AMERICAN ACADEMY OF OPHTHALMOLOGY®

Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs

Hans G. Lemij, MD, PhD,[1] *Coen de Vente, MSc,*[2,3] *Clara I. Sánchez, PhD,*[2,3] *Koen A. Vermeer, PhD*[4]

Purpose: Significant visual impairment due to glaucoma is largely caused by the disease being detected too late.

Objective: To build a labeled data set for training artificial intelligence (AI) algorithms for glaucoma screening by fundus photography, to assess the accuracy of the graders, and to characterize the features of all eyes with referable glaucoma (RG).

Design: Cross-sectional study.

Subjects: Color fundus photographs (CFPs) of 113 893 eyes of 60 357 individuals were obtained from EyePACS, California, United States, from a population screening program for diabetic retinopathy.

Methods: Carefully selected graders (ophthalmologists and optometrists) graded the images. To qualify, they had to pass the European Optic Disc Assessment Trial optic disc assessment with $\geq$ 85% accuracy and 92% specificity. Of 90 candidates, 30 passed. Each image of the EyePACS set was then scored by varying random pairs of graders as "RG," "no referable glaucoma (NRG)," or "ungradable (UG)." In case of disagreement, a glaucoma specialist made the final grading. Referable glaucoma was scored if visual field damage was expected. In case of RG, graders were instructed to mark up to 10 relevant glaucomatous features.

Main Outcome Measures: Qualitative features in eyes with RG.

Results: The performance of each grader was monitored; if the sensitivity and specificity dropped below 80% and 95%, respectively (the final grade served as reference), they exited the study and their gradings were redone by other graders. In all, 20 graders qualified; their mean sensitivity and specificity (standard deviation [SD]) were 85.6% (5.7) and 96.1% (2.8), respectively. The 2 graders agreed in 92.45% of the images (Gwet's AC2, expressing the inter-rater reliability, was 0.917). Of all gradings, the sensitivity and specificity (95% confidence interval) were 86.0 (85.2–86.7)% and 96.4 (96.3–96.5)%, respectively. Of all gradable eyes (n = 111 183; 97.62%) the prevalence of RG was 4.38%. The most common features of RG were the appearance of the neuroretinal rim (NRR) inferiorly and superiorly.

Conclusions: A large data set of CFPs was put together of sufficient quality to develop AI screening solutions for glaucoma. The most common features of RG were the appearance of the NRR inferiorly and superiorly. Disc hemorrhages were a rare feature of RG.

Financial Disclosure(s): Proprietary or commercial disclosure may be found after the references.

Glaucoma is a leading cause of irreversible blindness and visual impairment in the world. [1–4] The number of people with glaucoma worldwide is expected to grow to ~ 112 million by the year 2040.[4] The visual impairment may lead to loss of quality of life, loss of income, decreased mobility,[5] falls,[6,7] and dependence on others. There are ethnic and global differences in prevalence and type of glaucoma.[8–21] The main reason for severe visual impairment and blindness in glaucoma is that the disease is detected too late.[22–27] One reason for the late detection of glaucoma is that, in its early stages, the patient is typically asymptomatic.[28] In addition, even eye health professionals often fail to identify the disease.[21] As a result, in the developed world, only ~ 50% of all cases with glaucoma are detected.[10–12,29–32] In developing countries, the proportion of detected disease is considerably lower, maybe as low as ~ 5% to 10%.[32–36] Population screening and subsequent therapy may reduce the incidence of bilateral low vision and blindness.[37,38] Care costs increase fourfold when late disease is managed, leading to a significant financial burden in most countries.[39] Color fundus photography may be the best option for population-based screening programs for those at an increased risk of glaucoma, because it is the simplest, least expensive, and most widely used way of optic disc imaging. However, the workload to manually grade all these images makes it very costly and the differences between graders may yield a

https://doi.org/10.1016/j.xops.2023.100300
ISSN 2666-9145/23

questionable accuracy.[40] Several glaucoma screening programs have been shown to not be cost-effective, although eliminating the need for expert graders with better technologies that allow inexpensive imaging of the eye fundus; glaucoma screening has been shown to be cost-effective in India and China.[38,41–43] To reduce the cost of screening, automated detection of glaucoma may play an important role, because artificial intelligence (AI) based on supervised learning and classification of labeled fundus photographs alone has yielded promising results, often reaching the level of experienced clinicians.[40,44–55] However, when validated on other data sets, the performance of the AI models often drops, at times severely limiting the utility of AI under screening conditions, an issue that the current study wishes to address. Why may the external validation of AI be so disappointing?

The various studies show significant differences in their design and in the utilized data sets[40,44]; they differ, e.g., in the size of the training, test, and validation sets. In addition, the populations may differ significantly between studies in ethnicity and geographical location. Often, different camera systems were used. Differences also occur in whether the fundus photographs had been carefully selected in clinics (with "supernormal" [i.e., without any other significant disease or risk factors] and superglaucomatous [i.e., without any other eye disease]) or whether they were taken from a factual screening scenario. In addition, in some studies, the ungradable (UG) images were left out of the data sets, whereas in factual situations, perhaps most notably in screening situations, UG images are likely to occur. It is often unclear if the fundus photographs came from populations that were representative of the target population. Perhaps more importantly, the studies differ in their definition of glaucoma (from unconfirmed and unclearly defined "referable glaucoma [RG]" to glaucoma confirmed by 1 or more additional tests, such as [subjective grading of] visual fields or [subjective grading of] OCT scans). In addition, studies differ in the number of graders, the experience of the graders, and the amount of agreement between graders. In some studies, the data sets were "enriched" with fundus photographs of eyes with suspect glaucoma, although this category was not clearly defined. The grading accuracy of the graders (in terms of sensitivity and specificity, for instance) has also not been reported in all studies. In general, agreement between graders is poor.[56] Studies also differed significantly in the prevalence of disease per data set, from a low prevalence to be expected under screening circumstances to up to ~ 50%, which will have a significant effect on the pretest probability for the classification of disease. Taken together, the performance dropped when the AI was externally validated on different labeled data sets, perhaps not surprisingly given the described differences between studies and data sets. In addition, the prediction performance of the developed AI dropped in the presence of coexisting ocular disease.[49]

The purpose of the current study was to put together a large, labeled data set of color fundus photographs (CFPs) to be used for the training and validation of AI that could detect RG, even in the presence of coexisting disease, under factual screening conditions in many parts of the world. To that end, we aimed for a multiethnic data set, obtained with a wide array of fundus cameras by multiple operators in various locations, without the exclusion of UG images that occur under factual screening conditions. The photographs were to be classified by trained, qualified, and experienced graders, whose performance was continuously monitored. The scope of the current manuscript is to describe the methods used to obtain the labeled data set, as well as to describe the features of those fundus photographs that were labeled as RG. These features *per se* could be of clinical interest, especially for the design of glaucoma screening programs, because they offer the opportunity to see the relative importance of each of the features in RG. In addition, the results of this study may also be used to develop, improve, and refine (explainable) AI algorithms.

Methods

Data Set

Color fundus photographs of 113 893 eyes of 60 357 individuals were obtained from EyePACS, California, United States, from a population screening program for diabetic retinopathy.[57] They provided a large data set that was labeled as "random" and a small set labeled as "glaucoma suspects." The CFPs had been taken in ~ 500 screening centers across the United States on a large variety of cameras. Per eye, 3 images were taken, to reduce the risk that an eye could not be properly judged because the image was poorly aligned, over- or underexposed, or otherwise UG because of a closed eyelid or media opacities. The data set was multiethnic and included people of African descent (6%), Whites (8%), Asians (4%), Latin Americans (52%), native Americans (1%), people from the Indian subcontinent (3%), people of mixed ethnicity (1%), and people of unspecified ethnicity (25%). The participants' mean age (standard deviation [SD]) was 57.1 (10.4) years. All the CFPs were anonymized. The entire project was approved by the Institutional Review Board of the Rotterdam Eye Hospital.

Graders; Training and Selection

The basis for training and selection of the graders was the European Optic Disc Assessment Trial (EODAT), which was a trial in which several hundred European ophthalmologists from 11 countries were asked to grade 110 stereoscopic optic nerve photographs (slides) as either normal or glaucomatous.[56] All glaucomatous eyes showed reproducible visual field defects on standard automated perimetry. The results showed a large variation between observers, with an average accuracy of ~ 80%. Glaucoma specialists showed a slightly better accuracy of about 86%. Shortly after the EODAT had been completed, the top 3 European graders were asked to annotate every slide with as many features as they judged to be most typically glaucomatous. The annotated slides, after consensus had been reached between the 3 graders, served as the basis for a stereoscopic Optic Nerve Evaluation, 3-hour teaching program, that one of the current authors (H.L.) has been giving regularly ever since. Some key points of the course will be highlighted to illustrate the choice of glaucomatous feature options in the current study. During the course, special attention is given to the importance of disc size (and how it affects the size of the cup, the width of the neuroretinal rim [NRR], and the cup/disc ratio). In addition, the course points out another limitation of the cup/disc ratio, i.e., that the diameter of the cup is

difficult to determine with a flat sloping NRR, whereas it is much easier with an, albeit rarer, "punched out" steep sloping rim. The clinician should therefore pay attention to the appearance of the NRR rather than to the diameter of the cup or the cup/disc ratio. Relevant features of the rim are color (e.g., the presence of pallor), focal notching, more generalized thinning or narrowing of the rim, and bayonetting of the vessels, the latter being a sign of underlying NRR tissue loss.

For the present study, all candidate graders (90 experienced ophthalmologists and optometrists) had to take the EODAT test, $\geq$ 3 months after the Optic Nerve Evaluation teaching program. To qualify, they had to pass the EODAT test with $\geq$ 85% overall accuracy and 92% specificity. Of all 90 candidates, 30 passed.

Grading Tool

To facilitate the grading of all CFPs, Deepdee (Nijmegen, The Netherlands) was commissioned to put together a web-based grading tool, that was quick, easy to use, and that allowed the efficient grading of photo batches of 200 eyes at a time by randomly paired qualified graders (Fig 1). To avoid grading fatigue, each grader was allowed to grade a maximum of 2 batches (i.e., 400 eyes) per day. Each grader was oblivious to who the fellow, paired grader was. Per presented CFP, fitted to fill the screen, there were 3 available options, 1 of which the grader could select by clicking a button: "Referable glaucoma", "No referable glaucoma (NRG)," or "Ungradable"; if UG was selected, another photo of the same eye was presented immediately, up to a maximum of 2 additional photos. The grader could not return to a previous photo of the same eye. If RG was selected, 10 additional buttons appeared so that the grader could select up to 10 of the typically glaucomatous features of that eye. These features were as follows: "appearance NRR superiorly," "appearance NRR inferiorly," "baring of the circumlinear vessel superiorly," "baring of the circumlinear vessel inferiorly," "disc hemorrhage(s)," "retinal nerve fiber layer defect superiorly," "retinal nerve fiber layer defect inferiorly," "nasalization (nasal displacement) of the vessel trunk," "laminar dots," and "large cup." The (vertical) cup-to-disc (CDR) ratio was not an option, because it was never annotated by the 3 top graders of the EODAT. In addition, it strongly depends on disc size, and the borders of the cup are difficult to determine with sloping NRRs. To allow comparison with the fellow eye, its photo was presented at the click of a button. Several of these features will be further discussed below.

Grading Procedure

Each grader could log into the grading tool at any time and from wherever they wished. Before the first grading took place, the graders were encouraged to familiarize themselves with the grading tool in a sandbox environment. The actual grading procedure started with a new batch of 200 CFPs or from whichever photos remained from a previous, unfinished batch. Per presented CFP, the grader could select from 1 of 3 options that are as follows: RG, NRG, or UG. To select RG, the grader expected the glaucomatous signs to be associated with glaucomatous visual field defects on standard automated perimetry. If no glaucomatous visual field defects were expected (in case of, e.g., normal eyes or preperimetric glaucoma), the grader had been instructed to select NRG. Any signs of coexisting eye disease, e.g., diabetic retinopathy, were to be ignored. If the 2 graders scored identically on any of the 3 main categories (RG, NRG, or UG), their classification became the final label of the CFP. In case they disagreed, the photo was judged by a third grader, i.e., 1 of 2 glaucoma specialists (who had passed the EODAT test with a minimum accuracy of 95% each). One of the third graders is an author of the current paper (H.G.L.). The classification of the third grader then became the final label. The glaucomatous features that were provided by the graders in case they graded an eye as RG were recorded without adjudication.

The evaluation of the graders' performances by sensitivity and specificity was determined by comparing their score with the final label. To push the graders to not give up too easily in case of poor image quality, we applied a penalty if they classified a CFP as UG, while the final label was different: in case of a final label of NRG, their specificity was adversely affected whereas in case of a final label of RG, their sensitivity went down. In case the final label was UG, no penalties were given, regardless of the classification by any of the graders.

Grader Monitoring and Guidance

The accuracy of each grader was periodically monitored. If the sensitivity and specificity dropped below 80% or 95%, respectively (the final label served as reference), the grader exited the study and all their gradings were redone by other, randomly selected graders. In all, 20 graders qualified. All graders were encouraged to ask questions (by e-mail) throughout the grading process. Their questions were often related to either purely technical issues or to more medical ones. The one-to-one emails were answered as quickly as possible; technical issues were dealt with by an employee of Deepdee, while medical questions were answered by 1 of the authors (H.L.). Graders that were not proficient in English were encouraged to ask questions in their native tongue. In case specific questions were raised repeatedly, virtual meetings were called to discuss matters with all concerned. The graded data set, i.e., the images together with their annotations is referred to as REGAIS – Rotterdam EyePACS Glaucoma AI Screening data set.

Statistics

To determine the inter-rater reliability or intergrader agreement, several statistics may be produced. The most straightforward approach is to determine the percentages of agreement, which simply expresses the percentage of samples on which the graders agree. Because agreement may happen because of chance, Cohen's kappa was developed to correct for agreement by chance. However, Cohen's kappa can be low when agreement is high.[58] This so-called paradox may occur in unbalanced data sets such as the present one. Gwet's AC (agreement coefficient) addresses this problem and provides an improved inter-rater reliability metric for such cases.[59] Gwet's AC1 is limited to nominal data; AC2 may be used for ordinal and interval measurements as well and was calculated in this paper to express the inter-rater reliability. Gwet's AC2 calculations were performed based on the agreement of the classification of each image between the 2 graders. In addition, to aid the comparison with other reports, pooled results were produced for the agreement between a grading and the final label. These results will be displayed in a table and are further summarized by percentage of agreement and by Cohen's kappa.

Results

Graders

Of the 30 graders that had passed the EODAT entry examination, 10 failed the periodic monitoring process that took place throughout the grading procedure. Of the remaining 20 graders, the mean sensitivity and specificity (SD) were 85.6 (5.7)% and 96.1 (2.8)%, respectively (the

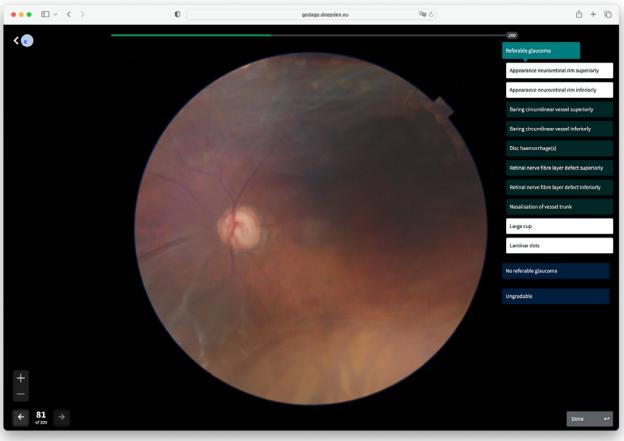

Figure 1. Screenshot of grading tool. The grading tool provided a large color fundus photograph, together with several displays and buttons for navigation purposes. To grade the image, there were 3 main buttons on the right, marked Referable glaucoma," "No referable glaucoma," and "Ungradable." If the button "Referable glaucoma" was selected, 10 additional buttons were presented for glaucomatous feature selection (up to 10 allowed). In this example, 4 features have been selected (highlighted).

final label served as reference). The 2 graders agreed in 92.45% of the images (Gwet's AC2 was 0.917). The third graders graded ~ 11 250 CFPs. Of all gradings, the sensitivity and specificity (95% confidence interval) were 86.0 (85.2−86.7)% and 96.4 (96.3−96.5)%, respectively. The (pooled) agreement between graders and the final label is shown in Table 1. Cohen's kappa was 0.709. Percentage of agreement was 96.0%.

UG Images

Of all 113 897 eyes, the CFPs of 2714 (2.38%) were classified as UG. Although we did not systematically score the reason for their ungradability, it was noticed that common causes were media opacities (including cataract and vitreous hemorrhage or asteroid hyalosis), overexposure (image too bright), underexposure (image too dark), region of interest (optic nerve head [ONH] and peripapillary region) outside the image, and a closed eyelid.

Referable Glaucoma

Within the subset of gradable eyes (n = 111 183), the prevalence of the final label RG was 4.38%. Figure 2 shows the prevalence of glaucomatous features in images graded as RG.

Table 1. Pooled Agreement of Each Grader Compared with the Final Label

		Final Label		
		NRG	*RG*	*UG*
Grader	NRG	203 478	907	155
	RG	3888	7986	42
	UG	3690	398	3765

NRG = no referable glaucoma; RG = referable glaucoma; UG = ungradable.

The appearance of the NRR inferiorly was the most common glaucomatous feature in the RG eyes, followed by the appearance of the NRR superiorly. A large cup was observed in approximately half of all glaucomatous eyes and a nasal displacement of the vessel trunk in about a third of eyes. Baring of the circumlinear vessels was more frequently observed than retinal nerve fiber layer defects. Disc hemorrhages were a rare finding (3% of eyes).

Some features were often observed together with other features (Figure 3). If, for instance, baring of the circumlinear vessel(s) was scored (either superiorly or inferiorly) as glaucomatous, chances were 90% that the matching NRR was flagged as glaucomatous as well. Conversely, when appearance of the NRR was flagged, baring of the circumlinear vessel(s) was scored in the corresponding region in 28%. In case of a retinal nerve fiber layer defect inferiorly, the probability of a glaucomatous appearance of the NRR inferiorly was, perhaps not surprisingly, 91%. If the appearance of the NRR superiorly was considered as glaucomatous, chances were 70% that the inferior NRR was also considered glaucomatous. Disc hemorrhages occurred quite rarely, but if they were present, they were often (72%) associated with a glaucomatous appearance of the inferior NRR.

Discussion

Artificial intelligence with deep learning has recently sparked off enormous interest in various disciplines in ophthalmology, including glaucoma.[40,44,54,55,60] One promising application is the large scale, population-based screening for glaucoma, based on relatively inexpensive CFPs, since AI algorithms have been shown to be more capable of accurately classifying glaucoma than human graders.[61] This might in turn reduce the enormous burden of glaucomatous visual impairment and blindness across the world. However, several studies have shown that the performance of promising AI dropped when externally

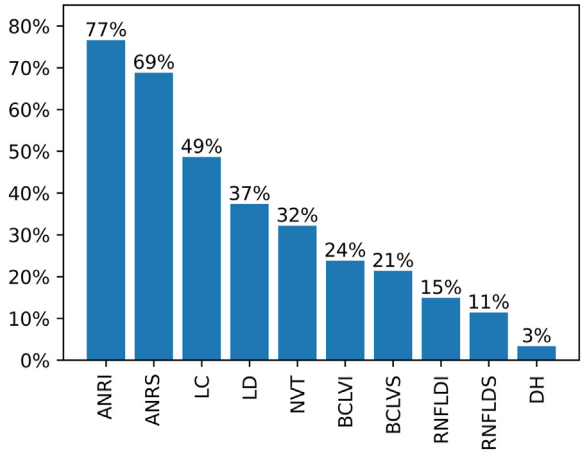

Figure 2. Prevalence of glaucomatous features in images graded as referable glaucoma. ANRI = appearance neuroretinal rim inferiorly; ANRS = appearance neuroretinal rim superiorly; BCLVI = baring of the circumlinear vessel(s) inferiorly; BCLVS = baring of the circumlinear vessel(s) superiorly; DH = disc hemorrhage(s); LC = large cup; LD = laminar dots; NVT = nasalization (nasal displacement) of the vessel trunk; RNFLDI = retinal nerve fiber layer defect inferiorly; RNFLDS = retinal nerve fiber layer defect superiorly.

validated on entirely different data sets,[52] sometimes to unacceptably low levels for screening, thereby probably discouraging initiatives to apply AI for screening on a large scale. Although a poor sensitivity would obviously lead to the undesirable situation of missed cases, a poor specificity would overburden the healthcare system because of too many false-positive referrals and the costs they would incur, especially with a relatively rare disease like glaucoma. In addition, too many false-positive referrals would soon lead to a general loss of trust, and interest, in such screening programs. It is therefore better for population-based screening to sacrifice a little sensitivity in exchange for a high specificity, as long as significantly more cases are detected than with the current design of the health care system. The missed cases would probably be detected in a next screening round, when the disease was likely to have progressed, hopefully still within an asymptomatic range.

We would argue that the disappointing validations of promising AI algorithms might, in part, be explained by the differences in size and composition between the used data sets, caused by, e.g., different definitions of RG, differences in ethnicities, the source of the CFP sets (e.g., clinic based or population based; a form of selection bias), other forms of data selection bias, differences in the prevalence of RG, any presence of comorbidity and differences in the grading process. A strong point of the current study is that we acquired a large, labeled set of CFPs that is multiethnic, and that was obtained on multiple cameras by multiple operators, designed for developing AI for glaucoma screening programs under realistic conditions. It is currently unclear to what extent the performance of AI algorithms, trained on our data set, may drop when validated on other data sets, e.g., in the presence of other coexisting disease.

Another strong point of the current study is the rigorous grader qualification process that we used. This process entailed initial training, followed by testing and selection by means of an examination. In addition, the performance of the graders was monitored throughout the grading process. In many glaucoma AI studies, the "ground truth" labeling and annotation were performed by only a small number of graders, often with limited clinical experience, and without validation of their gradings.

A limitation of our study was that we did not have visual fields available to support or refute the label of RG, which was defined as glaucoma with expected glaucomatous visual field defects (i.e., perimetric glaucoma). One reason for defining RG in such a way was that we wanted to exclude early, notably preperimetric glaucoma, because that would yield an undesirably low specificity. More importantly, we wanted to use the labeled data set to develop AI that would detect those individuals at a high risk of becoming seriously visually impaired. Undetected, probably early glaucoma cases, we reasoned, would likely be detected in follow-up testing. In addition, we think that careful clinical examination, in which the assessment of the ONH is only part of several tests (such as OCT imaging, visual field testing, tonometry, pachymetry, and slit lamp examination of the anterior and posterior segment, including gonioscopy) is required to make a diagnosis. Optic disc assessment by AI of CFPs would only serve as a first, but important and relatively inexpensive, screen for glaucoma detection. In addition, all our graders had to pass the EODAT test, in which all glaucomatous eyes showed reproducible visual field damage.

Another limitation of our study was that we had no OCT imaging of the fundus available for each eye. OCT imaging plays an ever-greater role in the clinical management of glaucoma, both for making a diagnosis and for monitoring progression. OCT imaging might therefore have supported the classification of our CFPs. In addition, the availability of such a combined data set might have served the development of AI for OCT images to screen the population at large for glaucoma. Our aim, however, was to provide a high-quality labeled data set for the development of AI for the low-cost screening for glaucoma. OCT-based AI has certainly shown promise for glaucoma detection,[50,62–70] but screening would arguably be too costly for many countries. We also think that there are too many differences between devices in scan protocols, preprocessing, processing, segmentation, normative data, etc. to allow the development of a universal AI, applicable across all (major) OCT devices for population-based screening programs. Importantly, OCT images seem not to yield higher performance for detecting glaucoma than CFPs. [71] Taken together, we think that OCT-based AI is not yet ready for widespread glaucoma screening, as opposed to its unquestionably important role in the clinic.

A potential limitation of the study was that our CFPs were obtained from a screening program for diabetic retinopathy and therefore did not fully represent the population at large. We would argue, however, that this is an advantage since AI screening for glaucoma is likely to be used quite soon in diabetic retinopathy screening programs because these are already in place and require the same equipment, i.e., (nonmydriatic) fundus cameras. The use of AI for the

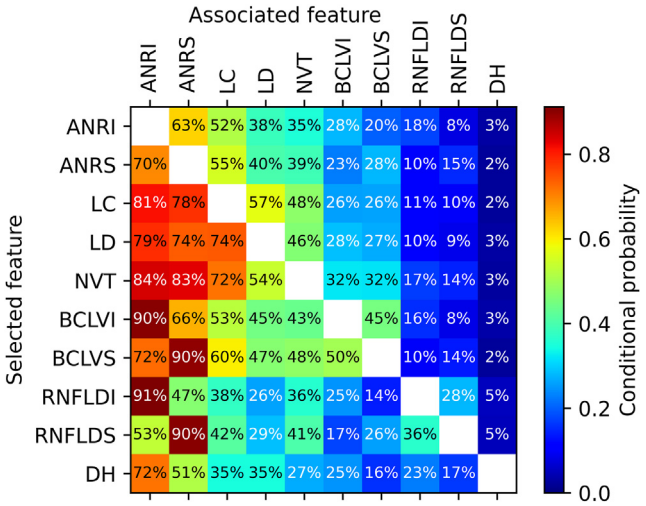

Figure 3. Conditional probabilities of glaucomatous features in referable glaucoma eyes, expressed as the likelihood of an associated feature given a selected feature. ANRI = appearance neuroretinal rim inferiorly; ANRS = appearance neuroretinal rim superiorly; BCLVI = baring of the circumlinear vessel(s) inferiorly; BCLVS = baring of the circumlinear vessel(s) superiorly; DH = disc hemorrhage(s); LC = large cup; LD = laminar dots; NVT = nasalization (nasal displacement) of the vessel trunk; RNFLDI = retinal nerve fiber layer defect inferiorly; RNFLDS = retinal nerve fiber layer defect superiorly.

detection of glaucoma could then be seen as a relatively simple add-on, thereby keeping costs low.

Perhaps another limitation of our study was that we graded single-field (monoscopic) CFPs instead of stereoscopic images. In clinical practice, stereoscopic viewing of the ONH is generally considered as superior to monoscopic assessment. However, when expert graders were put to the test, no difference in grading accuracy was found between their monoscopic and stereoscopic assessment of fundus photographs in identifying glaucoma.[72]

This study provides for the first time, as far as we know, an overview of the typical features of CFPs of eyes with RG in a population-based only screening program. A somewhat similar overview was obtained from a mixed data set (both population-based and clinic-based).[51] Overviews such as these may serve the design of future glaucoma screening programs, because they offer the opportunity to see which of them matter most in RG. In addition, the results of this study may also be used to develop, improve, and refine explainable AI algorithms, especially because we have made our labeled data set publicly available. In a recent challenge (Artificial Intelligence RObust Glaucoma Screening challenge [AIROGS]: https://airogs.grand-challenge.org; the AIROGS train set [a subset of the currently described REGAIS data set], which is publicly available, may be used at no cost under the CC BY-NC-ND 4.0 license for nonprofit use, and can be downloaded from https://zenodo.org/record/5793241#.Yybjci-Qlu0) the winning solution, which was trained on ~ 101 000 images from our REGAIS data set and tested on a separate and secretly kept test set of ~ 11 000 images (also from the REGAIS data set), showed a sensitivity for detecting RG of ~ 85% at a specificity of 95%.

A limitation of our study, which applies to many studies, is that human gradings of the ONH have limited reproducibility and poor inter-rater agreement.[55,73–75] We tried to address that problem by having each CFP graded by 2 independent, experienced, trained, qualified, and periodically monitored graders, and, in case of disagreement between the 2 in their final classification, have the CFP graded by a third grader who had passed the EODAT entry examination with an exceedingly high score of $\geq$ 95% accuracy. Nevertheless, there is a real risk, albeit small, that the first 2 graders both classified a CFP identically, but incorrectly. We suspect that these errors rarely occurred, but we cannot rule them out entirely.

Yet another limitation of our current study was that, although agreement between graders was required for the 3 main classifications (RG, NRG, or UG), or else the final classification was determined by the third grader, no such agreement was required for feature selection. When we designed the study, this was considered to make the grading process too complicated and time consuming. What's more, the main purpose of the study was to provide high-quality labels of the 3 main classifications (namely, RG, NRG, and UG) for the development of AI for population-based screening. In addition, 1 of the authors (H.L.), who graded ~ 10 000 CFPs as a third grader, got the impression that the RG features were carefully and correctly selected in the vast majority of cases. Tighter study designs, calling for stronger agreement between graders on the typical features of RG CFPs, may be required in the future.

In case of RG, the graders could select up to 10 glaucomatous features per CFP. The choice of these features was based on the consensus reached on the most relevant features by the 3 top scorers of the EODAT trial. [55] These features were arbitrary and might be open to discussion. A number of evaluation methods of the ONH have been proposed, the commonest perhaps being the CDR ratio, although it is limited by large measurement variability between and within graders.[73–76] In addition, there is wide variation in the CDR in the normal population.[77–79] More

importantly, the sensitivity of a large CDR goes down with the size of the disc.[79] Other schemes and classification methods have been proposed, some limiting the number of features, others expanding their number, sometimes focusing on several old features and adding new ones.[78,80,81] Over the years, the focus has moved away from the cup to the NRR.[78,80,82] This is meaningful, especially when a relatively small cup (with small CDR) is associated with typically glaucomatous damage to the NRR, e.g., in case of a notch. Therefore, the features we offered to our graders to choose from did not contain CDR as an option. Bayonetting of vessels over the NRR was considered an abnormal NRR and was therefore not explicitly presented as an option. The quite popular so-called ISNT rule, which relates to the width of the NRR and assumes that, in healthy eyes, the Inferior NRR is broader than the Superior rim, followed in width by the Nasal and Temporal rim, respectively; hence ISNT, has been shown to be insufficiently accurate for discriminating between healthy and glaucomatous eyes with high sensitivity and specificity.[83–85] The 3 top scorers of the EODAT trial never used the ISNT rule for classifying the RG eyes. Peripapillary atrophy (PPA) was not presented as an option, because its presence *per se* is not typical of glaucoma, although glaucomatous eyes tend to have larger areas of PPA (both alpha and beta zones) than normal eyes.[86] Beta zone PPA has been shown to be a relatively unimportant feature in predicting glaucoma.[51] Not surprisingly, the top 3 graders of the EODAT trial attributed little weight to PPA in discriminating between healthy and glaucomatous eyes.

While disc hemorrhages often get a lot of attention in training programs, they were observed quite rarely in our RG eyes, which confirms the earlier observation of low sensitivity, with high specificity, for detecting glaucoma.[87]

Conclusion

We have put together a large, labeled data set of CFPs for the development of an AI solution for low-cost population-based screening for glaucoma in multiple ethnic settings. Approximately 90% of the entire set is publicly available and may be used for training under the Creative Commons BY-NC-ND 4.0 license. The entire set was acquired on a wide array of fundus cameras in many locations by multiple operators. The grading was done by a pool of experienced, trained, certified, and periodically monitored graders. In case of RG, the characteristic features were noted. It turned out that the appearance of the NRR (notably notching, thinning, or narrowing) was the most predominant feature. Retinal nerve fiber layer defects were observed in ~ 13% of eyes and disc hemorrhages were a rare feature of RG. Other common features included visible laminar dots, baring of the circumlinear vessels, and nasal displacement of the vessel trunk. The labeled data set will be made publicly available. Our results and data may be of use for developing and refining future screening programs. The estimated sensitivity and specificity were above our target of 80% and 95%, respectively, and the annotated data set should therefore be of sufficient quality to develop AI screening solutions.

Footnotes and Disclosures

Originally received: October 15, 2022.
Final revision: February 12, 2023.
Accepted: March 13, 2023.
Available online: March 17, 2023. Manuscript no. XOPS-D-22-00217R3.

[1] The Rotterdam Eye Hospital, Rotterdam, the Netherlands.

[2] Quantitative Healthcare Analysis (QurAI) Group, Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands.

[3] Department of Biomedical Engineering and Physics, Amsterdam UMC, Amsterdam, the Netherlands.

[4] The Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, the Netherlands.

Presented at the 18th meeting of the Imaging and Morphometry Association for Glaucoma in Europe (IMAGE) from May 18 to 20, 2022 in Porto, Portugal.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

C. de V.: Support — Aeon Astron, Rotterdamse Stichting Blindenbelangen, Stichting Wetenschappelijk Onderzoek Oogziekenhuis (SWOO-Flieringa), Glaucoomfonds, Théa Pharma, Lameris Ootech, Allergan, Carl Zeiss Vision, Hoornvlies Stichting Nederland, Novartis, Revoir Recruitment, Oogfonds, Stichting Blindenhulp, Visu Farma, Stichting voor Ooglijders, Private donations (crowd funding); Grant — Eurostars; Payment — Théa Pharma.

H.G.L.: Support — Aeon Astron, Rotterdamse Stichting Blindenbelangen, Stichting Wetenschappelijk Onderzoek Oogziekenhuis (SWOO-Flieringa, Glaucoomfonds), Théa Pharma, Lameris Ootech, Allergan, Carl Zeiss Vision, Hoornvlies Stichting Nederland, Novartis, Revoir Recruitment, Oogfonds, Stichting Blindenhulp, Visu Farma, Stichting voor Ooglijders, Private donations (crowd funding); Payment — Théa Pharma.

C.I.S.: Support — Aeon Astron, Rotterdamse Stichting Blindenbelangen, Stichting Wetenschappelijk Onderzoek Oogziekenhuis (SWOO-Flieringa), Glaucoomfonds, Théa Pharma, Lameris Ootech, Allergan, Carl Zeiss Vision, Hoornvlies Stichting Nederland, Novartis, Revoir Recruitment, Oogfonds, Stichting Blindenhulp, Visu Farma, Stichting voor Ooglijders, Private donations (crowd funding); Payment — Novartis, Bayern.

K.A.V.: Support — Aeon Astron, Rotterdamse Stichting Blindenbelangen, Stichting Wetenschappelijk Onderzoek Oogziekenhuis (SWOO-Flieringa), Glaucoomfonds, Théa Pharma, Lameris Ootech, Allergan, Carl Zeiss Vision, Hoornvlies Stichting Nederland, Novartis, Revoir Recruitment, Oogfonds, Stichting Blindenhulp, Visu Farma, Stichting voor Ooglijders, Private donations (crowd funding).

Supported by unrestricted grants from Stichting Blindenbelangen, Glaucoomfonds, Stichting Wetenschappelijk Onderzoek Oogziekenhuis (SWOO—Flieringa), Hoornvlies Stichting Nederland, Stichting Blindenhulp, Stichting voor Ooglijders and by (unrestricted) donations from Aeon Astron, Théa Pharma, Laméris Ootech, Allergan, Carl Zeiss Vision, Visu Farma, Revoir Recruitment and Novartis as well as from many private donors (crowd funding).

HUMAN SUBJECTS: Human subjects were included in this study. The entire project was approved by the Institutional Review Board of the Rotterdam Eye Hospital. The study adhered to the Declaration of Helsinki. The patients had provided their informed consent.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Lemij, Vermeer.

Data collection: Lemij, Vermeer.

Analysis and interpretation: Lemij, de Vente, Sánchez, Vermeer.

Obtained funding: Lemij, de Vente, Sánchez, Vermeer

Overall responsibility: Lemij, de Vente, Sánchez, Vermeer.

Abbreviations and Acronyms:
AI = artificial intelligence; **CDR** = cup-to-disc ratio; **CFP** = color fundus photograph; **EODAT** = European Optic Disc Assessment Trial; **NRG** = no referable glaucoma; **NRR** = neuroretinal rim; **ONH** = optic nerve head; **PPA** = Peripapillary atrophy; **REGAIS** = Rotterdam EyePACS Glaucoma AI Screening data set; **RG** = Referable glaucoma; **UG** = Ungradable.

Keywords:
Artificial intelligence, Clinical features, color fundus photographs, glaucoma screening, labeled data set.

Correspondence:
Hans G. Lemij, The Rotterdam Eye Hospital, Rotterdam, the Netherlands. E-mail: h.lemij@oogziekenhuis.nl.

References

1. Burton MJ, Ramke J, Marques AP, et al. The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. *Lancet Global Health*. 2021;9:e489−e551.
2. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5:e1221−e1234.
3. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90:262−267.
4. Tham YC, 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;121:2081−2090.
5. Turano KA, Rubin GS, Quigley HA. Mobility performance in glaucoma. *Invest Ophthalmol Vis Sci*. 1999;40:2803−2809.
6. Ramrattan RS, Wolfs RC, Panda-Jonas S, et al. Prevalence and causes of visual field loss in the elderly and associations with impairment in daily functioning: the Rotterdam Study. *Arch Ophthalmol*. 2001;119:1788−1794.
7. Ramulu PY, Mihailovic A, West SK, et al. Predictors of falls per step and falls per year at and away from home in Glaucoma. *Am J Ophthalmol*. 2019;200:169−178.
8. Tielsch JM, Sommer A, Katz J, et al. Racial variations in the prevalence of primary open-angle glaucoma. The Baltimore Eye Survey. *Jama*. 1991;266:369−374.
9. Klein BE, Klein R, Sponsel WE, et al. Prevalence of glaucoma. The Beaver Dam Eye Study. *Ophthalmology*. 1992;99:1499−1504.
10. Leske MC, Connell AM, Schachat AP, Hyman L. The Barbados Eye Study. Prevalence of open angle glaucoma. *Arch Ophthalmol*. 1994;112:821−829.
11. 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.
12. Wensor MD, McCarty CA, Stanislavsky YL, et al. The prevalence of glaucoma in the Melbourne Visual Impairment Project. *Ophthalmology*. 1998;105:733−739.
13. Coffey M, Reidy A, Wormald R, et al. Prevalence of glaucoma in the west of Ireland. *Br J Ophthalmol*. 1993;77:17−21.
14. Quigley HA, West SK, Rodriguez J, et al. The prevalence of glaucoma in a population-based study of Hispanic subjects: Proyecto VER. *Arch Ophthalmol*. 2001;119:1819−1826.
15. Mason RP, Kosoko O, Wilson MR, et al. National survey of the prevalence and risk factors of glaucoma in St. Lucia, West Indies. Part I. Prevalence findings. *Ophthalmology*. 1989;96:1363−1368.
16. Wolfs RC, Borger PH, Ramrattan RS, et al. Changing views on open-angle glaucoma: definitions and prevalences–the Rotterdam Study. *Invest Ophthalmol Vis Sci*. 2000;41:3309−3321.
17. Bonomi L, Marchini G, Marraffa M, et al. Prevalence of glaucoma and intraocular pressure distribution in a defined population. The Egna-Neumarkt Study. *Ophthalmology*. 1998;105:209−215.
18. Foster PJ, Baasanhu J, Alsbirk PH, et al. Glaucoma in Mongolia. A population-based survey in Hövsgöl province, northern Mongolia. *Arch Ophthalmol*. 1996;114:1235−1241.
19. Foster PJ, Oen FT, Machin D, et al. The prevalence of glaucoma in Chinese residents of Singapore: a cross-sectional population survey of the Tanjong Pagar district. *Arch Ophthalmol*. 2000;118:1105−1111.
20. Salmon JF, Mermoud A, Ivey A, et al. The prevalence of primary angle closure glaucoma and open angle glaucoma in Mamre, western Cape, South Africa. *Arch Ophthalmol*. 1993;111:1263−1269.
21. Wong EY, Keeffe JE, Rait JL, et al. Detection of undiagnosed glaucoma by eye health professionals. *Ophthalmology*. 2004;111:1508−1514.
22. Peters D, Bengtsson B, Heijl A. Lifetime risk of blindness in open-angle glaucoma. *Am J Ophthalmol*. 2013;156:724−730.
23. Ang GS, Eke T. Lifetime visual prognosis for patients with primary open-angle glaucoma. *Eye (Lond)*. 2007;21:604−668.
24. Ernest PJ, Busch MJ, Webers CA, et al. Prevalence of end-of-life visual impairment in patients followed for glaucoma. *Acta Ophthalmol*. 2013;91:738−743.
25. Goh YW, Ang GS, Azuara-Blanco A. Lifetime visual prognosis of patients with glaucoma. *Clin Exp Ophthalmol*. 2011;39:766−770.
26. Forsman E, Kivelä T, Vesti E. Lifetime visual disability in open-angle glaucoma and ocular hypertension. *J Glaucoma*. 2007;16:313−319.
27. Saunders LJ, Russell RA, Kirwan JF, et al. Examining visual field loss in patients in glaucoma clinics during their predicted remaining lifetime. *Invest Ophthalmol Vis Sci*. 2014;55:102−109.
28. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *Jama*. 2014;311:1901−1911.
29. Quigley HA. Glaucoma. *Lancet*. 2011;377:1367−1377.
30. Dielemans I, de Jong PT, Stolk R, et al. Primary open-angle glaucoma, intraocular pressure, and diabetes mellitus in the general elderly population. The Rotterdam Study. *Ophthalmology*. 1996;103:1271−1275.
31. Friedman DS, Wolfs RC, O'Colmain BJ, et al. Prevalence of open-angle glaucoma among adults in the United States. *Arch Ophthalmol*. 2004;122:532−538.

32. Quigley HA. Number of people with glaucoma worldwide. *Br J Ophthalmol*. 1996;80:389–393.
33. Ramakrishnan R, Nirmalan PK, Krishnadas R, et al. Glaucoma in a rural population of southern India: the Aravind comprehensive eye survey. *Ophthalmology*. 2003;110:1484–1490.
34. Leite MT, Sakata LM, Medeiros FA. Managing glaucoma in developing countries. *Arq Bras Oftalmol*. 2011;74:83–84.
35. Rotchford AP, Kirwan JF, Muller MA, et al. Temba glaucoma study: a population-based cross-sectional survey in urban South Africa. *Ophthalmology*. 2003;110:376–382.
36. Budenz DL, Barton K, Whiteside-de Vos J, et al. Prevalence of glaucoma in an urban West African population: the Tema Eye Survey. *JAMA Ophthalmol*. 2013;131:651–658.
37. Aspberg J, Heijl A, Bengtsson B. Screening for open-angle glaucoma and its effect on blindness. *Am J Ophthalmol*. 2021;228:106–116.
38. Heijl A, Leske MC, Bengtsson B, et al. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol*. 2002;120:1268–1279.
39. Lee PP, Walt JG, Doyle JJ, et al. A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma. *Arch Ophthalmol*. 2006;124:12–19.
40. Mursch-Edlmayr AS, Ng WS, Diniz-Filho A, et al. Artificial intelligence algorithms to diagnose glaucoma and detect glaucoma progression: translation to clinical practice. *Transl Vis Sci Technol*. 2020;9:55.
41. *Glaucoma Screening*. Amsterdam: Kugler Publications; 2008.
42. John D, Parikh R. Cost-effectiveness of community screening for glaucoma in rural India: a decision analytical model. *Public Health*. 2018;155:142–151.
43. Tang J, Liang Y, O'Neill C, et al. Cost-effectiveness and cost-utility of population-based glaucoma screening in China: a decision-analytic Markov model. *Lancet Glob Health*. 2019;7:e968–e978.
44. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol*. 2020;9:42.
45. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8:14665.
46. 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;125:1199–1206.
47. Ahn JM, Kim S, Ahn KS, et al. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS One*. 2018;13:e0207982.
48. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353–1360.
49. Li F, Yan L, Wang Y, et al. Deep learning-based automated detection of glaucomatous optic neuropathy on color fundus photographs. *Graefes Arch Clin Exp Ophthalmol*. 2020;258:851–867.
50. Girard MJA, Schmetterer L. Artificial intelligence and deep learning in glaucoma: current state and future prospects. *Prog Brain Res*. 2020;257:37–64.
51. Phene S, Dunn RC, Hammel N, et al. Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs. *Ophthalmology*. 2019;126:1627–1639.
52. Christopher M, Belghith A, Bowd C, et al. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Sci Rep*. 2018;8:16685.
53. Rogers TW, Jaccard N, Carbonaro F, et al. Evaluation of an AI system for the automated detection of glaucoma from stereoscopic optic disc photographs: the European Optic Disc Assessment Study. *Eye (Lond)*. 2019;33:1791–1797.
54. Devalla SK, Liang Z, Pham TH, et al. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol*. 2020;104:301–311.
55. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology*. 2022;129:e14–e32.
56. Reus NJ, Lemij HG, Garway-Heath DF, et al. Clinical assessment of stereoscopic optic disc photographs for glaucoma: the European Optic Disc Assessment Trial. *Ophthalmology*. 2010;117:717–723.
57. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol*. 2009;3(3):509–516.
58. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43:543–549.
59. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61:29–48.
60. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167–175.
61. Tan NYQ, Friedman DS, Stalmans I, et al. Glaucoma screening: where are we and where do we need to go? *Curr Opin Ophthalmol*. 2020;31:91–100.
62. Huang ML, Chen HY. Development and comparison of automated classifiers for glaucoma diagnosis using Stratus optical coherence tomography. *Invest Ophthalmol Vis Sci*. 2005;46:4121–4129.
63. Burgansky-Eliash Z, Wollstein G, Chu T, et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Invest Ophthalmol Vis Sci*. 2005;46:4147–4152.
64. An G, Omodaka K, Tsuda S, et al. Comparison of machine-learning classification models for glaucoma management. *J Healthc Eng*. 2018;2018:6874765.
65. Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. *PLoS One*. 2017;12:e0177726.
66. Barella KA, Costa VP, Gonçalves Vidotti V, et al. Glaucoma diagnostic accuracy of machine learning classifiers using retinal nerve fiber layer and optic nerve data from SD-OCT. *J Ophthalmol*. 2013;2013:789129.
67. Christopher M, Belghith A, Weinreb RN, et al. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. *Invest Ophthalmol Vis Sci*. 2018;59:2748–2756.
68. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26:1086–1094.
69. Devalla SK, Chin KS, Mari JM, et al. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. *Invest Ophthalmol Vis Sci*. 2018;59:63–74.
70. Maetschke S, Antony B, Ishikawa H, et al. A feature agnostic approach for glaucoma detection in OCT volumes. *PLoS One*. 2019;14:e0219126.

71. Wu JH, Nishida T, Weinreb RN, Lin JW. Performances of machine learning in detecting glaucoma using fundus and retinal optical coherence tomography images: a meta-analysis. *Am J Ophthalmol*. 2022;237:1−12.
72. Chan HH, Ong DN, Kong YX, et al. Glaucomatous optic neuropathy evaluation (GONE) project: the effect of monoscopic versus stereoscopic viewing conditions on optic nerve evaluation. *Am J Ophthalmol*. 2014;157:936−944.
73. Varma R, Steinmann WC, Scott IU. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*. 1992;99:215−221.
74. Abrams LS, Scott IU, Spaeth GL, et al. Agreement among optometrists, ophthalmologists, and residents in evaluating the optic disc for glaucoma. *Ophthalmology*. 1994;101:1662−1667.
75. Jampel HD, Friedman D, Quigley H, et al. Agreement among glaucoma specialists in assessing progressive disc changes from photographs in open-angle glaucoma patients. *Am J Ophthalmol*. 2009;147:39−44.e1.
76. Lichter PR. Variability of expert observers in evaluating the optic disc. *Trans Am Ophthalmol Soc*. 1976;74:532−572.
77. Wolfs RC, Ramrattan RS, Hofman A, de Jong PT. Cup-to-disc ratio: ophthalmoscopy versus automated measurement in a general population: the Rotterdam study. *Ophthalmology*. 1999;106:1597−1601.
78. Jonas JB, Gusek GC, Naumann GO. Optic disc, cup and neuroretinal rim size, configuration and correlations in normal eyes. *Invest Ophthalmol Vis Sci*. 1988;29:1151−1158.
79. Garway-Heath DF, Ruben ST, Viswanathan A, Hitchings RA. Vertical cup/disc ratio in relation to optic disc size: its value in the assessment of the glaucoma suspect. *Br J Ophthalmol*. 1998;82:1118−1124.
80. Spaeth GL, Henderer J, Liu C, et al. The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma. *Trans Am Ophthalmol Soc*. 2002;100:181−185.
81. Read RM, Spaeth GL. The practical clinical appraisal of the optic disc in glaucoma: the natural history of cup progression and some specific disc-field correlations. *Trans Am Acad Ophthalmol Otolaryngol*. 1974;78: Op255−Op274.
82. Jonas JB, Gusek GC, Naumann GO. Optic disc morphometry in chronic primary open-angle glaucoma. II. Correlation of the intrapapillary morphometric data to visual field indices. *Graefes Arch Clin Exp Ophthalmol*. 1988;226:531−538.
83. Maupin E, Baudin F, Arnould L, et al. Accuracy of the ISNT rule and its variants for differentiating glaucomatous from normal eyes in a population-based study. *Br J Ophthalmol*. 2020;104:1412−1417.
84. Harizman N, Oliveira C, Chiang A, et al. The ISNT rule and differentiation of normal from glaucomatous eyes. *Arch Ophthalmol*. 2006;124:1579−1583.
85. Law SK, Kornmann HL, Nilforushan N, et al. Evaluation of the "IS" rule to differentiate glaucomatous eyes from normal. *J Glaucoma*. 2016;25:27−32.
86. Jonas JB, Nguyen XN, Gusek GC, Naumann GO. Parapapillary chorioretinal atrophy in normal and glaucoma eyes. I. Morphometric data. *Invest Ophthalmol Vis Sci*. 1989;30: 908−918.
87. Jonas JB, Nguyen NX, Naumann GO. Non-quantitative morphologic features in normal and glaucomatous optic discs. *Acta Ophthalmol (Copenh)*. 1989;67:361−366.