tvst

Special Issue

A Review of Deep Learning for Screening, Diagnosis, and Detection of Glaucoma Progression

Atalie C. Thompson[1], **Alessandro A. Jammal**[1], **and Felipe A. Medeiros**[1]

[1] Vision, Imaging and Performance Laboratory (VIP), Duke Eye Center, Duke University, Durham, NC, USA

Correspondence: Felipe A. Medeiros, Duke Eye Center, Department of Ophthalmology, Duke University, 2351 Erwin Rd, Durham, NC 27705, USA. e-mail: felipe.medeiros@duke.edu

Received: April 6, 2020
Accepted: May 21, 2020
Published: July 22, 2020

Keywords: glaucoma; deep learning; optical coherence tomography; visual fields

Citation: Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. Trans Vis Sci Tech. 2020;9(2):42, https://doi.org/10.1167/tvst.9.2.42

Because of recent advances in computing technology and the availability of large datasets, deep learning has risen to the forefront of artificial intelligence, with performances that often equal, or sometimes even exceed, those of human subjects on a variety of tasks, especially those related to image classification and pattern recognition. As one of the medical fields that is highly dependent on ancillary imaging tests, ophthalmology has been in a prime position to witness the application of deep learning algorithms that can help analyze the vast amount of data coming from those tests. In particular, glaucoma stands as one of the conditions where application of deep learning algorithms could potentially lead to better use of the vast amount of information coming from structural and functional tests evaluating the optic nerve and macula. The purpose of this article is to critically review recent applications of deep learning models in glaucoma, discussing their advantages but also focusing on the challenges inherent to the development of such models for screening, diagnosis and detection of progression. After a brief general overview of deep learning and how it compares to traditional machine learning classifiers, we discuss issues related to the training and validation of deep learning models and how they specifically apply to glaucoma. We then discuss specific scenarios where deep learning has been proposed for use in glaucoma, such as screening with fundus photography, and diagnosis and detection of glaucoma progression with optical coherence tomography and standard automated perimetry.

Translational Relevance: Deep learning algorithms have the potential to significantly improve diagnostic capabilities in glaucoma, but their application in clinical practice requires careful validation, with consideration of the target population, the reference standards used to build the models, and potential sources of bias.

Introduction

Despite the availability of effective treatments, glaucoma remains the leading cause of irreversible blindness worldwide.[1] Current projections estimate that 111.8 million people will have glaucoma by the year 2040, with people in Asia and Africa disproportionately affected.[1] Early detection and intervention can help prevent vision loss from glaucoma, but a majority of patients do not know they have the disease[2,3] because it is generally asymptomatic in early stages.[4–6] Thus early detection of glaucoma is important and may be improved by introducing novel approaches for screening, diagnosis, and detection of change over time.

Recent progress in artificial intelligence (AI) and the collation of large medical datasets have spurred great interest in the development of deep learning algorithms that would more quickly and accurately identify glaucomatous damage on diagnostic tests compared to subjective evaluation and other traditional methods.[7–10] The purpose of this article is to critically review recent applications of deep learning models in glaucoma, discussing their advantages but also focusing on the challenges inherent to the development of such models for screening, diagnosis and detection of progression. After a brief general overview of deep learning and how it compares to traditional machine learning classifiers, we discuss issues related to the training and validation of deep learning models and how they specifically apply to glaucoma. We then

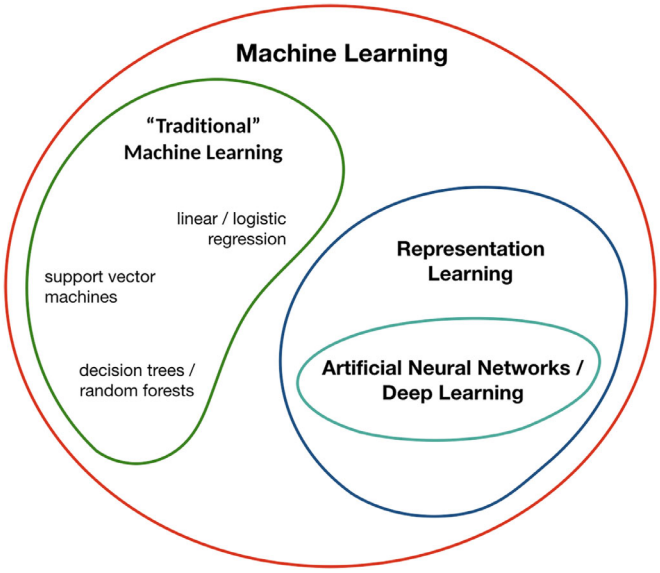

Figure 1. A diagram showing the organization of the classification of machine learning algorithms.

discuss specific scenarios where deep learning has been used in glaucoma, such as screening with fundus photography and diagnosis and detection of glaucoma progression with optical coherence tomography (OCT) and standard automated perimetry (SAP).

Artificial Intelligence, Machine Learning and Deep Learning

Artificial intelligence (AI) is a branch of computer science dealing with the simulation of intelligent behavior in computers, but in practice, and particularly in the popular press, "AI" has been used to describe any cutting-edge machine capability. Machine learning is a subset of AI that is concerned with setting up computer algorithms to recognize patterns in data, without human programmers having to dictate all aspects of this recognition.

In its most traditional form, machine learning algorithms still require human-designed code to transform raw data into input features, as these algorithms are not particularly good at learning features directly from raw data. Examples of these more traditional algorithms include logistic regression, k-Nearest Neighbor, decision trees, random forests, support vector machines (SVMs), among others (Fig. 1). The process of creating these initial features, however, can be a highly specialized task, requiring substantial subject-matter expertise, and there is no guarantee that the human-extracted features are optimal for use by the classifier. As an example, previous studies have used SVMs to improve detection of glaucoma damage from imaging data.[11–13] The SVMs used features such as global and sectoral parameters of retinal nerve fiber layer (RNFL) thickness, and measurements such as rim and cup area, cleverly combining them to reach a final glaucoma classification. However, although satisfactory performance has been reported for SVMs and other traditional algorithms in this scenario, there is no guarantee that the parameters used as initial features make the best use of the vast information produced by imaging. In fact, although many of these traditional machine learning techniques have been applied to enhance the diagnostic performance of imaging and perimetry in glaucoma,[13–22] they have not been widely incorporated into clinical practice. It should be noted, however, that depending on the type of data and application, traditional machine learning techniques may still provide the best solution to a particular problem.

In contrast to the traditional techniques described above, deep learning belongs to a class of machine learning algorithms that use "representation learning" (Fig. 1). These algorithms learn features (or representations) from data automatically, as long as enough data are given to them. A primary benefit of deep learning is that it eases the requirement for subject matter expertise. Instead of manually trying to curate relevant features from the data, one can feed the raw data directly to a deep leaning model, which will then automatically learn the most relevant features from the data. These features may be more subtle and comprehensive than those that would have been manually curated. As a trade-off, however, these automatically learned features may not be as straightforward to understand or explain, leading to the perception that deep learning models are a "black-box."

Deep learning models are a type of artificial neural network composed of several layers of artificial "neurons." These neurons are simple algorithms inspired by biological brain cells, in the sense that they receive input from other neurons, perform computations, and produce an output (Fig. 2). An artificial neural network is a collection of interconnected artificial neurons. Data are fed to the network and processed in some way with the goal of producing a desired outcome. Neural networks have been known for decades. Goldbaum et al.[22] used them to interpret perimetry results in glaucoma almost 30 years ago. However, only recently the advances in computational power have allowed the buildup of networks of several layers, that is, deep learning networks, which are able to process much more complex data, resulting in far better performance compared to the shallow artificial neural networks.

A type of deep learning network called *convolutional neural network* (CNN) has been the main one responsible for the explosion of deep learning applications in

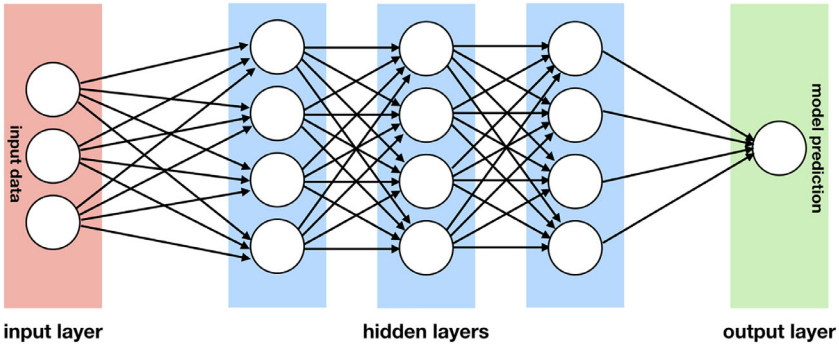

Figure 2. Schematic representation of "neurons" on an artificial neural network. The input data corresponds to the data one it trying to classify. The number of neurons in the input layer will depend on the input data (e.g., number of pixels in an image). These input neurons are then connected to neurons in hidden layers. There may be many hidden layers, which can be quite complex depending on the type of model. For convolutional neural networks, the hidden layers are of the convolutional type, specializing in spatial patterns. Finally, all calculations will converge to a final model prediction in the output layer.

computer vision, with performance sometimes surpassing that of humans for a variety of tasks. Nowadays, these networks are ubiquitous in their applications ranging from face recognition in smartphones to self-driving cars. CNNs have one or more convolutional layers, which consist of sets of filters, and are ideally suited to process spatial patterns and perform tasks such as image classification and object detection. These filters can be used to automatically extract features from images, obviating the need for manual labor in curating relevant features, a major limitation of traditional machine learning algorithms, as described above. As one of the medical fields that is highly dependent on ancillary imaging tests, ophthalmology has been a prime area to witness the application of CNN algorithms to help analyze data coming from these tests.

Training, Validation, and Testing of Deep Learning Models

Before a deep learning network can be used for a specific task, it needs to be trained, so that the specific computations needed at each artificial neuron (i.e., weights) and their interconnections can be determined to produce the desired result. In general terms, this training process involves feeding the network with data, observing the results, making modifications to the model, and repeating the process iteratively, until a desired level of accuracy is achieved.

There are essentially three ways to train a deep learning algorithm: supervised learning, unsupervised learning, and semisupervised learning. Supervised learning entails training of the algorithm with a completely labeled dataset. For example, if an algorithm is being purposed to identify glaucoma on fundus photographs, it can be initially trained by feeding the network with labeled photos of glaucoma and normal eyes. The network then "learns" the best features and the combination of them that will lead to the best discrimination of a glaucomatous from a normal photo. This learning process is done by comparing the algorithm's predictions to the actual labels and readjusting the weights of the artificial neurons, in a process known as backpropagation.[23] Numerous studies have been published using supervised learning algorithms to improve glaucoma detection.[7–9,24–31] Unsupervised learning, on the other hand, involves training the algorithm with unlabeled data. The goal is to have the model discover some hidden underlying structure or pattern in the data, without being told a priori what the task should be. For example, one can train a model to identify patterns of visual field damage in glaucoma with a large unlabeled set of visual fields from patients with the disease, in the hope that the model will "learn" the different patterns in those fields. This approach has previously been used to classify fields in glaucoma, as well as to detect progressive change over time.[17–19,32–36] Finally, semisupervised learning uses a combination of the two approaches, where one has a relatively small set of labeled data and generally a much larger amount of unlabeled data. The labeled dataset is used to obtain a reasonable initial model, which is then used to perform predictions on the unlabeled dataset. Such new predictions can then be used to retrain the model and the process is repeated until a final satisfactory model is obtained. This situation can occur, for example, when the process of data labeling is time-consuming or expensive. Application of semisupervised models in glaucoma has been rare,[37] but this is a promising approach that deserves more investigation.

The process of training requires the investigator to have an amount of data that will be used to train the model and a separate dataset that will be used to check the model's predictions (i.e, validation). If the predictions are unsatisfactory, then certain parameters of the model can be changed—for example, the number of hidden layers of neurons—and the training is repeated. It is important to note that this process of training and validating the model is highly iterative and time-consuming and still requires substantial human input. Therefore the notion of AI as being a fully self-programming intelligent algorithm is still not the day-to-day reality in deep learning. Importantly, the validation dataset needs to be independent from the training dataset, so that the predictions of the model are unbiased. Cross-validation may also be used for parameter search during training, notably in the presence of relatively small samples. Even more importantly, because the results of predictions on the validation dataset are used to fine-tune the model, there is a need for yet another set of data, independent of the training and validation sets, to be used for final testing of the model. This test set should be used only at the very end of the process, when the final model has been obtained and one needs to obtain the final estimates of accuracy. There is considerable confusion in the literature in the naming of these datasets, with the test set sometimes called the validation set. However, the important point is to acknowledge that the final estimate of accuracy of the model needs to be obtained on a test sample that has never been used in any part of the development of the model. Also, it is important to prevent "leakage" among the datasets. For example, if multiple datapoints or images are acquired on the same patient, then images from that patient should not exist in both the training and test dataset. Otherwise the predictions would be biased and would severely overestimate the algorithm's performance. In addition to ensuring that there is no data leakage between the test and training datasets, it is also important to ensure that the test set contains a representative sample of the target population to which the test is planned to be applied. Very often one sees that the test sample has characteristics (e.g., prevalence and severity of glaucoma) that are quite different than those of the target population.

One challenge to the development of a deep learning algorithm is the general requirement of very large datasets for training, which can be on the order of thousands or even millions of images. This occurs because of the very large number of parameters in these models. Some state-of-the-art CNNs have dozens of layers, resulting in millions of parameters that need to be trained. Ophthalmic image datasets of this size are not typically available, especially labeled datasets. However, transfer learning techniques have been applied to overcome this limitation.[38] In transfer learning, a CNN (e.g., ResNet, Inception)[39,40] previously trained on a very large general image dataset can be used as a general feature extractor and undergo additional training so it can be refined to perform a more specific task (e.g., distinguishing glaucoma from nonglaucoma) using a much smaller dataset. Transfer learning techniques are now ubiquitously applied to train CNNs that detect glaucoma on imaging datasets of more limited size.

Deep Learning Models in Glaucoma

There have been several publications using deep learning models for screening, diagnosis, and detection of progression in glaucoma (Table 1). To fully appreciate their impact, it is important not only to concentrate on the benefits afforded by advanced methods such as deep learning networks, but also to have a critical understanding of the context in which the diagnostic test is going to be applied.

Glaucoma Screening with Fundus Photography

An important misconception concerns what constitutes early glaucoma diagnosis from a screening standpoint. This is often meant to imply diagnosis at a very early stage, before any significant visual field loss is detectable by perimetry or sometimes even before the appearance of clear signs of optic nerve damage. However, focusing on this early stage of disease for screening is not only unnecessary, but also leads to problems related to the uncertainty in diagnosis. From a public health standpoint, an early diagnosis means diagnosing a patient at a stage earlier than the patient would have presented symptomatically. Given that symptomatic presentation of glaucoma generally occurs only at a late stage, almost any stage of glaucoma is early disease from the point of view of screening. Given the relatively low prevalence of glaucoma and the difficulties related to discriminating early glaucoma from normal variation, attempting to focus on screening programs for detecting very early disease will likely lead to failure. However, moving the focus to well-established cases of glaucoma that would still be asymptomatic will lead to much improved diagnostic accuracy and effectiveness. With this concept in mind, we can review recent studies proposing the use of deep learning methods to screen for glaucoma using a variety of diagnostic tests, from fundus photographs to OCT.

Fundus photography represents a relatively low-cost option for screening of certain eye diseases and has been successfully incorporated into teleophthalmology

Table 1. Summary of Studies Using Deep Learning Models in Glaucoma

Citation	Training/Validation Dataset	Test Dataset	Reference	Network	Data type	Output	Results
Ting et al.[27]	Train: 125,189	Test: 71,896	Subjective grading of photographs	Custom deep learning system	Color Fundus Photos	"Referable for glaucoma" vs. not	AUC 0.942; Sensitivity 96.4%, Specificity 87.2%
Li et al.[9]	Train: 31,745	8000	Subjective grading of photographs	Inception-v3	Color Fundus Photos	"Referable for glaucoma" vs. not	AUC 0.986; Sensitivity 95.6%, Specificity 92.0%
Christopher et al.[36]	9189 healthy, 5633 GON: divided randomly into multiple folds for 10-fold cross-validation.	10% test	Subjective grading of photographs	VGG6, Inception-v3, ResNet50	Color Fundus Photos	"GON" vs. healthy	ResNet50 AUC 0.91; Sensitivity 85% at 80% Specificity
Liu et al.[28]	Train: 29,865 GON, 11,046 probable GON, 200,121 unlikely GON	Validation: 4514 GON, 571 Probable GON, 23,484 unlikely GON	Subjective grading of photographs	ResNet	Color Fundus Photos	"Referable GON" vs. not	AUC 0.996, Sensitivity 96.2%, Specificity 97.7%
Ahn et al.[29]	Train: 228 Advanced glaucoma, 131 Early glaucoma, 385 Normal; Validation: 98 Advanced glaucoma, 61 Early glaucoma, 165 Normal	Test: 141 Advanced glaucoma, 87 Early glaucoma, 236 Normal	Subjective grading of visual field, OCT and RNFL photographs	Inception-v3; Custom 3-layer CNN	Color Fundus Photos	Glaucoma vs. Normal	Inception-v3 model: AUC 0.93; Average accuracy 84.5%; Custom 3-layer CNN: AUC 0.94, Average accuracy 87.9%
Phene et al.[31]	Train: 35,877 Non-glaucomatous, 20,740 Low-risk GS, 13,180 High-risk GS, 5307 Likely glaucoma, 18,487 Referable glaucoma; Tuning: 849 Non-glaucomatous, 259 Low-risk GS, 268 High-risk GS, 110 Likely glaucoma, 378 Referable glaucoma	Validation set A: 687 Non-glaucomatous, 290 Low-risk GS, 170 High-risk GS, 48 Likely glaucoma, 218 Referable glaucoma; Validation set B: 8753 Non-glaucomatous, N/A Low-risk GS, N/A High-risk GS, 890 Likely glaucoma, 890 Referable glaucoma; Validation set C: 63 Non-glaucomatous, N/A Low-risk GS, 175 High-risk GS, 108 Likely glaucoma, 283 Referable glaucoma	Validation set A: Referable GON based on subjective gradings of photographs; Validation set B: Referable GON based on glaucoma-related International Classification of Diseases codes; Validation set C: referable GON based on full glaucoma workup by glaucoma specialists including clinical exam, history, VF assessment, and OCT	Inception-v3	Color Fundus Photos	"Referable glaucoma" vs. Not	Validation set A: AUC 0.945; Validation set B: AUC 0.855; Validation set C: AUC 0.881
Shibata et al.[8]	Train: 1364 glaucomatous appearance vs. 1768 not glaucomatous appearance; 3-fold cross-validation	Test: 33 non-highly myopic glaucoma, 28 highly myopic glaucoma, 27 non-highly myopic normal, 22 highly myopic normal	Train: Subjective gradings of photographs Test: Subjective gradings of photographs and categorization of RNFL and macular inner retinal thickness measurements based on OCT normative database	ResNet	Color Fundus Photos	Glaucomatous vs. Not	AUC 96.5

Table 1. Continued

Citation	Training/Validation Dataset	Test Dataset	Reference	Network	Data type	Output	Results
Li et al.[30]	Train 20,793/Validation 2,311:11,176 GON-confirmed, 599 GON-suspected, 11,329 Normal; 10-fold cross-validation with a random selection of 9:1 for participants within each fold	Test: 1442 GON-confirmed, 515 GON-suspected, 1524 Normal	Subjective grading of photographs	ResNet101	Color Fundus Photos	GON-confirmed vs. GON-suspected vs. Normal; Referrals (GON-confirmed and GON-suspected) vs. Observation (Normal)	Comparison of GON-confirmed vs. GON-suspected vs. Normal: Accuracy 0.941, Sensitivity 0.957, Specificity 0.929. AUC 0.992 for Referrals (GON-confirmed and GON-suspected) vs. Observation (Normal)
Medeiros et al.[24]	Train + validation (80% train, 20% validation), 9,136 Glaucoma, 13,410 Suspect, 3982 Healthy	Test: 2070 Glaucoma,3345 Suspect, 877 Healthy	SDOCT global RNFL value; Abnormal (Glaucoma) vs. Normal (Normal + Bordeline) RNFL based on classification of global RNFL by SDOCT normative database	ResNet34	Color Optic Disc Photos paired to SDOCT global RNFL	SDOCT global RNFL value; Abnormal (Glaucoma) vs. Normal RNFL	Pearson $r = 0.832$, $P < 0.001$ between DL predicted and actual SDOCT value; Mean absolute error 7.39 µm; AUC 0.944 for DL vs 0.94 for SDOCT ($P = 0.72$); 90% sensitivity at 80% specificity for both.
Thompson et al.[26]	Train + validation (80% train, 20% validation): 4,570 Glaucoma, 1924 Suspect, 1046 Healthy	Test: 970 Glaucoma, 432 Suspect, 340 Healthy	Global and sector BMO-MRW thickness values; Abnormal (Glaucoma) vs. Normal (Suspect + Normal) based on classification of BMO-MRW global and sector values by SDOCT normative database	ResNet34	Color Optic Disc Photos paired to SDOCT global BMO-MRW	Global and sector BMO-MRW thickness values; Abnormal (Glaucoma) vs. Normal	Global BMO-MRW Pearson r = 0.88 ($P < 0.001$) between DL predicted and actual SDOCT value; Mean absolute error 27.8; AUC for DL 0.945 vs. actual SDOCT 0.933 ($P = 0.59$).
Devalla et al.[67]	40 control/60 glaucoma; training on datasets of 10, 20, 30 or 40 B-scans, with equal number of glaucoma and healthy scans in each cross-validation experiment	Cross-validation experiments with test sets of 90, 80, 70, or 60.	Manual segmentation of ONH OCT	Custom eight-layer CNN	Horizontal B-scan through ONH	Digital stain of RNFL+prelamina, RPE, all other retinal layers, choroid, peripapillary sclera, lamina cribrosa	Dice coefficient 0.84, Sensitivity 92%, specificity 99%, accuracy 94%

Table 1. Continued

Citation	Training/Validation Dataset	Test Dataset	Reference	Network	Data type	Output	Results
Mariottoni et al.[65]	Train 10,520/Validation 2742	Test Set 1 (images without segmentation errors or artifacts) 11,010; Test Set 2 (low-quality images with segmentation errors) 237; Test Set 3 (images with other artifacts) 776	Global RNFL thickness value	ResNet34	SDOCT raw B-scans of peripapillary RNFL	Global RNFL thickness value	Test set 1: Pearson r 0.983 ($P < 0.001$) between predicted segmentation-free and actual SDOCT global RNFL; MAE 2.41; Test set 2: DL correlation with BAE $r = 0.972$ vs. with conventional algorithm 0.94, $P < 0.001$. Test set 3: DL correlation with BAE $r = 0.94$ vs. with conventional algorithm $r = 0.64$, $P < 0.001$.
Thompson et al.[25]	Train + Validation (50%+20%): 4828 Glaucoma, 9638 Normal	Test (30%): 3897 Glaucoma, 2443 Normal	Glaucoma (based on GON and reproducible glaucomatous visual field defects) vs. Healthy	ResNet34	SDOCT raw B-scans of peripapillary RNFL	Glaucoma vs. Healthy	AUC 0.96 for DL algorithm vs. AUC 0.87 for global RNFL thickness ($P < 0.001$)
Maetschke et al.[66]	Train (80%): 672 POAG, 216 Healthy; Validation (10%): 30 Healthy, 82 POAG	Test (10%): 93 POAG, 17 Healthy	Glaucoma (based on glaucomatous VF defects on 2 consecutive tests) vs. Healthy	Custom 5-layer CNN	OCT of the ONH	Glaucoma vs. Healthy	AUC 0.94
Asaoka et al.[7]	Pretraining: 1371 Open angle glaucoma, 193 Healthy; Training: 94 Open angle glaucoma, 84 Healthy	Test: 114 Open angle glaucoma and MD > -5 dB, 82 Healthy	Glaucoma (based on GON and glaucomatous VF defects) vs. Healthy	Custom 6-layer CNN	8 $\times$ 8 macular grid	Glaucoma vs. Healthy	AUC 0.937
Xu et al.[69]	Cross-validation (85%: 80% training/20% validation) 1632 OAG, 1764 closed	Test (15%): 311 open, 329 closed	Angle Closed vs. Open based on gonioscopic grade	ResNet18; Inception v3	Anterior Segment-OCT	Angle closed vs. Open	AUC 0.928
Fu et al.[70]	7375 open angle, 895 angle closure: 5-fold cross-validation - four groups, each with 1654 angle closure tests for training, and one group of 1654 angle closure for testing	1654 angle closure for testing within each fold	Angle closed vs. open based on gonioscopic grade	VGG-16	Anterior Segment-OCT	Angle closed vs. open	AUC 0.96, sensitivity 90%, specificity 92%
Mariottoni et al.[76]	Training/Validation:3980 Glaucoma, 3732 Normal	Test:1061 Glaucoma, 1057 Normal	GON vs. GON suspects vs. Normal based on SAP and OCT objective criteria (see Table 2)	ResNet50	Optic Disc Photos	GON vs. Normal	AUC 0.92, Sensitivity 77% at Specificity 95%

Table 1. Continued

Citation	Training/Validation Dataset	Test Dataset	Reference	Network	Data type	Output	Results
Li et al.[71]	Overall: 2389 Glaucoma, 1623 Non-glaucoma: Train: 3712	Test: 300	Glaucoma (based on glaucomatous damage to ONH and reproducible glaucomatous VF defects) vs. Healthy	VGG	Pattern Deviation plots from Humphrey Field Analyzer 30-2 or 24-2 visual field tests	Glaucoma vs. Healthy	AUC 0.966, Sensitivity 93.2%, Specificity 82.6%
Kucur et al.[72]	1979 control (Rotterdam 244; Budapest 1735), 2811 Early glaucoma (Rotterdam 2,279; Budapest 532)– 10-fold cross- validation	10-fold cross-validation; unclear if separate test and validation datasets were used	Early glaucoma (based on glaucomatous neuroretinal rim loss, reproducible VF defects, and IOP) vs. Healthy	Custom 7-layer CNN	OCTOPUS 101 G1 and Humphrey Field Analyzer 24-2 visual field tests	Early Glaucoma vs. Healthy	Average Precision: Rotterdam 87.4%, Budapest 98.6%
Asaoka et al.[10]	171 Preperimetric glaucoma vs. 108 Normal and 63 artificially generated Normal—leave one out cross-validation	Leave one out cross-validation; a separate test dataset was not used	Preperimetric OAG (based on ONH changes, VF preceding perimetric field changes) vs. Healthy	Custom DL feed-forwardneural network	Humphrey Field Analyzer 24-2	Preperimetric glaucoma vs. Healthy	AUC 0.926
Berchuck et al.[89]	Train (81%): 768 Glaucoma, 1793 Glaucoma suspects, 547 Normal Validation (9%): 83 Glaucoma, 222 Glaucoma suspect, 58 Normal5-fold cross-validation	Test (9%): 93 Glaucoma, 206 Glaucoma suspect, 62 Normal	Glaucoma (repeatable glaucomatous VF defect and corresponding optic nerve damage) vs. Glaucoma suspect (high IOP or suspicious optic nerve but no VF defect) vs. Normal (No visual field or optic nerve defect)	Deep variational autoencoder	Humphrey Field Analyzer 24-2	Rates of VF progression compared to SAP MD; Prediction of future VF compared to point-wise regression predictions	Rate of progression significantly higher for VAE than MD at 2 years (25% vs. 9%) and 4 years (35% vs. 15%) from baseline. MAE for prediction of 4^{th}, 6^{th}, and 8^{th} visits significantly smaller for VAE than PW ($P < 0.001$)
Wen et al.[91]	Train + validation (80%): 25,723 and 10-fold cross-validation	Test (20%): 6720	Actual HFA points and Mean Deviation from HVF	CascadeNet- 5	Humphrey Field Analyzer 24-2	HFA points and Mean Deviation	PMAE 2.47; Mean difference in MD between predicted and actual MD = 0.41 dB, Pearson $r = 0.92$, $P < 0.001$

BAE, best available estimate; DL, deep learning; GON, glaucomatous optic neuropathy; VF, visual field; HFA, Humphrey Field Analyzer; HVF, Humphrey Visual Field; IOP, intraocular pressure; MAE, mean absolute error; PMAE, point-wise mean absolute error; POAG, primary open angle glaucoma; OAG, open angle glaucoma; ONH, optic nerve head; RPE, retina pigmented epithelium.

programs to detect diabetic retinopathy.[41] There are a number of inexpensive, portable nonmydriatic fundus cameras that can be used to acquire imaging in a low-resource setting, making this method particularly attractive for community-based or opportunistic screening.[42] Once a deep learning model is successfully trained to recognize the presence of disease on fundus photographs, it can then be easily deployed to provide gradings on previously unseen photos in real time. Ting and colleagues[27] proposed that a deep learning algorithm could be developed to screen for glaucoma in existing teleretinal imaging. Using a large database of 494,661 teleretinal photographs acquired in diabetics, 125,189 of which had been labeled by human graders in the training set, they developed an algorithm capable of detecting images that were considered "referable" for glaucoma, based on subjective grading of the photographs by ophthalmologists or professional graders. In the test dataset, their algorithm detected "referable" glaucoma on photographs with an area under the receiver operating characteristic (ROC) curve of 0.942, sensitivity of 96.4%, and specificity of 87.2%. It is important to note that, although Ting and colleagues[27] proposed that their approach could be used to screen for glaucoma, such application would not be appropriate at the level of specificity reported. A specificity of 87.2% would translate into approximately 13% of false-positive results. When applied in the context of screening, this would result in an enormous number of healthy individuals being unnecessarily referred for evaluation, if all those with positive tests were to be referred. As a matter of fact, simple calculations of disease probability based on diagnostic likelihood ratios[43–46] show that their proposed model would generally be of little utility if applied in the context of screening. Their estimates of sensitivity and specificity would result in a positive likelihood ratio of 7.5. So, for example, if one were to suppose a prevalence of 5% for glaucoma, a positive test result would bring the post-test probability of disease (i.e., the new probability after the test result is known) to 40%. Therefore, even if an individual were to test positive, his/her chance of disease would still be relatively low. A negative test result would only serve to decrease an already very low pre-test probability of disease of 5%. Therefore both positive, as well as negative test results would not do much to change the probability of glaucoma if applied in a screening setting.

In another study, Li et al.[9] were able to derive somewhat better results. Using a similar approach, they labeled 48,116 color fundus images as "referable" (yes vs. no) for glaucoma based on human graders, then trained a deep learning algorithm using 31,745 of the images and applied the algorithm to a random subset of 8000 images that had been separated for final testing. The ROC curve area, sensitivity, and specificity were 0.986, 95.6%, and 92%, respectively. However, it is not clear from the methodology whether the same subject could have had an image in both the training and validation datasets. It was also not clear in their study whether a completely separate test set was used for final evaluation of the model. As discussed before, this could lead to biased estimates of accuracy.

In contrast to diabetic retinopathy, the approach of training deep learning models to replicate human gradings of fundus photographs for glaucoma raises numerous potential problems. Previous studies have shown that human gradings have limited reproducibility[47–49] and poor interrater reliability.[48–50] Ophthalmologists tend to undercall glaucoma in small optic discs but overcall it in physiologically enlarged cups.[9] Thus, if human graders are used as the reference standard, then the algorithms can only perform as well as the human gradings and will essentially learn to replicate these common mistakes. For example, in the study by Li et al.,[9] the deep learning algorithm tended to underdiagnose glaucoma in high myopes, thus increasing the false-negative rate, but overcalled glaucoma in physiologically enlarged cups, thus inflating the false-positive rate. If such models are to be used in the context of screening for the disease, the graders should be trained to detect cases of well-established nerve damage, not dubious, potentially "referable" or suspect cases. As described before, by targeting well-defined cases, diagnostic accuracy could be improved, leading to more effective screening tests.

An alternative approach for training deep learning models for evaluation of fundus photographs in glaucoma has been proposed by Medeiros et al.[24] and called machine-to-machine (M2M). In the M2M model, a deep learning algorithm was trained on color fundus photographs that were labeled with an objective quantitative reference standard, the corresponding global retinal nerve fiber layer thickness measurement from spectral-domain optical coherence tomography (SDOCT). Because of its high reproducibility and accuracy,[51,52] SDOCT has become the de facto standard for objective quantification of structural damage in glaucoma.[52] However, unlike color fundus photographs, SDOCT technology is expensive and not easily portable, which limits the feasibility of widespread adoption in screening efforts. By training the M2M deep learning algorithm to predict the RNFL thickness value when assessing a color fundus photograph, the degree of glaucomatous damage could be quantified rather than just "qualified." There was a strong correlation between the predicted RNFL value from the deep learning algorithm's interpretation of

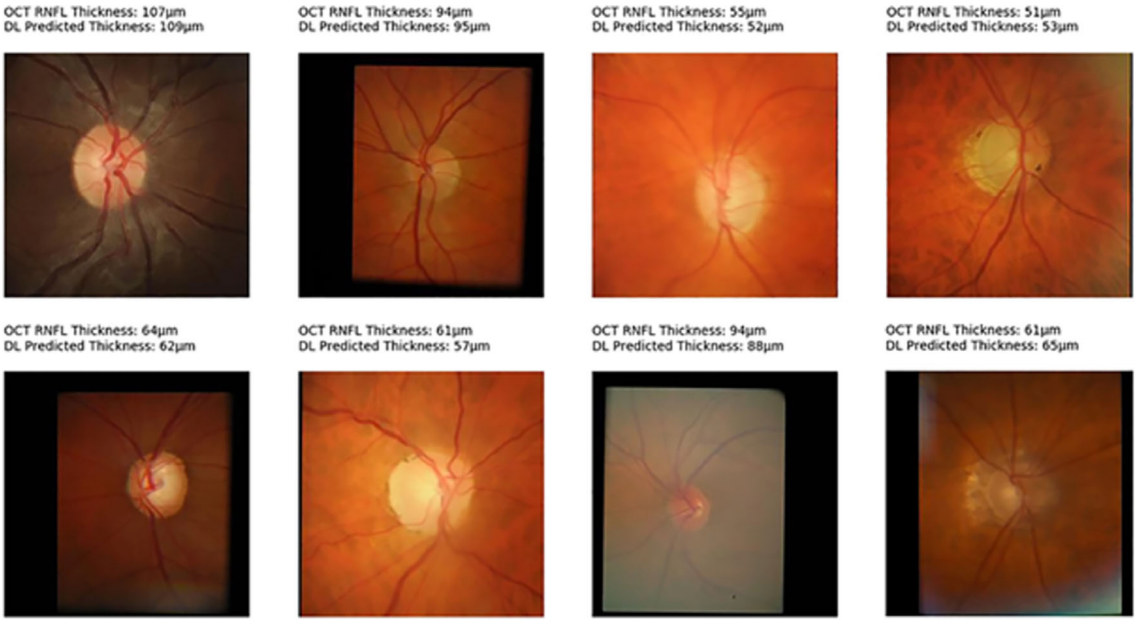

Figure 3. Examples of optic disc photographs and corresponding actual SDOCT measurements of average RNFL. Above each photo are also shown the DL prediction of average RNFL thickness from the optic disc photograph by the M2M algorithm. Note that the predictions from the DL algorithm can be quite close to actual SDOCT RNFL thickness measurements for a variety of photos. Adapted from Medeiros et al.[24].

the fundus image and the actual RNFL thickness value from the corresponding SDOCT ($r = 0.832$, $P < 0.001$), with a mean absolute error of approximately 7 µm (Fig. 3). The M2M model had a similar performance to that of SDOCT RNFL thickness to discriminate glaucomatous from normal eyes, as defined based on visual field loss, with ROC curve areas of 0.940 versus 0.944, respectively ($P = 0.724$). The authors used class activation maps or heatmaps to highlight the areas of the photographs that were most important to the deep learning model's predictions, and, as shown in Figure 4A, these maps showed that the model was correctly targeting the area of the optic disc and adjacent RNFL.

Thompson et al.[26] published a follow-up study using a similar approach in which the SDOCT Bruch's membrane opening-minimum rim width (BMO-MRW) parameter served as a reference standard for labeling optic disc photographs. BMO-MRW may be particularly useful in images where the optic disc is difficult to grade, such as cases of high myopia.[53] The DL predictions were again highly correlated with the actual BMO-MRW values (Pearson's $r = 0.88$, $P < 0.001$), and the ROC curve areas for discriminating between glaucomatous and healthy eyes were 0.945 for the DL predictions and 0.933 for the actual measurements ($P = 0.587$). Similarly, class activation maps confirmed that the neuroretinal rim was critical to the algorithm's classification (Fig. 4B). In a subsequent study, Jammal et al.[54] demonstrated that the M2M DL algorithm performed at least as well as and often better than human graders for detecting eyes with reproducible glaucomatous visual field loss.

Compared to training using subjective human labeling as the reference standard, the M2M approach may offer a distinct advantage, because the output is quantitative rather than qualitative, of allowing cutoffs to be established in order to optimize its application in a screening setting. Also, the ability to have quantification of the amount of neural loss raises the possibility that fundus photographs could be used to detect change over time in low-resource settings where SDOCT is unavailable, although this still needs confirmation.

It should be acknowledged that notable challenges remain before any of these algorithms are ready for application in real-world settings. For example, color fundus images can exhibit a wide range of photographic quality, especially when acquired in less controlled settings. It is possible that some deep learning algorithms may underperform if applied to images captured on different cameras from those used in the training dataset. The impact of co-morbid pathologies on the diagnostic performance of these algorithms is also uncertain, since they have so far been mostly trained and tested on datasets that eliminated images with other ocular pathologies (e.g., retinal diseases, high myopia). In that sense, the Pegasus (Visulytix Ltd,

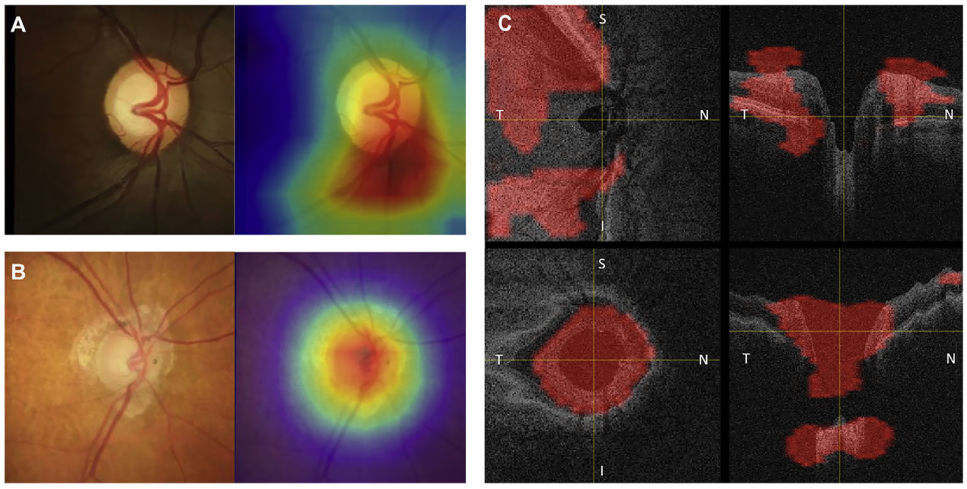

Figure 4. Class activation maps (CAM) for several examples of deep learning models. **(A)** Gradient-weighted CAM from the M2M model to predict RNFL thickness from fundus photographs. It can be seen that the heatmap correctly highlights the area of the optic nerve and adjacent RNFL as most relevant for the predictions. (adapted from Medeiros et al.[24]). **(B)** Gradient-weighted CAM map from the M2M model used to predict rim width in an eye with glaucoma. Note that the heatmap strongly highlights the cup and rim regions. (adapted from Thompson et al.[26]). **(C)** CAM showing the regions in a spectral-domain optical coherence tomography volume identified as the most important for the classification of the scan into healthy versus glaucoma. For glaucoma eyes the map generally highlighted regions that agree with established clinical markers for glaucoma diagnosis, such as the optic disc cup and neuroretinal rim. It should be noted, however, that the highlighted areas are often very broad, sometimes extending even to the vitreous (adapted from Maetschke et al.[66]).

London, UK) is a cloud-based AI system for evaluation of fundus photography that uses a collection of CNNs, each specializing on a different task as part of the image assessment, such as identification of key landmarks (optic disc, macula), clinical features, and pathology classification. The system is designed to generalize to any fundus photograph that contains the optic disc, by first using a CNN to find and extract the optic nerve and then feeding a standardized image to another CNN that performs the classification. In a study by Rogers and colleagues,[55] the Pegasus AI system was compared to 243 European Ophthalmologists and 208 British optometrists in grading photographs for the presence of glaucomatous damage, achieving an overall accuracy of 83.4% and an area under the receive operating characteristic curve of 0.871, which was comparable to that of average ophthalmologists and optometrists.

Glaucoma Diagnosis with Optical Coherence Tomography

SDOCT has become the most widespread diagnostic tool for detecting glaucomatous structural damage.[51,52] Measurements of the optic nerve head, macula, and RNFL are routinely used in clinical practice for disease diagnosis and detection of progression.[51,56–58] However, conventional assessment of structural damage with SDOCT requires that the structures of interest be segmented, so that appropriate measurements, such as RNFL thickness, can be extracted. This segmentation process is done automatically by the machine's software, but it is still largely imperfect. Studies have documented segmentation errors and artifacts on 19.9% to 46.3% of SDOCT scans of the RNFL.[59–63] Manual review and correction of segmentation errors, although possible, are time-consuming and difficult to perform in a busy clinical practice. Another difficulty in the interpretation of SDOCT scans arises from the multiple parameters and regions that are analyzed. It can be difficult for the clinician to integrate all the information derived from global and sectoral RNFL thickness measurements, as well as topographic optic nerve head parameters and macular assessment. The large number of parameters increases the chance of committing what is known as a type I error, in which an abnormality is found just by chance. This has led to the concept of "red disease," in which a diagnosis of glaucoma is erroneously made based solely on the finding of a "red" result on one or a few of the parameters given in the SDOCT printout, without other corroborating clinical features.[64]

Given these limitations in the interpretation of OCT, deep learning models may provide alternative ways to quantify structural damage without reliance on previously defined features derived from the automated segmentation software. As noted before, deep learning algorithms can learn features from data automatically, as long as enough data are given to

them. Therefore these models can make use of raw SDOCT images without requiring the input of predefined features. Along those lines, Mariottoni et al.[65] recently demonstrated that a segmentation-free deep learning algorithm could be trained to predict RNFL thickness when assessing a raw OCT B-scan. The segmentation-free predictions were highly correlated with the conventional RNFL thickness ($r = 0.983$, $P < 0.001$), with mean absolute error of approximately 2 µm in good-quality images. Most importantly, in images where the conventional segmentation failed, the deep learning model still extracted reliable RNFL thickness information. In a more general approach, Thompson et al.[25] showed that a deep learning algorithm could be trained using the raw SDOCT B-scan to directly discriminate glaucomatous from healthy eyes. The proposed algorithm achieved a better diagnostic performance than the conventional RNFL thickness parameters from the instrument's printout, with area under the ROC curve of 0.96 vs. 0.87 for the global peripapillary RNFL thickness ($P < 0.001$). Another study by Maetschke et al.[66] similarly developed a deep learning algorithm that could distinguish between glaucomatous and healthy eyes using raw, unsegmented OCT volumes of the optic nerve head. The algorithm also performed better than conventional SDOCT parameters, with an area under the ROC curve of 0.94 versus 0.89 for a logistic regression model combining SDOCT parameters. As illustrated in Figure 4C, the class activation maps (heatmaps) appeared to highlight regions in the OCT volume that have been clinically well established as important to glaucoma diagnosis, particularly the neuroretinal rim, optic disc cupping, and the lamina cribrosa and its surrounding area. Heatmaps can help us better understand a CNN by highlighting the most relevant pixels in the image used for the predictions. Highlighted regions can thus be subjected to more detailed analysis. It should be noted, however, that class activation maps usually do not have enough resolution to be able to precisely pinpoint small areas that were relevant for the classification. This lack of precision occurs because of the way deep learning models with convolutional layers are built, leading to a down-sampling of the final layers from which the maps are created. Also, the efficiency of a heatmap largely depends on the model used and the amount and quality of available training data. As such, one can see from the heatmaps shown in Figure 4C that they highlight very broad areas, which sometimes seem to include even the vitreous as relevant to the discrimination of glaucoma from normal. Although the deep learning algorithm may indeed be capturing information that is not yet clear to human eyes, the resolution limitations of these heatmaps need to be kept in mind.

In addition to RNFL[65] and optic nerve head scans,[66,67] deep learning models have also been used to investigate macular scans.[7,68] Asaoka et al.[7] showed that a deep learning model built from an 8 × 8 macular grid showed superior performance for detecting glaucoma damage compared to macular RNFL thickness or ganglion cell complex measurements. Of interest, the deep learning model also performed better than traditional techniques of SVM and random forest applied to the macular measurements. In another study, Muhammad and colleagues[68] attempted to build a "hybrid" deep learning system to detect glaucoma from wide-field swept-source OCT. In their approach, a pretrained CNN was initially used to extract features from probability map images, which were then used as input to a random forest model for classification. Their model performed better than conventional summary OCT parameters. However, the study included a very small sample of only 57 glaucoma patients and 45 healthy subjects. Even though the authors claim to have used cross-validation to assess the performance of the model on a different sample than that used for training, such a small sample is unlikely to allow for enough variation and generalizability.

In addition to analysis of posterior segment OCT, deep learning models have also been applied to anterior segment OCT images for diagnosing narrow angles or angle closure.[69,70] Fu et al.[70] found an area under the ROC curve of 0.96 for a deep learning system trained to detect angle closure from Visante OCT images, with sensitivity of 90% and specificity of 92%, compared to clinician gradings of the same images as reference standard. In another work, Xu et al.[69] tested three different multiclass CNNs in Chinese-American eyes, and the best-performing classifier (ResNet18 architecture) detected gonioscopic angle closure with an area under the ROC curve of 0.928 in the test dataset. Given the difficulties in the subjective interpretation of anterior segment OCT images, such models offer great promise in automating the evaluation of those images for detecting the presence of narrow angles.

Glaucoma Diagnosis with Standard Automated Perimetry

Visual field data have also been harnessed to train various deep learning algorithms to detect glaucomatous damage, often showing a similar, if not better, level of performance relative to expert graders.[15,21,22] Li et al[71] showed that a deep learning algorithm trained with the probability map of the pattern deviation image was better able to distinguish normal from glaucomatous visual fields (accuracy 87.6%) than either human graders (62.6%), the Glaucoma Staging

System 2 (52.3%), or the Advanced Glaucoma Intervention Study criteria (45.9%). Work by Asaoka and colleagues[10] suggested that deep learning algorithms may be preferable to other traditional machine learning classifiers for diagnosis of glaucoma in visual fields that still appear to be normal based on standard parameters. However, their study did not show convincing data on how much earlier the deep learning model could detect damage before conventional parameters. CNNs have also been shown to discriminate between controls and early glaucoma on visual fields with a higher accuracy than use of standard perimetry mean deviation (MD) or neural networks without convolutional features.[72]

Several machine learning approaches have been used to attempt to classify visual field data following unsupervised approaches, with methods such as independent component analysis and its variations.[17–19,34,35] In a more recent study, Elze et al[73] proposed a technique of "archetypal analysis" to classify patterns of visual field loss in glaucoma. The authors showed that the patterns detected by their technique, such as arcuate, partial arcuate, etc., corresponded well to classification by human graders in the Ocular Hypertension Treatment Study. In a follow-up study, Wang et al.[32] proposed to use archetypal analysis to classify central visual field patterns in glaucoma. It should be noted, however, that archetypal analysis is a statistical technique closely resembling traditional factor analysis and bearing no relationship to deep learning artificial neural networks. This fact, however, does not negate its potential benefit and future studies should evaluate whether this technique may provide clinically relevant information to be used in practice.

An important consideration of applying DL models to SAP interpretation is that these models are usually trained only with reliable tests and may not be able to identify unreliable exams which are often seen in clinical practice. Thus, before such approaches can be applied in real-world settings, DL models may also need to be trained to learn which tests are of poor or good quality and reliability.

Glaucoma Diagnosis with Structure and Function

Previous studies have shown that machine learning classifier models trained with both structural and functional tests may have improved discriminatory power compared to those trained with either structure or function alone.[7,14,16,74,75] In a similar way, it is likely that deep learning models trained with a combination of structural and functional tests may also show improved performance. It should be noted, however, that there are challenges related to the development of such models. Most notable is the lack of a perfect reference standard, or "gold standard," for glaucoma diagnosis. Given the fact that the reference standard may use a combination of structural and functional tests, it becomes difficult to set up a proper unbiased study to evaluate the diagnostic accuracy of a new diagnostic method proposing to also use a combination of structure and function. In these circumstances, it is important to understand the clinical purpose and the settings to which the new method is intended for application. For example, suppose that one wishes to develop a deep learning model that can replicate in a clinical setting the performance of glaucoma experts in diagnosing the disease. It is then reasonable to set up a study where the experts will produce the reference standard by grading a combination of SDOCT images and visual field printouts, perhaps accompanied by other clinical information, and a deep learning model will be trained to attempt to replicate such standard as much as possible, based on all available information as well. Such a model could have tremendous impact in clinical practice by bringing general practitioners to a level comparable to those of experts in diagnosing the disease in a clinical setting.

In a recent study, Mariottoni and colleagues[76] proposed a set of relatively simple structural and functional parameters that could be combined in an objective way to be used as a robust reference standard for the development of AI models for glaucoma diagnosis. The criteria proposed that a diagnosis of glaucomatous optic neuropathy should involve corresponding structural and functional damage, based on RNFL assessment by SDOCT and visual field assessment by standard automated perimetry. The set of criteria are summarized in Table 2 and uses well-established global and localized parameters with the requirement for topographic correspondence between structural and functional damage, which greatly enhances specificity. The authors then developed a deep learning model that used fundus photographs to discriminate glaucoma from normal eyes, which had been classified based on the objective reference standard. The model achieved an overall area under the ROC curve of 0.92. Of note, an objective reference standard combining SDOCT and SAP data may obviate the need for laborious and time-consuming expert gradings, and may increase the comparability of diagnostic studies across devices and populations.

Glaucoma Progression

Diagnosis of glaucoma progression remains a considerable challenge in clinical practice. The tests used to assess change over time, such as SDOCT and SAP, suffer from considerable test-retest

Table 2. Summary of Proposed Objective Criteria for Definition of GON

	SDOCT	SAP
GON		
Global loss	Global RNFL thickness outside normal limits	GHT outside normal limits or PSD, $P < 5\%$
Localized loss	RNFL thickness outside normal limits in at least one superior sector (temporal superior and/or nasal superior)	Inferior MD, $P < 5\%$
	RNFL thickness outside normal limits in at least one inferior sector (temporal inferior and/or nasal inferior)	Superior MD, $P < 5\%$
Normal	RNFL thickness within normal limits for all sectors and global	PSD probability not significant ($P > 5\%$) and GHT within normal limits

To be considered glaucomatous optic neuropathy, it was necessary to meet the criteria for global or localized loss. To be considered normal, it was required that both SDOCT and SAP results were normal. SDOCT-SAP pairs that do not meet the criteria for GON or normal are considered suspects. GHT, glaucoma hemifield test; PSD, pattern standard deviation; GON, glaucomatous optic neuropathy; SDOCT, spectral-domain optical coherence tomography; SAP, standard automated perimetry.

variability, making it difficult to discriminate true change from variability. In addition, there is no consensus on specific criteria to diagnose visual field or structural progression in glaucoma. Such a lack of consensus has hampered progress in the field and has made it difficult to compare the results of different approaches.

Despite these challenges, several groups have applied traditional machine learning methods to attempt to improve the ability to diagnose glaucomatous progression on clinical tests.[17–20,77–81] In one of the earliest works of this type, Brigatti et al.[78] trained a shallow artificial neural network using visual field indices (i.e., mean defect, threshold points, corrected loss, variance, false-positive ratio, false-negative ratio), along with patient age, and showed a sensitivity of 73% and specificity of 88% when human gradings were used as the reference standard. Given the lack of a perfect reference standard, several authors have proposed using unsupervised techniques to attempt to detect visual field progression. Sample et al.[17] proposed the use of independent component analysis to identify patterns of glaucoma damage and their change over time. Subsequent studies along the same line used variations of the methodology, including variational Bayesian independent component mixture model[18] and Gaussian mixture-model with expectation maximization.[19,20] Several other approaches have used Bayesian modeling in some way to improve prediction and detection of glaucoma progression and also to combine structural and functional measurements.[12,82–88] Of note, although promising, none of these techniques have been widely incorporated in clinical practice. The reason may rely on difficulties of implementation to clinical workflow, which may eventually be overcome with widespread adoption of electronical health records and clinical decision support systems. It is important to note, however, that not all of these methods have shown consistent and substantial advantages compared to relatively simple, well-established, and intuitive methods such as guided progression analysis (GPA; Carl-Zeiss Meditec, Inc, Jena, Germany) or trend-based analysis of mean deviation over time.

As for deep learning applications in detecting progression, there have been only very few studies. Berchuck et al.[89] proposed a deep learning variational autoencoder (VAE) model to learn a low-dimensional representation of SAP visual fields using 29,161 fields from 3832 patients. The model was then applied to predict rates of change and future visual field observations. The authors found that at four years of follow-up, the model identified 35% of the eyes as progressing versus only 15% for MD. In another study, Park et al.[90] used a recurrent neural network and showed that it achieved better prediction of future visual field observations compared to ordinary least squares linear regression. Wen et al.[91] also attempted to set up a deep learning model to predict future visual field observations based on the first visual field test only. Their model confirmed the patterns of progression that we know from clinical practice (e.g., a nasal step becomes an arcuate) based on a single Humphrey Visual Field. However, because only the baseline field was used in the prediction, the model was not able to provide information regarding when or how quickly progression would occur.

Conclusions

Deep learning is an exciting technique that holds enormous promise in glaucoma. Deep learning models have consistently been shown to detect and quantify glaucomatous damage using simple color fundus photographs, opening the potential for low-cost screening tests for the disease. In addition, deep learning has been shown to improve assessment of damage on raw SDOCT images and visual field data, which could improve the use of these tests in clinical practice. However, it should be noted that no matter how exciting AI technologies can be, validation of new diagnostic tests should be based on rigorous methodology with particular attention paid to how the reference standards are defined and the settings where the tests are going to be applied in practice. This is especially true for a disease such as glaucoma, where no litmus test exists for diagnosis or detection of change over time. The reference standards to be used may differ significantly, depending on how the test is going to be applied and its purpose. Similarly, the requirements for diagnostic accuracy may vary considerably depending on whether the test is being considered for community-based or opportunistic screening versus detection or monitoring of disease in a tertiary care center. Although significant progress has been made with AI and deep learning in glaucoma, a lot of work remains to be done.

Acknowledgments

Supported by National Institutes of Health/National Eye Institute grants number EY029885 (FAM).

Disclosure: **A.C. Thompson,** None; **A.A. Jammal,** None; **F.A. Medeiros,** Aeri Pharmaceuticals (C); Allergan (C, F), Annexon (C); Biogen (C); Carl Zeiss Meditec (C, F), Galimedix (C); Google Inc. (F); Heidelberg Engineering (F), IDx (C); nGoggle Inc. (P), Novartis (F); Stealth Biotherapeutics (C); Reichert (C, F)

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121:2081–2090.
2. Budenz DL, Barton K, Whiteside-de Vos J, et al. Prevalence of glaucoma in an urban West African population: the Tema Eye Survey. *JAMA Ophthalmol*. 2013;131:651–658.
3. Hennis A, Wu SY, Nemesure B, Honkanen R, Leske MC, Barbados Eye Studies G. Awareness of incident open-angle glaucoma in a population study: the Barbados Eye Studies. *Ophthalmology*. 2007;114:1816–1821.
4. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA*. 2014;311:1901–1911.
5. Harwerth RS, Carter-Dawson L, Barnes G, III, Holt WF, Crawford MLJ. Neural Losses Correlated with Visual Losses in Clinical Perimetry. *Invest Ophthalmol Vis Sci*. 2004;45:3152–3160.
6. Harwerth RS, Carter-Dawson L, Shen F, Crawford ML, 3rd. Ganglion cell losses underlying visual field defects from experimental glaucoma. *Invest Ophthalmol Vis Sci*. 1999;40:2242–2250.
7. Asaoka R, Murata H, Hirasawa K, et al. Using deep learning and transform learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. 2019;198:136–145.
8. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8:14665.
9. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125:1199–1206.
10. Asaoka R, Murata H, Iwase A, Araie M. Detecting Preperimetric Glaucoma with Standard Automated Perimetry Using a Deep Learning Classifier. *Ophthalmology*. 2016;123:1974–1980.
11. Burgansky-Eliash Z, Wollstein G, Chu T, et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Invest Ophthalmol Vis Sci*. 2005;46:4147–4152.
12. Belghith A, Bowd C, Medeiros FA, Balasubramanian M, Weinreb RN, Zangwill LM. Learning from healthy and stable eyes: A new approach for detection of glaucomatous progression. *Artif Intell Med*. 2015;64:105–115.
13. Shigueoka LS, Vasconcellos JPC, Schimiti RB, et al. Automated algorithms combining structure and function outperform general ophthalmologists in diagnosing glaucoma. *PLoS One*. 2018;13:e0207784.

14. Bowd C, Hao J, Tavares IM, et al. Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes. *Invest Ophthalmol Vis Sci*. 2008;49:945–953.
15. Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN, Sejnowski TJ. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. *IEEE Trans Biomed Eng*. 2002;49:963–974.
16. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. *PLoS One*. 2017;12:e0177726.
17. Sample PA, Boden C, Zhang Z, et al. Unsupervised machine learning with independent component analysis to identify areas of progression in glaucomatous visual fields. *Invest Ophthalmol Vis Sci*. 2005;46:3684–3692.
18. Goldbaum MH, Lee I, Jang G, et al. Progression of patterns (POP): a machine classifier algorithm to identify glaucoma progression in visual fields. *Invest Ophthalmol Vis Sci*. 2012;53:6557–6567.
19. Yousefi S, Balasubramanian M, Goldbaum MH, et al. Unsupervised Gaussian mixture-model with expectation maximization for detecting glaucomatous progression in standard automated perimetry visual fields. *Translational Vision Science & Technology*. 2016;5:2.
20. Yousefi S, Kiwaki T, Zheng Y, et al. Detection of longitudinal visual field progression in glaucoma using machine learning. *Am J Ophthalmol*. 2018;193:71–79.
21. Goldbaum MH, Sample PA, Chan K, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. *Invest Ophthalmol Vis Sci*. 2002;43:162–169.
22. Goldbaum MH, Sample PA, White H, et al. Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci*. 1994;35:3362–3373.
23. Chollet F. *Deep Learning with Python*. Shelter Island, NY: Manning Publications Co.; 2018.
24. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*. 2019;126:513–521.
25. Thompson AC, Jammal AA, Berchuck SI, Mariottoni EB, Medeiros FA. Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmol*. 2020;138:333–339.
26. Thompson AC, Jammal AA, Medeiros FA. A deep learning algorithm to quantify neuroretinal rim loss from optic disc photographs. *Am J Ophthalmol*. 2019;201:9–18.
27. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.
28. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmology*. 2019;137:1353–1360.
29. 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:e0207982.
30. Li F, Yan L, Wang Y, et al. Deep learning-based automated detection of glaucomatous optic neuropathy on color fundus photographs. *Graefes Arch Clin Exp Ophthalmol*. 2020;258:851–867.
31. 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;126:1627–1639.
32. Wang M, Tichelaar J, Pasquale LR, et al. Characterization of central visual field loss in end-stage glaucoma by unsupervised artificial intelligence. *JAMA Ophthalmol*. 2020;138:190–198.
33. Sample PA, Chan K, Boden C, et al. Using unsupervised learning with variational bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci*. 2004;45:2596–2605.
34. Goldbaum MH. Unsupervised learning with independent component analysis can identify patterns of glaucomatous visual field defects. *Trans Am Ophthalmol Soc*. 2005;103:270–280.
35. Goldbaum MH, Sample PA, Zhang Z, et al. Using unsupervised learning with independent component analysis to identify patterns of glaucomatous visual field defects. *Invest Ophthalmol Vis Sci*. 2005;46:3676–3683.
36. 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;8:16685.
37. Zhao R, Chen X, Xiyao L, Zailiang C, Guo F, Li S. Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. *IEEE J Biomed Health Inform*. 2020;24:1104–1113.

38. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3:9.
39. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *ArXiv e-prints*. 2015. https://arxiv.org/abs/1512.03385. Accessed 12.01.15.
40. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *ArXiv e-prints*. 2015. arXiv:1512.00567. Accessed 08.01.16.
41. Owsley C, McGwin G, Jr., Lee DJ, et al. Diabetes eye screening in urban settings serving minority populations: detection of diabetic retinopathy and other ocular findings using telemedicine. *JAMA Ophthalmol*. 2015;133:174–181.
42. Miller SE, Thapa S, Robin AL, et al. Glaucoma screening in nepal: cup-to-disc estimate with standard mydriatic fundus camera compared to portable nonmydriatic camera. *Am J Ophthalmol*. 2017;182:99–106.
43. Lisboa R, Mansouri K, Zangwill LM, Weinreb RN, Medeiros FA. Likelihood ratios for glaucoma diagnosis using spectral-domain optical coherence tomography. *Am J Ophthalmol*. 2013;156:918–926.e912.
44. Medeiros FA, Zangwill LM, Bowd C, Weinreb RN. Comparison of the GDx VCC scanning laser polarimeter, HRT II confocal scanning laser ophthalmoscope, and stratus OCT optical coherence tomograph for the detection of glaucoma. *Arch Ophthalmol*. 2004;122:827–837.
45. Medeiros FA, Zangwill LM, Bowd C, Sample PA, Weinreb RN. Use of progressive glaucomatous optic disk change as the reference standard for evaluation of diagnostic tests in glaucoma. *Am J Ophthalmol*. 2005;139:1010–1018.
46. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA*. 1994;271:703–707.
47. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215–221.
48. Abrams LS, Scott IU, Spaeth GL, Quigley HA, Varma R. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. *Ophthalmology*. 1994;101:1662–1667.
49. Jampel HD, Friedman D, Quigley H, et al. Agreement among glaucoma specialists in assessing progressive disc changes from photographs in open-angle glaucoma patients. *Am J Ophthalmol*. 2009;147:39–44 e31.
50. Chan HH, Ong DN, Kong YX, et al. Glaucomatous optic neuropathy evaluation (GONE) project: the effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am J Ophthalmol*. 2014;157:936–944.
51. Leung CK, Cheung CY, Weinreb RN, et al. Retinal nerve fiber layer imaging with spectral-domain optical coherence tomography: a variability and diagnostic performance study. *Ophthalmology*. 2009;116:1257–1263, 1263.e1251-1252.
52. Tatham AJ, Medeiros FA. Detecting structural progression in glaucoma with optical coherence tomography. *Ophthalmology*. 2017;124:S57–S65.
53. Reznicek L, Burzer S, Laubichler A, et al. Structure-function relationship comparison between retinal nerve fibre layer and Bruch's membrane opening-minimum rim width in glaucoma. *Int J Ophthalmol-Chi*. 2017;10:1534–1538.
54. Jammal AA, Thompson AC, Mariottoni EB, et al. Human versus machine: comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *Am J Ophthalmol*. 2020;211:123–131.
55. Rogers TW, Jaccard N, Carbonaro F, et al. Evaluation of an AI system for the automated detection of glaucoma from stereoscopic optic disc photographs: the European Optic Disc Assessment Study. *Eye (Lond)*. 2019;33:1791–1797.
56. Medeiros FA, Zangwill LM, Bowd C, Vessani RM, Susanna R, Jr., Weinreb RN. Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *Am J Ophthalmol*. 2005;139:44–55.
57. Roh KH, Jeoung JW, Park KH, Yoo BW, Kim DM. Long-term reproducibility of cirrus HD optical coherence tomography deviation map in clinically stable glaucomatous eyes. *Ophthalmology*. 2013;120:969–977.
58. Dong ZM, Wollstein G, Schuman JS. Clinical utility of optical coherence tomography in glaucoma. *Invest Ophthalmol Vis Sci*. 2016;57:OCT556–OCT567.
59. Asrani S, Essaid L, Alder BD, Santiago-Turla C. Artifacts in spectral-domain optical coherence tomography measurements in glaucoma. *JAMA Ophthalmol*. 2014;132:396–402.
60. 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.

61. Liu Y, Simavli H, Que CJ, et al. Patient characteristics associated with artifacts in Spectralis optical coherence tomography imaging of the retinal nerve fiber layer in glaucoma. *Am J Ophthalmol*. 2015;159:565–576.e562.
62. Miki A, Kumoi M, Usui S, et al. Prevalence and associated factors of segmentation errors in the peripapillary retinal nerve fiber layer and macular ganglion cell complex in spectral-domain optical coherence tomography images. *J Glaucoma*. 2017;26:995–1000.
63. Hardin JS, Taibbi G, Nelson SC, Chao D, Vizzeri G. Factors affecting cirrus-HD OCT optic disc scan quality: a review with case examples. *J Ophthalmol*. 2015;2015:746150.
64. Chong GT, Lee RK. Glaucoma versus red disease: imaging and glaucoma diagnosis. *Curr Opin Ophthalmol*. 2012;23:79–88.
65. Mariottoni EB, Jammal AA, Urata CN, et al. Quantification of retinal nerve fibre layer thickness on optical coherence tomography with a deep learning segmentation-free approach. *Scientific Reports*. 2020;10:402.
66. Maetschke S, Antony B, Ishikawa H, Wollstein G, Schuman J, Garnavi R. A feature agnostic approach for glaucoma detection in OCT volumes. *PLoS One*. 2019;14:e0219126.
67. Devalla SK, Chin KS, Mari JM, et al. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. *Invest Ophthalmol Vis Sci*. 2018;59:63–74.
68. 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;26:1086–1094.
69. Xu BY, Chiang M, Chaudhary S, Kulkarni S, Pardeshi AA, Varma R. Deep learning classifiers for automated detection of gonioscopic angle closure based on anterior segment OCT images. *Am J Ophthalmol*. 2019;208:273–280.
70. Fu H, Baskaran M, Xu Y, et al. A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. *Am J Ophthalmol*. 2019;203:37–45.
71. Li F, Wang Z, Qu G, et al. Automatic differentiation of Glaucoma visual field from non-glaucoma visual filed using deep convolutional neural network. *BMC Med Imaging*. 2018;18:35.
72. Kucur SS, Hollo G, Sznitman R. A deep learning approach to automatic detection of early glaucoma from visual fields. *PloS One*. 2018;13:e0206081.
73. Elze T, Pasquale LR, Shen LQ, Chen TC, Wiggs JL, Bex PJ. Patterns of functional vision loss in glaucoma determined with archetypal analysis. *J R Soc Interface*. 2015;12:20141118.
74. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol*. 1996;121:511–521.
75. Racette L, Chiou CY, Hao J, et al. Combining functional and structural tests improves the diagnostic accuracy of relevance vector machine classifiers. *J Glaucoma*. 2010;19:167–175.
76. Mariottoni EB, Jammal AA, Berchuck SI, Tavares IM, Medeiros FA. An objective structural and functional reference standard for diagnostic studies in glaucoma. medRxiv preprint. 2020, https://www.medrxiv.org/content/10.1101/2020.04.10.20057836v2. Accessed 04.19.20.
77. 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;59:2748–2756.
78. Brigatti L, Nouri-Mahdavi K, Weitzman M, Caprioli J. Automatic detection of glaucomatous visual field progression with neural networks. *Arch Ophthalmol*. 1997;115:725–728.
79. Lee J, Kim YK, Jeoung JW, Ha A, Kim YW, Park KH. Machine learning classifiers-based prediction of normal-tension glaucoma progression in young myopic patients. *Jpn J Ophthalmol*. 2020;64:68–76.
80. Yousefi S, Goldbaum MH, Balasubramanian M, et al. Glaucoma progression detection using structural retinal nerve fiber layer measurements and functional visual field points. *IEEE Trans Biomed Eng*. 2014;61:1143–1154.
81. Garcia GP, Nitta K, Lavieri MS, et al. Using Kalman filtering to forecast disease trajectory for patients with normal tension glaucoma. *Am J Ophthalmol*. 2019;199:111–119.
82. Medeiros FA, Zangwill LM, Weinreb RN. Improved prediction of rates of visual field loss in glaucoma using empirical Bayes estimates of slopes of change. *J Glaucoma*. 2012;21:147–154.
83. Medeiros FA, Zangwill LM, Girkin CA, Liebmann JM, Weinreb RN. Combining structural and functional measurements to improve estimates of rates of glaucomatous progression. *Am J Ophthalmol*. 2012;153:1197–1205.e1191.
84. Medeiros FA, Zangwill LM, Mansouri K, Lisboa R, Tafreshi A, Weinreb RN. Incorporating risk factors to improve the assessment of rates of glaucomatous progression. *Invest Ophthalmol Vis Sci*. 2012;53:2199–2207.
85. Medeiros FA, Weinreb RN, Moore G, Liebmann JM, Girkin CA, Zangwill LM. Integrating

event- and trend-based analyses to improve detection of glaucomatous visual field progression. *Ophthalmology*. 2012;119:458–467.

86. Zhu H, Crabb DP, Ho T, Garway-Heath DF. More accurate modeling of visual field progression in glaucoma: ANSWERS. *Invest Ophthalmol Vis Sci*. 2015;56:6077–6083.
87. Murata H, Zangwill LM, Fujino Y, et al. Validating variational Bayes linear regression method with multi-central datasets. *Invest Ophthalmol Vis Sci*. 2018;59:1897–1904.
88. Murata H, Araie M, Asaoka R. A new approach to measure visual field progression in glaucoma patients using variational bayes linear regression. *Invest Ophthalmol Vis Sci*. 2014;55:8386–8392.
89. Berchuck SI, Mukherjee S, Medeiros FA. Estimating rates of progression and predicting future visual fields in glaucoma using a deep variational autoencoder. *Scientific Reports*. 2019;9:18113.
90. Park K, Kim J, Lee J. Visual field prediction using recurrent neural network. *Sci Rep*. 2019;9:8385.
91. Wen JC, Lee CS, Keane PA, et al. Forecasting future Humphrey visual fields using deep learning. *PLoS One*. 2019;14:e0214875.