Assessment of a Segmentation-Free Deep Learning Algorithm for Diagnosing Glaucoma From Optical Coherence Tomography Scans

Atalie C. Thompson, MD, MPH; Alessandro A. Jammal, MD; Samuel I. Berchuck, PhD; Eduardo B. Mariottoni, MD; Felipe A. Medeiros, MD, PhD

IMPORTANCE Conventional segmentation of the retinal nerve fiber layer (RNFL) is prone to errors that may affect the accuracy of spectral-domain optical coherence tomography (SD-OCT) scans in detecting glaucomatous damage.

OBJECTIVE To develop a segmentation-free deep learning (DL) algorithm for assessment of glaucomatous damage using the entire circle B-scan image from SD-OCT.

DESIGN, SETTING, AND PARTICIPANTS This cross-sectional study at a single institution used data from SD-OCT images of eyes with glaucoma (perimetric and preperimetric) and normal eyes. The data set was randomly split at the patient level into a training (50%), validation (20%), and test data set (30%). Data were collected from March 2008 to April 2019, and analysis began April 2018.

EXPOSURES A convolutional neural network was trained to discriminate glaucomatous from normal eyes using the SD-OCT circle B-scan without segmentation lines.

MAIN OUTCOMES AND MEASURES The ability to discriminate glaucoma from healthy eyes was evaluated by comparing the area under the receiver operating characteristic curve and sensitivity at 80% or 95% specificity for the DL algorithm's predicted probability of glaucoma vs conventional RNFL thickness parameters given by SD-OCT software. The performance was also assessed in preperimetric glaucoma, as well as by visual field severity using Hodapp-Parrish-Anderson criteria.

RESULTS A total of 20 806 SD-OCT images from 1154 eyes of 635 individuals (612 [53%] with glaucoma and 542 normal eyes [47%]) were included. The mean (SD) age at SD-OCT scan was 70.8 (10.4) years in individuals with glaucoma and 55.8 (14.1) years in controls. There were 187 women (53.3%) in the glaucoma group and 165 (59.8%) in the control group. Of 612 eyes with glaucoma, 432 (70.4%) had perimetric and 180 (29.6%) had preperimetric glaucoma. The DL algorithm had a significantly higher area under the receiver operating characteristic curve than global RNFL thickness (0.96 vs 0.87; difference = 0.08 [95% CI, 0.04-0.12]) and each RNFL thickness sector for discriminating between glaucoma and controls (all *P* < .001). At 95% specificity, the DL algorithm (81%; 95% CI, 64%-97%) was more sensitive than global RNFL thickness (67%; 95% CI, 58%-76%). The areas under the receiver operating characteristic curve were also significantly greater for the DL algorithm compared with RNFL thickness at each stage of disease, especially preperimetric and mild perimetric glaucoma.

CONCLUSIONS AND RELEVANCE A segmentation-free DL algorithm performed better than conventional RNFL thickness parameters for diagnosing glaucomatous damage on OCT scans, especially in early disease. Future studies should investigate how such an approach contributes to diagnostic decisions when combined with other relevant clinical information, such as risk factors and perimetry results.

JAMA Ophthalmol. 2020;138(4):333-339. doi:10.1001/jamaophthalmol.2019.5983
Published online February 13, 2020.

Invited Commentary page 339

Supplemental content

Author Affiliations: Vision, Imaging and Performance Laboratory (VIP), Duke Eye Center, Duke University, Durham, North Carolina (Thompson, Jammal, Berchuck, Mariottoni, Medeiros); Department of Statistical Science and Forge, Duke University, Durham, North Carolina (Berchuck).

Corresponding Author: Felipe A. Medeiros, MD, PhD, Duke Eye Center, Department of Ophthalmology, Duke University, 2351 Erwin Rd, Durham, NC 27710 (felipe.medeiros@duke.edu).

In clinical practice, glaucoma is usually diagnosed by a combined analysis of several clinical parameters, including risk factors, such as age and intraocular pressure; tests for evaluation of structural damage to the optic nerve; and visual function assessment with perimetry. Among tests for structural evaluation,[1,2] spectral-domain optical coherence tomography (SD-OCT) is the most commonly used one, providing objective quantification of damage to the optic nerve head and retinal nerve fiber layer (RNFL).[3] To provide quantitative RNFL thickness measurements, conventional SD-OCT software applies segmentation algorithms to delineate the RNFL and extract its thickness. Although SD-OCT RNFL thickness measurements can be generally accurate for diagnosing glaucoma,[3,4] they may fail in the presence of segmentation errors.[5,6] Such errors have been reported in 20% to more than 40% of scans[5] and significantly affect the diagnostic accuracy of SD-OCT in glaucoma.[5,6]

Even in the absence of segmentation errors, the interpretation of conventional SD-OCT RNFL printouts may be difficult given the presence of a large number of summary parameters, in addition to maps and plots. The evaluation of multiple parameters increases the risk of making a type I error, ie, finding an abnormality just by chance. This has led to the phenomenon of red disease in which some patients receive a diagnosis of glaucoma based on SD-OCT in the absence of true disease.[7]

Recent advances in artificial intelligence have led to the development of deep learning (DL) algorithms that can accurately detect complex patterns in images, achieving levels of accuracy in image classification tasks that can sometimes surpass those of humans.[8-12] Deep learning algorithms can be trained to analyze an entire SD-OCT image, potentially providing more information related to the presence of glaucomatous damage than individual SD-OCT parameters. Segmentation-free analysis of the SD-OCT image may also eliminate the need for manual refinement of conventionally segmented retinal layers. By interpreting the whole image, use of DL algorithms may further minimize false positives, or red disease, that arise when clinicians assess multiple individual parameters.

The purpose of this study was to develop a segmentation-free DL algorithm to assess glaucomatous structural damage using the whole peripapillary SD-OCT scan image and to compare its performance to that of conventional RNFL thickness parameters.

Methods

This cross-sectional study used data from the Duke Glaucoma Repository, a database of electronic research and medical records developed by the Duke University Vision, Imaging and Performance Laboratory. The study protocol adhered to the tenets of the Declaration of Helsinki[13] and was conducted in accordance with the Health Insurance Portability and Accountability Act on approval by the Duke University institutional review board. A waiver of informed consent was granted owing to the retrospective nature of this research.

Key Points

Question Does a segmentation-free deep learning algorithm using the entire circle B-scan image from optical coherence tomography perform better than retinal nerve fiber layer for detecting glaucomatous damage?

Findings In this cross-sectional study of 1154 eyes of 635 individuals, the deep learning algorithm had a greater area under the curve than retinal nerve fiber layer global and sector parameters. This appeared to be even more likely in early disease.

Meaning These findings suggest a deep learning algorithm using the entire B-scan may be better able to detect glaucomatous disease than conventional retinal nerve fiber layer parameters from optical coherence tomography.

The database included information on ophthalmic diagnoses, medical history, and results from comprehensive ophthalmic examination including visual acuity, intraocular pressure, slitlamp biomicroscopy, gonioscopy, and dilated fundus examination. In addition, stereoscopic optic disc photographs (Nidek 3DX) and Spectralis SD-OCT (version 5.4.7.0.; Heidelberg Engineering) images and associated data were reviewed. Standard automated perimetry (SAP) using the 24-2 test pattern and Swedish interactive thresholding algorithm (Carl Zeiss Meditec) was included if the test was reliable, containing fewer than 33% fixation losses and 15% false-positive errors. All eyes with glaucoma had primary open-angle glaucoma based on open angles on gonioscopy, clinical examination, and grading of stereophotographs and visual fields. Patients with other ocular or systemic diseases that could adversely affect the optic nerve or visual field were excluded. Eyes with a refractive error greater than or equal to +6.0 or −6.0 diopters were excluded.

Two experienced graders masked to the participant's identity and any other test information graded the photographs for the presence of signs of glaucomatous optic neuropathy as well as for change over time. Disagreements between graders were resolved by a third experienced grader. Eyes were categorized with perimetric glaucoma if they had evidence of glaucomatous optic neuropathy (ie, cupping, diffuse or focal rim thinning, optic disc hemorrhage, or RNFL defects) and a reproducible visual field defect on at least 2 consecutive SAP tests with pattern standard deviation less than 5% or glaucoma hemifield test outside normal limits. In addition, eyes with glaucomatous optic neuropathy whose contralateral eye had evidence of perimetric glaucoma were also included but categorized as preperimetric glaucoma. Eyes that had a history of documented optic disc progression on stereophotographs (ie, progressive rim thinning or enlargement of RNFL defects) in the absence of visual field loss were also categorized as having preperimetric glaucoma.

Eyes with perimetric glaucoma were further classified into mild, moderate, and severe visual field loss by applying the Hodapp-Parrish-Anderson criteria.[14] Healthy control eyes had to have a normal optic disc stereophotograph with no evidence of glaucomatous optic neuropathy, ocular hypertension (ie, intraocular pressure >21 mm Hg), or SAP abnormality in either eye. A normal SAP test result was required to have mean deviation and pattern standard deviation with $P > .05$ and a glaucoma hemifield test within normal limits.

Spectral-Domain Optical Coherence Tomography

Circumpapillary 12° scans were acquired using Spectralis SD-OCT (version 5.4.7.0.; Heidelberg Engineering).[15] Corneal curvature and axial length measurements were entered into the instrument's software. The Spectralis Anatomic Positioning System (Heidelberg Engineering) was used to adjust for eye movements. All of the images were manually reviewed for image quality, scan centration, and artifacts. Those with signal strength less than 15 dB or with artifacts such as inversion or clipping of the image were excluded.

The accuracy of segmentation was reviewed by a reading center. Segmentation errors were corrected whenever possible. If correction was not possible, the image was discarded. Global and sectoral RNFL thicknesses parameters were automatically computed by the SD-OCT software.

Development of the DL Algorithm

We trained a segmentation-free DL algorithm to assess glaucomatous damage from the raw SD-OCT peripapillary B-scan image (ie, without segmentation lines). The algorithm was trained to differentiate glaucomatous from normal eyes, as defined above, and provide a probability of glaucoma as output. Spectral-domain optical coherence tomography images were randomly separated at the participant level into a training (50%), validation (20%), and test sample (30%). This approach prevented leakage and biased estimates of test performance by ensuring that no data of any participant were present in both the training and test samples.

The images used in the present study did not contain segmentation lines in them so that the DL algorithm could identify which features were most relevant to predict the presence of glaucoma without relying on conventional segmentation. The SD-OCT B-scans were preprocessed first by downsampling the images to 496 × 496 pixels followed by scaling of the pixel values to range from 0 to 1. The heterogeneity of the images was improved by augmenting the data through random lighting adjustment of image balance and contrast of up to 5%, random horizontal image flips, and random image rotations of up to 10°. These subtle image transformations were only applied to the training set; they helped to prevent overfitting and allowed the DL algorithm to appreciate the most relevant features of each image.[16]

A residual deep convolutional neural network (ResNet34) architecture was used for the DL algorithm, which had been previously trained on the ImageNet data set.[17,18] Training was performed by first unfreezing the final 2 layers. Subsequently, all layers were unfrozen and the network was fine-tuned with differential learning rates and Adam optimizer. Gradient-weighted class activation maps were built over the SD-OCT images and helped identify the most important parts of the image for the DL algorithm's classification.[19,20]

Statistical Analyses

Receiver operating characteristic (ROC) curves were used to evaluate the diagnostic accuracies of the different parameters investigated in the study. A Probit ROC regression model with maximum likelihood estimator was used to adjust for the potentially confounding effects of age at the time of scan acquisition.[21-23] The area under the ROC curve (AUC) was used to summarize diagnostic accuracy, with 1.0 representing perfect discrimination and 0.5 representing chance discrimination. The difference in the AUC of 2 curves was compared using a Wald test based on the bootstrap covariance.[22] In addition, sensitivities at fixed specificities of 80% and 95% were calculated.

To maximize the data for the study, we included in the analyses all images and the corresponding diagnosis (ie, normal vs preperimetric vs perimetric glaucoma) that were available for each eye included in the study at the time of imaging. To account for the correlation between observations from the same eye, a bootstrap resampling procedure was used to derive 95% CIs and P values, where the eye-level clusters were considered as the units of resampling. This procedure is commonly used to account for the presence of multiple correlated measurements within the same participant.[21] Deep learning models were implemented using Keras (version 2.1.4.; MIT), an open-source Python library. Statistical analyses used Stata (version 15, StataCorp). The α level (type 1 error) was set at .05. Analysis began April 2019.

Results

The data set included 20 806 RNFL circle B-scans from SD-OCT from 1154 eyes of 635 participants, divided into training and validation (14 466 [70%]) and test (6340 [30%]) samples. The test sample consisted of 6340 SD-OCT scans acquired in 348 eyes of 191 participants. The mean (SD) age at SD-OCT scan was 70.8 (10.4) years in individuals with glaucoma and 55.8 (14.1) years in controls. There were 187 women (53.3%) in the glaucoma group and 165 (59.8%) in the control group. **Table 1** displays the demographic and clinical characteristics of the participants and eyes in the training/validation vs test samples.

Table 2 reports AUCs and sensitivities at 80% or 95% specificity for the DL algorithm's predicted probability of glaucoma and RNFL thickness parameters. The DL algorithm had a significantly higher AUC than global RNFL thickness (0.96 vs 0.87; difference = 0.08 [95% CI, 0.04-0.12]) and each of the RNFL sectors for discriminating between glaucoma and controls (all $P < .001$). In addition, the DL algorithm was more sensitive at 80% specificity (94% [95% CI, 87%-100%]) and 95% specificity (81% [95% CI, 64%-97%]) than global or sectoral RNFL thickness parameters (Table 2). The **Figure**, A, shows the ROC curves for the DL segmentation-free algorithm vs global RNFL thickness for discriminating glaucomatous from healthy eyes.

The eTable in the Supplement shows global RNFL thickness measurements, SAP mean deviation, and DL probability of glaucoma for the different diagnostic categories. As expected, SAP mean deviation was greater for perimetric (median, −4.41 dB; interquartile range, −10.13 dB to −2.09 dB) than preperimetric glaucoma (median, −0.30 dB; interquartile range, −1.32 dB to 0.43 dB) and both groups had greater values than normal eyes (median, 0.22 dB; interquartile range, −0.64 dB

Table 1. Demographic and Clinical Characteristics of Eyes and Patients in the Training and Validation vs Test Samples

	No. (%)					
	Training + Validation Sample			Test Sample		
Characteristic	Overall	Normal	Glaucoma	Overall	Normal	Glaucoma
Individuals	444 (100)	185 (41.7)	259 (58.3)	191 (100)	91 (47.6)	100 (52.4)
Eyes	806 (100)	363 (45.0)	443 (55.0)	348 (100)	179 (51.4)	169 (48.6)
SD-OCT scans	14 466 (100)	9638 (66.6)	4828 (33.4)	6340 (100)	2443 (38.5)	3897 (61.5)
Age at SD-OCT scan, mean (SD), y	66.6 (13.5)	56.6 (13.9)	71.6 ± 10.1)	63.3 (14.2)	54.3 (14.5)	68.9 (10.7)
Women	241 (54.3)	110 (59.5)	131 (50.6)	114 (59.7)	55 (60.4)	59 (59.0)
African American	93 (21.0)	37 (20.0)	56 (21.6)	45 (23.6)	18 (19.8)	27 (27.0)
SAP MD, median (IQR), dB	−1.15 (−3.52 to 0.16)	0.2 (−0.65 to 0.91)	−2.31 (−5.25 to −0.7)	−1.17 (−4.2 to 0.2)	0.22 (−0.64 to 1.12)	−3.23 (−7.66 to −1.01)
Global RNFL thickness, mean (SD), µm	81.8 (16.9)	96.7 (10.2)	74.3 (14.5)	82.3 (18.5)	97.4 (9.3)	72.8 (16.4)
Preperimetric	NA	NA	131 (29.6)	NA	NA	49 (29.0)
Visual field defect						
Mild	NA	NA	133 (30.0)	NA	NA	45 (26.6)
Moderate	NA	NA	78 (17.6)	NA	NA	24 (14.2)
Severe	NA	NA	101 (22.8)	NA	NA	51 (30.2)

Abbreviations: IQR, interquartile range; MD, mean deviation; NA, not applicable; RNFL, retinal nerve fiber layer; SAP, standard automated perimetry; SD-OCT, spectral-domain optical coherence tomography.

Table 2. Diagnostic Accuracies of the Deep Learning Algorithm's Probability of Glaucoma vs the Global and Sectoral RNFL Thicknesses in the Test Sample

	(95% CI)				
Characteristic	ROC Curve Area	Difference in ROC Curve Areas[a]	*P* Value[b]	Sensitivity at 95% Specificity, %	Sensitivity at 80% Specificity, %
Deep learning probability	0.96 (0.92-1.00)	NA	NA	81 (64-97)	94 (87-100)
Global RNFL thickness	0.87 (0.83-0.92)	0.08 (0.04-0.12)	<.001	67 (58-76)	80 (74-87)
Temporal RNFL thickness	0.62 (0.54-0.70)	0.33 (0.26-0.40)	<.001	19 (12-27)	40 (31-49)
Superior temporal RNFL thickness	0.86 (0.81-0.91)	0.09 (0.05-0.14)	<.001	62 (51-72)	78 (70-86)
Inferior temporal RNFL thickness	0.87 (0.83-0.92)	0.08 (0.04-0.13)	<.001	67 (58-77)	80 (73-88)
Nasal RNFL thickness	0.74 (0.68-0.80)	0.22 (0.16-0.27)	<.001	24 (14-35)	53 (42-64)
Superior nasal RNFL thickness	0.79 (0.72-0.85)	0.17 (0.10-0.23)	<.001	41 (30-52)	64 (54-74)
Inferior nasal RNFL thickness	0.82 (0.75-0.88)	0.14 (0.08-0.20)	<.001	37 (24-51)	67 (56-79)

Abbreviations: NA, not applicable; RNFL, retinal nerve fiber layer; ROC, receiver operating characteristic.

[a] Difference areas under the ROC curves in comparison with the deep learning algorithm.

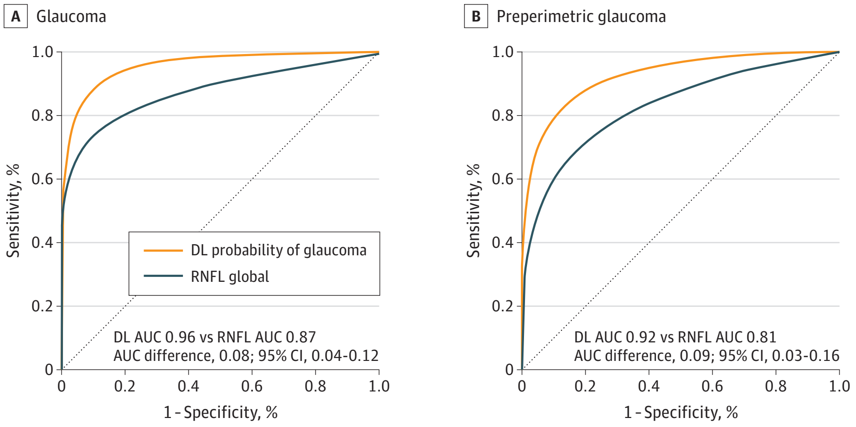
[b] *P* values for the comparison of ROC curve areas of each parameter in association with the deep learning algorithm.

to 1.12 dB). The mean (SD) DL probability of glaucoma was 0.87 (0.24) in perimetric glaucoma, 0.75 (0.28) in preperimetric glaucoma, and 0.17 (0.24) in healthy eyes.

Table 3 demonstrates the performance of the DL algorithm for detection of preperimetric glaucoma, as well as detection of perimetric glaucoma stratified by mild, moderate, and severe visual field loss by Hodapp-Parrish-Anderson criteria. The AUC for the DL probability of glaucoma was significantly greater than that of global RNFL thickness for discriminating preperimetric glaucoma from healthy eyes (0.92 vs 0.83; difference = 0.09 [95% CI, 0.03-0.16]; *P* = .002). The Figure, B, shows the ROCs for discriminating preperimetric glaucoma from healthy eyes for the DL algorithm and global RNFL thickness. For 95% specificity, the DL algorithm had sensitivity of 70% compared with only 49% for global RNFL thickness. The DL algorithm also exhibited a significantly larger AUC for detection of mild as well as moderate and advanced glaucoma.

eFigure 1 in the Supplement shows an example eye with perimetric glaucoma that exhibited a superior arcuate visual field defect with corresponding inferior rim loss. The conventional SD-OCT inferior temporal thickness parameter is outside normal limits. The class activation heat map overlying the SD-OCT circle B-scan shows in red the area that had the greatest association with the algorithm prediction, which, as expected, corresponded to the inferior temporal region. eFigure 2 in the Supplement shows an example eye with preperimetric glaucoma. Although the conventional SD-OCT RNFL thickness parameters were mostly in the normal range, with only a borderline global RNFL thickness, the DL segmentation-free algorithm estimated a probability of 1.0 that the eye had glaucoma. eFigure 3 in the Supplement shows that the superior temporal and inferior temporal portions of the scan were the most important areas influencing the algorithm's prediction in a healthy eye.

Figure. Receiver Operating Characteristic Curves Comparing the Deep Learning (DL) Algorithm's Probability of Glaucoma and the Global Retinal Nerve Fiber Layer (RNFL) for All Glaucoma and Preperimetric Glaucoma



AUC indicates area under the receiver operating characteristic curve.

Table 3. Diagnostic Accuracy Measures for the Deep Learning Algorithm's Probability of Glaucoma vs the Global RNFL Thickness in the Test Sample[a]

	ROC Curve Areas (95% CI)				Sensitivity at 95% Specificity (95% CI)		Sensitivity at 80% Specificity (95% CI)	
Characteristic	Global RNFL Thickness	DL Probability	Difference	*P* Value[b]	Global RNFL Thickness	DL Probability	Global RNFL Thickness	DL Probability
Preperimetric	0.83 (0.75-0.91)	0.92 (0.86-0.99)	0.09 (0.03-0.16)	.002	49 (33-65)	70 (48-91)	71 (58-84)	88 (76-99)
Perimetric	0.89 (0.85-0.94)	0.97 (0.93-1.0)	0.07 (0.04-0.11)	<.001	71 (62-80)	85 (68-100)	83 (77-90)	96 (90-100)
Mild	0.82 (0.75-0.89)	0.92 (0.85-0.99)	0.09 (0.03-0.16)	.002	50 (37-63)	69 (49-90)	71 (60-82)	87 (74-99)
Moderate	0.93 (0.89-0.97)	0.99 (0.97-1.0)	0.06 (0.02-0.09)	<.001	73 (61-85)	93 (80-100)	89 (82-96)	99 (97-100)
Severe	0.96 (0.92-1.0)	0.99 (0.98-1.0)	0.03 (0.01-0.08)	.058	88 (78-97)	98 (92-100)	94 (88-100)	99 (99-100)

Abbreviations: DL, deep learning; RNFL, retinal nerve fiber layer; ROC, receiver operating characteristic.

[a] Stratified by glaucoma disease severity in preperimetric and perimetric, and according to the Hodapp-Parrish-Anderson criteria.

[b] *P* values for the comparison of ROC curve areas of global RNFL thickness vs DL probability for each category of disease severity.

Discussion

We developed a DL algorithm to estimate the probability of glaucomatous damage from evaluation of the entire circular B-scan from SD-OCT. The algorithm had greater accuracy for detecting structural glaucomatous damage compared with conventional RNFL thickness parameters, notably for eyes with preperimetric glaucoma and mild visual field defects.

Other groups have also used DL to detect glaucoma from OCT data.[9,11] For example, Asaoka et al[9] demonstrated that early perimetric glaucoma could be accurately diagnosed using a DL algorithm trained with SD-OCT parameters extracted from the macula. However, their method still required conventional segmentation to extract thicknesses of the RNFL and ganglion cell complex. In a small sample study, Muhammad et al[10] trained a neural network to extract features from maps derived from conventional automated segmentation of wide-field swept-source SD-OCT, which were then used in a random forest model to predict glaucomatous damage.[11] Similar to the work of Asaoka et al,[8,9] Muhammad et al's approach[10] also required conventional segmentation and, therefore, would still be problematic in the presence of segmentation errors. In contrast, our approach used raw B-scans without requiring any segmentation of the retinal layers. As segmentation errors are very common in OCT scans, a segmentation-free approach is likely to provide results that are more robust when applied in a clinical practice scenario.

Besides outperforming all conventional RNFL thickness parameters for detecting glaucoma, the DL approach presented in this study may have additional advantages. The single probabilistic output may afford a simpler interpretation. Integration of information from the plethora of parameters given by the conventional SD-OCT printout may sometimes be difficult. In addition, the use of multiple parameters may increase the incidence of false-positive test results, as commonly seen in cases of red disease. The class activation maps may also help to highlight the areas of the scan that had the greatest contribution to the algorithm's output (eFigures 1 and 2 in the Supplement). Interestingly, the maps seemed to include other layers beyond the RNFL, which may also be important in assessing glaucomatous damage. However, it should be noted that these maps have limited resolution due to downsampling of the final convolutional layers in a DL model, which limits their accuracy in pinpointing the areas of damage.[19,20]

Limitations

This study has limitations. Although we used an independent test set for final assessment of diagnostic accuracy, external validation in populations from other clinical settings is desirable. It should be noted that although the ROC curve areas were statistically significantly different between the DL model and the conventional OCT parameters, some overlap was seen in the 95% CIs. Of note, in the overall comparison with global RNFL thickness, the difference in ROC curve areas in relation to the DL model had a lower limit of the 95% CI of 0.04 (Table 2). As ROC curve areas range from 0.5 to 1.0, this number would represent approximately 8% of the range, still a meaningful difference. As more data and studies accumulate, future meta-analyses could be done to obtain even more precise CIs around the point estimates of differences in diagnostic accuracy, helping clarify further the clinical relevance of DL applications on OCT data. In addition, it should be noted that the diagnosis of glaucoma is not based on the results of a single test, but rather on a combined interpretation of information on risk factors, such as age and intraocular pressure, and results of structural and functional tests. Therefore, it remains to be seen how the incorporation of such an algorithm would affect clinical diagnosis when combined with these other pieces of information that are acquired in clinical practice, as well as for assessing change over time.

Conclusions

In summary, we developed a segmentation-free DL algorithm that can predict the probability of glaucomatous structural damage from the SD-OCT circle B-scan. The algorithm performed better than global and sectoral RNFL thickness parameters for discriminating glaucomatous from control eyes, especially in cases of preperimetric or early perimetric glaucoma. Application of this DL algorithm in a clinical setting may improve the accuracy and sensitivity of SD-OCT for diagnosing glaucoma, while obviating the need for error-prone segmentation of retinal layers.

ARTICLE INFORMATION

Accepted for Publication: December, 2, 2019.

Published Online: February 13, 2020. doi:10.1001/jamaophthalmol.2019.5983

Author Contributions: Drs Thompson and Medeiros 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.
Concept and design: Thompson, Jammal, Berchuck, Medeiros.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Thompson, Jammal, Mariottoni, Medeiros.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: All authors.
Obtained funding: Medeiros.
Administrative, technical, or material support: Medeiros.
Supervision: Medeiros.

Conflict of Interest Disclosures: Dr Thompson is a recipient of the Heed Ophthalmic Fellowship. Dr Jammal reports other support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior outside the submitted work. Dr Medeiros reports grants from National Institutes of Health/National Eye Institute, Carl Zeiss Meditec, and Google during the conduct of the study; nonfinancial support (equipment) from Heidelberg Engineering during the conduct of the study; and personal fees from Bausch + Lomb, Merck, Sensimed, Topcon, Reichert, Novartis, Allergan, Galimedix Therapeutics, Stealth BioTherapeutics, and Biogen outside the submitted work. No other disclosures were reported.

Funding/Support: This work is supported in part by National Institutes of Health/National Eye Institute (grants EY027651 [Dr Medeiros], EY029885 [Dr Medeiros], and EY021818 [Dr Medeiros]).

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.

REFERENCES

1. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. 2014;311(18):1901-1911. doi:10.1001/jama.2014.3192

2. 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.

3. Lisboa R, Paranhos A Jr, Weinreb RN, Zangwill LM, Leite MT, Medeiros FA. Comparison of different spectral domain OCT scanning protocols for diagnosing preperimetric glaucoma. *Invest Ophthalmol Vis Sci*. 2013;54(5):3417-3425. doi:10.1167/iovs.13-11676

4. Ervin A-M, Boland MV, Myrowitz EH, et al. Screening for glaucoma: comparative effectiveness. *AHRQ Comparative Effectiveness Reviews*. 2012;:59.

5. Mansberger SL, Menda SA, Fortune BA, Gardiner SK, Demirel S. Automated segmentation errors when using optical coherence tomography to measure retinal nerve fiber layer thickness in glaucoma. *Am J Ophthalmol*. 2017;174:1-8. doi:10.1016/j.ajo.2016.10.020

6. Asrani S, Essaid L, Alder BD, Santiago-Turla C. Artifacts in spectral-domain optical coherence tomography measurements in glaucoma. *JAMA Ophthalmol*. 2014;132(4):396-402. doi:10.1001/jamaophthalmol.2013.7974

7. Chong GT, Lee RK. Glaucoma versus red disease: imaging and glaucoma diagnosis. *Curr Opin Ophthalmol*. 2012;23(2):79-88. doi:10.1097/ICU.0b013e32834ff431

8. Asaoka R, Murata H, Hirasawa K, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. 2019;198:136-145. doi:10.1016/j.ajo.2018.10.007

9. Asaoka R, Hirasawa K, Iwase A, et al. Validating the usefulness of the "random forests" classifier to diagnose early glaucoma with optical coherence tomography. *Am J Ophthalmol*. 2017;174:95-103. doi:10.1016/j.ajo.2016.11.001

10. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26(12):1086-1094. doi:10.1097/IJG.0000000000000765

11. An G, Omodaka K, Hashimoto K, et al. Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. *J Healthc Eng*. 2019;2019:4061313. doi:10.1155/2019/4061313

12. Christopher M, Belghith A, Weinreb RN, et al. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. *Invest Ophthalmol Vis Sci*. 2018;59(7):2748-2756. doi:10.1167/iovs.17-23387

13. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-2194. doi:10.1001/jama.2013.281053

14. Hodapp E, Parrish RK, Anderson DR. *Clinical Decisions In Glaucoma*. St Louis, Mo: Mosby; 1993.

15. Vizzeri G, Balasubramanian M, Bowd C, Weinreb RN, Medeiros FA, Zangwill LM. Spectral domain-optical coherence tomography to detect localized retinal nerve fiber layer defects in glaucomatous eyes. *Opt Express*. 2009;17(5):4004-4018. doi:10.1364/OE.17.004004

16. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. preprint. Posted online December 13, 2017. arXiv 1712.04621.

17. Deng J, Dong W, Socher R, Li L-J, Li K, Imagenet LF-F. A large-scale hierarchical image database. Paper presented at: Computer Vision and Pattern

Recognition. Computer Vision and Pattern Recognition, IEEE Conference 2009; June 20-25, 2009; Miami, FL. https://ieeexplore.ieee.org/document/5206848. Accessed January 6, 2020.

18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Published online December 10, 2015. arXiv:1512.03385

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Published online October 7, 2016. Revised December 3, 2019. arXiv:1610.02391

20. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: why did you say that? Published online November 22, 2016. Revised January 25, 2017. arXiv:1611.07450

21. Medeiros FA, Sample PA, Zangwill LM, Liebmann JM, Girkin CA, Weinreb RN. A statistical approach to the evaluation of covariate effects on the receiver operating characteristic curves of diagnostic tests in glaucoma. *Invest Ophthalmol Vis Sci*. 2006;47(6):2520-2527. doi:10.1167/iovs.05-1441

22. Alonzo TA, Pepe MS. Distribution-free ROC analysis using binary regression techniques. *Biostatistics*. 2002;3(3):421-432. doi:10.1093/biostatistics/3.3.421

23. Pepe MS. A regression modeling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*. 1997;84:595-608. doi:10.1093/biomet/84.3.595

Invited Commentary

Data-Driven, Feature-Agnostic Deep Learning vs Retinal Nerve Fiber Layer Thickness for the Diagnosis of Glaucoma

Christine A. Petersen, MD; Parmita Mehta, MS; Aaron Y. Lee, MD, MSCI

Related article page 333

In this issue of *JAMA Ophthalmology*, Thompson et al[1] report that a deep learning (DL) model using unsegmented spectral-domain optical coherence tomography (SD-OCT) scans to detect glaucoma performs better than retinal nerve fiber layer (RNFL) thickness parameters extracted by automated segmentation. The authors used data from the Duke Glaucoma Repository, which included 20 806 SD-OCT images from 1154 eyes of 635 individuals. A convolutional neural network DL model was trained on unsegmented, raw SD-OCT peripapillary B-scan images in a fully data-driven manner. This DL model was compared with conventional RNFL thickness measurements in its ability to discriminate glaucomatous from control eyes based on the area under the receiver operating curve (AUC) and prespecified sensitivity cutoffs of 80% and 95%. Testing of the model was stratified by preperimetric glaucoma and for mild, moderate, and severe glaucoma. The AUC for the DL model (0.96) on the entire test data was significantly higher than the AUC for the global RNFL thickness–based model (0.87) and than the AUCs of each of the sectoral RNFL thickness models (ranging from 0.62 for the temporal sector to 0.87 for the inferior temporal sector). The DL model did well for preperimetric, mild, and moderate glaucoma. However, for the severe glaucoma subset, the difference in the performance of the DL model and the global RNFL thickness–based model was not statistically significant.

The ground truth of glaucoma was rigorously established by expert review of stereoscopic optic disc photographs by masked experts, and the severity of glaucoma was determined by the presence of a reproducible visual field defect on standard automated perimetry using the 24-2 test pattern. Severity of perimetric glaucoma was stratified as mild, moderate, or severe based on Hodapp-Parrish-Anderson criteria. If there was progression of optic disc changes associated with glaucoma on stereophotography and no visual field defect, the eye was determined to have preperimetric glaucoma. Both eyes of the same individual were eligible for inclusion, and all data were partitioned at the patient level for training and testing. Control eyes had a normal intraocular pressure and no abnormalities on stereophotography or standard automated perimetry. Although only allowing the experts to consider the disc photographs may have limited their diagnostic ability, excluding OCT RNFL data in the establishment of ground truth was an important step, since the DL model was to be compared with OCT RNFL thickness measurements.

One major strength of the study by Thompson et al[1] is that it compares the DL model with the clinically relevant baseline of RNFL thickness. Studies have reported a high error rate in automated segmentation of RNFL thickness,[2] and Thompson et al[1] minimized this error by having a reading center review and manually correct segmentation errors and discarding images when correction was not possible to create a fair baseline comparison. This step is not always performed in clinical practice; thus, a DL model that does not rely on information provided by segmentation will have higher accuracy.

The stratification of glaucoma by severity is an important aspect of this study. Accurate early detection of glaucoma is critical because it allows for initiation of therapeutic interventions to slow disease progression at an earlier stage. Although the receiver operating curves are only shown for the DL model compared with glaucoma and preperimetric glaucoma, the listed AUCs are significantly better for preperimetric glaucoma as well as mild and moderate glaucoma. The sensitivity for detection of glaucoma is also superior for the DL model compared with the global RNFL thickness–based model at both the 80% and 95% sensitivity thresholds.

While the comparison of the DL model's performance to that of RNFL thickness measurements in detecting glaucoma is a strength of this study, the consideration of each quadrant only independently is also a potential weakness. Global RNFL thickness was found to have higher sensitivity at both 80% and 95% specificity and a larger AUC than any of the 6 sectoral RNFL thicknesses, but inferior temporal quadrant thickness in particular was nearly as good as global thickness. The question remains as to whether a combination of sectors, such as the inferior temporal and superior temporal sectors, may have had a superior diagnostic ability.

A 2019 study by Zheng et al[3] evaluating the diagnostic criteria of RNFL thickness and other neuroretinal rim findings on