tvst

Special Issue

A 3D Deep Learning System for Detecting Referable Glaucoma Using Full OCT Macular Cube Scans

Daniel B. Russakoff[1], Suria S. Mannil[2], Jonathan D. Oakley[1], An Ran Ran[3], Carol Y. Cheung[3], Srilakshmi Dasari[4], Mohammed Riyazzuddin[4], Sriharsha Nagaraj[4], Harsha L. Rao[4], Dolly Chang[2], and Robert T. Chang[2]

[1] Voxeleron LLC, San Francisco, CA, USA
[2] Byers Eye Institute, Stanford University, Palo Alto, CA, USA
[3] Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong
[4] Narayana Nethralaya Foundation, Bangalore-India

Correspondence: Robert T. Chang, Department of Ophthalmology, Byers Eye Institute, Stanford University, 2452 Watson Ct, Palo Alto, CA 94303, USA. e-mail: viroptic@gmail.com

Received: October 1, 2019
Accepted: December 10, 2019
Published: February 18, 2020

Keywords: machine learning; glaucoma; suspects

Citation: Russakoff DB, Mannil SS, Oakley JD, Ran AR, Cheung CY, Dasari S, Riyazzuddin M, Nagaraj S, Rao HL, Chang D, Chang RT. A 3D deep learning system for detecting referable glaucoma using full OCT macular cube scans. Trans Vis Sci Tech. 2020;9(2):12, https://doi.org/10.1167/tvst.9.2.12

Purpose: The purpose of this study was to develop a 3D deep learning system from spectral domain optical coherence tomography (SD-OCT) macular cubes to differentiate between referable and nonreferable cases for glaucoma applied to real-world datasets to understand how this would affect the performance.

Methods: There were 2805 Cirrus optical coherence tomography (OCT) macula volumes (Macula protocol 512 $\times$ 128) of 1095 eyes from 586 patients at a single site that were used to train a fully 3D convolutional neural network (CNN). Referable glaucoma included true glaucoma, pre-perimetric glaucoma, and high-risk suspects, based on qualitative fundus photographs, visual fields, OCT reports, and clinical examinations, including intraocular pressure (IOP) and treatment history as the binary (two class) ground truth. The curated real-world dataset did not include eyes with retinal disease or nonglaucomatous optic neuropathies. The cubes were first homogenized using layer segmentation with the Orion Software (Voxeleron) to achieve standardization. The algorithm was tested on two separate external validation sets from different glaucoma studies, comprised of Cirrus macular cube scans of 505 and 336 eyes, respectively.

Results: The area under the receiver operating characteristic (AUROC) curve for the development dataset for distinguishing referable glaucoma was 0.88 for our CNN using homogenization, 0.82 without homogenization, and 0.81 for a CNN architecture from the existing literature. For the external validation datasets, which had different glaucoma definitions, the AUCs were 0.78 and 0.95, respectively. The performance of the model across myopia severity distribution has been assessed in the dataset from the United States and was found to have an AUC of 0.85, 0.92, and 0.95 in the severe, moderate, and mild myopia, respectively.

Conclusions: A 3D deep learning algorithm trained on macular OCT volumes without retinal disease to detect referable glaucoma performs better with retinal segmentation preprocessing and performs reasonably well across all levels of myopia.

Translational Relevance: Interpretation of OCT macula volumes based on normative data color distributions is highly influenced by population demographics and characteristics, such as refractive error, as well as the size of the normative database. Referable glaucoma, in this study, was chosen to include cases that should be seen by a specialist. This study is unique because it uses multimodal patient data for the glaucoma definition, and includes all severities of myopia as well as validates the algorithm with international data to understand generalizability potential.

Introduction

Glaucoma is a chronic, optic neuropathy characterized by visual field defects and progressive vision loss.[1,2] Optical coherence tomography (OCT) based thickness measurements of retinal nerve fiber layer (RNFL) and the ganglion cell with inner plexiform layer (GCIPL) are commonly used scanning protocols by both ophthalmologists and optometrists in the screening, diagnosis, and monitoring of glaucoma.[3] Improved segmentation algorithms built into spectral domain OCT (SD-OCT) facilitate the assessment of macular parameters for glaucoma evaluation, the region of the retina with the highest concentration of retinal ganglion cells (RGCs).[4] Some studies have shown a stronger relationship between GCIPL thickness and progression of structural glaucomatous loss compared to peripapillary RNFL thickness.[5] Therefore, measurements of the macular GCIPL may reflect axonal loss earlier.[5] However, these GCIPL parameters are extracted from a segmented slice and compared to a normative database. Thus, many false positives can occur in high myopes or due to other scan artifacts during real-world scan acquisition (compared to clinical trials).[6,7] The Cirrus normative database is relatively small to be representative of the entire range of healthy population, and, thus, the authors hypothesize that deep learning applied to the macular cube in larger datasets of referable and nonreferable disease may create a better algorithm than normative data. In addition, a more accurate ground truth diagnosis of glaucoma using multimodal longitudinal clinical data may help reduce human diagnosis label variation from just interpreting the processed OCT report summary as to which cases to refer.

Even though 3D OCT images are readily available, clinicians often do not have time to view every slice in either the macula or the optic nerve. For example, only 512 samplings along a 3.4 mm circle of the 40,000 samplings (1.28%) are used in the current RNFL thickness analysis.[7] The Ganglion Cell Analysis (GCA) algorithm uses the data obtained by the Cirrus Macular Cube 512 × 128 scan protocol (a total of 65,536 sampled points) within a 14.13 mm^2 elliptical annulus area with the fovea at the center in over 1024 samplings to detect and measure macular GCIPL thickness.[8] The measurements are compared to a normative database and color coded into four categories (white, green, yellow, and red) for rapid clinical use.[7] Due to this relatively small sampling and the wide natural variance of RNFL and GCIPL parameters, especially from axial length and high myopia, results obtained by SD-OCT may be incorrectly flagged as abnormal but are not necessarily due to the occurrence of real disease.[9,10] In addition, early signs of pathologic damage may go unnoticed, because the majority of the 3D dataset is not summarized in the report. Moreover, subtle pathologic changes are difficult to detect using predefined sectors because all the data in each sector is summarized by a single index, which is not a sensitive method to assess early disease damage.[9] One of the important benefits of 3D volumes is the 3D spatial contextual information available, which can be a tremendous help in disease characterization that are ambiguous in an individual 2D B-scan[11] and, thus, algorithms may find important patterns that humans may not see. Whereas the normative database for the present Cirrus SD-OCT algorithm consists of 284 healthy individuals,[7] deep learning algorithms applied to the entire cube scan can learn over thousands (or even millions of cubes if available) to overcome poor representation of a small normative database.

One of the problems in diagnosing glaucoma is that there is no single test with a high sensitivity and high specificity to confirm the diagnosis; thus, a referable definition may be more appropriate, particularly for reducing false positives. Machine learning techniques are used to automatically recognize complex patterns in a given dataset (unsupervised learning), or creating a classifier predicting group membership of new cases (supervised learning), where a group label, such as a disease, is available for each case.[12] To ensure good performance of the machine learning techniques in a given dataset, all possible sources of bias should be removed or minimized. Miguel Caixinha and Sandrina Nunes introduced conventional machine learning (CML) techniques and reviewed applications of CML for diagnosis and monitoring of multimodal ocular disease.[12] According to their study,[12] by assessing patients based on the summation of all the major risk factors, and based on multimodal investigations, predictive models will be beneficial in diagnosis and treatment. In a study by Bowd et al.,[13] combining OCT and visual field measurements using machine learning classifiers resulted in a trend of increased area under the receiver operating characteristic (AUROC) curves for discriminating between healthy and glaucomatous eyes, compared to using each measurement technique alone. Therefore, a more accurate ground truth diagnosis of glaucoma using multimodal longitudinal data, including fundus photograph, visual field, including intraocular pressure (IOP), and treatment data may help reduce human diagnosis label variation from just interpreting the processed OCT report summary.

Recent research has looked at various machine learning approaches as applied to images, grouped into two categories: classical machine learning and

deep learning methods. Classical machine learning typically involved extracting handcrafted features from segmented OCT volumes and relying on established classifiers, such as support vector machines and random forests for final arbitration.[14] Deep learning methods, such as convolutional neural networks (CNNs), directly operate on OCT volumes without any human-designed disease markers and can be considered feature-agnostic.[14,15]

Most of the recent studies[15–17] have looked into optic nerve head OCT scans to detect glaucoma in well curated datasets, which exclude the challenging cases, such as glaucoma suspects, cases with high refractive error, or low signal strength. In this work, we include raw OCT macula scans from real-world scenarios for CNN training, and attempt to differentiate between high-risk cases, which require referral and evaluation by a glaucoma specialist and low-risk cases (split across referable vs. nonreferable disease) that can be observed without frequent testing.

Methods

This study adhered to the tenets of the Declaration of Helsinki, and the protocols were approved by the institutional review board of Stanford School of Medicine (USA), Chinese University of Hong Kong (CUHK) (Hong Kong), and Narayana Nethralaya Foundation, Bangalore (India). Informed consent was waived based on the study's retrospective design, anonymized dataset of OCT images, minimal risk, and confidentiality protections.

Training and Primary Validation Set

The 3D OCT cube (volume) of macula images (Cirrus HD-OCT; Carl Zeiss Meditec, Dublin, CA, USA) of 1957 eyes evaluated at the Byers Eye Institute, Stanford School of Medicine, from March 2010 to December 2017 were exported for the study. Scanning with the Cirrus (Carl Zeiss Meditec) OCT was performed using the 512 × 128 scan pattern (Macular Cube protocol) where a 6 × 6 mm area on the retina was scanned with 128 horizontal lines, each consisting of 512 A-scans per line.

Next, the dataset was cleaned to remove any cases of nonglaucomatous optic nerve head pathologies, such as nonglaucomatous optic neuropathy/optic nerve head hypoplasia and optic nerve pit, and other retinal pathologies, such as retinal detachment, age-related macular degeneration, myopic macular degeneration, macular hole, diabetic retinopathy, and arterial and venous obstruction. A total of 749 eyes were excluded during screening due to the presence of these associated pathologies based on chart review.

The number of eyes after exclusion was 1208. Following this, 93 eyes were excluded due to quality assessment and 20 eyes were excluded after arbitration as described below. Therefore, in total, 1095 eyes of 586 patients were obtained for training and primary validation from Stanford. Additionally, we reviewed the clinical history and testing of each scan to categorize those who are true glaucoma, preperimetric glaucoma, high-risk suspects, low-risk suspects, or normal, as per criteria on Table 1. To become more clinically relevant, we combined referables as definite true glaucoma cases, preperimetric glaucoma cases plus high-risk glaucoma suspects together, and nonreferables as true normal and low-risk suspects.

Included subjects performed visual field (VF) by static, automated, white-on-white threshold perimetry using the Humphrey Field Analyzer III (Carl Zeiss Meditec).

The inclusion criteria were (1) age equal to or older than 18 years old; (2) reliable VF tests (acceptable results defined below); and (3) availability of SD-OCT macula scans (acceptable results defined below).

A reliable VF report is defined as (a) fixation losses <33%; (b) false positive rate <25%; (c) false negative rate <25%; and (d) no appearance of lid or lens rim artifacts, and no appearance of cloverleaf patterns.

SD-OCT scans with signal strength (SS) <3, or any artifact within a 14.13mm^2 elliptical annulus area centered on the fovea, were excluded from the study. Artifacts included blink, motion, registration, and mirroring. SS of three was chosen because the qualitative maps were used and not the quantitative cutoff values.

Cases were labeled according to the criteria mentioned in Table 1 by a glaucoma fellowship trained ophthalmologist with >2 years' experience (S.S.M.) based on fundus image, VF, OCT RNFL and GCIPL parameters, and IOP lowering treatment (based on chart review). In cases where labeling needed arbitration, a senior glaucoma specialist with >10 years' experience (R.T.C.) reviewed the cases and his diagnoses were considered final. Twenty conflicting cases out of 36 were eliminated based on insufficient data on chart review. To compute intergrader agreement for diagnosis, a third glaucoma fellowship-trained specialist (D.C.) adjudicated the labeling of randomly selected 50 high- and low-risk cases. Following this, Cohen's k value was calculated. Intergrader agreement calculations resulted in a Light's k (arithmetic mean of Cohen's k) of 0.415 considered to represent moderate agreement.[18]

Table 1. Criteria for Classification of Input Data

Labels	Criteria	No. of Eyes	No. of Patients	Classification
True glaucoma	• Clinical glaucomatous disc changes (as per ISGEO classification[34]) *AND* • 2 repeatable VF defects as per Anderson's criteria[35] (reliably measured data were used (i.e., with a fixation loss <20%, false-positive errors <15%, and false-negative errors <33%) OR total cupping of the optic nerve and unable to perform VF evaluation AND • On treatment for glaucoma or has undergone surgery/SLT-ALT *AND* • OCT glaucomatous defects on deviation maps + not all green on OCT RNFL and/or OCT GCIPL maps)	514	287	**REFERABLE**
Preperimetric glaucoma	• Clinical glaucomatous disc changes (as per ISGEO definition)[34] *AND* • No VF defects *AND* • On treatment for glaucoma or has undergone surgery/SLT-ALT AND • OCT glaucomatous defects on deviation maps+ (not all green OCT RNFL and/or OCT GCIPL)	41	26	**REFERABLE**
High-risk suspects	• Disc changes suspicious for glaucoma (ISGEO disc definition[34] for suspect) *AND* • Without any VF defects OR VF defect not fulfilling Anderson's criteria[35] AND • on prophylactic therapy or follow up advise less than 1 year AND • Not all green OCT RNFL and/or OCT GCIPL	112	68	**REFERABLE**
Low-risk suspects	• Disc changes suspicious for glaucoma (as per ISGEO[34] classification for suspects) *AND* • No VF defects *AND* • Not on treatment /advised no review or review after 1 year or more *AND* • Not all green OCT RNFL and or OCT GCIPL	108	66	**NONREFERABLE**
True normal	• No VF defects *AND* • No disc changes for glaucoma (few cases have high cup disc ratio >0.6 but no other glaucomatous disc changes) *AND* • No treatment/no review *AND* • All green OCT RNFL and OCT GCIPL	320	183	**NONREFERABLE**

A patient can have eyes in two different categories. International Society of Geographical and Epidemiological Ophthalmology (ISGEO); Selective Laser Trabeculoplasty (SLT); Argon Laser Trabeculoplasty (ALT).

The final training and primary validation dataset consisted of 2805 scans from 586 patients and were placed into one of two categories:

- **REFERABLE**: requiring referral for evaluation by a glaucoma specialist and including:
  - True glaucoma (TG)
  - Preperimetric glaucoma (PPG)
  - High-risk suspects
- **NONREFERABLE**: not requiring referral for further evaluation by a glaucoma specialist and including:
  - True normals (TNs)
  - Low-risk suspects

In total, there were 667 eyes of 381 patients labeled as REFERABLE cases comprised of 514 eyes from 287 patients with a diagnosis of glaucoma (TG), 41 eyes from 26 patients with a diagnosis of PPG, 112 eyes from 68 patients with a diagnosis of being high-risk glaucoma suspects, and a total of 428 eyes of 249 patients labeled as NONREFERABLE cases comprised of 320 eyes from 183 definitive normal patients (TNs), and 108 eyes from 66 low-risk glaucoma suspects were included from Byers Eye Institute at Stanford. In the case of multiple visits of a given patient, we used all of the visits for training (1710 additional scans in total), but only tested on that patient's first visit. When performing the training, however, as we will discuss below, a single patient never had images in both the training and testing sets for a given run. The reason for testing only on a patient's first visit is to focus our attention on catching referable glaucoma as early as possible, where it is the most clinically relevant.

External Validation/Test Sets

Once internal validation was achieved, two datasets were used for external validation. Cirrus SD-OCT macular cubes of 505 eyes of 264 patients from CUHK (Hong Kong) and 336 eyes of 199 patients from Nayayana Nethralaya (India) were used for external validation. All the volumes contained in the external datasets were also obtained from the Cirrus SD-OCT machine (Carl Zeiss Meditec) according to the 512 × 128 Macula cube scanning protocol.

The first dataset was composed of OCT 3D cube images of macula from CUHK and the second dataset was composed of OCT 3D cube images of macula from Narayana Nethralaya Foundation, India.

Regarding the glaucoma definitions, for the external validation dataset from Hong Kong, two glaucoma specialists worked separately to label all the eyes into "True Glaucoma" or "True Normal" combined with VF tests. In this dataset, true glaucoma was defined as those cases with RNFL defects on thickness or deviation maps that correlated in position with the VF defects. Most of the images were labeled as "True Glaucoma" or "True Normal" when the two graders arrived at the same categorization separately. The few cases with disagreement were reviewed by a senior glaucoma specialist to make the final decision. This dataset consisted of 305 "True Glaucoma" cases and 200 "True Normal" cases.

For the external evaluation set from India, an experienced glaucoma specialist labeled the cases into "True Glaucoma" and "True Normal." The definitions used to label cases in this dataset were similar to those used to label cases in the Stanford (US) dataset (Table 1). This dataset consisted of 163 "True Glaucoma" cases and 173 "True Normal" cases. Further comparisons of these datasets are available in Table 2. All external datasets were screened for exclusion of other optic nerve head, macula, or retinal pathologies.

Development of Deep Learning System

For the CNN, we explored a number of architectures before settling on gNet3D (Fig. 1). We tried both deeper, more complex networks as well as shallower, simpler ones and eventually chose the latter. The gNet3D consists of just three convolutional layers and two fully connected layers with dropout regularization.[19] We used the newly introduced AdaBound optimizer for the optimization.[20] AdaBound features the ability to get the results of stochastic gradient descent while converging at the speed of Adam.[21] Perhaps most importantly, however, is our preprocessing step of homogenizing the data. Considering only the raw OCT cubes by themselves ignores a great deal of spatial context, such as the location, orientation, and scale of the retina. To account for this spatial information, we first homogenized the data by extending the technique we presented in Russakoff et al.[22] to 3D using automated layer segmentation software (Fig. 2; Orion, Voxeleron). The homogenization also allows for the analysis of the textures of the images at the same scale across the entire dataset. Recent work has focused attention on the inherent bias of CNNs[23] to texture over other types of features. The homogenization allows for the CNN to focus its learning on a smaller domain of textures, which has the effect of improving the generalizability of the learned results. The gNet3D was then trained on the data above to return the probability of "REFERABLE" vs. "NONREFERABLE" that may be converted to a binary classification. We evaluated its performance using five-fold cross validation with no patient's scans split across the training, validation, or testing sets.

We compared our results to those from Maetschke et al.[15] by implementing their reported deep learning framework and running it over our data in the same way, namely using five-fold cross-validation. We review this work in more detail in the discussion below. The five-fold cross-validation is performed by randomly partitioning the data into five equal subsets, or folds. One fold is used as a test set, whereas the other four folds are used to train a model. We created each partition taking care that no one patient's data ever appeared in both the training and the testing sets. By performing this procedure on each of the five folds in

Table 2. Demographic Data

	Referable (US) Training, Testing, and Primary Validation Dataset	NonReferable (US) Training, Testing, and Primary Validation Dataset	Referable (Hong Kong)	Non Referable (Hong Kong)	Referable (India)	NonReferable (India)
Age (in years)	68.25 (±15.5)	67.59 (±15.40)	65.9 (±9.30)	65.27 (±11.30)	63.84 (±11.72)	54.763 (± 14.95)
***P* value**	*Reference**	*P* = 0.5098* *Reference***	*P* = 0.0147*	*P* = 0.0603**	*P* < 0.001*	*P* < 0.001*
Asian	43.5%	48.60%	100%	100%	100%	100%
***P* value**	*Reference**	*P = 0.1138* *Reference***	*P < 0.001**	*P < 0.001***	*P < 0.001**	*P < 0.001***
White	38.6%	33.40%	0	0	0	0
***P* value**	*Reference**	*P = 0.0956* *Reference***	*P < 0.001**	*P < 0.001***	*P < 0.001**	*P < 0.001***
African American	5.04%	6.40%	0	0	0	0
***P* value**	*Reference**	*P = 0.3592* *Reference***	*P < 0.001**	*P < 0.001***	*P < 0.001**	*P < 0.001***
Hispanic	9.14%	7.20% *Reference***	0	0	0	0
***P* value**	*Reference**	*P = 0.2804** *Reference***	*P < 0.001**	*P < 0.001***	*P < 0.001**	*P < 0.001***
Data of ethnicity unavailable	3.68%	4.4%	0	0	0	0
***P* value**	*Reference**	*P = 0.5683* *Reference***	*P < 0.001**	*P < 0.001***	*P < 0.001**	*P < 0.001***
Average MD	−7.175 (±7.451)	−1.24 (±2.09)	−8.005 (±6.81)	−0.900 (±1.30)	−12.74 (±9.22)	−1.20 (±1.30)
***P* value**	*Reference**	*P < 0.001 ** *Reference***	*P = 0.1209**	*P = 0.0363***	*P < 0.001**	*P = 0.8693***
Mean refractive error	−1.67 (±3.2)	−0.4713 ±2.53)	−0.85 (± 2.57)	−0.51 (±2.15)	−0.483 (± 2.25)	−0.440 (±2.19)
***P* value**	*Reference**	*P < 0.001** *Reference***	*P < 0.001**	*P = 0.8538***	*P < 0.001**	*P = 0.8882***

Demographic data such as age, ethnicity distribution (in %), mean values with SDs for VF parameter in terms of mean field defects and refractive error in terms of spherical equivalent in referable and nonreferable groups from the United States, Hong Kong, and India datasets. The statistical analysis was performed with the Statistical Package for Social Sciences (SPSS) 10.1 (SPSS Inc., Chicago, IL, USA). Results are expressed as mean (±SD) and paired Student's *t*-test was used to evaluate the level of significance. A *P* value of 0.001 or less was considered significant. Chi-square test was used for comparisons of categorical demographic data for proportions. US, United States.

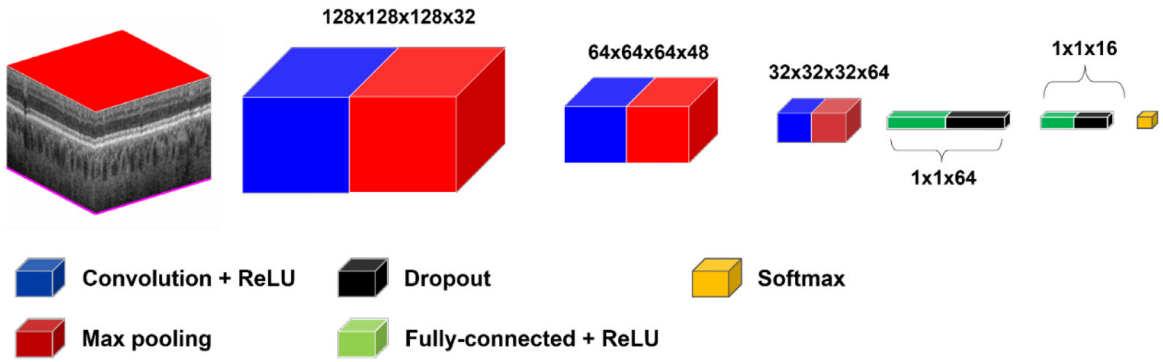

Figure 1. Schematic of gNet3D. The model consists of three 3D convolutional layers together with two fully connected layers. Rectified Linear Unit (ReLU).

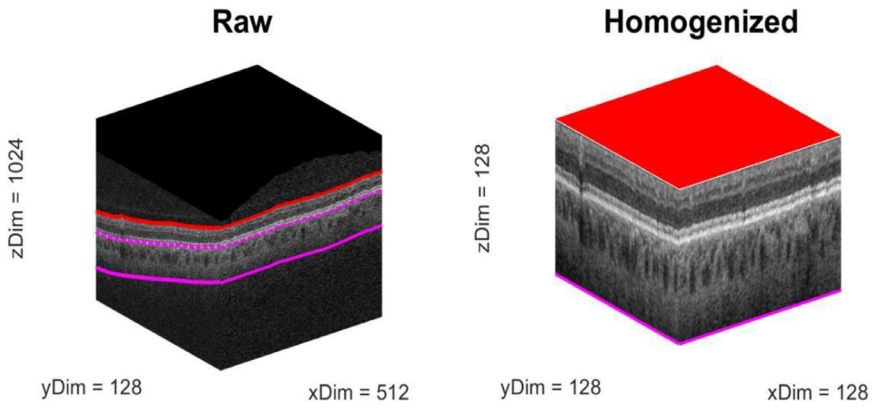

Figure 2. Example of the preprocessing used to homogenize the images (all axes are in pixels). The segmentation allows for a simple normalization to a standardized size: Note that the lower limit is a fixed offset (390 µm) from Bruch's membrane, which is itself estimated as a baseline to the retinal pigment epithelium. This homogenization step helps add spatial context to the classifier by factoring out position and scale variations in the images.

turn, we can generate a prediction for each data point. In addition, for the entire set of predictions, we generate a receiver operating characteristic (ROC) curve as well as the corresponding area under that curve (AUC). We also generated precision-recall curves for this data to complement the AUC ROC results. This analysis was performed five times with five different cross-validation partitions.

We also applied the gNet3D trained on the Stanford data to two independent test sets from outside institutions. In particular, we used each of the five models generated during the initial cross-validation and applied them as a classifier ensemble, taking the median value as the final output. We also evaluated the performance of the model across myopia severity in the dataset from the United States (Table 3). Finally, we performed an occlusion sensitivity analysis to investigate which areas of the volumes were most useful in discriminating the two classes. We ran this analysis on a random subset of 40% of the original data using the models trained from the first cross-validation split.

Results

Demographic data, such as age, ethnicity distribution, mean values with SDs for VF parameter in terms of mean field defects (MDs) and refractive error in terms of spherical equivalent in the referable and nonreferable groups from the United States, Hong Kong, and India datasets are given in Table 2. In the training, testing, primary validation referable dataset, and referable dataset from India, there were significant differences in age, mean deviation, and mean refractive error in terms of spherical equivalent, whereas there was only significant difference in mean refractive error in terms of spherical equivalent in the referable dataset from Hong Kong.

In all cases, we measured the classifier performance for referable versus nonreferable glaucoma using the AUROC curve. The results are summarized in Figures 3 and 4. The AUC for classification of referable glaucoma using gNet3D is 0.88 versus 0.81 for the

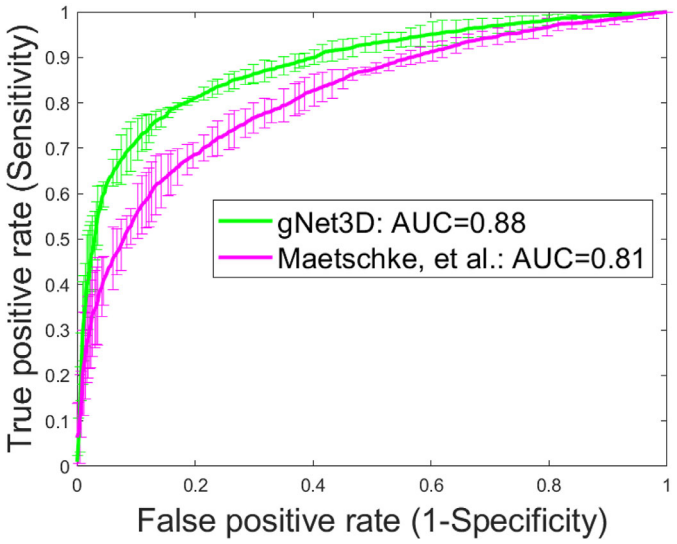

Figure 3. Illustration of the improved performance of gNet3D over the framework of Maetschke et al.[15] on this dataset.

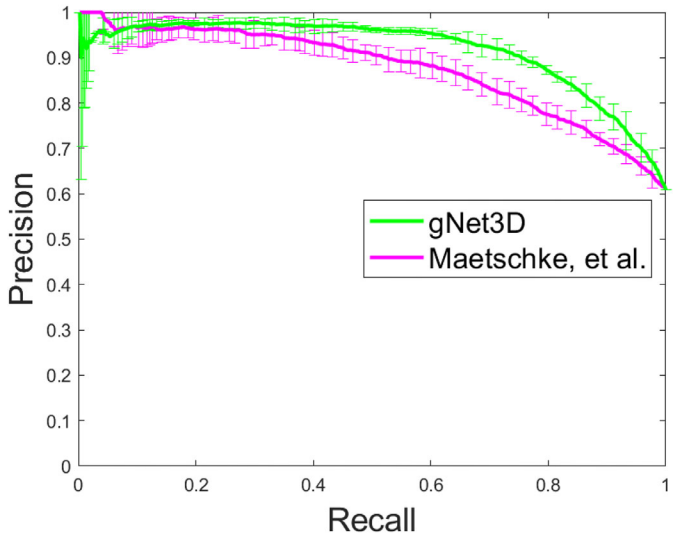

Figure 4. Precision-Recall curve illustrating the same improvement in Figure 3.

framework reported by Maetschke et al.[15] We report this primarily as a way of demonstrating how much more challenging our current dataset is with suspects included. The other main difference between our data and theirs is that they use the optical nerve head (ONH) cube scans versus our macular volumes. There is nothing, however, in their framework to suggest its performance is tuned more to ONH versus macular anatomy, so we would expect a similar performance improvement on ONH cubes. Additionally, to further investigate the importance of the preprocessing, we trained our gNet3D framework on the raw macular cubes. The AUC for that experiment was 0.82, comparable to Maetschke et al.[15] and suggests that the architecture is likely less important than the preprocessing.

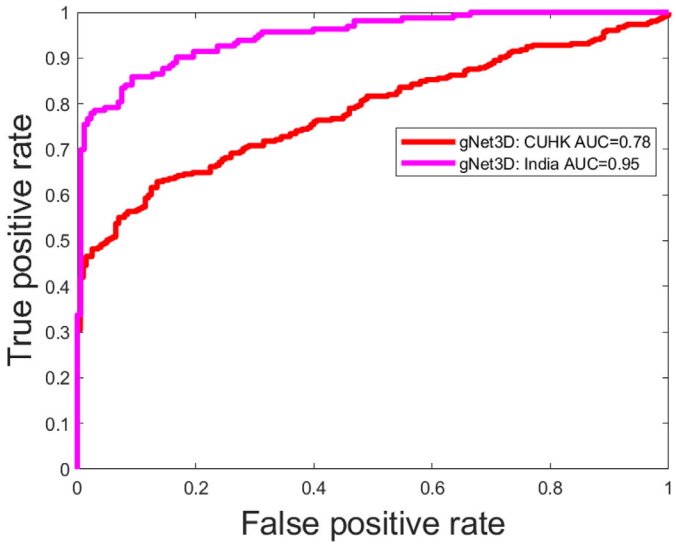

Figure 5. Results from application of the gNet3D trained on United States (Stanford) data applied to two outside institutions. The discrepancy here is likely due to the differing characteristics of the two datasets. For example, the India data had referral cases with significantly lower mean deviation and the Hong Kong (CUHK) data consists exclusively of Chinese Asian eyes (Table 2).

We have also used the models trained on the Stanford data and attempted to validate them on external data from CUHK and Narayana Nethralaya Eye Hospital, India. These results are shown in Figure 5.

Finally, the occlusion sensitivity analysis results are shown in Figure 6. In brief, it demonstrates that, as expected, the inferior regions have the highest response in the classifier in the "Referable" cases.

Further, the performance of the model across myopia severity distribution was assessed in the dataset from the United States and was found to have an AUC of 0.85, 0.92, and 0.95 in severe, moderate, and mild myopia, respectively (Table 4).

Discussion

OCT is now one of the most common imaging procedures with 5.35 million OCT scans performed in the US Medicare population in 2014 alone.[24] Most of the recent studies[15–17] have been on utilization of deep learning in detecting glaucoma, excluding the cases that require identification beyond a nonspecialist-level interpretation and excluding real-world scenarios, which include high myopes who are glaucoma suspects. This, despite several large population-based studies using different definitions of high myopia, including Los Angeles Latino Eye Study and the Blue Mountain

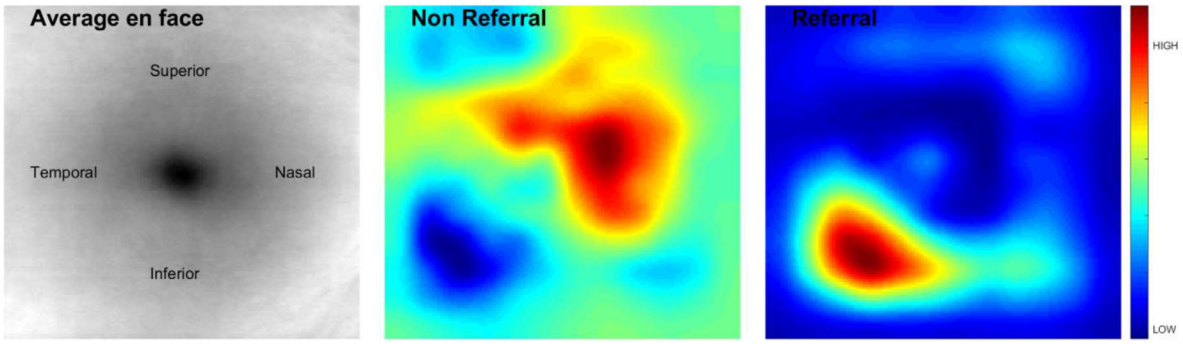

Figure 6. Illustration of the occlusion sensitivity analysis.

Eye Study (BMES), which have found that myopes are more likely to have glaucoma.[25,26] So far, ours is the first study that has included glaucoma suspects who are the most enigmatic subset of patients presenting in any ophthalmology clinic. Our training dataset included SD-OCT scans of all degrees of myopia (Table 3), glaucoma suspects, and scans with low SS (because we included >3) and, hence, arguably represents the challenge, as it exists in the clinics today. Although strict exclusion criteria, such as VF defect thresholds, low corrected visual acuity are common, our cohort did not exclude such patients and, hence, was more challenging.

Even though recent studies have reported similar diagnostic capability of macula parameter based on GCIPL and ONH parameter based on RNFL,[27] as of now, ours is the first study utilizing unsegmented OCT macula cube volumes in glaucoma and suspect detection using deep learning. A recent study using deep learning on macula scans using GCIPL thickness maps to detect glaucoma had an AUC of 0.9307.[28] The differences from our study were that they defined glaucoma based on the presence of glaucomatous disc changes with RNFL defects on clinical examination along with corresponding VF defects and did not use multimodal longitudinal imaging for ground truth definitions. They used segmented data to estimate GCIPL thickness, did not use raw 3D cube scans, and did not include glaucoma suspects.

It is known that high myopia can be associated with elongation of the eye and resultant stretching effect, which can lead to decreasing GCIPL thickness identified as defects on OCT GCIPL deviation and thickness maps that may resemble glaucomatous damage. We included all ranges of myopia in both our referable and nonreferable datasets to account for these variations.

We defined severity of myopia by slightly modifying the BMES.[29] We modified the BMES category of moderate to severe myopia (spherical equivalent [SE] >-3D) by further subdividing it into mild myopia (SE up to -3D), moderate myopia (SE -3 up to -6D), and severe myopia (SE lesser than -6D), using cutoffs established in the Beijing Eye Study.[30] The myopia severity distribution has been used to analyze the distribution of cases across the dataset and to show that severe myopia cases have been included in the training, primary validation, and test datasets (Table 3).

It is known that diagnosing glaucoma and high-risk glaucoma suspects in the setting of myopia is a common challenge due to alteration of the appearance of the optic nerve and macula. Myopic refractive error impacts RNFL and macular thickness measurements due to stretching and thinning of these layers due to an increased axial length and optical projection artifact of the scanning area.[31] This often results in many false-positive diagnoses, also known as "Red Disease." The performance of the model across myopia severity distribution has been assessed in the dataset from the United States and was found to have an AUC of 0.85, 0.92, and 0.95 in the severe, moderate, and mild myopia, respectively (Table 4). The performance of our model in detecting referable cases in severe myopia with an AUC of 0.85 is a promising tool to show the normative database limitations can be overcome with ever increasing datasets. Our performance with external test sets demonstrated good results across geographical and ethnicity distribution. Results from application of the gNet3D trained on US data applied to two outside institutions was, interestingly, better in the dataset from India. A possible explanation for this could be significantly lower mean deviation in this dataset compared to that from the United States and Hong Kong (Table 2). Another possible reason for better performance in the India dataset could be because the referable and nonreferable datasets from India included only true glaucoma and true normal, respectively, and did not include suspects, and these cases would likely be easier to differentiate.

Table 3. Myopia Severity Distribution

	US (Referable) $N =$ 667 Eyes Training, Testing, and Primary Validation Dataset)	US (Non Referable) $N = 428$ Eyes Training, Testing, and Primary Validation Dataset)	Hong Kong (Referable) $N = 305$ Eyes	Hong Kong (Non Referable) $N = 200$ Eyes	India (Referable) $N = 163$ Eyes	India (NonReferable) $N = 173$ Eyes
Severe myopia	6.45% ($N = 43$)	1.87% ($N = 8$)	4.7%	0%	4.59%	4.58%
P value	*Reference**	*$P < 0.001$* Reference***	*$P = 0.283$**	*$P < 0.001$***	*$P = 0.371$**	*$P = 0.652$***
Moderate myopia	8.55% ($N = 57$)	5.14% ($N = 22$)	12.50%	21.01%	6.32%	9%
Mild myopia	20.84% ($N = 139$)	19.86% ($N = 85$)	37.50%	15.70%	35.63%	29%
Emmetropia	3.45% ($N = 23$)	4.91% ($N = 21$)	5.90%	10.50%	10.9%	17.74%
Hypermetropia	19.04% ($N = 127$)	32.24% ($N = 138$)	39.20%	47.3%	42.5%	39.69%
No data	41.68%	35.98%	N/A	4.59%	N/A	N/A

χ^2 test was used for severe myopia distribution analysis (myopia severity distribution: severe: D ≤ -6; moderate: $-6 <$ D ≤ -3, mild: D > -3, where D is diopter).

Table 4. Results of the Proposed Model on the Dataset from the United States for Each Myopia Severity Level

Myopia Severity	Number of Cases	AUC
Severe	51	0.85
Moderate	79	0.95
Mild	224	0.92

AUC, area under the curve.

The external validation performance was lower in the dataset from Hong Kong. The distribution of ethnicity in this dataset, which consisted exclusively of Chinese Asian eyes, could be a reason for the disparity in performance. Our training, primary validation, and primary test sets included subjects of white, Asian (which included Chinese Asians, Non-Chinese Asians, and Indians), African American, and Hispanic origin (Table 2). Another reason for the difference in performance on the test set from Hong Kong could be the inclusion of only gradable images with SS $\geq$5. The data from the United States included cases with SS $\geq$3 and excluded images with artifacts, which obscured imaging of measuring GCIPL within a 14.13 mm^2 elliptical annulus area centered on the fovea.

The variation in performance between the datasets from India and Hong Kong versus that from the United States could also be attributed to the significant difference in the mean refractive error between the referable group from the United States and that from India and Hong Kong ($P < 0.001$). Data from the United States had significantly higher mean myopic refractive error compared to India and Hong Kong in the referable group (Table 2). Another reason for the difference in the performance in the Hong Kong dataset could be due to the fact that structural defects in true glaucoma cases in this dataset were based only on OCT RNFL deviation and thickness maps and did not include GCIPL maps, which might have possibly categorized low-risk cases with RNFL defects as true glaucoma. This is an example where generalization with external dataset is influenced by referable definitions, which can be difficult to reach a consensus standard.

In the recent study by An Ran et al.,[16] the 3D deep learning system for optic nerve cubes had an AUC of 0.969, sensitivity of 89%, specificity of 96% (92–99), and accuracy of 91% (89–93) in the primary validation set for detection of glaucomatous optic neuropathy. This study had good performance across an external dataset from the United States with an AUC of 0.893. The major difference in their study from ours was that although their study used ONH cube scans, we used macula SD OCT scans. In their study, only gradable images with SS $\geq$5 were included for training and validation, but in our study we used lower intensity scores (SS $\geq$3) and excluded images with artifacts that obscured imaging of measuring GCIPL within a 14.13 mm^2 elliptical annulus area centered on the fovea for training and validation, hence representing real-world data. Unlike in our study, this study did not include eyes with suspected glaucoma and those with preperimetric glaucoma in the training dataset.

Maetschke et al.'s[15] work used 3D CNN to classify eyes as healthy or glaucomatous directly from raw, unsegmented OCT volumes of the ONH and achieved a high AUC of 0.94, and is the most similar to ours. The cases included in training and validation sets in their study were ONH scans of well-defined cases of glaucoma versus normals and did not include high- or low-risk suspects. As an attempt at comparison, we applied their published deep learning framework to our dataset of macular cubes and saw, as can be expected from a more challenging dataset, a large drop in performance. We matched this drop in performance by applying our own deep learning architecture on the raw cubes. This result underscores the importance of adding spatial context to the CNNs via our data homogenization process. Their feature agnostic framework has merit but, to work on the domain of all possible scans in all positions and orientations would likely require orders of magnitude more data, more akin to what is seen for ImageNet[32] where there is little attempt to restrict the domain. Our homogenization effectively restricts the domain of input images allowing the network to focus its discriminative power on a smaller subset of image characteristics. The homogenization also allowed us a way to easily compare scans in a standardized reference frame for the occlusion sensitivity analysis. This analysis shows the regions of the image that are most discriminative with respect to the input classes. The results in Figure 6 suggest that areas of interest in determining referables from nonreferables are inferior and temporal. This concurs with what we know about glaucoma in the macula currently, where the inferior region is commonly affected (first) with inferior temporal bundle defects. Prior publications have found that glaucoma most often affects the inferior temporal followed by superior temporal fundus regions, followed by the temporal horizontal sector.[33] On histopathology, this finding has been attributed to the thicker nerve fiber layer in the inferior and superior peripapillary areas compared to the temporal and nasal peripapillary regions; a wider neuro retinal rim in the inferior and superior disc regions compared with the nasal and temporal disc sectors; and the morphology of the lamina cribrosa with the largest pores in the inferior and superior disc regions and the smallest pores in the temporal and nasal areas.[33] Our occlusion

sensitivity analysis highlights infero-temporal regions as an area of interest in identifying referables, which includes high-risk suspects along with preperimetric and perimetric glaucoma cases. This analysis shows the regions of the image that are most discriminative with respect to the input classes.

Glaucoma suspects are controversial clinical dilemmas. One of the most difficult challenges faced by clinicians in an ophthalmology clinic is in identifying high- and low-risk glaucoma suspects and thereby deciding whether to treat prophylactically or not treat a suspect. This is mainly because there is no consensus for identifying risk among glaucoma suspect cases even among experts. Ours is one of the first studies to use machine learning in risk stratification. A major highlight of our study was inclusion of these cases based on broad criteria, which are routinely used by glaucoma specialists in risk stratification based on longitudinal chart review. We also have included preperimetric glaucoma cases, which are often excluded from recent studies while training algorithms.[13,14] Another strength of our study was multinational external dataset validation across geographical and ethnicity distribution from Hong Kong and India.

Our study had a few limitations. We have included SD-OCT scans with low SS. This is because many of times, clinicians are deprived of high-quality OCT images for diagnosis and evaluation of glaucoma, as many of the elderly glaucoma patients have associated age-related cataract, corneal decompensations, or vitreous degenerations. Our aim was to train the algorithm to be able to identify available feature agnostic representations to detect glaucoma even on low-quality images, hence replicate real-world presentations. Despite the generalizability of the performance of our algorithm across the datasets, one major limitation of our study was the lack of inclusion of suspects in our external datasets and we are in the process of acquiring images of suspects from these centers, which would be included in our future databases. Our training and internal validation sets had consistent definitions and the differences in performance in the external sets can be attributed due to this difference in the data.

Another limitation of our study was that we have not included risk factors like pseudo-exfoliation, pigment dispersion, or any secondary mechanisms during risk categorization. Cases were labeled only depending on structural and functional defects and requirement for therapy or frequency of advised follow-up based on longitudinal chart review as referable versus nonreferable. Another drawback of our study was that we have not yet inspected cases with false predictions and have not correlated it with myopia severity, axial length, disc size, or cup-to-disc ratio. Even though we have not excluded any cases based on axial length or disc sizes and have included extremes of these characters in our datasets, we have not looked into the performance of our model on different subsets of disc sizes or axial length. One other major limitation is that AUC, which has been used to analyze the performance of the model, does not demonstrate what a normal clinician's performance would have been, given the same data. This study is being carried out separately.

Going forward, we plan to include raw OCT macula scans along with ONH scans for better algorithm development (ensemble techniques to improve performance). Lastly, we look forward to larger training cohorts of high- and low-risk suspects with better consensus among more glaucoma specialists for the CNN to achieve optimal performance on these challenging cases.

Conclusion

A deep CNN accounting for spatial context is capable of accurately referring patients for glaucoma in a dataset representative of a real-world clinical setting. This work demonstrates that a referable versus nonreferable definition can still be applied across different datasets with reasonable performance.

Acknowledgments

Supported by Santen grant, Stanford Center for Innovation in Global Health (CIGH) Seed Grant, Research to Prevent Blindness-National Eye Institute P30-EY026877

Disclosure: **D.B. Russakoff,** None; **S.S. Mannil,** None; **J.D. Oakley**, None; **A.R. Ran,** None; **C.Y. Cheung,** None; **S. Dasari,** None; **M. Riyazzuddin**, None; **S. Nagaraj,** None; **H.L. Rao,** None; **D. Chang,** None; **R.T. Chang,** None

References

1. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet*. 2004;363:1711–1720.
2. Coleman AL, Brigatti L. The glaucomas. *Minerva Med*. 2001;92:365–379.
3. Medeiros FA, Zangwill LM, Alencar LM, et al. Detection of glaucoma progression with stratus

OCT retinal nerve fiber layer, optic nerve head, and macular thickness measurements. *Invest Ophthalmol Vis Sci*. 2009;50:5741–5748.
4. Bussel II, Wollstein G, Schuman JS. OCT for glaucoma diagnosis, screening and detection of glaucoma progression. *Br J Ophthalmol*. 2014;98(suppl. 2):ii15–ii19.
5. Chauhan BC, Vianna JR, Differential effects of aging in the macular retinal layers, neuroretinal rim, and peripapillary retinal nerve fiber layer ophthalmology. 2019. pii: S0161-6420(19)32077-9. Available at: https://doi.org/10.1016/j.ophtha.2019.09.013.
6. Kim CY, Jung JW, Lee SY, Kim NR. Agreement of retinal nerve fiber layer color codes between Stratus and Cirrus OCT according to glaucoma severity. *Invest Ophthalmol Vis Sci*. 2012;53:3193–3200.
7. 510(k) Premarket Notification - Cirrus HD-OCT with RNIFL, Macular, Optic Nerve Head and Ganglion Cell Normative Databases Cirrus 6.0 Software.
8. Kim KE, Park KH, Yoo BW, et al. Topographic localization of macular retinal ganglion cell loss associated with localized peripapillary retinal nerve fiber layer defect. *Invest Ophthalmol Vis Sci*. 2014;55:3501–3508.
9. Xu J, Ishikawa H, Wollstein G, Schuman JS. 3D optical coherence tomography super pixel with machine classifier analysis for glaucoma detection. *Conf Proc IEEE Eng Med Biol Soc*. 2011;2011:3395–3398. doi:10.1109/IEMBS.2011.6090919.
10. Chong GT, Lee RK. Glaucoma versus red disease: imaging and glaucoma diagnosis. *Curr Opin Ophthalmol*. 2012;23:79–88.
11. Miri Mohammad Saleh. "A multimodal machine-learning graph-based approach for segmenting glaucomatous optic nerve head structures from SD-OCT volumes and fundus photographs." PhD (Doctor of Philosophy) thesis, University of Iowa, 2016. Available at: https://doi.org/10.17077/etd.23kdq0ph.
12. Caixinha M, Nunes S. Machine learning techniques in clinical vision sciences. *Curr Eye Res*. 2017;42:1–15.
13. Bowd C, Hao J, Tavares IM, et al. Bayesian machine learning classifiers for combining structural and functional measurements to classify healthy and glaucomatous eyes. *Investig Ophthalmol Vis Sci*. 2008;49:945–953.
14. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall Press; 2009:1152.
15. 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.
16. Ran AR, Cheung CY, Wang X, et al. Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep- learning analysis. *The Lancet Digital Health*. 2019;1:e172–e182.
17. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
20. Luo L, Xiong Y, Liu Y, Sun X. Adaptive gradient methods with dynamic bound of learning rate. *arXiv e-prints*; 2019.
21. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv e-prints*; 2014.
22. Russakoff DB, Lamin A, Oakley JD, Dubis AM, Sivaprasad S. Deep learning for prediction of AMD progression: A pilot study. *Invest Ophthalmol Vis Sci*. 2019;60:712–722.
23. Geirhos R, Rubisch P, Michaelis C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. International Conference on Learning Representations (ICLR), May 2019. Available at: https://openreview.net/forum?id=Bygh9j09KX.
24. Centers for Medicare & Medicaid Services. Medicare Provider Utilization and Payment Data: Physician and Other Supplier. Available at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.
25. Varma R, Paz SH, Azen SP, et al. The Los Angeles Latino Eye Study: Design, methods, and baseline data. *Ophthalmology*. 2004;111:1121–1131.
26. Mitchell P, Smith W, Attebo K, Healey PR. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. *Ophthalmology*. 1996;103:1661–1669.
27. Mwanza JC, Durbin MK, Budenz DL, et al. Glaucoma diagnostic accuracy of ganglion cell- inner plexiform layer thickness: comparison with nerve

fiber layer and optic nerve head. *Ophthalmology*. 2012;119:1151–1158.
28. 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.
29. Mitchell P, Hourihan F, Sandbach J, Wang JJ. The relationship between glaucoma and myopia: The Blue Mountains Eye Study. *Ophthalmology*. 1999;106:2010–2015.
30. Xu L, Wang Y, Wang S, Wang Y, Jonas, JB. High myopia and glaucoma susceptibility: The Beijing Eye Study. *Ophthalmology*. 2007;114:216–220.
31. Mwanza JC, Sayyad FE, Aref AA, et al. Rates of abnormal retinal nerve fiber layer and ganglion cell layer oct scans in healthy myopic eyes: Cirrus versus rtvue. *Ophthalmic Surg Lasers Imaging Retina*. 2012;43: S67–S74.
32. Deng J,, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. CVPR09. 2009. Available at: http://www.image-net.org/papers/imagenet_cvpr09.bib.
33. Jonas JB, Schiro D. Localised wedge shaped defects of the retinal nerve fibre layer in glaucoma. *Br J Ophthalmol*. 1994;78:285–290. doi:10.1136/bjo.78.4.285.
34. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol*. 2002;86:238–242.
35. Anderson DR, Patella VM. *Automated Static Perimetry* (2nd Ed). St. Louis, MO: Mosby and Co; 1999:152–153.