change based on the optic disc (AUROC, 0.91; 95% CI, 0.88-0.94), VF (AUROC, 0.86; 95% CI, 0.76-0.93) or either (AUROC, 0.88; 95% CI, 0.82-0.92) by OHTS Endpoint Committee members who incorporated the participant's medical history and both fundus photographs and VF information in their decision-making process. Given the challenging and subjective nature of POAG determination, these results suggest a role for artificial intelligence in improving the accuracy and consistency of the process at lower cost.[5] Moreover, models used here provide a probability of glaucoma as output, and classification thresholds can be adjusted. Specificity and sensitivity of the diagnostic classification can be adjusted to reflect clinical trial goals. With this in mind, we presented sensitivity and precision results at various levels of specificity.

The DL models tested on the OHTS Endpoint Committee determination of POAG generally performed better than those tested on the OHTS ODRC and VFRC determinations when identifying glaucoma from fundus photographs. This is likely

Figure. False-positive Rates by Ocular Hypertension Treatment Study (OHTS) Endpoint Committee Determination of Primary Open-Angle Glaucoma



These figures illustrate the higher probability of observing a false-positive throughout the course of the Ocular Hypertension Treatment Study (OHTS) phase 1 and 2 trials in eyes that developed a primary open-angle glaucoma (POAG) end point (converting eyes) compared with those that did not (nonconverting eyes). The results are calculated at 90% specificity for the OHTS Endpoint Committee determination based on optic disc changes (model 1), OHTS Endpoint Committee determination based on visual field changes (model 2), and OHTS Endpoint Committee determination based on either optic disc or visual field changes (model 3).

in part due to the OHTS design. Reading center personnel reviewed either photographs or VFs alone and did not have access to other clinical information to help determine if changes were attributable to POAG or other causes. As expected, models trained using optic disc changes determined by the ODRC as ground truth performed better than models trained using VF changes determined by the VFRC as ground truth. The high diagnostic accuracy of the current model suggests that DL can be used to automate the determination of POAG for clinical trials and management.

The reported higher false-positive rate in early photographs of eyes that later developed POAG compared with eyes that did not develop POAG (Figure; eTable 1 in the Supplement) suggests that DL models detected POAG in some eyes earlier than the OHTS Endpoint Committee or reading centers. These false-positives likely were true-positives detecting disease-related change on average more than 4 years earlier in eyes with ocular hypertension; this was, in part, a result of the OHTS study design that emphasized high specificity for glaucomatous determination.[4] The models used here provide a probability of glaucoma as output, allowing changes in sensitivity and specificity to desired levels by adjusting the cutoffs used to define POAG. This makes DL models adaptable to different study goals. For instance, one may wish to adjust the desired specificity for the purpose of attempting to detect moderate to severe glaucoma, where a false-negative may result in delayed treatment, leading to a preventable loss of vision.

In the current study, we also reported the generalizability of results from OHTS DL models to 3 external data sets, an important step in assessing model performance. The current DL models showed somewhat better generalizability to the LAG data set than the DIGS/ADAGES and ACRIMA data sets. Poorer performance in independent test sets likely is affected by differences among test sets in ground truth determination, types of cameras, and technician experience as well as differences in study populations. There is also considerable evidence that assessment of optic disc photographs for glaucoma determination is highly variable, even among glaucoma experts.[23-26] Given the variability in assessment of photographs for glaucoma detection, it is likely that there are differences in the criteria used to detect glaucoma in the different external test data sets. Differences in labeling and study populations have been shown to affect DL model performance.[10] A strength of the OHTS is that the POAG determination and study population are very well documented. However, even during the OHTS POAG determination by 3 glaucoma specialist OHTS Endpoint Committee members, there was initial consensus in only 61% of eyes evaluated; 39% of eyes required regrading and/or discussion to reach consensus on POAG status.

OHTS is a multicenter study with a wide range of cameras, technician expertise, and participants; the current DL model was able to incorporate these differences into its classification as evidenced by the strong performance in the OHTS independent test set (optic disc and VF changes: AUROC, 0.91; 95% CI, 0.88-0.94). Given that clinical trials require standardization of end points and few use end point committees for their review of reading center results, we be-

Table 3. Diagnostic Accuracy of the Model Developed Using the Ocular Hypertension Treatment Study Endpoint Committee Determination of End Points Based on Optic Disc Changes in Independent Test Sets[a]

Source	Country	Ground truth determination	Test set size (images), No.		AUROC (95% CI)		Sensitivity			
			Healthy	Glaucoma	All eyes	Mild glaucoma (VF MD ≥−6 dB)	80% Specificity	85% Specificity	90% Specificity	95% Specificity
DIGS[18]	US	Agreement by 2 expert graders from the University of California, San Diego, reviewing photographs only	5184	4289	0.74 (0.69-0.79)	0.71 (0.65-0.76)	0.59	0.52	0.43	0.30
ACRIMA[16]	Spain	1 Expert grader reviewing photographs only	309	396	0.74 (0.70-0.77)	Not available	0.55	0.46	0.38	0.29
LAG[17]	China	Glaucoma specialists from Beijing Tongren Hospital reviewing photographs, IOP, and VF	3143	1711	0.79 (0.78-0.81)	Not available	0.66	0.59	0.53	0.42

Abbreviations: ADAGES, African Descent and Glaucoma Evaluation Study; AUROC, area under the receiver operating characteristic curve; DIGS, Diagnostic Innovations in Glaucoma Study; IOP, intraocular pressure; LAG, Large-scale Attention-Based Glaucoma; MD, mean deviation; VF, visual field.

[a] The DIGS/ADAGES data set consisted of photographs of 227 women and 151 men with a mean age of 64.0 years (95% CI, 63.5-64.6). Demographic characteristics of the LAG and ACRIMA study population are not readily available.

lieve DL can help provide standardized assessment across study centers, thereby increasing the efficiency and reducing the cost of clinical trials.

Thakur and colleagues[27] also used DL to detect glaucoma in OHTS fundus photographs and reported somewhat higher AUROCs than the current study. Specifically, the AUROC for classifying eyes with and without glaucoma based on the OHTS Endpoint Committee was 0.94 (95% CI, 0.94-0.96) compared with 0.88 (95% CI, 0.82-0.92) in the current study. This discrepancy in classification performance may be due in part to the fact that Thakur et al[27] excluded approximately 24% of the available photographs due to extreme artifacts. In contrast, we included all available OHTS photographs to better reflect clinical practice. Our analysis of the association of photograph quality with model performance suggests that removing poor-quality photographs can improve the diagnostic accuracy in some but not all models. Thus, decreased performance in our model could be in part due to less-than-ideal images that were included in the DL model training, validation, and tests sets. These less-than-ideal images cannot be avoided in clinical settings.

The current study also investigated the relative performance of DL models in a subset of eyes with early glaucoma. Although AUROCs were up to 0.03 lower, the general pattern of performance was similar to that observed when all eyes with glaucoma were included; AUROCs generally were greater for the OHTS Endpoint Committee POAG determination tests sets and AUROCs were greater for ODRC than VFRC POAG determination.

Limitations

There are several possible limitations to this study. First, the number of eyes that developed POAG is much smaller than the number of eyes that did not develop POAG, resulting in an imbalanced data set. To address this common problem, we implemented additional class weights into the model. Second, the class activation maps identified regions on the ONH and neuroretinal rim for clear model decisions, both accurate and inaccurate. The areas used by the model in the borderline results (those photographs without a clear model decision on whether the eye had POAG or not) were less focused on the ONH and the generally accepted parapapillary regions where characteristics of POAG are most frequently observed (eFigure 3 in the Supplement). Third, we included all photographs in our modeling to better reflect real-world scenarios for clinical management and clinical trials in which poor-quality photographs are sometimes the only ones available. Removing poor-quality photographs in a post hoc analysis resulted in small to moderate increases in AUROCs. Fourth, we cropped all photographs, which may have reduced model performance if informative information is located in the peripheral retina.

Conclusions

In conclusion, the high diagnostic accuracy of the current DL model based on OHTS photographs suggests that DL models have the potential to standardize and automate the determination of POAG end points for clinical trials and management. We believe integration of DL analyses of photographic images and other test results in clinical trials could reduce the cost and improve the consistency and accuracy of end point assessments, either by decreasing or replacing the personnel required to complete the task. Moreover, given the performance of the DL analysis in comparison with expert human observation, this approach may be promising to provide decision support in clinical settings.

ARTICLE INFORMATION

Accepted for Publication: January 19, 2022.

Published Online: March 17, 2022. doi:10.1001/jamaophthalmol.2022.0244

Author Affiliations: Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California, San Diego, La Jolla (Fan, Bowd, Christopher, Brye, Proudfoot, Rezapour, Belghith, Goldbaum, Chuter, Fazio, Weinreb, Zangwill); Department of Computer Science and Engineering, University of California, San Diego, La Jolla (Fan, Kriegman); Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai, China (Fan); Department of Ophthalmology, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz,

Rheinland-Pfalz, Germany (Rezapour); Department of Ophthalmology, School of Medicine, The University of Alabama at Birmingham (Girkin, Fazio); Department of Biomedical Engineering, School of Engineering, The University of Alabama at Birmingham (Fazio); Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Medical Center, New York, New York (Liebmann); Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, Washington University in St Louis, St Louis, Missouri (Gordon, Kass).

Author Contributions: Drs Fan and Zangwill had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Fan and Bowd contributed equally to this work.
Study concept and design: Fan, Bowd, Christopher, Belghith, Liebmann, Gordon, Kriegman, Zangwill.
Acquisition, analysis, or interpretation of data: Fan, Christopher, Brye, Proudfoot, Rezapour, Goldbaum, Chuter, Girkin, Fazio, Liebmann, Weinreb, Gordon, Kass, Kriegman, Zangwill.
Drafting of the manuscript: Fan, Bowd, Fazio, Liebmann, Kriegman, Zangwill.
Critical revision of the manuscript for important intellectual content: Fan, Bowd, Christopher, Brye, Proudfoot, Rezapour, Belghith, Goldbaum, Chuter, Girkin, Liebmann, Weinreb, Gordon, Kass, Zangwill.
Statistical analysis: Fan, Brye, Proudfoot, Belghith, Zangwill.
Obtained funding: Belghith, Girkin, Fazio, Liebmann, Weinreb, Gordon, Kass, Zangwill.
Administrative, technical, or material support: Christopher, Chuter, Girkin, Liebmann, Weinreb, Gordon, Kass, Zangwill.
Study supervision: Bowd, Christopher, Belghith, Liebmann, Gordon, Kriegman, Zangwill.

Conflict of Interest Disclosures: Dr Christopher has received grants from National Eye Institute and has a patent for estimating the likelihood of primary open-angle glaucoma issued to the University of California, San Diego. Dr Rezapour has received grants from the German Research Foundation and the German Ophthalmological Society. Dr Fazio has received grants from the National Institutes of Health and nonfinancial support from Heidelberg Engineering. Dr Weinreb has received nonfinancial support from Carl Zeiss Meditec, Heidelberg, Konan Medical, Optovue, and Centervue; grants from Bausch & Lomb and the National Eye Institute; and personal fees from Topcon, Allergan, and Equinox; and has a patent issued to Carl Zeiss Meditec, with royalties paid. Dr Gordon has received grants from the National Eye Institute. Dr Zangwill has received grants from the National Eye Institute and Heidelberg Engineering, personal fees from AbbVie, and nonfinancial support from Heidelberg Engineering, Carl Zeiss Meditec, Optovue, and Topcon; and has a patent issued to Carl Zeiss Meditec with royalties paid and a patent issued to the University of California, San Diego. No other disclosures were reported.

Funding/Support: This study was supported by grants R01EY029058, R21EY027945, K99EY030942, R01EY011008, R01EY19869, R01EY027510, and R01EY026574 and Core Grant P30EY022589 from the National Eye Institute; grants from the National Center on Minority Health and Health Disparities; Horncrest Foundation; grants EY09341 and EY09307 and Vision Core Grant P30EY02687 from the National Institutes of Health to the Department of Ophthalmology and Visual Sciences at Washington University; Merck Research Laboratories; Pfizer; White House Station; an unrestricted grant from Research to Prevent Blindness; research fellowship grant RE 4155/1-1 from the German Research Foundation; and grants from the German Ophthalmological Society.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Meeting Presentation: This manuscript was presented virtually at the 2021 annual Association for Research in Vision and Ophthalmology meeting; May 5, 2021.

Additional Information: Our source code and well-trained models are available on request at https://sites.google.com/view/jamaoph.

REFERENCES

1. Gordon MO, Kass MA. *The Ocular Hypertension Treatment Study (OHTS)*. National Technical Information Services; 1997.

2. Gordon MO, Kass MA. The Ocular Hypertension Treatment Study: design and baseline description of the participants. *Arch Ophthalmol*. 1999;117(5):573-583. doi:10.1001/archopht.117.5.573

3. Kass MA, Heuer DK, Higginbotham EJ, et al. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120(6):701-713. doi:10.1001/archopht.120.6.701

4. Gordon MO, Higginbotham EJ, Heuer DK, et al; Ocular Hypertension Treatment Study. Assessment of the impact of an endpoint committee in the Ocular Hypertension Treatment Study. *Am J Ophthalmol*. 2019;199:193-199. doi:10.1016/j.ajo.2018.11.006

5. Lee CS, Lee AY. How artificial intelligence can transform randomized controlled trials. *Transl Vis Sci Technol*. 2020;9(2):9. doi:10.1167/tvst.9.2.9

6. Ahn JM, Kim S, Ahn KS, Cho SH, Lee KB, Kim US. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS One*. 2018;13(11):e0207982. doi:10.1371/journal.pone.0207982

7. Asaoka R, Tanito M, Shibata N, et al. Validation of a deep learning model to screen for glaucoma using images from different fundus cameras and data augmentation. *Ophthalmol Glaucoma*. 2019;2(4):224-231. doi:10.1016/j.ogla.2019.03.008

8. Chai YD, Liu HY, Xu J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowledge-Based Syst*. 2018;161:147-156. doi:10.1016/j.knosys.2018.07.043

9. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8(1):16685. doi:10.1038/s41598-018-35044-9

10. Christopher M, Nakahara K, Bowd C, et al. Effects of study population, labeling and training on glaucoma detection using deep learning algorithms. *Transl Vis Sci Technol*. 2020;9(2):27. doi:10.1167/tvst.9.2.27

11. Jammal AA, Thompson AC, Mariottoni EB, et al. Human versus machine: comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *Am J Ophthalmol*. 2020;211:123-131. doi:10.1016/j.ajo.2019.11.006

12. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8(1):14665. doi:10.1038/s41598-018-33013-w

13. Phene S, Dunn RC, Hammel N, et al. Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs. *Ophthalmology*. 2019;126(12):1627-1639. doi:10.1016/j.ophtha.2019.07.024

14. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. Accessed December 20, 2016. doi:10.1109/CVPR.2016.90

15. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL. Accessed March 8, 2020. doi:10.1109/CVPR.2009.5206848

16. Diaz-Pinto A, Morales S, Naranjo V, Köhler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online*. 2019;18(1):29. doi:10.1186/s12938-019-0649-y

17. Li L, Xu M, Wang X, Jiang L, Liu H. Attention based glaucoma detection: a large-scale database and CNN model. Paper presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 15-20, 2019; Long Beach, CA. Accessed June 12, 2020. doi:10.1109/CVPR.2019.01082

18. Sample PA, Girkin CA, Zangwill LM, et al; African Descent and Glaucoma Evaluation Study Group. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127(9):1136-1145. doi:10.1001/archophthalmol.2009.187

19. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997;53(2):567-578. doi:10.2307/2533958

20. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432

21. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. Paper presented at: 2018 IEEE Winter Conference on Applications of Computer Vision; March 12-15, 2018; Lake Tahoe, NV. Accessed August 30, 2020. doi:10.1109/WACV.2018.00097

22. Chuter B, Christopher M, Fan R, et al. A deep learning model to assess fundus photograph image quality and improve predictive value of deep learning models of glaucoma detection. *Invest Ophthalmol Vis Sci*. 2021;62:1016.

23. Reus NJ, Lemij HG, Garway-Heath DF, et al. Clinical assessment of stereoscopic optic disc photographs for glaucoma: the European Optic Disc Assessment Trial. *Ophthalmology*. 2010;117(4):717-723. doi:10.1016/j.ophtha.2009.09.026

24. Lichter PR. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc*. 1976;74:532-572.

25. Tielsch JM, Katz J, Quigley HA, Miller NR, Sommer A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology*. 1988;95(3):350-356. doi:10.1016/S0161-6420(88)33177-5

26. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99(2):215-221. doi:10.1016/S0161-6420(92)31990-6

27. Thakur A, Goldbaum M, Yousefi S. Predicting glaucoma before onset using deep learning. *Ophthalmol Glaucoma*. 2020;3(4):262-268. doi:10.1016/j.ogla.2020.04.012

Supplemental Online Content

Fan R, Bowd C, Christopher M, et al. Detecting glaucoma in the Ocular Hypertension Study using deep learning. *JAMA Ophthalmol*. Published online March 17, 2022. doi:10.1001/jamaophthalmol.2022.0244

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods. Implemented class weights added to loss function to address POAG vs non-POAG end point imbalance

Dataset Preparation

Because the 22 OHTS sites used different fundus cameras, resulting in inherent variability in image quality and resolution, the training of DL models was much more challenging than if photographs came from a single site or camera. To provide consistent model inputs, we first extracted a region centered on the ONH using a semantic segmentation DL model, DeepLabv3+ (with a ResNet-18 backbone network trained for ONH extraction). A square region surrounding the extracted ONH was then automatically cropped for input to the DL model, where the side length of each cropped image is approximately two times larger than the ONH diameter. The cropped fundus images were then resized to 224×224 pixels. A single reviewer manually reviewed each cropped image to ensure it was correctly centered on the ONH (see Supplemental eFigure 1). The OHTS stereophotographs included both simultaneous and sequential photographs – which cover different areas of the optic nerve head. The images were cropped to provide consistent inputs to the deep learning model.

Data Augmentation

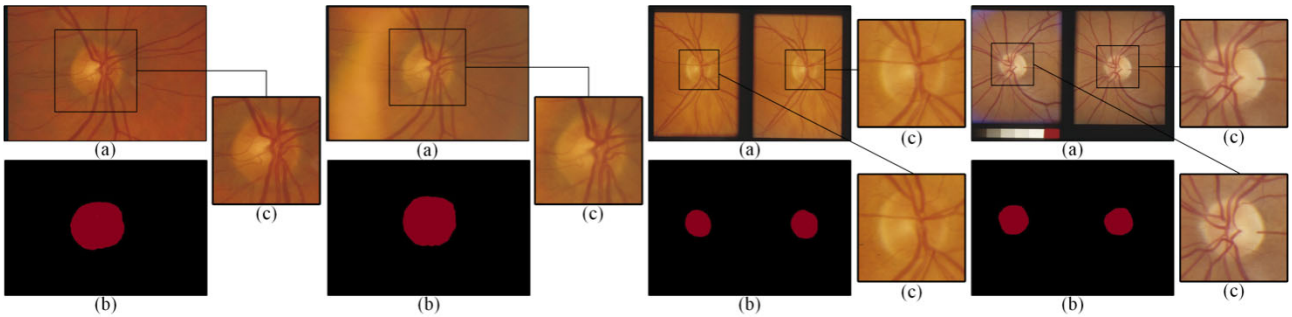
Several data augmentation strategies were applied to increase variation in the training set. To mimic the inclusion of both OD and OS orientations, horizontally mirrored versions of all photographs were added. In addition, we completed horizontal and vertical translation (≤40 pixels) and rotation (≤5°), in which the ONH center of each photograph was randomly perturbed by a small amount to reflect the common situation in which photographs are not well-centered. Each augmented image was assigned the same label (healthy or POAG) as the original input image from which it was derived.

Supplement highlights the conventional cross entropy loss *L* as follows:

$$L = -\frac{n_1}{n_0 + n_1} y \log p - \frac{n_0}{n_0 + n_1} (1 - y) \log(1 - p)$$

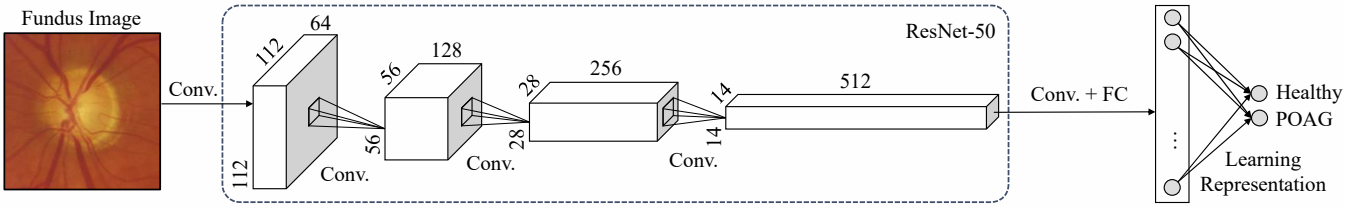
where y denotes the class label ($y = 0$ for healthy images and $y = 1$ for POAG images), p represents POAG prediction probability output by the network, and n_0 and n_1 denote the number of healthy and POAG images, respectively. We utilized the stochastic gradient descent with momentum (SGDM) optimizer too minimize (1), where the learning rate is set to 0.001 and the batch size is set to 30. The DL models we used were initially trained on the ImageNet database. In addition, due to the class imbalance of the OHTS dataset, we selected the best parameters of each DL model based on its achieved F-scores on the validation set, as this metric is better to use for seeking a balance between precision and recall, especially when the class distribution is uneven. Furthermore, we adopt the early stopping mechanism on the validation set to avoid over-fitting, where the tolerance is 5 epochs.

eFigure 1. Examples of optic nerve head detection and fundus photograph cropping



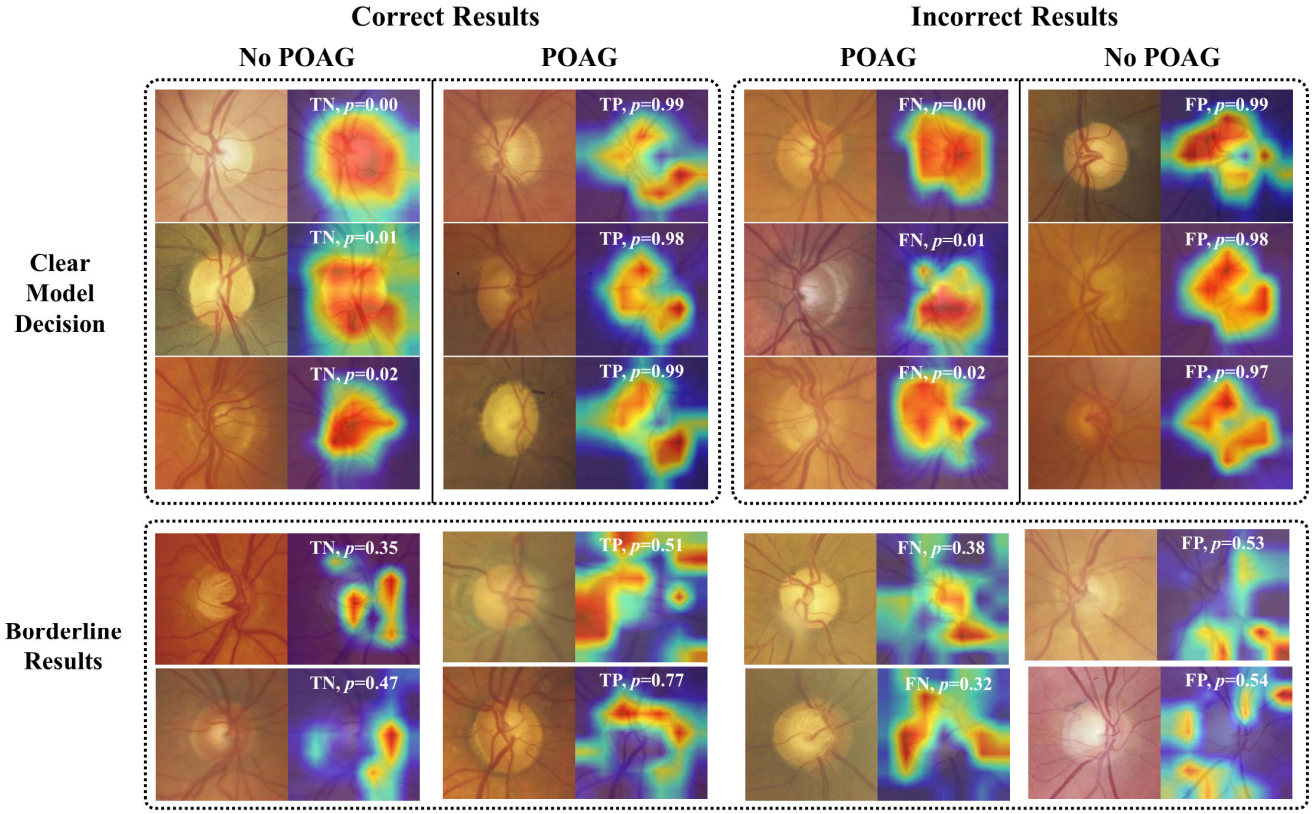
Examples of optic nerve head (ONH) detection and fundus photograph cropping: (a) raw fundus photographs; (b) ONHs (in red) detected by our trained DeepLabv3+ model; (c) cropped regions centered on the ONHs. The Ocular Hypertension Treatment Study (OHTS) dataset contains sequential (left two groups) and stereo (right two groups) fundus photographs.

eFigure 2. ResNet-50 architecture



ResNet-50 architecture, where Conv. represents a convolution layer and FC represents a fully-connected layer. A Softmax layer was added in the last to produce two scales indicating the probability distribution of healthy and primary open angle glaucoma (POAG) classes, respectively.

eFigure 3. Examples of deep learning model visualizations of the optic disc changes attributable to primary open-angle glaucoma (POAG) by the OHTS Endpoint Committee



Examples of deep learning model visualizations of the optic disc changes attributable to primary open-angle glaucoma (POAG) by the OHTS Endpoint Committee. Both clear model decisions (top) and borderline results (bottom) are shown. These results suggest that the region within the ONH had the greatest impact on clear model decisions (*p*, the probability of POAG estimated by the model, between 0.0 and 0.1 or between 0.9 and 1.0). The borderline results (*p* between 0.3 and 0.7) were less consistent with respect to the location on the disc in which the model based its decisions. TP refers to true positive, TN refers to true negative, FP refers to false positive, FN refers to false negative.

eTable 1. Time elapsed between the earliest false-positive photograph for each case and later OHTS Endpoint Committee determination of POAG

Ground Truth Determined by	POAG Detection Modality	Number of false positive (FP) results Participants (Eyes)/Visits	Mean (95% CI) number of years between model first false positive results and OHTS Endpoint Committee detection of primary open angle glaucoma (POAG)
Endpoint Committee	Optic Disc Photograph and/or Visual Field	31 (38) /198	5.2 (4.1, 6.3)
	Optic Disc Photograph	27 (32) /158	4.5 (3.4, 5.6)
	Visual Field	22 (26) /125	4.6 (3.3 , 5.8)

eTable 2. Diagnostic accuracy of deep learning model performance in identifying POAG using only 3 randomly selected visits/images in the test set

Ground Truth Determined by	POAG Detection Modality	POAG (n)	Area Under the Receiver Operating Characteristic Curve (95% CI)		Sensitivity at Specificity of:			
		Participants (Eyes)/Visits	All eyes	Early glaucoma VF MD ≥ -6 dB	80%	85%	90%	95%
Endpoint Committee	Optic Disc Photograph and/or Visual Field	47 (60) /102	0.89 (0.84, 0.94)	0.87 (0.79, 0.92)	0.83	0.81	0.69	0.56
	Optic Disc Photograph	36 (46) / 76	0.91 (0.85, 0.95)	0.91 (0.87, 0.95)	0.86	0.83	0.76	0.57
	Visual Field	31 (35) / 61	0.85 (0.75, 0.92)	0.84 (0.72, 0.93)	0.74	0.70	0.69	0.48
Reading Centers	Optic Disc Photograph	49 (69) / 93	0.89 (0.83, 0.93)	0.89 (0.83, 0.93)	0.85	0.77	0.73	0.55
	Visual Field	42 (49) / 76	0.83 (0.75, 0.89)	0.81 (0.71, 0.89)	0.68	0.63	0.61	0.49