Detecting Glaucoma in the Ocular Hypertension Treatment Study Using Deep Learning: Implications for clinical trial endpoints

Rui Fan^{1,7*}, Christopher Bowd^{1*}, Mark Christopher¹, Nicole Brye¹, James A. Proudfoot¹, Jasmin Rezapour^{1,2}, Akram Belghith¹, Michael H. Goldbaum¹, Benton Chuter¹, Christopher A. Girkin³, Massimo A. Fazio^{1,3,4}, Jeffrey M. Liebmann⁵, Robert N. Weinreb¹, Mae O. Gordon⁶, Michael A. Kass⁶, David Kriegman⁷, and Linda M. Zangwill¹

1. Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, CA, United States.
2. Department of Ophthalmology, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz, Rheinland-Pfalz, Germany.
3. Department of Ophthalmology, School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
4. Department of Biomedical Engineering, School of Engineering, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
5. Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Medical Center, New York, New York, United States.
6. Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, Washington University, St. Louis, Mo. United States.
7. Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, United States.

* These two authors contributed equally to this work.

Grant support:

National Eye Institute R01EY029058, R21EY027945, K99EY030942, R01EY011008, R01EY19869, R01EY027510, R01EY026574, Core Grant P30EY022589; the National Center on Minority Health and Health Disparities; Horncrest Foundation; awards to the Department of Ophthalmology and Visual Sciences at Washington University (NIH grants EY09341, EY09307 and NIH Vision Core Grant P30EY02687); Merck Research Laboratories; Pfizer, Inc.; White House Station (New Jersey); an unrestricted grant from Research to Prevent Blindness, Inc. (New York, NY); German Research Foundation research fellowship grant (RE 4155/1-1); and a German Ophthalmological Society Grant.

Financial disclosures:

- C.A.Girkin: Heidelberg Engineering, GmbH (F); National Eye Institute: (F), Amydis (C);
- M.A.Fazio: Heidelberg Engineering, GmbH (F), Equinox Ophthalmic (F), National Eye Institute (F);
- J.M. Liebmann: Alcon (C), Allergan (C), Bausch & Lomb (C, F), Carl Zeiss Meditec (C, F), Heidelberg Engineering (C, F), Optovue (F), Reichert (C, F), Topcon (F), Valeant Pharmaceuticals (C);
- R.N.Weinreb: Toromedes, Carl Zeiss Meditec (P, F), Allergan (C), Heidelberg Engineering (F), Konan Medical (F), Optovue (F), Centervue (F), Bausch & Lomb (F), Topcon (C, F), National Eye Institute (F), Equinox (C);
- L.M.Zangwill: Heidelberg Engineering GmbH (F), Optovue Inc. (F), National Eye Institute (F), Carl Zeiss Meditec Inc. (F, P), Topcon Medical Systems Inc. (F);
- None: R.Fan, C.Bowd, M.Christopher, N.Brye, J.A.Proudfoot, J.Rezapour, A.Belghith, B.Chuter, M.A.Kass, D.Kriegman.

ABSTRACT

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

Study Design: Cohort study.

Participants: 1,636 OHTS participants with a mean follow-up (range) of 10.7 (0.0, 14.3) years.

Methods: 66,715 photographs from 3,272 eyes were used to train and test a ResNet-50 model to detect the OHTS Endpoint Committee POAG determination based on optic disc (n=287 eyes, 3,502 photographs) and/or visual field (n=198 eyes, 2,300 visual fields) changes. OHTS training, validation and testing sets were randomly determined using an 85-5-10 percentage split by subject. Three independent test sets were used to estimate the generalizability of the model: UCSD Diagnostic Innovations in Glaucoma Study (DIGS, USA), ACRIMA (Spain) and Large-scale Attention-based Glaucoma (LAG, China).

Main Outcome 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 at 90% specificity was used to determine whether the DL model detected POAG before the Endpoint Committee POAG determination.

Results: The DL model achieved an AUROC (95% CI) of 0.88 (0.82, 0.92) for the overall OHTS POAG endpoint. For the OHTS endpoints based on optic disc changes or visual field changes, AUROCs were 0.91 (0.88, 0.94) and 0.86 (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 (19.1%) compared to eyes that did not develop POAG (7.3%) during their OHTS follow-up. The diagnostic accuracy of the DL model developed on the OHTS optic disc endpoint applied to 3 independent datasets was lower with AUROC ranging from 0.74 to 0.79.

Conclusions: High diagnostic accuracy of the current model suggests that DL can be used to automate the determination of POAG 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.

When planning clinical trials for assessing the efficacy of disease treatments such as medical or surgical intervention, the most important consideration arguably is the primary endpoint. The primary endpoint serves as an indicator of the success of the clinical trial as it represents success or failure of the treatment being assessed.

The Ocular Hypertension Treatment Study^{1,2} (OHTS) began as a large 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 ocular hypertensive eyes. In the OHTS, the primary endpoint was the development of POAG in one or both patient eyes defined as a reproducible clinically significant optic disc changes pathognomic for POAG or a reproducible glaucomatous visual field (VF) defect.³ The assessment of optic disc and VF changes from baseline were determined by masked study-certified readers at the independent Optic Disc Reading Center (ODRC) and VF Reading Center (VFRC). The final attribution to POAG was decided by a three-member masked Endpoint Committee of glaucoma experts who reviewed both the photographs and VFs to determine whether observed changes were due to POAG or another disease that can gradually affect appearance of the optic disc or visual function (*e.g.*, macular degeneration, diabetic retinopathy, ischemic optic neuropathy).

The use of two independent reading centers and a masked Endpoint Committee of 3 glaucoma experts is a demanding, laborious and complicated task. In addition, as agreement among the Endpoint Committee member's assessment was required, several consensus grading sessions were necessary before a final endpoint was determined. The three committee members reached unanimity in 61% of the endpoints in the first round of masked 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 The use of independent reading centers and a three-person, three-round consensus process endpoint determination of POAG was employed to ensure high specificity; appropriate for the OHTS, but not necessarily needed for all clinical trials.

Recent improvements in machine learning methods have allowed automated detection of glaucoma (and other eye diseases) that could be useful for automating endpoint determination in clinical trials.⁵ Specifically, machine learning endpoints have the potential to reduce the need for manual assessment, thereby improving the reproducibility of the endpoint determinations. For instance, deep learning (DL) approaches, including deep convolutional neural networks (DCNNs), have been employed to classify fundus photographs from glaucoma eyes and to detect glaucoma and estimate structural and visual field defects in those eyes.⁶⁻¹³ Besides the increased consistency and potential cost saving of automation, 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 cut-offs.

The current study assessed the automated diagnosis of POAG by DL algorithms trained and validated on fundus stereophotographs from the OHTS to determine the classification accuracy and generalizability in independent samples from the OHTS and three external independent test sets. We hypothesized that DL models trained on images from the OHTS would successfully classify eyes as POAG or healthy in independent test sets at an acceptable level, suggesting that automated classification can supplant the need for multi-tiered expert assessment of optic disc images in clinical trials (*i.e.*, can function as an automated surrogate for multi-layered reading center and/or endpoint committee assessment). We also compared results from models trained on the OHTS VFRC and ODRC POAG assessment to describe the relative effectiveness of each stage of photograph and

visual field classification in OHTS to assess the accuracy for detecting conversion to POAG in ocular hypertensive eyes.

Methods

Data Collection

The OHTS^{1,2} was initiated in 1994 and is the first large randomized clinical trial to document the safety and efficacy of topical ocular hypertensive medication in preventing/delaying the onset of visual field and/or optic nerve damage in subjects with ocular hypertension at moderate risk for developing POAG. Details of the study methods have been reported previously.^{1,2} The OHTS recruited 1636 ocular hypertensive participants with elevated intraocular pressure (IOP) from 22 sites. Each participant was seen twice a year for Humphrey 30-2 visual field (VF) testing and once a year for stereoscopic optic nerve head (ONH) photographs. The demographic and clinical characteristics included age, ethnicity, gender, IOP, central corneal thickness, and refractive status. At study entry, all participants were required to have normal-appearing optic nerve heads based on review of stereoscopic optic discs and visual fields as determined by the ODRC and VFRC. After each visit, the ODRC compared the baseline test to the follow-up test to determine if there was evidence of glaucomatous changes. Specifically, if two consecutive sets of ONH photographs demonstrated change from the baseline as determined by the ODRC, the case was reviewed by the 3 masked glaucoma specialist members of the Endpoint Committee. Similarly, if the VFRC determined that three consecutive sets of VFs were abnormal, then the case was reviewed by the Endpoint Committee. Each member of the Endpoint Committee independently reviewed the subject's medical history and compared baseline and follow-up VFs and ONH photographs to determine whether the visual fields were due to POAG and whether the changes in the ONH photographs were clinically significant and due to POAG. The advantages of utilizing an Endpoint Committee in the OHTS have recently been reported.⁴ In brief, and most importantly, using an Endpoint Committee had a significant effect on the accuracy of POAG incidence rate with 16.3% of study participants reaching an unadjudicated all-cause endpoint but only 9.5% of participants developing a POAG Endpoint Committee adjudicated endpoint. As treatment is unlikely to affect a non-POAG study participant, removal of these unadjudicated, all-cause endpoints led to a more accurate estimate of the efficacy of treatment; treatment reduced the estimates of a POAG adjudicated endpoint by 56% (relative risk of 0.44), while it reduced all-cause endpoints by 33% (relative risk of 0.67).

For this report, we utilized OHTS ONH photographs collected during the randomized clinical OHTS Phase 1 (1994-2002) and the longitudinal follow-up OHTS Phase 2 (2002-2009) to determine whether deep learning algorithms can accurately classify eyes based on optic disc changes and visual field changes identified by the ODRC, VFRC and Endpoint Committees as the ground truth. All photographs, regardless of quality were included. Specifically, we trained 5 deep learning algorithms to identify each of the following five outcomes:

Endpoint Committee Determination

- ENPOAGDISC: Optic disc changes attributable to POAG by Endpoint Committee.
- ENPOAGVF: VF changes attributable to POAG by Endpoint Committee.
- ENPOAGANY: Optic disc or VF changes attributable to POAG by Endpoint Committee.

Reading Center Determination

- RCPOAGDISC: Optic disc changes attributable to POAG by ODRC.

- RCPOAGVF: VF changes attributable to POAG by VFRC.

Any photograph taken on or after the initial classification of POAG by the Endpoint Committee determinations was included as POAG for the 3 ENDPOAGDISC, ENDPOAGVF and ENPOAGANY DL models. For the Reading Center determinations, any photograph taken on the visit determined by the ODRC or VFRC as POAG was considered POAG. In contrast to the ground truth used in the DL models for the 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 Endpoint Committee.

Dataset Preparation

Because the 22 OHTS sites used different fundus cameras, resulting in inherent variability in image quality and resolution, the training of deep learning models was much more challenging than if photographs came from a single site or camera. To this end, prior to training DL models for POAG determination, we first extracted a region centered on the ONH from each raw fundus photograph using a semantic segmentation DL model, DeepLabv3+¹⁴ (with a ResNet-18 backbone network trained for ONH extraction). A square region surrounding the extracted ONH was then automatically cropped from each image for input in the DL model. Each cropped image was manually reviewed by a single reviewer to ensure that it was correctly centered on the ONH (Figure 1). The cropped fundus images were then resized to 224 x 224 pixels.

Figure 1. Examples of ONH detection and fundus photograph cropping: (a) raw fundus photographs; (b) ONHs (in red) detected by our trained DeepLabv3+⁴ model; (c) cropped region centered on the ONH. The OHTS dataset contains sequential (left two groups) and stereo (right two groups) fundus photographs.

Figure 2. ResNet-50⁶ 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 GON classes, respectively.

Data Augmentation

Several data augmentation strategies were applied to increase the amount and type of variation of 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 and rotation, in which the center of the ONH region of each photograph was randomly perturbed by a small amount to reflect the common situation in which ONH photographs are not always well-centered. Each augmented image was assigned the same label (healthy or POAG) as the original input image from which it was derived.⁹

Deep Learning Models

In our experiments, a ResNet-50¹⁵ model pretrained on the ImageNet¹⁶ database was fine-tuned for POAG detection. As illustrated in Figure 2, we modified the FC layer of ResNet-50¹⁵ so that it could output two scalars indicating the probability distribution of healthy and GON classes, respectively, with respect to a given task, such as ENPOAGDISC, ENPOAGVF, ENPOAGANY, RCPOAGDISC, or RCPOAGVF classifications.

Model Training and Selection

The OHTS dataset was divided into training, validation, and testing sets, using an 85-5-10 percentage split by participant, so that all images from one participant were included in the same partition (training, validation, or testing). This ensured that the images from a given participant in the validation/testing set had not been encountered by the model during the training process. Each of the 5 models utilized the exact same training, validation, and test sets.

The DL model training was carried out on two NVIDIA Geforce RTX 2080 Super GPUs, each of which has an 8 GB GDDR6 memory. Because the OHTS dataset is imbalanced with the majority of eyes not developing POAG (1299/1636 (79.4%)), we implemented additional class weights into the loss function (see Supplement for details).

2.5 Performance Evaluation

The trained DL model was evaluated on the OHTS test set as well as three additional independent test datasets of optic disc photographs labeled as glaucoma or healthy: (a) ACRIMA¹⁷, (b) LAG¹⁸, and (c) DIGS/ADAGES.¹⁹ Performance in distinguishing between healthy and glaucoma eyes was evaluated using sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC). To evaluate the models in detecting early glaucoma, we conducted a subset analysis in eyes with a VF mean deviation (MD) better than -6 dB. For each model, AUROC scores for classifying healthy versus all GON eyes and mild GON eyes were also computed. The AUROC scores of different models were statistically compared using a clustered bootstrap approach to address the correlation between eyes and between visits.²⁰ To help evaluate clinical utility, the sensitivity of each model at four fixed levels of specificity (80%, 85%, 90%, and 95%) was evaluated. Furthermore, Grad-CAM++²¹, a common network explanation and image classification visualization technique, was employed to help understand the model decision-making process.

Results

A total of 1,636 OHTS 1,2 participants (3,272 eyes) provided 66,715 fundus images for DL model training, validation and testing (Table 1). The participants had a mean age of 55.9 years at study entry, and approximately 25% were of African Descent.

Table 1. Characteristics of OHTS training, validation and test Sets: n (%) or mean (95% CI)

	Training	Validation	Test	p-value
Number of Subjects (Eyes)	1147 (2294)	167 (334)	322 (644)	
Number of Eye Visits	25899	3745	7088	
Age (Years)	57.2 (56.6, 57.8)	57.1 (55.6, 58.7)	57.2 (56.1, 58.2)	0.995
Race (% European Descent)	75.9%	71.9%	73.9%	0.433
Baseline VF MD (dB)	-0.02 (-0.10, 0.06)	0.02 (-0.20, 0.23)	-0.12 (-0.27, 0.04)	0.504
No Glaucoma				
Number of Subjects (Eyes)	994(918)	131 (94)	250 (240)	
Number of Eye Visits	24463	3503	6675	
VF MD (dB)	-0.18 (-0.25, -0.12)	-0.09 (-0.25, 0.07)	-0.16 (-0.28, -0.05)	0.531
Developed a POAG Endpoint by Visual Field or Photo				
Number of Subjects (Eyes)	229 (286)	36 (44)	72 (96)	
Number of Eye Visits	1390	44	96	

Table 2. Diagnostic Accuracy of Deep Learning Model Performance in Identifying POAG by the Ocular Hypertensive Treatment Study Endpoint Committee and Optic Disc and Visual Field Reading Centers

Ground Truth Determined by	POAG Detection Modality	POAG (n)	Area Under the Receiver Operating Characteristic Curve (95% CI)		Sensitivity at Specificity of:			
		Subjects (Eyes)/Visits	All eyes	Early glaucoma VF MD $\geq$ -6 dB	80%	85%	90%	95%
Endpoint Committee	Optic Disc	52 (71) / 352	0.88 (0.82, 0.92)	0.86 (0.79, 0.91)	0.79	0.75	0.65	0.50
	Photograph and/or Visual Field							
	Optic Disc Photograph	41 (56) / 262	0.91 (0.88, 0.94)	0.90 (0.87, 0.94)	0.84	0.79	0.69	0.52
	Visual Field	35 (41) / 195	0.86 (0.76, 0.93)	0.83 (0.70, 0.91)	0.74	0.69	0.63	0.47
Reading Centers	Optic Disc Photograph	60 (77) / 318	0.89 (0.85, 0.92)	0.87 (0.83, 0.91)	0.80	0.73	0.68	0.51
	Visual Field	61 (78) / 242	0.83 (0.76, 0.88)	0.80 (0.72, 0.86)	0.64	0.58	0.51	0.42

Identifying POAG Based on Optic Disc or Visual Field Changes provided by the OHTS Endpoint Committee or Reading Center

The fundus photograph-based DL models detected conversion to POAG with good accuracy. Specifically, the best diagnostic accuracy of the deep learning model was achieved for the Endpoint Committee POAG attribution based on optic disc changes (ENPOAGDISC) followed by either optic disc or VF changes (ENDPOAGANY) and VF only change (ENPOAGVF), with AUROCs (95% CI) of 0.91 (0.88, 0.94), 0.88 (0.82, 0.92), 0.86 (0.76, 0.93), respectively (Table 2). The diagnostic accuracy (AUROCs (95% CI)) of the Reading Center's POAG attribution by optic disc photographs and by VFs was 0.89 (0.85, 0.92), and 0.83 (0.76, 0.88), respectively. The diagnostic accuracy of detecting early POAG (VF MD $\geq$ -6 dB) was also generally high for Endpoint Committee determinations the OHTS ODRC POAG determinations with AUROCs ranging from 0.82 to 0.89. Model performance was lower for the VFRC POAG determination of early glaucoma (AUROC = 0.78).

The false-positive (FP) rates (at 90% specificity) were higher in photographs of hypertensive eyes acquired before they reached a POAG endpoint compared to the false positive rate of hypertensive eyes that did not develop POAG (25.4% vs. 8.0% for the overall OHTS Endpoint Committee (ENPOAGANY), 22.0% vs. 6.5% for the OHTS endpoints based on optic disc changes (ENPOAGDISC), and 17.7% vs. 4.2% for the OHTS endpoints based on visual field changes (ENDPOAGVF) for eyes that did vs. did not develop POAG, respectively). Figure 3 illustrates the increasing predicted probability of observing a false positive over the course of OHTS I and II in eyes that eventually developed POAG compared to eyes that did not.

Figure 3. These figures illustrate the higher probability of observing a false positive throughout the course of Ocular Hypertension Treatment Study I and II in eyes that developed a POAG endpoint (converting eyes) compared to those that did not (non-converting eyes). The results are calculated at 90% specificity for top: POAG endpoint committee determination based on optic disc, middle: POAG endpoint committee determination based on visual fields, and bottom: POAG endpoint committee determination based on either optic disc or visual field.

Table 3 shows the diagnostic accuracy of the DL models trained based on the OHTS optic disc endpoints on the three independent clinical datasets, which was lower compared to the OHTS test set (AUROC (95% CI) for DIGS/ADAGES of 0.74 (0.69, 0.79), ACRIMA of 0.74 (0.70, 0.77) and LAG of 0.79 (0.78, 0.81)).

Table 3. Diagnostic Accuracy in Independent Test Sets

	Test set size		AUROC (95% CI)		Sensitivity at Specificity of:			
	Healthy	GON	All eyes	Mild glaucoma VF MD $\geq$ -6 dB	80%	85%	90%	95%
DIGS	5184	4289	0.744 (0.695, 0.789)	0.709 (0.651, 0.763)	0.59	0.52	0.43	0.30
ACRIMA	309	396	0.736 (0.701, 0.773)	—	0.55	0.46	0.38	0.29
LAG	3143	1711	0.794 (0.780, 0.807)	—	0.66	0.59	0.53	0.42

Figure 4. Examples of deep learning model visualizations of the optic disc changes attributable to primary open-angle glaucoma by Endpoint Committee. Both clear model decisions (top) and borderline results (bottom) are shown. These results suggest that the region within the optic nerve head had the greatest impact on clear model decisions (p 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=true positive, TN=true negative, FP=false positive, FN=false negative

Model Visualization

We utilized Grad-CAM++²¹ to determine which regions of the photographs were saliently important for the deep learning models decision making (Figure 4). These results suggest that the region within the optic nerve head had the greatest 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 and GON eyes, in both the 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 optic nerve head regions.

Discussion

These results suggest that the DL models can provide good accuracy for the determination of glaucomatous change based on the optic disc (AUROC=0.91), visual field (AUROC=0.86) or either (AUROC=0.88) by OHTS Endpoint Committee members and/or Reading Centers (ODRC AUROC=0.89, VFRC AUROC=0.83). Given the challenge of POAG determination by reading centers and endpoint committees because of its subjective nature, these results suggest a role for AI in improving the accuracy and consistency of the process, at lower cost.⁵ Moreover, the specificity of the diagnostic classification can be adjusted to reflect clinical trials that are designed with high specificity or high sensitivity in mind by adjusting the cut-off probably accordingly. In this study, these results are presented by the sensitivities at various levels of specificity.

Specifically, the DL models tested on the 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 due in part by design, as the reading center personnel were masked to review either photographs or visual fields alone and did not have other clinical information necessary to determine whether changes could be attributable to POAG or to other causes. Furthermore, models using optic disc changes as determined by ODRC as the ground truth performed better than models using visual field changes as determined by VFRC as the ground truth for training deep learning models to identify glaucoma images, as expected. The high diagnostic accuracy of the current deep learning model suggests that deep learning 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 to non-POAG eyes (Figure 3, Table 3) suggests that deep learning models detected POAG in some eyes earlier than the OHTS POAG Endpoint Committee or Reading Centers. These false positives likely were true positives detecting disease related change earlier in OHT eyes; this was in part a result of the OHTS study design that emphasized high specificity for glaucomatous determination.²² The CNNs used herein provide a probability of glaucoma as output allowing changes in sensitivity and specificity to desired levels by adjusting the cut-offs used to define POAG. This makes deep-learning models adaptable to different study goals. For instance, one may wish to relax 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 DL models trained and validated on OHTS data to several independent data sets, an important concern in assessing model

usefulness. The current deep learning models showed somewhat better generalizability to the LAG dataset than the DIGS/ADAGES and ACRIMA datasets. Poorer performance in independent test sets likely is affected by differences in ground truth determination among test sets as well as differences in study populations. There is considerable evidence that assessment of optic disc photographs for glaucoma determination is highly variable, even among glaucoma experts.²³⁻²⁴⁻²⁶ 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 independent test datasets. Differences in labeling and study populations have been shown to affect deep learning 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 determination of POAG by 3 glaucoma specialist members of the Endpoint Committee, there was initial consensus in only 61% of eyes evaluated; 39% of eyes required regrading and/or discussion to reach consensus on POAG status.

The current study also investigated the relative performance of DL models in a subset of early glaucoma eyes with MD better than -6.0 dB. Although AUROCs were up to 0.03 lower, the general relative pattern of performance was similar to that observed when all glaucoma eyes were included; AUROCs generally were greater for the Endpoint Committee POAG determination tests sets and AUROCs were greater for ODRC POAG determination than VFRC determination.

A recent study by Thakur and colleagues²⁷ also used a DCNN to detect glaucoma onset in fundus photographs from the OHTS reported somewhat better results than those reported herein. For instance, the reported AUROC for classifying non-glaucomatous and glaucomatous eyes based on classifications from the independent endpoint committee (ENPOAGANY here) was 0.95 compared to 0.87 in the current study. This discrepancy in classification performance may be due in part to the fact that these authors determined that approximately 24% of the available fundus photographs contained extreme artifacts and were excluded from their study. In contrast, we included all available OHTS fundus images to better reflect clinical practice. Thus, decreased performance in our model could be due in part to misclassification of less-than-ideal images. An additional reason for the discrepancy in results likely is that different images were included in the DL model training, validation, and tests sets among studies, which cannot be avoided. To address the former possibility, we employed an objective deep learning algorithm to assign a quality metric shown to improve classification success in a subset of OHTS fundus photographs²⁸. Including only the highest quality images (approximately 73% of the test eyes contributing at least one photograph to the analysis) in a post hoc analysis, the algorithm increased the model accuracy (AUROC) from 0.86 to 0.90 for ENPOAGVF, and 0.83 to 0.87 for RCPOAGVF. No improvement was found for ENPOAGDISC, ENPOAGANY and RCPOAGDISC.

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 dataset. To address this common problem, we implemented additional class weights into the model. In addition, the class activation maps identified regions on the optic nerve head and neuroretinal rim for

clear model decisions, both accurate and inaccurate. The areas utilized by the model in the borderline results (those photographs without a clear model decision on whether the eye was POAG or not) were less focused on the optic nerve head and the generally accepted parapapillary regions where characteristics of POAG are most frequently observed (Figure 4, bottom panel). 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 depending on the endpoints used, as described above.

Conclusion

In conclusion, the high diagnostic accuracy and generalizability of the current deep learning model suggest that DL can be used to automate the determination of POAG for clinical trials and management. We believe integration of DCNN analyses of photographic images and other test results in clinical trials could reduce the cost and improve the consistency and accuracy of endpoint assessments. If not replacing these traditional clinical trial endpoint scenarios, DCNN analyses could decrease the personnel required to complete the task. Moreover, given the performance of the DCNN analysis in comparison with expert human observation, this approach may be promising to provide diagnostic assistance in the clinical setting.

References

1. Gordon MO, Kass MA. The Ocular Hypertension Study Group (OHTS), Manual of Procedures. In: Washington, DC National Technical Information Service; 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.
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; discussion 829-730.
4. Gordon MO, Higginbotham EJ, Heuer DK, et al. Assessment of the Impact of an Endpoint Committee in the Ocular Hypertension Treatment Study. *Am J Ophthalmol*. 2019;199:193-199.
5. Lee CS, Lee AY. How artificial intelligence can transform randomized controlled trials. *Trans Vis Sci Tech*. 2020;9(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).
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.
8. Chai YD, Liu HY, Xu J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowl-Based Syst*. 2018;161:147-156.
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.
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.
11. Jammal AA, Thompson AC, Mariottoni EB, et al. Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gratings for Detecting Glaucoma on Fundus Photographs. *American journal of ophthalmology*. 2020;211:123-131.
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-Uk*. 2018;8.
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.
14. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable for convolution for semantic image segmentation. Paper presented at: Proceedings of the European Conference on Computer Vision (ECCV)2018.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
16. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Paper presented at: 2009 IEEE conference on computer vision and pattern recognition2009.

17. Diaz-Pinto A, Morales S, Naranjo V, Kohler T, Mossi JM, Navea A. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online*. 2019;18(1):29.
18. Li L, Xu M, Wang X, Jiang L, Liu H. Attention based glaucoma detection: A large-scale database and cnn model. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019.
19. Sample PA, Girkin CA, Zangwill LM, et al. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127(9):1136-1145.
20. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997;53(2):567-578.
21. Chattopadhyay 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 (WACV) 2018.
22. Gordon MO, Higginbotham EJ, Heuer DK, et al. Assessment of the Impact of an Endpoint Committee in the Ocular Hypertension Treatment Study. *American journal of ophthalmology*. 2019;199:193-199.
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.
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.
26. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99(2):215-221.
27. Thakur A, Goldbaum M, Yousefi S. Predicting Glaucoma before Onset Using Deep Learning. *Ophthalmol Glaucoma*. 2020;3(4):262-268.
28. Chuter B, Christopher M, Fan R, Belghith A, Bowd C, Brye N, Proudfoot J.A, Rezapour J, Fazio M.A, Goldbaum MH, Weinreb RN, Girkin, CA, Liebmann JM, De Moraes C, Zangwill LM. 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(8):1016 e-ARVO abstract

Supplemental Materials

Supplement highlights the conventional cross entropy loss L as follows:

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

where y denotes the class label ($y = 0$ for healthy images and $y = 1$ for GON images), p represents GON prediction probability output by the network, and n_0 and n_1 denote the number of healthy and GON 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 DCNNs we used were initially trained on the ImageNet¹⁶ database. In addition, due to the class imbalance of the OHTS dataset, we selected the best parameters of each DCNN 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.