AMERICAN ACADEMY OF OPHTHALMOLOGY®

Deep Learning and Glaucoma Specialists

The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs

Sonia Phene, BS,[1,] R. Carter Dunn, MS, MBA,[1,*] Naama Hammel, MD,[1,*] Yun Liu, PhD,[1] Jonathan Krause, PhD,[1] Naho Kitade, BA,[1] Mike Schaekermann, BS,[1] Rory Sayres, PhD,[1] Derek J. Wu, BS,[1] Ashish Bora, MS,[1] Christopher Semturs, MS,[1] Anita Misra, BTech,[1] Abigail E. Huang, MD,[1] Arielle Spitze, MD,[2,3] Felipe A. Medeiros, MD, PhD,[4] April Y. Maa, MD,[5,6] Monica Gandhi, MD,[7] Greg S. Corrado, PhD,[1] Lily Peng, MD, PhD,[1,‡] Dale R. Webster, PhD[1,‡]*

Purpose: To develop and validate a deep learning (DL) algorithm that predicts referable glaucomatous optic neuropathy (GON) and optic nerve head (ONH) features from color fundus images, to determine the relative importance of these features in referral decisions by glaucoma specialists (GSs) and the algorithm, and to compare the performance of the algorithm with eye care providers.

Design: Development and validation of an algorithm.

Participants: Fundus images from screening programs, studies, and a glaucoma clinic.

Methods: A DL algorithm was trained using a retrospective dataset of 86 618 images, assessed for glaucomatous ONH features and referable GON (defined as ONH appearance worrisome enough to justify referral for comprehensive examination) by 43 graders. The algorithm was validated using 3 datasets: dataset A (1205 images, 1 image/patient; 18.1% referable), images adjudicated by panels of GSs; dataset B (9642 images, 1 image/patient; 9.2% referable), images from a diabetic teleretinal screening program; and dataset C (346 images, 1 image/patient; 81.7% referable), images from a glaucoma clinic.

Main Outcome Measures: The algorithm was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity for referable GON and glaucomatous ONH features.

Results: The algorithm's AUC for referable GON was 0.945 (95% confidence interval [CI], 0.929–0.960) in dataset A, 0.855 (95% CI, 0.841–0.870) in dataset B, and 0.881 (95% CI, 0.838–0.918) in dataset C. Algorithm AUCs ranged between 0.661 and 0.973 for glaucomatous ONH features. The algorithm showed significantly higher sensitivity than 7 of 10 graders not involved in determining the reference standard, including 2 of 3 GSs, and showed higher specificity than 3 graders (including 1 GS), while remaining comparable to others. For both GSs and the algorithm, the most crucial features related to referable GON were: presence of vertical cup-to-disc ratio of 0.7 or more, neuroretinal rim notching, retinal nerve fiber layer defect, and bared circumlinear vessels.

Conclusions: A DL algorithm trained on fundus images alone can detect referable GON with higher sensitivity than and comparable specificity to eye care providers. The algorithm maintained good performance on an independent dataset with diagnoses based on a full glaucoma workup. *Ophthalmology 2019;■:1–13 © 2019 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.aaojournal.org.

Glaucoma, the so-called silent thief of sight, is the leading cause of preventable, irreversible blindness worldwide.[1,2] Glaucoma currently affects approximately 70 million individuals and is projected to affect approximately 112 million in 2040.[2] The disease can remain asymptomatic until severe, and an estimated 50% to 90% of people with glaucoma remain undiagnosed.[3–5] Thus, glaucoma screening is recommended for early detection and treatment.[6,7] However, this currently requires a clinical examination combined with quantitative functional and structural measurements, accessible only to a small percentage of the world's population. A cost-effective tool to detect glaucoma could expand screening access to a much larger patient population. However, such a tool is currently unavailable.[8]

Retinal fundus photography is a well-established diagnostic tool for eye diseases, enabling fundus images to be evaluated for the presence of retinal and optic nerve head (ONH) pathologic features.[8–10] Glaucoma, characterized by progressive degeneration of retinal ganglion cells and loss of their axons, results in

https://doi.org/10.1016/j.ophtha.2019.07.024
ISSN 0161-6420/19

characteristic structural ONH and retinal nerve fiber layer (RNFL) changes.[11] These changes can be detected in fundus images[12] and are the most important aspect of glaucoma diagnosis.[13]

Physical features that have been associated with glaucomatous optic neuropathy (GON) include neuroretinal rim thinning, notching, or both; increased or asymmetric cup-to-disc ratio (CDR); excavation of the cup; RNFL thinning; disc hemorrhages; parapapillary atrophy (PPA); laminar dots; nasalization of central ONH vessels; and baring of circumlinear vessels.[14,15] However, several of the features listed may appear in otherwise healthy optic nerves,[16,17] may result from non-glaucomatous pathology,[18–20] or are poorly defined in the literature.[21] The relative importance of these features for the diagnosis of glaucoma has not been validated, and practitioners may weigh features differently, often focusing on increased CDR, neuroretinal rim thinning and notching, disc hemorrhages, and RNFL defects.[22]

Deep learning (DL)[23] has been applied to produce highly accurate algorithms that can detect eye conditions such as diabetic retinopathy (DR) with accuracy comparable with that of human experts.[24–26] The use of this technology may aid screening efforts, and recent work demonstrates value in assisting ophthalmologists.[27] DL algorithms also have been developed for other diseases, including glaucoma.[26,28–31] However, the diagnosis of glaucoma poses several challenges that distinguish it from conditions such as DR. While DR has well-established clinical guidelines for diagnosis from fundus images, the diagnosis of glaucoma consists of the compilation of clinical examination data with functional and structural testing. There is no single agreed-upon set of guidelines for glaucoma detection from fundus images.[32] Teleretinal screening programs have worked around these limitations by focusing on subsets of the potential diagnostic features present in fundus images. However, these criteria vary between programs, and the relationship of the individual features to overall glaucoma assessment is not well characterized. This strongly limits the clinical value of these diagnoses and correspondingly, limits the value of DL algorithms developed around these heuristics.

The present study aimed to extend the usefulness of DL algorithms by explicitly bridging the gap between individual diagnostic features present in fundus imagery and referable GON assessment by glaucoma specialists. We developed an algorithm that is trained on both individual pathologic features and overall GON referability. Our study aimed to answer 3 main questions: Can we develop a DL algorithm that performs at or above expert level in referable GON detection? How do glaucoma specialists weigh the relative importance of individual features toward a referable GON decision? And do DL algorithms consider the same features with similar relative importance?

Methods

Datasets

Development Datasets. The development datasets for this study consisted of fundus images obtained from multiple sources: EyePACS,[33] Inoveon,[34] and the Age-Related Eye Disease Study[35] in the United States; 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya); and the United Kingdom Biobank.[36]

Validation Datasets. Three datasets were used for validation. Validation dataset A comprised a random subset of color fundus images from EyePACS, Inoveon, the United Kingdom Biobank, the Age-Related Eye Disease Study, and Sankara Nethralaya. Validation dataset B comprised macula-centered color fundus images from the Atlanta Veterans Affairs (VA) Eye Clinic diabetic teleretinal screening program. Validation dataset C comprised macula-centered and disc-centered color fundus images from the glaucoma clinic at Dr. Shroff's Charity Eye Hospital, New Delhi, India.

All images were de-identified according to the Health Insurance Portability and Accountability Act Safe Harbor provision before transfer to the study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review Institutional Review Board. From each source, only 1 image per patient was used in the final analysis, and each image appeared only in either the development or validation sets (see details in "Development of the Algorithm").

Development Dataset Grading

Gradability. Graders assessed the gradability of specific ONH features in every image. Gradability was measured using the degree to which each distinct feature was detected in the image and the ability of each feature to be identified (based on image quality, blurring, media opacity, or any other confounding reason). In addition, graders were asked to assess gradability for determining referable GON. If graders selected "ungradable" for a particular feature or referable GON, then no grade was collected for that aspect.

Feature-Level Grading. The American Academy of Ophthalmology's Primary Open-Angle Glaucoma Suspect Preferred Practice Patterns provides a list of features that may indicate GON: vertical elongation of the optic cup, excavation of the cup, thinning of the RNFL, notching of the neuroretinal rim, thinning of the inferior or superior neuroretinal rim or both, disc hemorrhage, PPA, nasalization of central ONH vessels, baring of circumlinear vessels, and absence of neuroretinal rim pallor. Additionally, the finding of a laminar dot sign (visible lamina cribrosa fenestrations in the cup) is viewed by some clinicians as a sign of glaucomatous damage to the ONH.[37,38] Violation of the inferior > superior > nasal > temporal (ISNT) neuroretinal rim area rule is viewed by some clinicians as a sign of glaucomatous rim loss. ISNT rule variants, focusing on the inferior, superior, and temporal quadrants, also have been validated.[39] To enable systematic training of graders, we aggregated these features (Table S1, available at www.aaojournal.org), developed grading guidelines for each, and iterated on the guidelines with a panel of 3 fellowship-trained glaucoma specialists (N.H., A.E.H., and A.S.) to increase inter-rater agreement.

Referable Glaucomatous Optic Neuropathy Grading. In addition to these features, we also developed guidelines for a 4-point GON assessment (as shown in Table S1, available at www.aaojournal.org) where the high-risk glaucoma suspect or likely glaucoma levels were considered referable, that is, the ONH appearance was worrisome enough to justify referral for comprehensive examination. Graders were asked to provide a referable GON grade after evaluating the image for each of the features on the list.

Graders. A total of 43 graders (14 fellowship-trained glaucoma specialists including N.H., A.E.H., and A.S., 26 ophthalmologists, and 3 optometrists) were trained on the grading guidelines and were required to pass a certification test before grading the development or validation A datasets.

Development of the Algorithm

Data preprocessing, algorithm design, and hyperparameters are detailed in the Supplemental Material (available at www.aaojournal.org). Using these data, a deep convolutional neural network with the Inception-v3 architecture[40] was developed and trained in TensorFlow.[41] To speed up training, the network was initialized using parameters from a network pretrained to classify objects in the ImageNet dataset.[42] This procedure is similar to that previously described by Krause et al.[25]

The input to the network is a single color fundus photograph resized to 587×587 resolution. The network was trained to output a probability for each of the possible values for each feature. For instance, for referable GON, the network outputs 4 probabilities corresponding to no-risk, low-risk, high-risk, and likely glaucoma. These probabilities sum to 1. The network is able to make multiple predictions simultaneously, such as referable GON and the presence or absence of various ONH features.

The development datasets (Table 1) consisted of 2 disjoint subsets: training and tuning. The training set of 86 618 images was used to optimize the network parameters. This dataset was enriched for referable GON images using active learning,[43] a machine learning technique, to preferentially increase the number of relevant examples. The tuning set consisted of 1508 images, each independently graded by 3 glaucoma specialists (from a group of 11 specialists, including N.H., A.E.H., and A.S.) and was used to determine thresholds for 3 operating points for the algorithm: high sensitivity, high specificity, and a balanced sensitivity and specificity point (see Supplemental Material for details). The tuning dataset and validation dataset A (described next) were chosen randomly from a pool that also was enriched, but using preliminary screening reviews by a separate panel of graders for images suspicious for a glaucomatous-appearing disc.

An ensemble of 10 networks[44] was trained on the same training set, and the outputs were averaged to yield the final prediction. Each network was built in an identical manner (see “Algorithm Design” in the Supplemental Material for details).

Clinical Validation Datasets

Three datasets were used for validation. For validation dataset A (1205 images), the reference standard was determined by a rotating panel of 3 fellowship-trained glaucoma specialists (from a cohort of 12 including N.H., A.E.H., and A.S.). Each fundus image was first graded independently by each glaucoma specialist in the panel (round 1). To resolve any disagreements between graders and thus reduce potential inadvertent grading errors and increase the quality of the reference standard, the same 3 glaucoma specialists (in a random order) reviewed the image a second time (round 2), this time with access to the annotations and comments from round 1 and previous graders in round 2. Grader identities were anonymized throughout the grading process. Each image was reviewed a maximum of 6 times (twice by each grader) or fewer if consensus was reached before the sixth review. By the completion of round 2, 49.3% (594 images) were fully adjudicated on all features and referable GON.

For images unresolved after round 2, the reference standard for gradability, each ONH feature, and referable GON were determined by the median of the 3 grades in round 2. The median corresponded to the majority vote if at least 2 graders agreed and to the intermediate grade if all 3 graders disagreed. This method was determined to closely approximate the full adjudication process described in Krause et al[25] based on a comparison of multiple methods over 100 images described in Tables S2 and S3 (available at www.aaojournal.org). Images for which the final grade was ungradable were excluded from further analysis. Finally, both to reflect the fact that referral decisions are binary (yes or no) and to facilitate comparisons between validation datasets with different types of glaucoma assessment, we binarized both the model predictions and the reference standard. Specifically, for validation dataset A, high-risk glaucoma suspect or likely glaucoma levels were considered referable, and ONH features were binarized according to Table 2.

To compare the performance of the algorithm with graders, a random subset of 411 images from validation dataset A was graded by 10 graders (3 fellowship-trained glaucoma specialists, 4 ophthalmologists, and 3 optometrists). None of these graders participated in creating the reference standard for those images. Assessments of graders were binarized in the same way as the reference standard described previously.

For further evaluation on an independent population, the algorithm was also applied to a second validation dataset, dataset B, comprising 9642 macula-centered color fundus images from the Atlanta VA Eye Clinic diabetic teleretinal screening program. Glaucoma-related International Classification of Diseases (ICD) codes on a per-patient basis and ONH referral codes recorded in the patient’s medical history at any point before the capture of the image or up to 1 year after the image was obtained were used as a reference standard for this dataset. Presence of such a code was considered referable because these are patients that either were referred for an in-person examination by the diabetic teleretinal screening program or were given such a code during an in-person examination. Ophthalmologists at the VA screening program recorded the ICD codes with access to 3 fundus images per eye and the patient’s clinical record. The specific codes used were ICD, Ninth Edition, Clinical Modification, codes for glaucoma (category 365) and ICD, Tenth Edition, Clinical Modification, codes for glaucoma (category H40). When a patient had an ONH referral code on a particular visit, the image from that visit was chosen. When a patient with a glaucoma ICD code underwent multiple visits, the visit closest to the recorded date of the glaucoma-related ICD code was chosen. To evaluate our algorithm’s image-level predictions against the patient-level codes, we applied the algorithm to the fundus images from that visit, typically 1 per eye. The image that produced the highest referable GON prediction was used in the final patient-level analysis; only 1 eye per patient was used in the final analysis.

Lastly, the algorithm was applied to validation dataset C, comprising 346 macula-centered and disc-centered fundus images from the glaucoma clinic at Dr. Shroff’s Charity Eye Hospital. The reference standard for this dataset was an eye-level diagnosis of glaucoma as determined by a glaucoma specialist based on a full glaucoma workup that included history, clinical examination, visual field (VF), and OCT assessment. For the final evaluation, we randomly included 1 eye per patient and selected an image from that eye at random. The reference standard for these images consisted of eye-level diagnoses on a 3-tiered scale of normal, suspect, and glaucoma. A diagnosis of suspect or glaucoma was considered referable for purposes of comparison against the algorithm because these patients were undergoing regular follow-up.

Evaluating the Algorithm

The algorithm’s performance was evaluated on each of the 3 validation datasets. In all cases, the algorithm only takes as input and makes predictions based on the fundus photograph, although the reference standard for that photograph may be determined using other clinical data as described previously. The prediction for referable GON was a continuous number

Table 1. Baseline Characteristics

	Development Datasets[*]		Validation Datasets		
Characteristics	*Training Dataset*	*Tuning Dataset*	*Validation Dataset A*	*Validation Dataset B (Veterans Affairs Atlanta)*	*Validation Dataset C (Dr. Shroff)*
No. of images (1 image per patient)	86 618	1508	1205	9642	346
No. of graders	43 graders: 14 glaucoma specialists, 26 ophthalmologists, and 3 optometrists	11 glaucoma specialists	12 glaucoma specialists	6 ophthalmologists	4 glaucoma specialists
No. of grades per image	1–2	3	3–6	1	1
Grades per grader, median (IQR)	861 (225–2952)	388 (96–553)	185 (93–731)	N/A[†]	NA[†]
Data used to define the reference standard	45° fundus photograph	45° fundus photograph	45° fundus photograph	Health factors and ICD codes	Full glaucoma workup including VF and OCT
Patient demographics					
No. of patients	86 618	1508	1205	9642	346
Age (yrs), median (IQR)	57 (49–65)	57 (49–64.8)	57 (49.5–64)	64 (68.7–57.5)	N/A[†]
No. of images for which age was available	64 861	1386	1115	9642	
Female gender, no./total no. (%) of images for which gender was known	32 413/62 178 (52.1)	731/1349 (54.2)	571/1075 (53.2)	458/9642 (4.7)	N/A[†]
Glaucoma/GON gradability distribution					
Images gradable for glaucoma, no./total (%) among images for which glaucoma gradability was assessed	74 973/86 127 (87.0)	1451/1508 (96.2)	1171/1204 (97.3)	9642/9642 (100)[‡]	346/346 (100)[‡]
Glaucoma/GON risk distribution					
No. nonglaucomatous (%)	35 877 (47.8)	849 (57.1)	687 (57.5)	8753 (90.8)	63 (18.2)
No. low-risk glaucoma suspect (%)	20 740 (27.6)	259 (17.4)	290 (24.3)	N/A[§]	N/A[§]
No. high-risk glaucoma suspect (%)	13 180 (17.5)	268 (18.0)	170 (14.1)	N/A[§]	175 (50.6)
No. likely glaucoma (%)	5307 (7.1)	110 (7.4)	48 (4.0)	890 (9.2)	108 (31.2)
No. referable glaucoma (%)	18 487 (24.6)	378 (25.4)	218 (18.1)	890 (9.2)	283 (81.7)

GON = glaucomatous optic neuropathy; ICD = International Classification of Diseases; IQR = interquartile range; N/A = Not available; VF = visual field.

[*] Prevalence of referable GON risk images is higher than in the general population in part because of active learning, a machine learning technique used to increase preferentially the number of relevant examples (see "Methods").

[†] Data not available.

[‡] All images in validation sets B and C were of overall adequate image quality, but not specifically labeled by graders for glaucoma gradability.

[§] Finer-grained categorization not available, see "Methods."

between 0 and 1, corresponding to the predicted likelihood of that condition being present in the image. Receiver operating characteristic (ROC) curves were plotted by varying the operating threshold.

Evaluating Optic Nerve Head Feature Importance

To assess which ONH features were associated most strongly with referable GON, a multivariable logistic regression analysis was carried out to compute the odds ratio and β coefficients (i.e., natural logarithms of odds ratios) for each feature on referable GON. Five logistic regression analyses were carried out. For validation dataset A, analyses were carried out using the reference standard, the round 1 median (see "Clinical Validation Datasets"), and the algorithm's predictions (Table 3). Analyses also were carried out on the training dataset median and the tuning dataset median (Table S4, available at www.aaojournal.org).

Statistical Analysis

To compute the confidence intervals (CIs) for the algorithm's area under the receiver operating characteristic curve (AUC), we used a nonparametric bootstrap procedure[45] with 2000 samples. Confidence intervals for sensitivities and specificities were calculated using the exact Clopper-Pearson interval.[46] To compare the algorithm's performance (sensitivity and specificity) with that of graders, we used the 2-tailed McNemar test.[45] To compare the distributions of vertical CDR, we used the Kolmogorov-Smirnov test.[47] We measured intergrader agreement using a weighted Krippendorff's α, which is robust to multiple graders.[48] The weighting function was the difference between grades (e.g., the distance between the first grade [non-glaucomatous] and third grade [high-risk glaucoma suspect] is 3 − 1 = 2).

Results

A DL algorithm was developed using a training set of 86 618 fundus images and tuning set of 1508 fundus images that were assessed for the presence of glaucomatous ONH features and referable GON. The images were graded by a panel of 43 graders using a set of detailed grading guidelines (Table S1). Patient demographics and image characteristics for the development and validation sets are summarized in Table 1.

Table 2. Evaluation of Algorithm Performance for Detecting Presence of Individual Features on Validation Dataset A

Feature	Area under the Receiver Operating Characteristic Curve (95% Confidence Interval)	No. of Labeled Images	Prevalence (%)	Binary Cutoffs
Rim width				
I vs. S	0.661 (0.594–0.722)	1162	8.2	I < S vs. I > S or I ≅ S
S vs. T	0.946 (0.897–0.981)	1156	1.6	S < T vs. S > T or S ≅ T
Notch	0.908 (0.852–0.956)	1162	2.6	Yes/possible vs. no
Laminar dot sign	0.950 (0.937–0.963)	1013	24	Yes/possible vs. no
Nasalization				
Emerging	0.973 (0.954–0.987)	1166	4.7	Yes vs. possible/no
Directed	0.957 (0.944–0.969)	1167	15.9	Yes vs. possible/no
Baring of circumlinear vessels	0.723 (0.688–0.755)	1154	22.7	Present and clearly bared vs. all else
Disc hemorrhage	0.758 (0.666–0.844)	1173	2.1	Yes/possible vs. no
β PPA	0.933 (0.914–0.948)	1170	16.9	Yes/possible vs. no
RNFL defect	0.778 (0.706–0.843)	973	6.5	Yes/possible vs. no
Vertical CDR	0.922 (0.869-0.963)	1154	4.6	≥0.7 vs. <0.7

CDR = cup-to-disc ratio; I = inferior; PPA = parapapillary atrophy; RNFL = retinal nerve fiber layer; S = superior; T = temporal.

We first evaluated the algorithm on validation dataset A in which referable GON was assessed by glaucoma specialists using the images alone and achieved an AUC of 0.945 (95% CI, 0.929–0.960), with sensitivity of 80.0% and specificity of 90.2% at the balanced operating point (Fig 1A). Figure 2 summarizes the performance of the algorithm in detecting referable GON on a subset of 411 images from validation dataset A. On this random subset, the algorithm achieved an AUC of 0.933 (95% CI, 0.899–0.961), with sensitivity of 76.1% and specificity of 92.3% at the balanced operating point. Ten eye care providers who were not involved in creating the reference standard graded this same subset. The performance of these 10 graders on this same subset is plotted for comparison on the ROC curve for algorithm performance (Fig 2; Table S5, available at www.aaojournal.org).

Graders' sensitivities ranged from 29.2% to 73.6% (37.5%–63.9% for glaucoma specialists, 29.2%–73.6% for ophthalmologists, and 29.2%–70.8% for optometrists), and specificities ranged from 75.8% to 92.6% (85.5%–90.6% for glaucoma specialists, 75.8%–90.6% for ophthalmologists, and 87.6%–92.6% for optometrists). The algorithm was significantly more sensitive than 7 of 10 graders (including 2 of 3 glaucoma specialists) with no statistically significant difference for the other 3 graders. The algorithm showed significantly higher specificity than 1 glaucoma specialist and 2 ophthalmologists (Table S5) and no statistically significant difference in specificity for the other 7 of 10 graders.

Next, we evaluated the performance of the algorithm relative to real clinical decisions as the reference standard. First, for validation dataset B, in which ophthalmologists made referral decisions with access to 3 fundus images and clinical data, the algorithm achieved

Table 3. Logistic Regression Models to Understand the Relative Importance of Individual Optic Nerve Head Features for Glaucomatous Optic Neuropathy Referral Decisions in Validation Dataset A: the Reference Standard, Algorithm Predictions, and Round 1 Majority (n = 1015)*

Feature	Reference Standard			Algorithm Predictions			Round 1 Majority		
	Odds Ratio	*P Value*	*Rank*	*Odds Ratio*	*P Value*	*Rank*	*Odds Ratio*	*P Value*	*Rank*
Vertical CDR ≥0.7	581.671[†]	<0.001	1	347.861[†]	<0.001	1	475.757[†]	<0.001	1
Notch: possible or yes	29.438	<0.001	2	9.564	0.021	3	4.158	0.218	4
RNFL defect: possible or yes	10.740	<0.001	3	13.098	<0.001	2	12.946	<0.001	2
Circumlinear vessels: present + bared	4.728	<0.001	4	6.241	<0.001	4	4.852	<0.001	3
Laminar dot: possible or yes	3.594	<0.001	5	3.320	<0.001	7	3.882	<0.001	6
Disc hemorrhage: possible or yes	3.221	0.043	6	2.178	0.369	9	1.649	0.508	9
Nasalization emerging: yes	3.162	0.008	7	4.253	0.001	5	4.014	<0.001	5
Rim comparison: I < S	2.560	0.004	8	3.512	<0.001	6	3.033	<0.001	7
Nasalization directed: yes	2.230	0.002	9	3.010	<0.001	8	2.239	0.001	8
Rim comparison: S < T	1.461	0.799	10	1.257	0.894	11	1.175	0.919	11
β PPA: possible or yes	1.319	0.226	11	1.584	0.076	10	1.357	0.192	10

CDR = cup-to-disc ratio; I = inferior; PPA = parapapillary atrophy; S = superior; RNFL = retinal nerve fiber layer; T = temporal.

*Some images in validation dataset A were excluded from analysis because they were ungradable on referral criteria or for a specific optic nerve head feature.

[†]Extreme odds ratios indicate almost perfect correlation between the feature and the final referral prediction or grade.

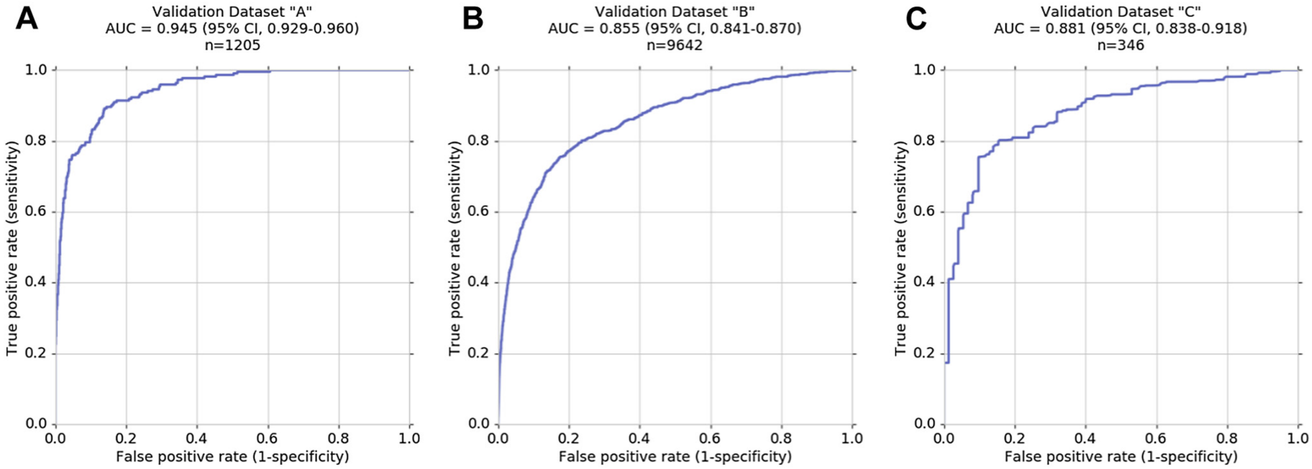

Figure 1. Receiver operating characteristic curve analyses for referable glaucomatous optic neuropathy risk in 3 independent validation datasets. **A**, Validation dataset A, in which the reference standard for all images was determined by a panel of 3 glaucoma specialists. **B**, Validation dataset B (Veterans Affairs Atlanta), in which the reference standard for all images was determined by glaucoma-related International Classification of Diseases codes, assigned to images by eye care providers at a screening program. **C**, Validation dataset C (Dr. Shroff's Charity Eye Hospital), in which the reference standard was determined by glaucoma specialists based on full glaucoma workups. AUC = area under the receiver operating characteristic curve; CI = confidence interval.

an AUC of 0.855 (95% CI, 0.841−0.870; Fig 1B). Second, for validation dataset C, which included glaucoma and glaucoma suspect diagnoses based on a full clinical workup including history, clinical examination, VF assessment, and OCT, the algorithm achieved an AUC of 0.881 (95% CI, 0.836−0.918; Fig 1C).

Additionally, we looked at the performance of the algorithm on binarized ONH features (Table 2). Our algorithm showed good performance in predicting the presence of the different ONH features, with AUCs ranging between 0.661 and 0.973. We also evaluated the relative importance of the different ONH features in the predictions of the algorithm and graders (Table 3; Fig S1A, B, available at www.aaojournal.org). A vertical CDR of 0.7 or more, presence of a neuroretinal rim notch, presence of a RNFL defect, and presence of bared circumlinear vessels showed the strongest correlation with an overall assessment of referable GON for both glaucoma specialists and the algorithm. Presence of a neuroretinal rim notch, presence of laminar dots, and presence of a disc hemorrhage (in descending order) showed a stronger correlation with the referable GON of the reference standard labels compared with the round 1 majority grades. See Table S4 (available at www.aaojournal.org) for similar analyses on the training and tuning datasets.

To better understand the association of ONH features with referable GON assessment, we plotted the distributions of ONH features in the refer and no-refer categories (Fig 3D; Fig S2, available at www.aaojournal.org). Most strikingly, the vertical CDR distributions were significantly different between the refer and no-refer categories ($P < 0.001$ for differences in the distribution; Fig 3A). We further plotted referral rate as a function of each ONH feature's grade, mimicking a positive predictive value analysis of each feature. Marked increases in referral rates were observed when RNFL defect, disc hemorrhage, laminar dot sign, and β PPA were present or possibly present. Figure 4A includes fundus examples in which the algorithm's referral predictions differed from the reference standard in validation dataset B. Figure 4B includes fundus examples in which the algorithm's referral predictions differed from the reference standard in validation dataset C.

We further sliced the results for validation dataset A based on self-reported sex: the algorithm achieved an AUC of 0.939 (95% CI, 0.916−0.960) for females (n = 571; Fig S3A, available at www.aaojournal.org) and an AUC of 0.947 (95% CI, 0.921−0.968) for males (n = 504; Fig S3B). Validation dataset B included too few females (4.7%) to make the analysis by self-reported sex meaningful, and self-reported sex was not available for validation dataset C at the time of study.

Discussion

In the present study, we developed a DL algorithm for the prediction of referable GON and the prediction of glaucomatous ONH features from fundus images. We validated this algorithm on multiple independent datasets from different patient populations using established adjudication methods, clinical referral decisions, and actual clinical diagnoses as reference standards. In addition, we analyzed the relative importance that glaucoma specialists attribute to the different ONH features in their overall assessment of referable GON and whether the DL algorithm learned a similar relationship.

Performance of the Algorithm

Our DL algorithm achieved an AUC of 0.945 for the detection of referable GON using a reference standard based on glaucoma specialists' assessment of fundus photographs alone, (validation dataset A). When evaluated against the referral decisions of eye care providers in a teleretinal screening program (validation dataset B), our algorithm achieved an AUC of 0.855. This decrease in performance relative to validation dataset A may be explained by the difference in reference standard and patient populations between the 2 validation datasets. Specifically, while both the algorithm and graders of validation dataset A had access

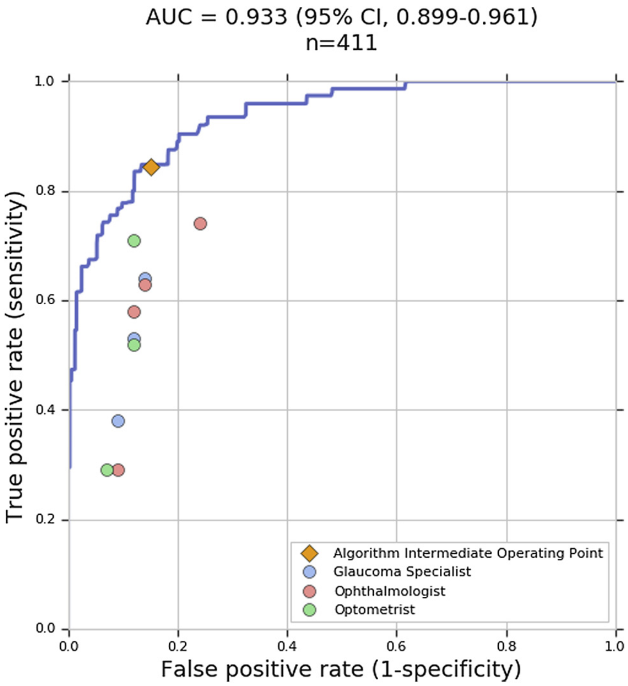

Figure 2. Receiver operating characteristic curve analysis for referable glaucomatous optic neuropathy risk in a subset of validation dataset A (n = 411) with comparison with clinicians. The algorithm is illustrated as a blue line, with 10 individual graders indicated by colored dots: glaucoma specialists (blue), ophthalmologists (red), and optometrists (green). The diamond corresponds to the balanced operating point of the algorithm, chosen based on performance on the tuning set. For each image, the reference standard was determined by a different set of 3 glaucoma specialists in an adjudication panel (see "Methods"). Images labeled by graders as ungradable for glaucoma were classified as "refer" (referable) to enable comparison on the same set of images. For a sensitivity and specificity analysis excluding the ungradable images on a per-grader basis, see Table S6. See Figure 1A for analysis on the entire validation dataset A. AUC = area under the receiver operating characteristic curve; CI = confidence interval.

to only a single fundus image from each patient, graders for validation dataset B had access to the electronic medical record, including patient history and previous notes containing clinical data. When evaluated on a validation dataset with a reference standard of glaucoma diagnosis based on a full glaucoma workup (validation dataset C), the model achieved an AUC of 0.881. Similar to validation dataset B, this decrease in performance relative to validation dataset A may be explained by differences in reference standard and patient population. Nonetheless, the fact that our algorithm, which was developed using eye care providers' grades based on fundus photographs alone, maintained good performance on an independent dataset with grades based on a full workup suggests that the algorithm's prediction of referable GON is fairly well correlated with the results of a full glaucoma workup.

Several previous studies reported performance of DL algorithms with AUCs ranging between 0.91 and 0.986.[26,28–31] Our work advances the field by demonstrating performance on par with or surpassing eye care providers after addressing the following limitations of prior works: (1) grading of images by non-glaucoma specialists, limiting the validity of GON identification[26,29,31]; (2) lack of a consistent definition of glaucoma for the reference standard, limiting the effectiveness of algorithm performance evaluation[28]; (3) exclusion of low-quality images from the training set, limiting the usefulness of the algorithm in real-life settings[30]; (4) exclusion of images on which graders had disagreed from the validation set, skewing the dataset toward easier cases[30]; and (5) use of fundus images zoomed in on the optic nerve, limiting assessment of the RNFL.[31]

Optic Nerve Head Features Analysis

We demonstrate herein that a DL algorithm trained only on fundus photographs containing ONHs can accurately predict glaucoma assessments made by glaucoma specialists with access to additional patient information (validation set C). This suggests that combining information about different ONH pathologic features has clinically relevant predictive power for overall glaucoma risk. Therefore, we wanted to quantitatively assess how glaucoma specialists make clinical decisions on suspicious ONHs. For each image, we asked our graders to indicate the presence or absence of certain glaucoma-associated ONH features, in addition to overall assessment of referable GON. It is worth noting that for referral decision, graders were guided to use their "gut feel" tied to the clinical actions they deemed necessary, rather than directly map features to overall risk (see Table S1). We have also trained an algorithm for the detection of most known glaucomatous ONH features, showing good performance in predicting the presence of the different ONH features, with AUCs ranging between 0.661 and 0.973.

Not surprisingly, we found that the presence of a vertical CDR of 0.7 or more was highly correlated with an overall assessment of an image as referable GON. Of several additional ONH features listed as possibly indicative of GON in the American Academy of Ophthalmology's Primary Open-Angle Glaucoma Suspect Preferred Practice Patterns,[14] the presence of a neuroretinal rim notch, RNFL defect, or a bared circumlinear vessel were most correlated with referable GON, with similar ranking order and coefficients for algorithm predictions and reference standard (as shown in Table 3).

Studies training DL algorithms for glaucoma detection based on color fundus images used slightly varying sets of ONH features to determine glaucoma risk, with different thresholds for CDR. Li et al[29] defined referable GON as any of the following: CDR of more than 0.7, rim thinning or notching, disc hemorrhage, or RNFL defect. Shibata et al[30] labeled images as glaucoma according to the following: focal rim notching or generalized rim thinning, large CDR with cup excavation with or without laminar dot sign, RNFL defects with edges at the ONH margin, disc hemorrhages, and peripapillary atrophy. Ting et al[26] identified possible glaucoma if any of the following were present: CDR of 0.8 or more, focal thinning or notching of the neuroretinal rim, disc hemorrhages, or RNFL defect. Other studies[28,31] did not report image grading guidelines.

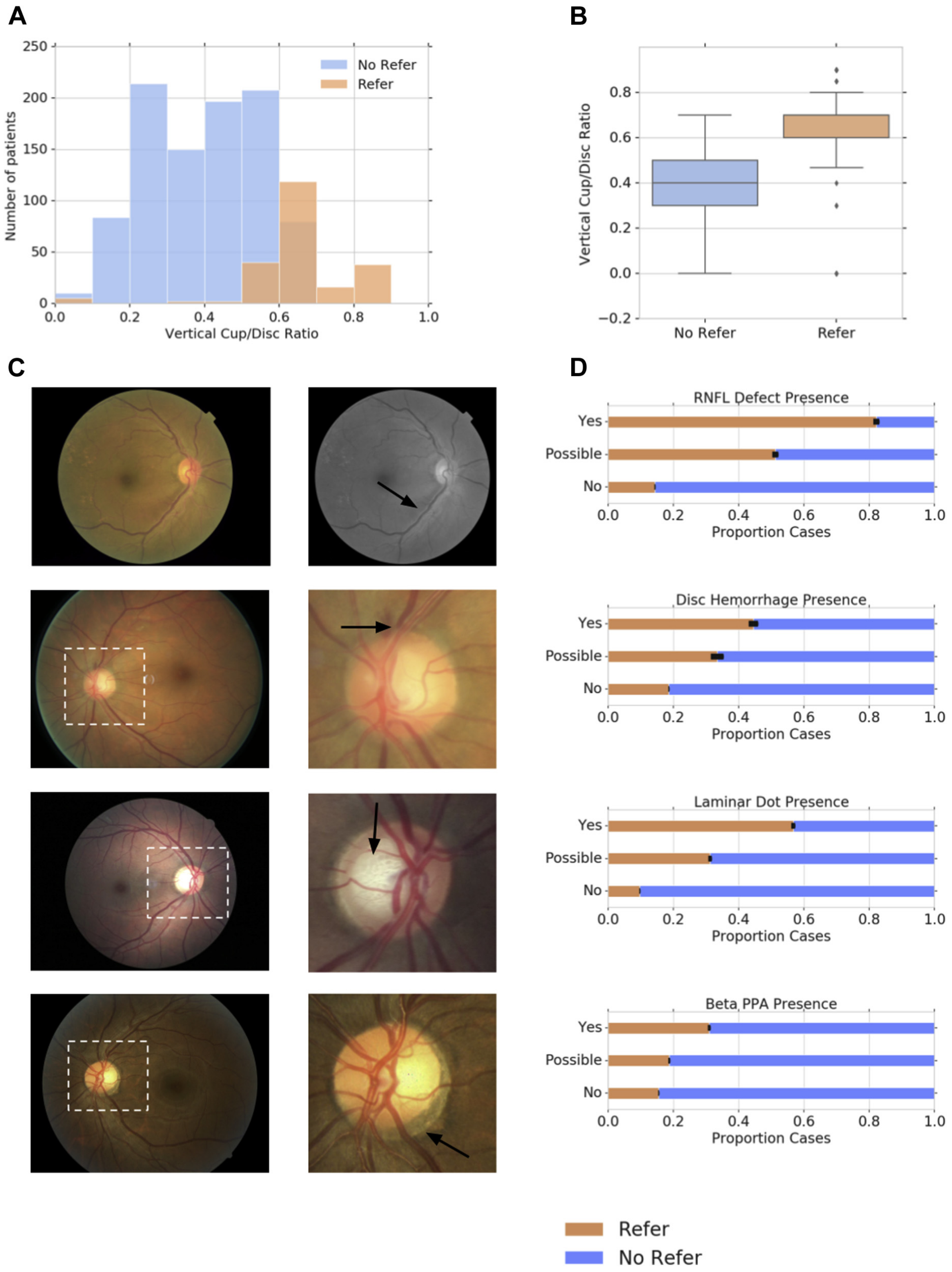

Figure 3. Proportions of selected optic nerve head (ONH) feature grades among the refer and no-refer categories in validation data set A. **A**, Bar graph showing distribution of images based on their vertical cup-to-disc-ratio, stratified by refer or no-refer categories of glaucomatous optic neuropathy (GON) risk. **B**, Box-and-whisker plot of vertical cup-to-disc-ratio by refer or no-refer categories of GON risk. **C**, Optic nerve head feature example images for retinal nerve fiber layer (RNFL) defect, disc hemorrhage, laminar dot sign, and β parapapillary atrophy (PPA). **D**, Corresponding distributions of ONH feature presence by refer or no-refer categories of GON risk. Error bars represent the 95% confidence intervals. **C**, Dotted squares indicate the part of the image that has been enlarged and presented on the right. Black arrows point out the features of interest corresponding to panel **D**.

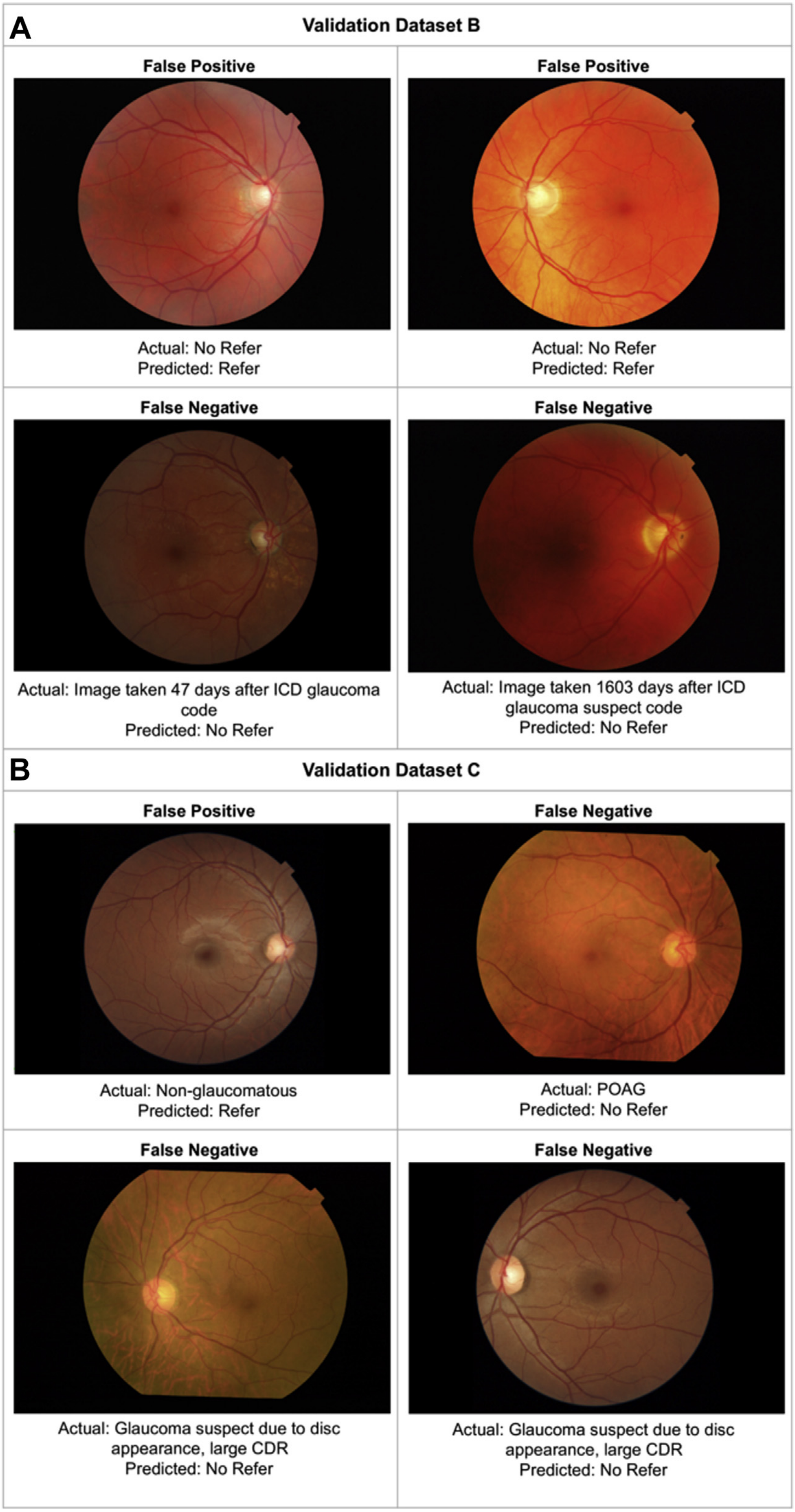

Figure 4. Examples of incorrect algorithm predictions on external validation datasets. **A**, Images showing a referral prediction different than the reference standard for validation set B. Discrepancies between optic nerve head appearance and Veterans Affairs (VA) diabetic teleretinal screening program glaucoma referral decisions may be explained by VA providers' access to clinical data available in patients' electronic medical records. **B**, Images showing a referral prediction different than the diagnosis reference standard for validation set C. Discrepancies between model referral decisions versus those based on clinical diagnoses may be explained by provider access to visual field and OCT findings as well as clinical data. CDR = cup-to-disc ratio; ICD = International Classification of Diseases; POAG = primary open-angle glaucoma.

Our model found the ONH features of CDR, notch, and RNFL defect to be correlated most often with referable GON, which is in line with the approaches described previously. Interestingly, features mostly not assessed in previous studies, that is, baring of circumlinear vessels, laminar dots (somewhat assessed in Shibata et al[30]), and nasalization of central ONH vessels, were significantly correlated with referable GON, whereas disc hemorrhage was not. The relatively low ranking of disc hemorrhage in our study may be explained by the study population. Images used in this study largely were derived from DR screening programs. We asked our graders to mark the presence of any disc hemorrhage, regardless of its presumed etiology. Therefore, some of the disc hemorrhages identified by our graders may have been the result of DR.

Our study is the only study to specifically ask graders to assess for the presence of baring of circumlinear vessels (an ONH feature first described by Herschler and Osher[49] in 1980) and nasalization of central ONH vessels (a feature included in the American Academy of Ophthalmology's Primary Open-Angle Glaucoma Suspect Preferred Practice Patterns), both associated with glaucomatous optic nerves yet rarely delineated separately by glaucoma specialists in clinical practice. Interestingly, the presence of bared circumlinear vessels was among the ONH features that were most highly correlated with referable GON, whereas nasalization of central ONH vessels was not. One reason for this distinction may be that the presence of bared circumlinear vessels is associated with adjacent rim thinning, making it more likely to be correlated with GON than nasalization of central ONH vessels, often seen in large discs.[50]

Our findings may be helpful in deciphering the decision-making process by which glaucoma specialists identify an ONH as glaucomatous, often previously described by many as a "gut feeling." Understanding which features are weighed most highly by glaucoma specialists for grading an ONH as more likely to be glaucomatous may help in the development of training programs to enable nonspecialists to recognize those features in a teleophthalmology or screening setting. Similarly, by training the algorithm to detect individual ONH features, we may be able to better explain what the algorithm's predictions rely on and thus, gain insight into what is referred to frequently as a black box of machine learning.

Limitations and Future Work

Our study has several limitations. First, although ONH assessment for detection of referable glaucoma risk is commonly accepted in screening and teleophthalmology settings, it is well established that subjective evaluation of optic disc photographs suffers from low reproducibility, even when performed by expert graders[51,52] (e.g., a panel of glaucoma specialists, as carried out in this study). Second, the diagnosis of glaucoma is not based on ONH appearance alone but also relies on the compilation of risk factors (such as race, age, and family history) and repeated clinical, functional, and structural testing over time, which were included in only 1 validation dataset in our study. In addition, our study did not provide graders with absolute disc size or subjects' race, 2 important factors in the assessment of an ONH for glaucoma.[53,54] Thus, the predictions of our algorithm are not diagnostic for glaucoma but rather are an assessment of the ONH appearance for referable GON. Nonetheless, we believe that our algorithm's generalization to an external dataset with glaucoma diagnoses based on a full glaucoma workup demonstrates the robustness of our algorithm for flagging suspicious discs and detection of referable GON, which can be useful in screening settings. A longitudinal dataset consisting of fundus images, clinical examination data, OCT scans, and VF tests could be used to train an algorithm that may be able to diagnose glaucoma, and perhaps even predict progression.

Regarding leveraging other imaging modalities, spectral-domain (SD) OCT has become a widespread tool for glaucoma diagnosis and follow-up over the last decade, enabling visualization and quantitative analyses of the RNFL and ganglion cell layers.[55] Recent efforts applying DL to SD and swept-source OCT images were able to detect early glaucoma and distinguish healthy from mild glaucoma with high accuracy.[56,57] However, although 3-dimensional SD OCT scans enable better structural analysis of the ONH than 2-dimensional color fundus photographs, optic disc fundus photography is the least expensive and globally, the most commonly used imaging modality for the structural assessment of the ONH. Fundus photography is also widely deployed in DR screening programs in which algorithms to detect risk of non-DR pathology such as age-related macular degeneration and glaucoma could be deployed. Thus, there will remain a role for fundus photography as an imaging modality for screening purposes, especially in low-resource settings.

In conclusion, we developed a DL algorithm with higher sensitivity and comparable specificity to eye care providers in detecting referable GON in color fundus images. The algorithm's prediction of referable GON maintained good performance on an independent dataset with diagnoses based on a full glaucoma workup. Additionally, our work provides insight into which ONH features drive GON assessment by glaucoma specialists. These insights may help to improve clinical decisions for referring patients to glaucoma specialists based on ONH findings during diabetic fundus image assessments. We believe that an algorithm such as this may also enable effective screening for glaucoma in settings where clinicians trained to interpret ONH features are not available, thus reaching underserved populations worldwide. The use of such a tool presents an opportunity to reduce the number of undiagnosed patients with glaucoma and thus, provides the chance to intervene before permanent vision loss occurs.

Acknowledgments

The authors thank Dr. Shroff's Charity Eye Hospital, Atlanta Veterans Affairs Medical Center, Inoveon, Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya for providing de-identified data; William Chen, BA, Quang Duong, PhD, Xiang Ji, MS, Jess Yoshimi, BS, Cristhian Cruz, MS, Olga Kanzheleva, MS, Miles Hutson, BS, and Brian Basham, BS, from Google Health, for software infrastructure contributions; Jorge Cuadros, OD, PhD, from EyePACS, for data access and helpful conversations; and the National Eye Institute Study of Age-Related Macular Degeneration Research Group and study participants for their valuable contribution to this research.

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90(3):262−267.
2. Tham Y-C, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121(11):2081−2090.
3. Leite MT, Sakata LM, Medeiros FA. Managing glaucoma in developing countries. *Arq Bras Oftalmol*. 2011;74(2): 83−84.
4. Rotchford AP, Kirwan JF, Muller MA, et al. Temba Glaucoma Study: a population-based cross-sectional survey in urban South Africa. *Ophthalmology*. 2003;110(2):376−382.
5. Hennis A, Wu S-Y, Nemesure B, et al. Awareness of incident open-angle glaucoma in a population study. *Ophthalmology*. 2007;114(10):1816−1821.
6. Prum Jr BE, Lim MC, Mansberger SL, et al. Primary Open-Angle Glaucoma Suspect Preferred Practice Pattern(®) Guidelines. *Ophthalmology*. 2016;123(1):P112−P151.
7. Weinreb RN. *Glaucoma Screening*. Amsterdam, The Netherlands: Kugler Publications; 2008.
8. Newman-Casey PA, Verkade AJ, Oren G, Robin AL. Gaps in glaucoma care: a systematic review of monoscopic disc photos to screen for glaucoma. *Expert Rev Ophthalmol*. 2014;9(6): 467−474.
9. Bernardes R, Serranho P, Lobo C. Digital ocular fundus imaging: a review. *Ophthalmologica*. 2011;226(4):161−181.
10. Shi L, Wu H, Dong J, et al. Telemedicine for detecting diabetic retinopathy: a systematic review and meta-analysis. *Br J Ophthalmol*. 2015;99(6):823−831.
11. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet*. 2004;363(9422):1711−1720.
12. Bowd C, Weinreb RN, Zangwill LM. Evaluating the optic disc and retinal nerve fiber layer in glaucoma. I: clinical examination and photographic methods. *Semin Ophthalmol*. 2000;15(4):194−205.
13. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma. *JAMA*. 2014;311(18):1901−1911.
14. Prum Jr BE, Rosenberg LF, Gedde SJ, et al. Primary Open-Angle Glaucoma Preferred Practice Pattern(®) Guidelines. *Ophthalmology*. 2016;123(1):P41−P111.
15. Hollands H, Johnson D, Hollands S, et al. Do findings on routine examination identify patients at risk for primary open-angle glaucoma? *JAMA*. 2013;309(19):2035.
16. Mardin CY, Horn F, Viestenz A, et al. [Healthy optic discs with large cups—a diagnostic challenge in glaucoma]. *Klin Monbl Augenheilkd*. 2006;223(4):308−314.
17. Jonas JB, Fernández MC. Shape of the neuroretinal rim and position of the central retinal vessels in glaucoma. *Br J Ophthalmol*. 1994;78(2):99−102.
18. Jonas JB, Schiro D. Localized retinal nerve fiber layer defects in nonglaucomatous optic nerve atrophy. *Graefes Arch Clin Exp Ophthalmol*. 1994;232(12):759−760.
19. Chihara E, Matsuoka T, Ogura Y, Matsumura M. Retinal nerve fiber layer defect as an early manifestation of diabetic retinopathy. *Ophthalmology*. 1993;100(8):1147−1151.
20. Chaum E, Drewry RD, Ware GT, Charles S. Nerve fiber bundle visual field defect resulting from a giant peripapillary cotton-wool spot. *J Neuroophthalmol*. 2001;21(4):276−277.
21. Sutton GE, Motolko MA, Phelps CD. Baring of a circumlinear vessel in glaucoma. *Arch Ophthalmol*. 1983;101(5):739−744.
22. Fingeret M, Medeiros FA, Susanna Jr R, Weinreb RN. Five rules to evaluate the optic disc and retinal nerve fiber layer for glaucoma. *Optometry*. 2005;76(11):661−668.
23. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436−444.
24. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402−2410.
25. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264−1272.
26. Ting DSW, Cheung CY-L, 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(22):2211−2223.
27. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. 2019;126(4): 552−564.
28. Liu S, Graham SL, Schulz A, et al. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. *Ophthalmol Glaucoma*. 2018;1(1): 15−22.
29. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125(8):1199−1206.
30. 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(1):14665.
31. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8(1):16685.
32. Thomas S-M, Jeyaraman MM, Hodge WG, et al. The effectiveness of teleglaucoma versus in-patient examination for glaucoma screening: a systematic review and meta-analysis. *PLoS One*. 2014;9(12):e113779.
33. EyePACS LLC. Welcome to EyePACS. http://www.eyepacs.org; 2018. Accessed 5.12.18.
34. Inoveon Corp. Home page. http://www.inoveon.com/; 2015. Accessed 5.12.18.
35. Age-Related Eye Disease Study Research Group. The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clin Trials*. 1999;20(6): 573−600.
36. UK Biobank. About UK Biobank. http://www.ukbiobank.ac.uk/aboutbiobankuk; 2019. Accessed 5.12.18.
37. Lopes FSS, Dorairaj S, Junqueira DLM, et al. Analysis of neuroretinal rim distribution and vascular pattern in eyes with presumed large physiological cupping: a comparative study. *BMC Ophthalmol*. 2014;14:72.
38. Susanna Jr R. The lamina cribrosa and visual field defects in open-angle glaucoma. *Can J Ophthalmol*. 1983;18(3): 124−126.

39. Poon LY-C, Valle DS-D, Turalba AV, et al. The ISNT rule: how often does it apply to disc photographs and retinal nerve fiber layer measurements in the normal population? *Am J Ophthalmol*. 2017;184:19–27.
40. Szegedy C, Vanhouke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. http://arxiv.org/pdf/1512.00567v3.pdf; 2015. Accessed 3.12.18.
41. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Available at: tensorflow.org.
42. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Bartlett P, ed. *Advances in Neural Information Processing Systems*. RedHook, NY: Curran Associates, Inc; 2012: 1097–1105.
43. Settles B; Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*. 2012;6(1):1–114.
44. Opitz D, Maclin R. Popular ensemble methods: an empirical study. *J Artif Intell Res*. 1999;11:169–198.
45. Chihara LM, Hesterberg TC. *Mathematical statistics with resampling and R*. Hoboken, NJ: Wiley; 2018.
46. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404.
47. Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc*. 1951;46(253):68.
48. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: SAGE Publications; 2018.
49. Herschler J, Osher RH. Baring of the circumlinear vessel. An early sign of optic nerve damage. *Arch Ophthalmol*. 1980;98(5):865–869.
50. Susanna R Jr, Medeiros FA. The Optic Nerve in Glaucoma. 2nd ed. Amsterdam, the Netherlands; 2006.
51. Tielsch JM, Katz J, Quigley HA, et al. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology*. 1988;95(3):350–356.
52. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99(2):215–221.
53. Zangwill LM, Weinreb RN, Berry CC, et al. Racial differences in optic disc topography: baseline results from the confocal scanning laser ophthalmoscopy ancillary study to the ocular hypertension treatment study. *Arch Ophthalmol*. 2004;122(1):22–28.
54. Lee RY, Kao AA, Kasuga T, et al. Ethnic variation in optic disc size by fundus photography. *Curr Eye Res*. 2013;38(11):1142–1147.
55. Tatham AJ, Medeiros FA. Detecting structural progression in glaucoma with optical coherence tomography. *Ophthalmology*. 2017;124(12S):S57–S65.
56. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26(12):1086–1094.
57. 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.

Footnotes and Financial Disclosures

Originally received: December 20, 2018.
Final revision: July 18, 2019.
Accepted: July 26, 2019.
Available online: ■■■. Manuscript no. 2018-2813.

[1] Google Health, Google LLC, Mountain View, California.

[2] Virginia Ophthalmology Associates, Norfolk, Virginia.

[3] Department of Ophthalmology, Eastern Virginia Medical School, Norfolk, Virginia.

[4] Department of Ophthalmology, Duke University, Durham, North Carolina.

[5] Department of Ophthalmology, Emory University School of Medicine, Atlanta, Georgia.

[6] Ophthalmology Section, Atlanta Veterans Affairs Medical Center, Atlanta, Georgia.

[7] Dr. Shroff's Charity Eye Hospital, New Delhi, India.

*These authors contributed equally as first authors.

‡Both authors contributed equally.

Financial Disclosure(s):
The author(s) have made the following disclosure(s): S.P.: Employee and equity owner – Google LLC (Mountain View, CA).

R.C.D.: Employee and equity owner – Google LLC (Mountain View, CA).

N.H.: Employee and equity owner – Google LLC (Mountain View, CA).

Y.L.: Employee and Equity owner – Google LLC (Mountain View, CA); Patent – WO2018035473A3 (Processing fundus images using machine learning models).

J.K.: Employee and equity owner – Google LLC (Mountain View, CA).

N.K.: Employee and equity owner – Google LLC (Mountain View, CA).

M.S.: Employee – Google LLC (Mountain View, CA).

R.S.: Employee and equity owner – Google LLC (Mountain View, CA).

D.J.W.: Employee and equity owner – Google LLC (Mountain View, CA).

A.B.: Employee and equity owner – Google LLC (Mountain View, CA).

C.S.: Employee and equity owner – Google LLC (Mountain View, CA).

A.M.: Employee and equity owner – Google LLC (Mountain View, CA).

A.E.H.: Consultant – Google LLC (Mountain View, CA); Employee – Advanced Clinical (Deerfield, IL).

F.A.M.: Consultant – Carl-Zeiss Meditec, Inc., Reichert, Inc., Allergan, Novartis, Quark Pharmaceuticals, Stealth Biotherapeutics, Galimedix Therapeutics, Inc., Biozeus, Inc., Ngoggle, Inc.; Financial support – Carl-Zeiss Meditec, Heidelberg Engineering, Reichert, Dyopsys, Inc., Google LLC (Mountain View, CA)

G.S.C.: Employee and equity owner – Google LLC (Mountain View, CA); Patent – WO2018035473A3 (Processing fundus images using machine learning models).

L.P.: Employee and equity owner – Google LLC (Mountain View, CA); Patent – WO2018035473A3 (Processing fundus images using machine learning models).

D.R.W.: Employee and equity owner – Google LLC (Mountain View, CA); Patent – WO2018035473A3 (Processing fundus images using machine learning models).

Google LLC, Mountain View, California, funded this study and had a role in its approval for publication. This research was conducted using the United Kingdom Biobank Resource under application number 17643. Some images used for the analyses described in this manuscript were obtained from the National Eye Institute Study of Age-Related Macular Degeneration (NEI-AMD) database (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1) through dbGaP accession number phs000001.v3.p1.c1. Funding support for the NEI-AMD was provided by the National Eye Institute, National Institutes of Health, Bethesda, Maryland (grant no.: N01-EY-0-2127).

HUMAN SUBJECTS: Human subjects were included in this study. All images were de-identified according to HIPAA Safe Harbor prior to transfer to the study investigators. Ethics review and Institutional Review Board exemption were obtained using Quorum Review Institutional Review Board. All research complied with the Health Insurance Portability and Accountability Act of 1996.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Phene, Dunn, Hammel, Liu, Krause, Huang, Spitze

Analysis and interpretation: Phene, Dunn, Hammel, Liu, Krause, Kitade, Sayres, Wu, Bora, Semturs, Schaekermann, Huang, Medeiros

Data collection: Phene, Dunn, Hammel, Liu, Kitade, Schaekermann, Misra, Huang, Spitze, Maa, Gandhi, Corrado, Peng, Webster

Obtained funding: Phene, Dunn, Hammel, Liu, Krause, Kitade, Schaekermann, Sayres, Wu, Bora, Semturs, Misra, Medeiros, Corrado, Peng, Webster

Overall responsibility: Phene, Dunn, Hammel, Liu, Krause, Kitade, Schaekermann, Sayres, Wu, Bora, Semturs, Misra, Huang, Spitze, Medeiros, Maa, Gandhi, Corrado, Peng, Webster

Abbreviations and Acronyms:

AUC = area under the receiver operating characteristic curve; **CI** = confidence interval; **DL** = deep learning; **DR** = diabetic retinopathy; **GON** = glaucomatous optic neuropathy; **ICD** = International Classification of Diseases; **ISNT** = inferior > superior > nasal > temporal; **ONH** = optic nerve head; **PPA** = parapapillary atrophy; **RNFL** = retinal nerve fiber layer; **VA** = Veterans Affairs; **VF** = visual field.

Correspondence:

Naama Hammel, MD, Google Health, Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043. E-mail: nhammel@google.com.