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

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

12-07-2021 / 19-01-2022

CITATION

Fan, Rui; Bowd, Christopher; Christopher, Mark; Brye, Nicole; Proudfoot, James A.; Rezapour, Jasmin; et al. (2021): Detecting Glaucoma in the Ocular Hypertension Treatment Study Using Deep Learning: Implications for clinical trial endpoints. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.14959947.v2>

DOI

[10.36227/techrxiv.14959947.v2](https://doi.org/10.36227/techrxiv.14959947.v2)

Detecting Glaucoma in the Ocular Hypertension Study Using Deep Learning: Implications for Clinical Trial Endpoints

Rui Fan PhD^{1,2,3}*, Christopher Bowd PhD¹, Mark Christopher PhD¹, Nicole Brye¹, James A. Proudfoot MS¹, Jasmin Rezapour MD^{1,4}, Akram Belghith PhD¹, Michael H. Goldbaum MD¹, Benton Chuter MS¹, Christopher A. Girkin MD⁵, Massimo A. Fazio PhD^{1,5,6}, Jeffrey M. Liebmann MD⁷, Robert N. Weinreb MD¹, Mae O. Gordon PhD⁸, Michael A. Kass MD⁸, David Kriegman PhD², and Linda M. Zangwill PhD^{1*}

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 Computer Science and Engineering, University of California San Diego, La Jolla, CA, United States.
3. Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China.
4. Department of Ophthalmology, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz, Rheinland-Pfalz, Germany.
5. Department of Ophthalmology, School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
6. Department of Biomedical Engineering, School of Engineering, The University of Alabama at Birmingham, Birmingham, Alabama, United States.
7. Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Columbia University Medical Center, New York, New York, United States.
8. Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, Washington University, St. Louis, Mo. United States.

***Rui Fan and Christopher Bowd contributed equally to this work.**
Corresponding Author Linda M. Zangwill

9500 Gilman Drive

La Jolla, CA 92093-0946

Shiley Eye Institute/Hamilton Glaucoma Center

Viterbi Family Department of Ophthalmology

University of California, San Diego

T: (858) 534-7686

Email: lzangwill@health.ucsd.edu

Key Points

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

Finding: In this cohort study of 1,636 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 endpoint determination and improve the efficiency of POAG clinical trials.

ABSTRACT

Importance: Automated deep learning (DL) analyses of fundus photographs potentially can reduce the cost and improve the efficiency of reading center assessment of endpoints 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).

Study Design: Diagnostic/prognostic Study

Setting and Participants: 1,636 OHTS participants from 22 sites with a mean follow-up (range) of 10.7 (0.0, 14.3) years. 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. Three independent test sets were used to evaluate the generalizability of the model.

Main Outcome 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 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 follow-up. The diagnostic accuracy of the DL model developed on the optic disc endpoint applied to 3 independent datasets was lower with AUROC ranging from 0.74 to 0.79.

Conclusion 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 Endpoint Committee, reflecting the OHTS design that emphasized a high specificity for POAG determination by requiring a clinically significant change from baseline.

In clinical trials 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 clinical trial success as it measures efficacy of the assessed treatment.

The Ocular Hypertension Treatment Study^{1,2} (OHTS) 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 ocular hypertensive eyes. In OHTS, the primary endpoint 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 VF Reading Center (VFRC). The final POAG determination was decided by a three-member Endpoint Committee of glaucoma experts who reviewed the photographs and VFs to determine whether observed changes were due to POAG or another disease (*e.g.* ischemic optic neuropathy).

The use of two reading centers and a masked Endpoint Committee 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 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 endpoint determination process 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 resulted in automated glaucoma detection methods 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, have been employed 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 cut-offs. An advantage of using the OHTS dataset 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 an 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 three external test sets. We hypothesized that DL models trained on OHTS photographs 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. 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 ocular hypertensive eyes.

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/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} The current report followed the Standards for Reporting Diagnostic accuracy studies (STARD) reporting guidelines.

The OHTS recruited 1,636 ocular hypertensive participants with elevated intraocular pressure (IOP) 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, ethnicity, self-reported sex, IOP, 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. 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 photographs demonstrated change from the baseline, the case was reviewed by the 3 masked glaucoma specialist Endpoint Committee members. Similarly, if the VFRC determined that three consecutive sets of VFs were abnormal, then the case was reviewed by the Endpoint Committee. Each Endpoint Committee member independently reviewed the participant's medical history and compared baseline and 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 utilizing an Endpoint Committee in the OHTS have recently been reported.⁴ In brief, 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 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 photographs collected during the randomized clinical OHTS Phase 1 (1994-2002) and the longitudinal follow-up OHTS Phase 2 (2002-2009) to determine whether DL algorithms can accurately classify eyes based on optic disc changes and VF changes identified by the ODRC, VFRC and 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 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.

Photographs taken on or after the initial classification of POAG by the Endpoint Committee determinations were included as POAG for the ENDPOAGDISC, ENDPOAGVF and ENPOAGANY DL models. For the Reading Center determinations, photographs taken on the visit determined by the ODRC or VFRC as POAG were 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.

Deep Learning Models, Training and Selection

Details of data set preparation, including data augmentation are described in the Supplementary Materials (eMethods 1).

In our experiments, a ResNet-50¹⁴ model pretrained on the ImageNet¹⁵ database was fine-tuned for POAG detection. As illustrated in Supplement eFigure 2 , we modified the fully connected layer so that it could output two scalars indicating the probability of healthy and POAG classes, with respect to ENPOAGDISC, ENPOAGVF, ENPOAGANY, RCPOAGDISC, or RCPOAGVF classifications. Our source code and well-trained models are available upon request at: <https://sites.google.com/view/jamaoph>

The OHTS dataset was divided into training, validation, and test sets, using an 85-5-10 percentage split by participant, so that all images from one participant were included in the same partition. 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 with 8 GB GDDR6 memory. Because the OHTS dataset is imbalanced with most eyes not developing POAG (1299/1636 (79.4%)), we implemented additional class weights into the loss function (see Supplement eMethods 1 for details).

Performance Evaluation

The trained DL models were evaluated on the OHTS test set as well as three additional independent test datasets of optic disc photographs labeled as glaucoma or healthy: (a) ACRIMA¹⁶, (b) LAG¹⁷ and (c) Diagnostic Innovations in Glaucoma Study (DIGS)/African Descent and Glaucoma Evaluation Study (ADAGES)¹⁸. The ACRIMA public dataset provides cropped fundus photographs. We applied the same cropping strategy to the LAG and DIGS/ADAGES datasets that we applied to the OHTS photographs (eMethods 1). Because these datasets are only used to test the generalizability of our DL model, we did not apply flipping and rotation operations to the external independent datasets.

Performance in distinguishing between healthy and glaucoma eyes 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 four 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 along with recall (also known as sensitivity) which are particularly informative in imbalanced datasets.²⁰ Furthermore, Grad-CAM++²¹, a common network explanation and visualization technique, was employed to help understand the model decision-making process. Compared to 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.

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.

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) also was generally high for Endpoint Committee and the 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).

Given that the OHTS dataset 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, (false positives), than classifying the smaller number of POAG eyes as without 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 (FP) rates (at 90% specificity) were higher in photographs of ocular 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 VF changes [ENDPOAGVF] for eyes that did vs. did not develop POAG, respectively). Figure 1 illustrates the increasing 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. The mean (95% CI) time between the occurrence of the first false positive result and the development of POAG ranged from 4.5 (3.4, 5.6) years for optic disc photograph POAG endpoint (ENPOAGDISC) to 5.2 (4.1, 6.3) years for optic disc or VF POAG endpoint (ENPOAGVF) (see supplementary eTable1).

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]).

We utilized Grad-CAM++²¹ to determine which regions of the photographs were saliently important for the DL models' decision-making (See Supplemental eFigure 3). 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 and POAG eyes 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), VF (AUROC=0.86) or either (AUROC=0.88) 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 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, 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 ODRC as ground truth performed better than models trained using VF changes determined by 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 to non-POAG eyes (Figure 1, Table 3) suggests that DL 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 on average over 4 years earlier in ocular hypertensive eyes; this was, in part, a result of the OHTS study design that emphasized high specificity for glaucomatous determination.²² The models used here 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 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 datasets, an important step in assessing model performance. The current DL 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 among test sets in ground truth determination, types of camera, 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 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 external test datasets. 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 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.

As 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 (AUROC = 0.91 for ENPOAGDISC). Given that clinical trials require standardization of endpoints and few use Endpoint Committees for their review of Reading Center results, we believe DL can help provide standardized assessment across study centers, thereby increasing the efficiency and reducing the cost of clinical trials.

Thakur and colleagues²⁷ also used DL to detect glaucoma in OHTS fundus photographs and reported somewhat higher AUROC than the current study. Specifically, the AUROC for classifying non-glaucomatous and glaucomatous eyes based on the endpoint committee (ENPOAGANY) 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 they excluded approximately 24% of the available photographs due to extreme artifacts. In contrast, we included all available OHTS photographs to better reflect clinical practice. Thus, decreased performance in our model could be due in part 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. To address the former possibility, we employed an objective DL algorithm to assign a quality metric 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.

The current study also investigated the relative performance of DL models in a subset of early glaucoma eyes. Although AUROCs were up to 0.03 lower, the general 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 than VFRC POAG determination. We also investigated the model's accuracy if fewer visits were included in the test set as is often observed in other studies. Instead of using all OHTS photographs; using 3 randomly chosen visits per eye the results were similar to those obtained using all photographs with AUROC (95% CI) ranging from 0.83 (0.75, 0.89) for RCPOAGVF determination to 0.91 (0.85, 0.95) for ENPOAGDISC (See Supplemental eTable 2).

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. Second, the class activation maps identified regions on the ONH 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 ONH and the generally accepted parapapillary regions where characteristics

of POAG are most frequently observed (See Supplemental eFigure 3, bottom panel). 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. Finally, we cropped all photographs which may have reduced model performance if informative information is located in the peripheral retina.

Conclusion

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 endpoints 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 endpoint assessments. If not replacing these traditional clinical trial endpoint scenarios, DL analyses could decrease 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.

Table 1. Characteristics of the Ocular Hypertension Treatment Study training, validation, and test sets. (*n* (%) or mean (95% CI))

	Training	Validation	Test	<i>p</i> -value
Number of Participants (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
Sex (Female) (Male)	661 (57.6%) 486 (42.4%)	97 (58.1%) 70 (41.9%)	173 (53.7%) 149 (46.3%)	0.436
Self-reported Race European Descent African Descent	871 (75.9%) 276 (24.1%)	120 (71.9%) 47 (28.1%)	238 (73.9%) 84 (26.1%)	0.433
Baseline Visual Field Mean Deviation (MD) (dB)	-0.02 (-0.10, 0.06)	0.02 (-0.20, 0.23)	-0.12 (-0.27, 0.04)	0.504
Baseline Photograph Based Vertical Cup Disc Ratio	0.39 (0.38, 0.40)	0.36 (0.33, 0.40)	0.39 (0.37, 0.41)	0.280
No Glaucoma				
Number of Participants (Eyes)	994 (918)	131 (94)	250 (240)	
Number of Eye Visits	24463	3503	6675	
Visual Field MD (dB)	-0.18 (-0.25, -0.12)	-0.09 (-0.25, 0.07)	-0.16 (-0.28, -0.05)	0.531
Developed a Primary Open Angle Glaucoma (POAG) Endpoint by Visual Field or Photograph				
Number of Participants (Eyes)	229 (286)	36 (44)	72 (96)	
Number of Eye Visits	1390	44	96	

Table 2. Diagnostic Accuracy of DL Model Performance in Identifying Primary Open Angle Glaucoma 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:			
		Participants (Eyes)/Visits	All eyes	Early glaucoma VF MD $\geq$ -6 dB		80%	85%	90%	95%
Endpoint Committee	Optic Disc Photograph and/or Visual Field	52 (71) / 352	0.88 (0.82, 0.92)	0.86 (0.79, 0.91)	Sensitivity Precision	0.81 0.21	0.77 0.26	0.69 0.32	0.56 0.43
	Optic Disc Photograph	41 (56) / 262	0.91 (0.88, 0.94)	0.90 (0.87, 0.94)	Sensitivity Precision	0.86 0.17	0.81 0.21	0.73 0.26	0.56 0.35
	Visual Field	35 (41) / 195	0.86 (0.76, 0.93)	0.83 (0.70, 0.91)	Sensitivity Precision	0.80 0.14	0.76 0.17	0.69 0.22	0.55 0.31
Reading Centers	Optic Disc Photograph	60 (77) / 318	0.89 (0.85, 0.92)	0.87 (0.83, 0.91)	Sensitivity Precision	0.83 0.20	0.77 0.23	0.71 0.30	0.56 0.40
	Visual Field	61 (78) / 242	0.83 (0.76, 0.88)	0.80 (0.72, 0.86)	Sensitivity Precision	0.69 0.15	0.64 0.18	0.58 0.22	0.48 0.32

Table 3. Diagnostic Accuracy of the model developed using the Endpoint Committee determination of endpoints based on optic disc changes (ENDPOAGDISC) in Independent Test Sets*

	Country	Ground Truth Determination	Test set size (images)		AUROC (95% CI)		Sensitivity at Specificity of:			
			Healthy	Glaucoma	All eyes	Mild glaucoma VF MD $\geq$ -6 dB	80%	85%	90%	95%
Diagnostic Innovations in Glaucoma Study (DIGS) ¹⁸	United States	Agreement by two 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	One 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 visual fields	3143	1711	0.79 (0.78, 0.81)	Not available	0.66	0.59	0.53	0.42

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

enpoagany **enpoagdisc** **enpoagvf**

Figure Legend

Figure 1. 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.

Acknowledgements:

A. Funding:

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.

B. Role of the Funder/Sponsor: The funding organizations 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.

C. Conflict of Interest 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.

D. Author Contributions: Drs. Rui Fan and Linda Zangwill had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Acquisition, analysis, or interpretation of data: All authors

Drafting of the manuscript: All authors

Critical revision of the manuscript for important intellectual content: All authors

Statistical Analysis: Proudfoot and Brye

Obtained Funding: Zangwill

Administrative, technical, or material support: Zangwill

Supervision: Zangwill

E. Meeting Presentation: This manuscript was presented at ARVO 2021.

F. Non-author contributions to data collection, analysis or writing/editing assistance: NONE

REFERENCES

- 1 Gordon, M. O. & Kass, M. A. (Washington, DC National Technical Information Service, 1997).
- 2 Gordon, M. O. & Kass, M. A. The Ocular Hypertension Treatment Study: design and baseline description of the participants. *Arch Ophthalmol* **117**, 573-583, doi:10.1001/archopht.117.5.573 (1999).
- 3 Kass, M. A. *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* **120**, 701-713; discussion 829-730, doi:ecs20045 [pii] (2002).
- 4 Gordon, M. O. *et al.* Assessment of the Impact of an Endpoint Committee in the Ocular Hypertension Treatment Study. *American journal of ophthalmology* **199**, 193-199, doi:10.1016/j.ajo.2018.11.006 (2019).
- 5 Lee, C. S. & Lee, A. Y. How artificial intelligence can transform randomized controlled trials. *Trans Vis Sci Tech* **9**, doi:doi:<https://doi.org/10.1167/tvst.9.2.9> (2020).
- 6 Ahn, J. M. *et al.* A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PloS one* **13**, doi:ARTN e0207982
10.1371/journal.pone.0207982 (2018).
- 7 Asaoka, R. *et al.* Validation of a Deep Learning Model to Screen for Glaucoma Using Images from Different Fundus Cameras and Data Augmentation. *Ophthalmol Glaucoma* **2**, 224-231, doi:10.1016/j.ogla.2019.03.008 (2019).
- 8 Chai, Y. D., Liu, H. Y. & Xu, J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. *Knowl-Based Syst* **161**, 147-156, doi:10.1016/j.knosys.2018.07.043 (2018).
- 9 Christopher, M. *et al.* Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. *Sci Rep* **8**, 16685, doi:10.1038/s41598-018-35044-9 (2018).
- 10 Christopher, M. *et al.* Effects of Study Population, Labeling and Training on Glaucoma Detection Using Deep Learning Algorithms. *Transl Vis Sci Technol* **9**, 27, doi:10.1167/tvst.9.2.27 (2020).
- 11 Jammal, A. A. *et al.* Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gratings for Detecting Glaucoma on Fundus Photographs. *American journal of ophthalmology* **211**, 123-131, doi:10.1016/j.ajo.2019.11.006 (2020).
- 12 Shibata, N. *et al.* Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep-Uk* **8**, doi:ARTN 14665
10.1038/s41598-018-33013-w (2018).
- 13 Phene, S. *et al.* Deep Learning and Glaucoma Specialists The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* **126**, 1627-1639, doi:10.1016/j.ophtha.2019.07.024 (2019).
- 14 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770-778.
- 15 Deng, J. *et al.* in *2009 IEEE conference on computer vision and pattern recognition*. 248-255.
- 16 Diaz-Pinto, A. *et al.* CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed Eng Online* **18**, 29, doi:10.1186/s12938-019-0649-y (2019).

- 17 Li, L., Xu, M., Wang, X., Jiang, L. & Liu, H. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10571-10580.
- 18 Sample, P. A. *et al.* The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol* **127**, 1136-1145, doi:127/9/1136 [pii]
10.1001/archophthalmol.2009.187 (2009).
- 19 Obuchowski, N. A. Nonparametric analysis of clustered ROC curve data. *Biometrics* **53**, 567-578 (1997).
- 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* **10**, e0118432, doi:10.1371/journal.pone.0118432 (2015).
- 21 Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 839-847.
- 22 Gordon, M. O. *et al.* Assessment of the Impact of an Endpoint Committee in the Ocular Hypertension Treatment Study. *American journal of ophthalmology* **199**, 193-199 (2019).
- 23 Reus, N. J. *et al.* Clinical assessment of stereoscopic optic disc photographs for glaucoma: the European Optic Disc Assessment Trial. *Ophthalmology* **117**, 717-723, doi:10.1016/j.ophtha.2009.09.026 (2010).
- 24 Lichter, P. R. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc* **74**, 532-572 (1976).
- 25 Tielsch, J. M., Katz, J., Quigley, H. A., Miller, N. R. & Sommer, A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology* **95**, 350-356, doi:10.1016/s0161-6420(88)33177-5 (1988).
- 26 Varma, R., Steinmann, W. C. & Scott, I. U. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology* **99**, 215-221, doi:10.1016/s0161-6420(92)31990-6 (1992).
- 27 Thakur, A., Goldbaum, M. & Yousefi, S. Predicting Glaucoma before Onset Using Deep Learning. *Ophthalmol Glaucoma* **3**, 262-268, doi:10.1016/j.ogla.2020.04.012 (2020).
- 28 Chuter, B. *et al.* A deep learning model to assess fundus photograph image quality and improve predictive value of deep learning models of glaucoma detection. *Invest Ophth Vis Sci* **62**, 1016-1016 (2021).

SUPPLEMENT

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

eFigure 2. ResNet-50 architecture.

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

eMethods 1. Implemented class weights added to loss function to address Primary Open Angle Glaucoma (POAG) vs. non-POAG endpoint imbalance.

eTable 1. The time (years) elapsed between the earliest false positive photo for each case and eventual endpoint.

eTable 2. Diagnostic Accuracy of Deep-Learning Model Performance in Identifying Primary Open Angle Glaucoma (POAG) for three randomly selected visits/images.

eFigure 1. 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, 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 by 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 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.

eMethods 1.

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+¹⁴ (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 ($\leq 5^\circ$), 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.⁹

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 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 ImageNet¹⁶ 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.

eTable 2. Diagnostic Accuracy of Deep-Learning Model Performance in Identifying Primary Open Angle Glaucoma (POAG) by the Ocular Hypertensive Treatment Study Endpoint Committee and Optic Disc and Visual Field Reading Centers using at most 3 visits per eye.

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 $\geq$ -6 dB	80%	85%.	90%.	95%.
Endpoint Committee	Optic Disc	47 (60) / 102	0.89 (0.84, 0.94)	0.87 (0.79, 0.92)	0.83	0.81	0.69	0.56
	Photograph and/or Visual Field							
	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

eTable 1. The mean (95% CI) time (in years) between the occurrence of the false positive result at 90% specificity and the development of POAG is summarized in eTable2.

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 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)