Title

Deep Learning Identifies High-Quality Fundus Photographs and Increases Accuracy in Automated Primary Open Angle Glaucoma Detection.

Journal

Translational Vision Science & Technology, 13(1)

ISSN

2164-2591

Authors

Chuter, Benton

Huynh, Justin

Bowd, Christopher

et al.

Publication Date

2024-01-29

DOI

10.1167/tvst.13.1.23

Peer reviewed

tvst

Artificial Intelligence

Deep Learning Identifies High-Quality Fundus Photographs and Increases Accuracy in Automated Primary Open Angle Glaucoma Detection

Benton Chuter[1], Justin Huynh[1], Christopher Bowd[1], Evan Walker[1], Jasmin Rezapour[1,2], Nicole Brye[1], Akram Belghith[1], Massimo A. Fazio[3], Christopher A. Girkin[3], Gustavo De Moraes[4], Jeffrey M. Liebmann[4], Robert N. Weinreb[1], Linda M. Zangwill[1], and Mark Christopher[1]

[1] Hamilton Glaucoma Center, Shiley Eye Institute, Viterbi Family Department of Ophthalmology, UC San Diego, La Jolla, California, United States
[2] Department of Ophthalmology, University Medical Center Mainz, Germany
[3] School of Medicine, Callahan Eye Hospital, University of Alabama-Birmingham, Birmingham, Alabama, United States
[4] Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Department of Ophthalmology, Columbia University Medical Center, New York, New York, United States

Correspondence: Mark Christopher, Hamilton Glaucoma Center, Shiley Eye Institute, Viterbi Family Department of Ophthalmology, UC San Diego, La Jolla, CA 92037, United States. e-mail: mac157@health.ucsd.edu

Received: April 10, 2023
Accepted: December 26, 2023
Published: January 29, 2024

Keywords: fundus; glaucoma; deep learning; quality

Citation: Chuter B, Huynh J, Bowd C, Walker E, Rezapour J, Brye N, Belghith A, Fazio MA, Girkin CA, De Moraes G, Liebmann JM, Weinreb RN, Zangwill LM, Christopher M. Deep learning identifies high-quality fundus photographs and increases accuracy in automated primary open angle glaucoma detection. Transl Vis Sci Technol. 2024;13(1):23, https://doi.org/10.1167/tvst.13.1.23

Purpose: To develop and evaluate a deep learning (DL) model to assess fundus photograph quality, and quantitatively measure its impact on automated POAG detection in independent study populations.

Methods: Image quality ground truth was determined by manual review of 2815 fundus photographs of healthy and POAG eyes from the Diagnostic Innovations in Glaucoma Study and African Descent and Glaucoma Evaluation Study (DIGS/ADAGES), as well as 11,350 from the Ocular Hypertension Treatment Study (OHTS). Human experts assessed a photograph as high quality if of sufficient quality to determine POAG status and poor quality if not. A DL quality model was trained on photographs from DIGS/ADAGES and tested on OHTS. The effect of DL quality assessment on DL POAG detection was measured using area under the receiver operating characteristic (AUROC).

Results: The DL quality model yielded an AUROC of 0.97 for differentiating between high- and low-quality photographs; qualitative human review affirmed high model performance. Diagnostic accuracy of the DL POAG model was significantly greater ($P < 0.001$) in good (AUROC, 0.87; 95% CI, 0.80–0.92) compared with poor quality photographs (AUROC, 0.77; 95% CI, 0.67–0.88).

Conclusions: The DL quality model was able to accurately assess fundus photograph quality. Using automated quality assessment to filter out low-quality photographs increased the accuracy of a DL POAG detection model.

Translational Relevance: Incorporating DL quality assessment into automated review of fundus photographs can help to decrease the burden of manual review and improve accuracy for automated DL POAG detection.

Introduction

Retinal fundus imaging plays a crucial role in the diagnosis and management of glaucoma, a leading cause of permanent visual impairment and blindness.[1–5] However, poor quality can decrease the usefulness of photographs in the clinical management of glaucoma. Differences in cameras and conditions as well as differences between patients and technicians can lead to variable fundus photograph quality. Common quality issues include incorrect brightness, diminished

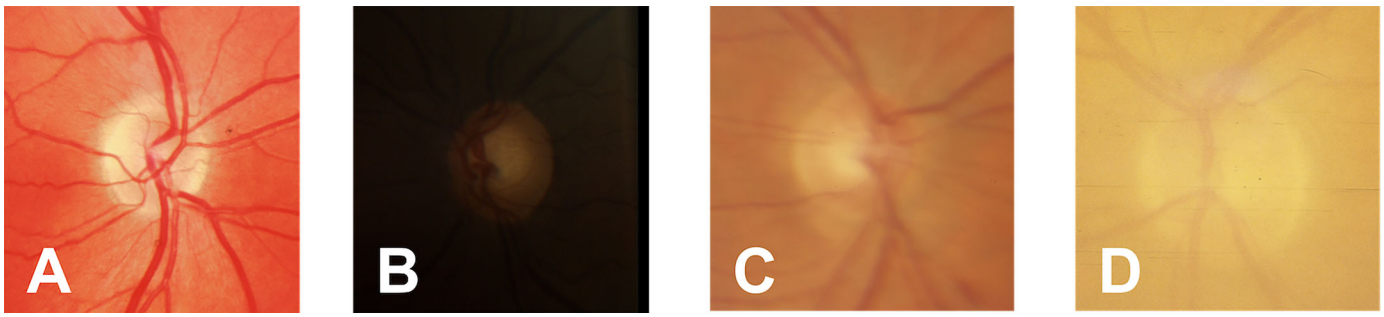

Figure 1. Sample images from training dataset: high-quality image (**A**) and low-quality images (**B**, **C**, **D**). Quality ratings assigned by human reviewers to train the quality model.

contrast, improper focus, blurring, poor image centering, and noise (Fig. 1).[6] resulting in images that cannot be accurately graded for POAG.[6] These issues remain prevalent, and a substantial portion of fundus photographs captured by trained technicians are only partially gradable or not gradable at all.[8,9]

Poor quality images can disrupt both manual and automated assessment of fundus photographs, with implications for clinical and research applications.[5,7–10] To help ensure capture of sufficiently good quality photographs, redundant images may be taken. However, this process requires additional technician and clinician time for image capture and review, potentially making the cost and labor required for telemedicine programs prohibitively high. Automated quality assessment models could provide real-time feedback and even help to direct image capture to ensure that gradable images are collected. In addition to clinical and telemedicine settings, automated quality assessment could prove to be a useful tool in research settings. Large, well-annotated image datasets have become a critical resource across ophthalmology to build and evaluate AI-based diagnostic and decision support systems. Compiling these large datasets requires substantial resources, including manually identifying and removing low-quality images.[11] In addition to the increased time and cost, low-quality images may also decrease the performance of automated review. Deep learning (DL)–based systems to review fundus photographs have already been deployed in clinical settings,[12] and a number of methods for photograph-based automated glaucoma detection have been described recently.[2,3,13] Given the increasing role automated review of fundus photographs will have in both research and clinical settings, standardized automated identification of quality issues in ophthalmic images is critical to improve clinical and research workflows by ensuring that good quality images are available for review.

Previous work has described automated approaches for fundus image quality assessment.[14–17] These approaches have achieved high accuracy in identifying images with serious quality issues. However, these approaches are typically designed to address general image quality issues in a disease-agnostic way. The goal of this work is to develop and evaluate a DL approach for automated quality assessment specifically in the context of review of fundus photographs for POAG. Our approach used independent, diverse datasets collected at more than 30 clinical centers and quantified the impact of automated quality assessment on the performance of a previously described DL POAG detection model.[2,3] For these analyses, fundus photograph quality was based on the ability of experts to review them for the presence of POAG.

Methods

Data Collection

Study participants were chosen from longitudinal studies investigating optic nerve structure and function in glaucoma as well as a randomized clinical trial evaluating the safety and efficacy of topical ocular hypotensive medication in POAG onset in hypertensive eyes. More specifically, this study used fundus photographs from the Diagnostic Innovations in Glaucoma Study (DIGS, clinicaltrials.gov identifier: NCT00221897), African Descent and Glaucoma Evaluation Study (ADAGES, clinicaltrials.gov identifier: NCT00221923), and Ocular Hypertension Treatment Study (OHTS, clinicaltrials.gov identifier: NCT00000125 phases I and II).[18] Data from these studies were used to create two datasets: (1) DIGS/ADAGES used to train the quality model and perform initial qualitative evaluation and (2) OHTS used as an independent, external dataset for additional quantitative evaluation.[19] For all studies, recruitment

and methodology were approved by each institution's institutional review board and adhered to the Declaration of Helsinki and the Health Insurance Portability and Accountability Act. As such, all participants provided informed consent at recruitment. Details of these studies' methods have been extensively described previously, but relevant aspects are briefly described in this article.[18,19]

The DIGS and ADAGES studies are a collaborative initiative between the UCSD Hamilton Glaucoma Center, University of Alabama at Birmingham Department of Ophthalmology (Birmingham, AL), and the Columbia University Medical Center Edward S. Harkness Eye Institute (New York, NY). The DIGS/ADAGES studies collected stereo fundus photographs and visual field (VF) testing semiannually as a part of their longitudinal design. Fundus photographs were taken via film as simultaneous stereoscopic optic nerve head (ONH) photographs. VF testing was performed using the Humphrey Field Analyzer II (Carl Zeiss Meditec, Dubin, CA) with a standard 24-2 testing pattern and the Swedish Interactive Thresholding Algorithm. VF tests exceeding 33% for fixation losses, false-negative (FN) errors, or false-positive (FP) errors were discarded. The mean deviation from VF testing closest in time to image capture, not surpassing 1 year, was calculated to estimate VF function at the time of imaging for all ONH images. For the images used in this study, VFs were graded for quality in accordance with protocols by the UCSD Visual Field Assessment Center.

OHTS began as a randomized clinical trial in 1994, and the associated imaging and data were used to provide an independent, external dataset to further evaluate our DL models. Written informed consent was obtained from all participants at enrollment at each of the 33 study centers. For inclusion in this work, the requirement for institutional review board approval was waived because only deidentified data were used, following the Standards for Reporting of Diagnostic Accuracy reporting guidelines. We recruited 1636 patients with ocular hypertension from 33 clinical sites to be evaluated biannually with Humphrey 30-2 VF testing and annually for stereoscopic optic nerve ONH photography. Photographs taken during the OHTS randomized clinical phase I trial from 1994 to 2002 and the longitudinal follow-up OHTS phase II trial from 2002 to 2009 were used in this study.

Quality Grading

All photographs from DIGS/ADAGES and OHTS datasets were preprocessed using an automated segmentation model to localize and crop a square image centered on the ONH, where the side length of each crop was approximately equal to two times the optic disc diameter. A standard window size was chosen to give the expert graders and models a consistent view of the ONH region. The 2× disc diameter size was selected because it is the largest field of view available across images in all datasets, given that they were captured using a variety different camera systems, settings, and fields of view. Although this view may not capture diagnostic information further from the disc, our previous work has shown high accuracy in POAG detection using this window size.[2,3] Details of this preprocessing have been described previously.[2] The cropped fundus images were then resized to 224 × 224 pixels, and an expert manually reviewed each cropped image to confirm correct ONH centering. The downsampled size of 224 × 224 was chosen to match the expected input size of the ResNet50 models that we previously used for POAG with high accuracy. Our own empirical testing does not show an improvement in performance when using a larger 512 × 512 resolution. An expert manually reviewed each cropped image to confirm correct ONH centering. Cropped ONH images were subsequently used for quality grading, as well as for input to all subsequent DL models. For these datasets, simultaneous stereo photos were available in some cases, sequential stereo photos (i.e., two photos captured using a monocular fundus camera in succession to produce a pseudo-stereo pair) in other cases, and monocular fundus photos in the remainder. In the analyses described here, stereo pairs were separated and treated as monocular fundus photos so that all data could be included.

These ONH images were reviewed for quality by two independent expert reviewers (C.B. and J.R.) at the UCSD Optic Disc Reading Center. For this work, image quality was defined in terms of gradeability for POAG. That is, an image was annotated as high quality or gradable it had sufficient quality to determine the presence or absence of POAG according to the expert reviewers. A label of low quality was assigned to images where quality issues prevented reviewers from being able to grade the images for POAG with confidence. Image quality ground truth for all selected images from both DIGS/ADAGES was determined by consensus between two independent reviewers (Fig. 1).

POAG Grading

In DIGS/ADAGES, two independent, masked graders reviewed film stereophotographs using a stereoscopic viewer for the presence of POAG. For grader disagreement, a third experienced grader adjudicated. In preparation for analysis, photograph

films were digitally converted by scanning 35-mm film slides and storing them in high resolution (~2200 × ~1500 pixels) with a TIFF format. Stereo image pairs were divided into separate images of the ONH. The combined dataset comprised 7411 stereo pairs split into 14,822 individual ONH images, captured from 4363 eyes of 2329 participants. For this study, the data was evaluated cross-sectionally with a binary label (healthy vs. POAG) assigned at the image level.

In OHTS, masked readers at the independent Optic Disc Reading Center and Visual Field Reading Center independently assessed optic disc photographs and VFs.[20] Although all participants were required to begin the study with normal-appearing ONHs and VFs as determined by the OHTS Optic Disc Reading Center and Visual Field Reading Center, if two consecutive sets of photographs or three consecutive sets of VFs demonstrated change from the baseline, the case was reviewed by the three glaucoma specialist OHTS Endpoint Committee members. Final POAG status was determined by a three-member Endpoint Committee of glaucoma experts who reviewed the photographs and VFs alongside medical history to identify POAG, the primary endpoint, or other potential pathophysiologies such as ischemic optic neuropathy. However, unanimity was required for label assignment, often requiring multiple consensus grading sessions.[20] By design, the criteria for the development of POAG was designed for high specificity to correctly identify individuals without glaucoma.

DL Quality Model

The clinical and demographic information of the study populations used in the training, validation and independent test is presented in Table 1. The DIGS/ADAGES dataset was used for model training, validation, and qualitative evaluation of DL model performance. The OHTS dataset was partitioned in OHTS Test Set 1 and OHTS Test Set 2. These were used as independent test datasets that had no overlap with the training data (or each other) and were collected as part of a separate study. These independent test sets were selected to help provide a better estimate of model generalizability. OHTS Test Set 1 was used to evaluate the accuracy of the DL quality model in distinguishing between good and poor quality images. OHTS Test Set 2 was used to measure the impact of DL quality predictions on DL POAG detection. The DIGS/ADAGES dataset was partitioned into training and validation sets using a 90%–10% split. Partitioning was done by patient and there was no overlap in the participants included between the datasets. The training set was used to train model weights, whereas the validation set was used for model selection (models were evaluated periodically on the validation during training and the highest performing model was selected for further evaluation on the test sets). The model was trained to distinguish between high-quality (gradable) and low-quality (ungradable) fundus photographs using the expert quality grades described elsewhere in this article as ground truth. The DL quality model produced a quantitative output indicating a likelihood of low image quality, with output near 0.0 indicating a low likelihood of low quality and output near 1.0 indicating a high likelihood of low quality. Similar to our previously described DL POAG model, we adopted a ResNet50 architecture and model weights were initialized based on training on a general image dataset (ImageNet).[21] We also evaluated additional architectures (Xception[22] and InceptionResNetV2[23]) for use in our DL quality model, but they achieved comparable or slightly worse results, although the differences were not statistically significant (results not shown). During training, data augmentation was also used to increase fundus photo variation seen by the model. Specifically, random horizontal flipping (to simulate other eye orientation), translation, and small image rotations and rescalings were applied to the training images.

Model Evaluation

Qualitative model evaluation was also performed using manual review of DL model quality assessment. The DL quality model outputs a quantitative score between 0.0 and 1.0. Fundus photographs from the DIGS/ADAGES dataset were binned into six quality categories based on quantitative DL quality predictions. Three experts then reviewed 100 randomly sampled images within each category to assess the level of quality associated with each category, with disagreements resolved by the third grader. This review was also used to empirically determine an appropriate quality threshold to distinguish between high- and low-quality images.

The DL quality model was quantitatively evaluated based on area under the receiver operating characteristic curve (AUROC) using the independent OHTS dataset. AUROC was computed using a bootstrapping approach to account for multiple images from the same patients and eyes.[24] Fundus images from OHTS used for this AUROC analysis were then binned into six quality categories based on the quantitative DL quality predictions and assessed by expert graders for quality, in the manner described previously with the DIGS/ADAGES images, for additional empirical evaluation.

Table 1. Clinical and Demographic Characteristics of the Training, Validation, and Independent Test Sets

	DIGS/ADAGES		OHTS	
Characteristic	(A) Training Set	(B) Validation Set	(C) Test Set 1	(D) Test Set 2
No. of participants	888	108	332	304
No. of eyes	1002	117	585	608
No. of images	2520	295	11350	12278
VF mean deviation (95% CI), dB	−3.13 (−3.53 to −2.74)	−4.16 (−5.60 to −2.73)	−0.34 (−0.40 to −0.28)	−0.30 (−0.48 to −0.12)
Mean age (95% CI), years	65.01 (64.22 to 65.81)	63.73 (61.40 to 66.06)	62.21 (61.95 to 62.48)	61.6 (60.5 to 62.7)
Sex (%)				
Female	519 (58.4%)	61 (56.5%)	163 (49.1%)	138 (45.4%)
Male	367 (41.3%)	47 (43.5%)	163 (49.1%)	166 (54.6%)
NA	2 (0.2%)	0 (0.0%)	6 (1.8%)	0 (0.0%)
Race (%)				
American Indian/Alaska Native	2 (0.2%)	0 (0.0%)	1 (0.3%)	0 (0.0%)
Asian	28 (3.2%)	3 (2.8%)	2 (0.6%)	0 (0.0%)
Black or African American	304 (34.2%)	47 (43.5%)	92 (27.7%)	81 (26.6%)
Pacific Islander	1 (0.1%)	0 (0.0%)	15 (4.5%)	0 (0.0%)
Unknown or not reported	29 (3.3%)	3 (2.8%)	9 (2.7%)	0 (0.0%)
White	524 (59.0%)	55 (50.9%)	213 (64.2%)	223 (73.4%)
Developed a POAG end point by VF or photograph				
No. of participants	399 (15.9%)	51 (17.3%)	86 (0.8%)	53 (0.5%)
No. of eyes	457 (18.2%)	54 (18.4%)	136 (1.2%)	72 (0.6%)
No. of images	1173 (46.6%)	137 (46.5%)	3194 (28.2%)	801 (6.6%)
Not known to develop glaucoma				
No. of participants	489 (19.5%)	57 (19.4%)	246 (2.2%)	251 (2.1%)
No. of eyes	545 (21.7)	63 (21.4%)	449 (4%)	536 (4.4%)
No. of high-quality images	2290 (90.9%)	269 (91.2%)	11265 (99.3%)	11729 (95.6%)
No. of low-quality images	230 (9.2%)	26 (8.9%)	85 (0.8%)	549 (4.5%)

DL model visualization techniques were also used to assess the DL quality model. Gradient weighted class activation mapping (Grad-CAM) is a technique to identify image regions that drive model predictions.[25] For a given DL model and input image, Grad-CAM outputs a coarse localization map highlighting regions of the input image most influential to the model's prediction, in this case, low or high quality.

In addition, a previously published DL POAG detection model developed on the DIGS/ADAGES dataset was used to measure the impact of DL predicted image quality on automated POAG detection.[2,3] The DL POAG detection model was applied to the OHTS dataset and model performance was assessed for both high and low-quality subsets of the data.

Results

DL Quality Model Performance

Expert review of the fundus images in each dataset resulted in good and poor quality labels assigned to each image in the DIGS/ADAGES training set (90.9% good, 9.1% poor), OHTS Test Set 1 (99.2% good, 0.8% poor), and OHTS Test Set 2 (95.5% good, 4.5% poor). Because these datasets have previously been vetted for quality, poor quality images were relatively rare in these

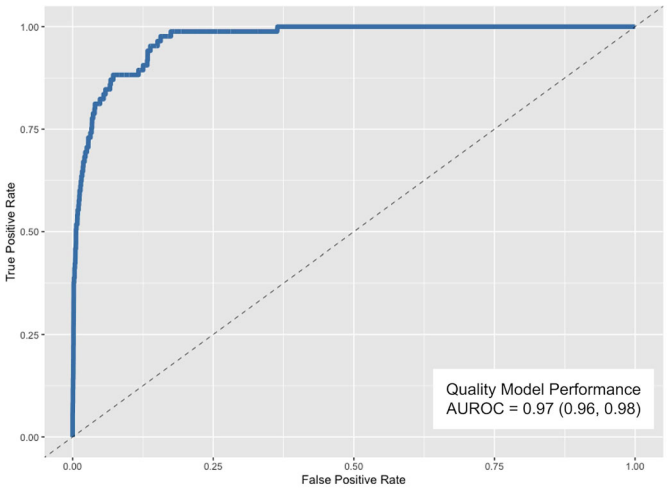

Figure 2. Performance (AUROC) of DL quality model on detection of high- and low-quality photographs.

datasets. The DL quality model achieved an AUROC of 0.97 (95% CI, 0.96–0.98) (Fig. 2) for detecting poor quality images on the OHTS Test Set 1. Based on the expert qualitative review of DL quality predictions and Youden's index, a quality threshold (0.10) was selected to assign binary gradable and ungradable scores to images.[26] Using this threshold, the model achieved accuracy of 0.97 (95% CI, 0.96–0.97), sensitivity of 0.73 (95% CI, 0.62–0.83) specificity of 0.97 (95% CI, 0.96–0.98), and precision of 0.16 (95% CI, 0.12–0.20) for detecting gradable and ungradable photographs on OHTS Test Set 1. Figure 2 presents the ROC curve illustrating DL quality model performance with the OHTS dataset.

Impact of Automated Quality Assessment on Automated POAG Detection

The diagnostic accuracy of the DL model for POAG detection performed better on good quality compared with poor quality OHTS photographs, as determined by the DL quality model. On OHTS Test Set 2, it achieved an AUROC of 0.87 (95% CI, 0.80–0.92) on good quality and 0.77 (95% CI, 0.67–0.88) on poor quality images in detecting POAG (Fig. 2, Table 2) and the difference was significant ($P < 0.001$). Sensitivity at 0.80, 0.85, 0.90, and 0.95 specificity was higher in the good quality compared with poor quality photographs (0.79 vs. 0.53, 0.75 vs. 0.47, 0.69 vs. 0.41, and 0.56 vs. 0.23).

Qualitative Review of DL Quality Model Predictions

Human review of images across the quality score range qualitatively showed that increased DL quality score reliably corresponded to lower quality images for both the DIGS/ADAGES and OHTS datasets (Fig. 3). Excessive brightness, low contrast and blurriness appeared to negatively affect glaucoma gradeability (Figs. 4J, 4K, 4L).

Human experts qualitatively reviewed all 23 FN (the model identified a photograph was good quality when it was poor) and 336 FP (the model identified a photograph as poor quality when it was good) examples. This review revealed consistent patterns in FN and FP model errors. When the model prediction resulted in a FN, it was often due to failures in identifying poor cropping of the images, often in eyes with larger discs as quality issues that preclude grading. However, in the majority of cases the quality model correctly identified images with poor cropping as low quality. Qualitative review of FPs, for which the image incorrectly identified high-quality images as poor, demonstrated undervaluing of the significance of obvious glaucomatous features that enabled grading in comparison to the general image quality defects that led to low-quality scores (Fig. 5). This factor could also help to account for the relatively high number of FPs compared with FNs (336 vs. 23). That is, these images had quality issues (e.g., blurring, artifacts, noise) and the model was recognizing them correctly. However, the clear present and extent of glaucomatous damage meant it was still gradable, despite quality issues.

Visualizing DL Quality Model Predictions

Grad-CAMs on several sample images from the OHTS dataset are shown in Figure 6. For low-quality images with large, visible artifacts, the model focused

Table 2. POAG Detection Performance on Low- vs. High-quality Images From OHTS

Quality Values Set	n	AUC (95% CI)	Sensitivity @ 80% Spec	85% Spec	90% Spec	95% Spec
High quality ($0 \leq q < 0.1$)	11729	0.87 (0.80–0.92)	0.79	0.75	0.69	0.56
Low quality ($q \geq 0.1$)	549	0.77 (0.67–0.88)	0.53	0.47	0.41	0.23

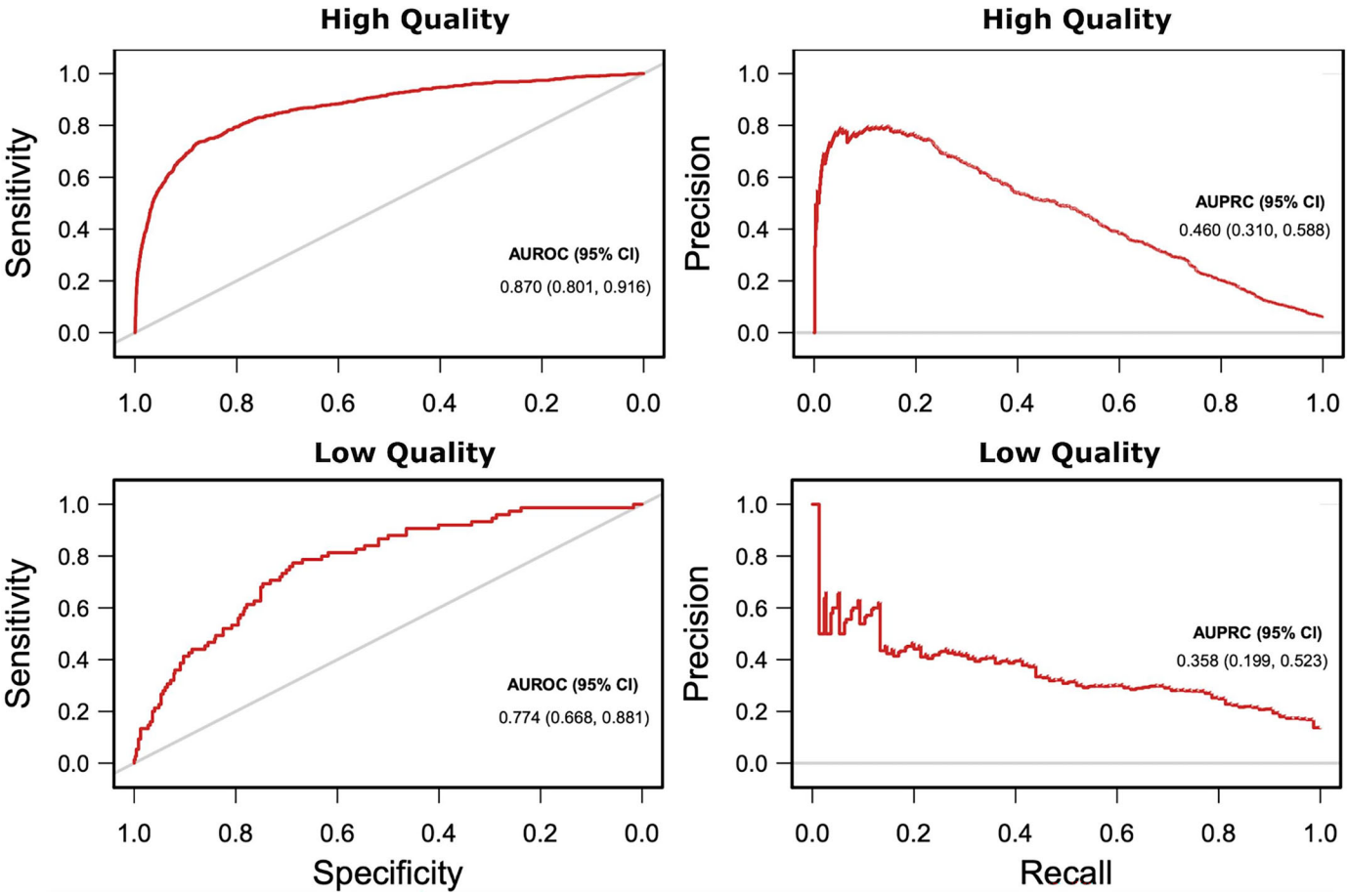

Figure 3. Performance (AUROC) of POAG detection model for high- and low-quality photographs.

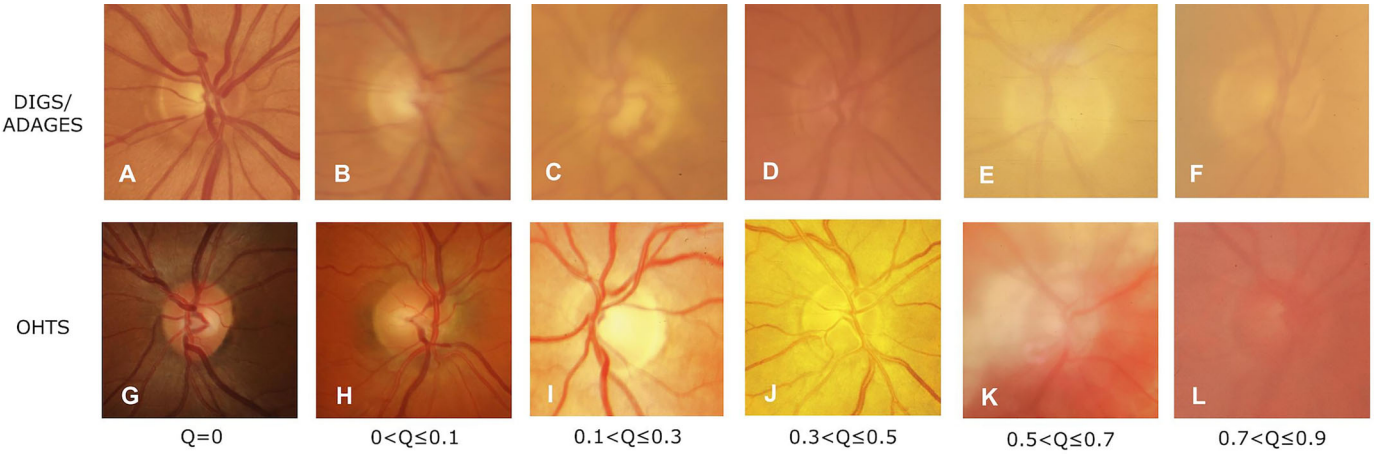

Figure 4. Sample images along with the quality score from the DIGS/ADAGES and OHTS datasets. Quality scores (*Q* in the figure) are generated by the DL quality model and estimate the likelihood of poor quality (0.0 = predicted good quality, 1.0 = predicted poor quality).

most strongly on those artifacts (Fig. 6, left). For low-quality images with severe blur throughout the image, the model broadened to larger regions of the image, prioritizing the ONH and some peripheral regions (Fig. 6, middle). For high-quality images, the model concentrated on superior and inferior regions of the neuroretinal rim area and retinal vasculature (Fig. 6, right).

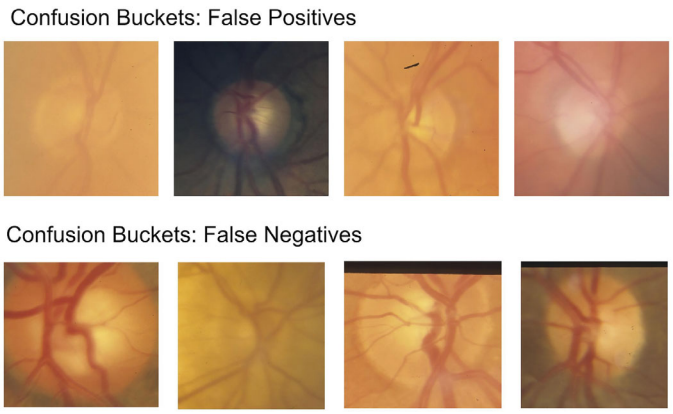

Figure 5. Example FP and FN images for identifying the quality of the photographs.

Discussion

The DL quality model developed in the current study accurately identified gradable photographs, the use of which impacted the performance of a previously described DL POAG detection model.[2,3] Automated quality assessment of fundus photographs may improve clinical and research workflows by quickly identifying images with sufficient quality and detail for disease grading. Recent work in automated fundus has also reported high performance in distinguishing low and high-quality fundus images, with AUROCs of 0.95 to 0.99.[14–17] Our AUROC on the OHTS data is within this range. One important distinction to note, however, is that our ground truth is explicitly based on gradeability for a particular disease of interest rather than general image quality, which is the typical approach taken in previous work. We also directly measured the impact of applying a DL-based quality filtering to an existing DL POAG model. Our use of a DL approach to estimate image quality based on gradeability also has an advantage over some traditional metrics of signal quality (e.g., signal-to-noise ratio), because it does not require a reference standard high-quality image to measure noise and directly addresses a central question relevant to diagnosis and screening—is this image gradable for glaucoma?

In recent years, numerous studies have applied DL models to retinal fundus photographs, achieving high accuracy for a variety of tasks such as disease detection and vessel segmentation.[1–3,7,27,28] However, most of these studies use only high-quality photographs; quality can vary greatly in real-world situations and significantly affect the performance of these models in clinical practice. Beede et al[29] deployed an AI-assisted diabetic retinopathy detection model across eleven clinics in Thailand and found poor lighting conditions to be a major factor leading to ungradable images and decreased performance.[30] To address the issue of quality, several investigators have proposed automated quality assessment models for fundus images.[31–33] However, the majority of these studies did not focus on image gradeability for any specific pathologies, and instead focused on general image quality. Further, most of the previous studies were trained and evaluated on photographs from the same source datasets, raising the question of whether such approaches will generalize to different patient populations, devices, or lighting conditions from a variety of cameras. Also, although these techniques showed high AUROC on test sets, they were not evaluated on relevant downstream diagnostic tasks, such as disease detection, limiting their evaluation as a clinical tool.[32] In addition, most prior work focused on

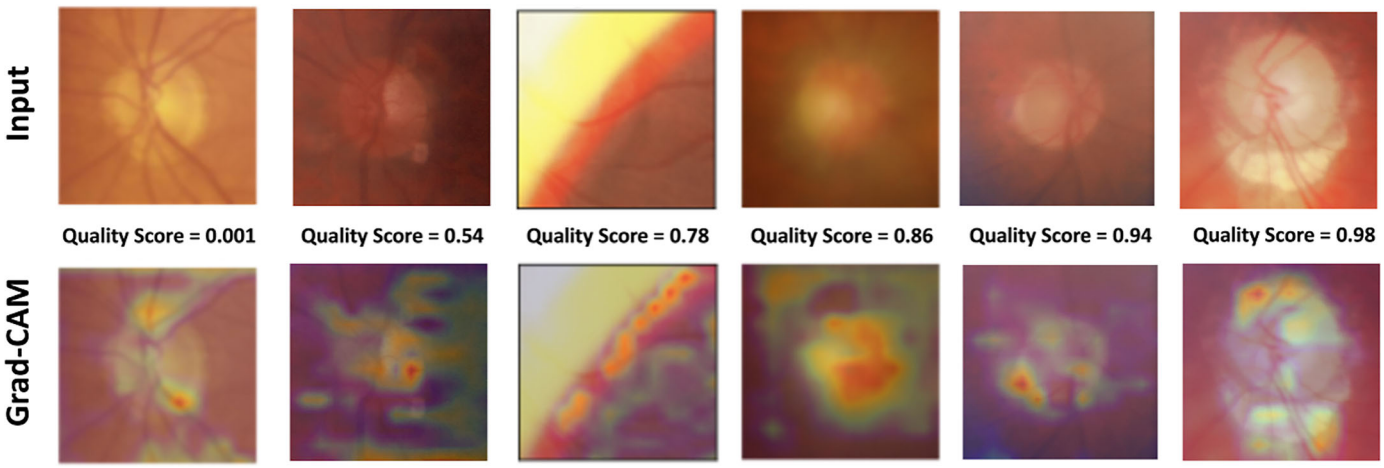

Figure 6. Grad-CAM of sample images. The red indicates areas that were of high importance to the model's prediction. Green and blue indicate progressively lower importance.

applications of photographs for detection of diabetic retinopathy, not glaucoma.

In contrast, we propose a DL quality assessment model trained on manual labels for glaucoma gradeability using photographs from more than 30 clinical sites. The model achieved strong performance on quality assessment (AUROC = 0.97) on an independent dataset. Expert review of DL quality predictions suggests that the model is identifying commonly seen quality issues in fundus photographs (e.g., lighting, blur, cropping). We further evaluated the model on the downstream task of glaucoma detection, where good quality photographs improved automated POAG detection performance compared with poor quality photographs (AUROC = 0.87 vs. 0.77). This finding suggests that the DL quality model may be an effective prescreening tool for automated POAG detection.

Grad-CAM was used to identify areas that drove model decisions regarding image quality. Unsurprisingly, in images with obvious, localized quality issues (i.e., imaging artifacts, blurred regions), the DL quality model focused on those areas. In good quality images, the model seemed to focus on the neuroretinal rim, especially the inferior and superior regions of the rim. This finding is similar to our previous results showing the DL POAG model also focused on inferior and superior rim regions[2] and also corresponds with general guidelines that are often applied to ONH review—namely, that glaucomatous damage is first visible in the inferior rim, followed by the superior rim.[34] This could indicate that POAG gradeability could be maintained even in the presence of quality issues, as long as these areas are preserved.

A strength of this study is that the DL model was developed using DIGS/ADAGES photographs from 3 independent study centers, and tested on the OHTS photographs representing 33 study centers, each with their unique cameras, technicians, and study populations. Another strength is that the images used to train and test the model were graded specifically for glaucoma gradeability as opposed to general image quality. Although photographs with overall good quality are ideal to aid in clinical decision-making, in the case of glaucoma, it is possible to identify characteristic defects in poorer quality photographs in cases of advanced glaucoma. Furthermore, the model was evaluated on the downstream task of glaucoma detection using a previously developed DL POAG detection model. To the best of our knowledge, we are the first to develop a DL quality assessment model specifically for glaucoma gradeability.

One limitation of this study is that our glaucoma gradeability training labels are binary labels, with no additional information about severity or annotations for specific artifacts that may affect gradeability. Although qualitative analysis shows that our model is able to quantify glaucoma gradeability on a spectrum of severity (Fig. 3) despite being trained on binary labels, and it focuses on relevant visual features such as the ONH, retinal vasculature, and artifacts (Fig. 6). Future work could focus on producing more informative and explainable quality assessments via ablation studies, more detailed training labels, or complex model architectures. Another limitation is the dataset imbalance of high- and low-quality images, with good quality photographs vastly outnumbering the poor. The studies in which these images were collected (DIGS/ADAGES and OHTS) took a number of approaches to help ensure good quality images, including patient selection, training and certification of photographers, the capture of multiple photos per visit, and the review and recapturing of photos in cases of insufficient quality, among other approaches. This circumstance is not representative of real-world clinical data, where poor quality images can be quite common. Even in the presence of this large imbalance, our review of FPs and FNs as well as overall model performance in other measures suggest that the model itself generally performs well, although it did lead to a low precision. Undersampling of high-quality images or data augmentation with artificial generation of more low-quality images could be explored to ameliorate this imbalance and improve the measured precision in subsequent research. Finally, this model may have overvalued general image quality to classify an image as ungradable in cases where glaucomatous defects were otherwise sufficiently severe for the image to remain gradable by human reviewers, leading to FPs for low image gradeability (Fig. 5).

In clinical practice, our model could be used as a tool for real-time quality assessment of retinal fundus images acquired from patients with suspected glaucoma: images assessed in real time as poor quality could be discarded and the photograph could be retaken in the same visit, ensuring glaucoma gradeability of acquired images and decreasing the risk of preventable complications and errors in clinical workflow. In addition, because our technique was evaluated directly on downstream glaucoma detection, it could be incorporated into a clinical pipeline for glaucoma detection as a quality screening tool, preceding an automated glaucoma detection model or human reader with a large workload. When applied to a downstream automated glaucoma detection model, our results indicate that glaucoma detection performance may improve. Application of this approach to clinical research and clinical trials that use qualitative assessment by an optic disc could enhance repro-

ducibility and efficiency, because the model can both ensure gradeability of images, improve reader performance, and decrease reader workload. In conclusion, the current results suggest that a DL model can accurately assess glaucoma gradeability of retinal funduscopic images from a wide variety of populations and conditions. Moreover, using the automated assessment of quality to filter out low-quality fundus photographs increased the accuracy of a downstream DL POAG detection model.

Acknowledgments

Supported by grants NEI: K99 EY030942, EY11008, P30 EY022589, EY026590, EY022039, EY021818, EY023704, EY029058, T32 EY026590, and R21 EY027945; German Research Foundation (DFG) (RE 4155/1-1); German Ophthalmological Society (DOG); and an unrestricted grant from Research to Prevent Blindness (New York, NY).

Disclosure: **B. Chuter**, UCSD Summer Research Fellowship, 2021; **J. Huynh**, T35: Short-Term National Research Service Award (NRSA); **C. Bowd**, None; **E. Walker**, None; **J. Rezapour**, German Research Foundation (DFG) (RE 4155/1-1) (F), German Ophthalmological Society (DOG) grant (F); **N. Brye**, None; **A. Belghith**, None; **M.A. Fazio**, National Eye Institute (F), EyeSight Foundation of Alabama (F), Research to Prevent Blindness (F), Heidelberg Engineering, GmbH (F), Topcon (F); **C.A. Girkin**, National Eye Institute (F), EyeSight Foundation of Alabama (F), Research to Prevent Blindness (F), Heidelberg Engineering, GmbH (F); **G. De Moraes**, Novartis (C), Galimedix (C), Belite (C), Reichert (C), Carl Zeiss (C), Perfuse Therapeutics (C), Heidelberg (R), Topcon (R), Ora Clinical (E); **J.M. Liebmann**, Alcon (C), Allergan (C), Bausch & Lomb (C, F), Carl Zeiss Meditec (C, F), Heidelberg Engineering (C, F), Reichert (C), Valeant Pharmaceuticals (C), National Eye Institute (F), Novartis (F), Optovue (F), Reichert Technologies (F), Research to Prevent Blindness (F); **R.N. Weinreb**, Abbvie (C), Aerie Pharmaceuticals (C), Alcon (C), Allergan (C), Amydis (C), Equinox (C), Eyenovia (C), Nicox (C), Santen (C), Topcon (C, F), Heidelberg Engineering (F), Carl Zeiss Meditec (F), Optovue (F), Centervue (F), NEI (F), NIMHD (F), RPB (F), Toromedes (P), Zeiss Meditec (P); **L.M. Zangwill**, National Eye Institute (F), Carl Zeiss Meditec Inc. (F), Heidelberg Engineering GmbH (F), Optovue Inc. (F), Topcon Medical Systems Inc. (F), Zeiss Meditec (P), Abbvie (C), Digital Diagnostics (C); **M. Christopher**, National Eye Institute (F), The Glaucoma Foundation (F)

References

1. Xiangyu C, Yanwu X, Damon Wing Kee W, Tien Yin W, Jiang L. Glaucoma detection based on deep convolutional neural network. *Annu Int Conf IEEE Eng Med Biol Soc*. 2015;2015:715–718.
2. 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.
3. Christopher M, Nakahara K, Bowd C, et al. Effects of study population, labeling and training on glaucoma detection using deep learning algorithms. *Transl Vis Sci Technol*. 2020;9:27.
4. Wang S, Jin K, Lu H, Cheng C, Ye J, Qian D. Human visual system-based fundus image quality assessment of portable fundus camera photographs. *IEEE Trans Med Imaging*. 2016;35:1046–1055.
5. Veiga D, Pereira C, Ferreira M, Goncalves L, Monteiro J. Quality evaluation of digital fundus images through combined measures. *J Med Imaging (Bellingham)*. 2014;1:014001.
6. Shen Z, Fu H, Shen J, Shao L. Modeling and enhancing low-quality retinal fundus images. *IEEE Trans Med Imaging*. 2021;40:996–1006.
7. Paulus J, Meier J, Bock R, Hornegger J, Michelson G. Automated quality assessment of retinal fundus photos. *Int J Comput Assist Radiol Surg*. 2010;5:557–564.
8. Coyner AS, Swan R, Brown JM, et al. Deep learning for image quality assessment of fundus images in retinopathy of prematurity. *AMIA Annu Symp Proc*. 2018;2018:1224–1232.
9. Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated quality assessment of colour fundus images for diabetic retinopathy screening in telemedicine. *J Digit Imaging*. 2018;31:869–878.
10. Coyner AS, Swan R, Campbell JP, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmol Retina*. 2019;3:444–450.
11. Li HH, Abraham JR, Sevgi DD, et al. Automated quality assessment and image selection of ultra-widefield fluorescein angiography images through deep learning. *Transl Vis Sci Technol*. 2020;9:52.
12. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopa-

thy in primary care offices. *NPJ Digit Med*. 2018;1: 39.
13. Fan R, Bowd C, Christopher M, et al. Detecting Glaucoma in the Ocular Hypertension Treatment Study using deep learning: implications for clinical trial endpoints. TechRxiv 2021.
14. Shi C, Lee J, Wang G, Dou X, Yuan F, Zee B. Assessment of image quality on color fundus retinal images using the automatic retinal image analysis. *Sci Rep*. 2022;12:10455.
15. Shen Y, Sheng B, Fang R, et al. Domain-invariant interpretable fundus image quality assessment. *Med Image Anal*. 2020;61:101654.
16. Karlsson RA, Jonsson BA, Hardarson SH, Olafsdottir OB, Halldorsson GH, Stefansson E. Automatic fundus image quality assessment on a continuous scale. *Comput Biol Med*. 2021;129: 104114.
17. Abramovich O, Pizem H, Van Eijgen J, et al. FundusQ-Net: a regression quality assessment deep learning algorithm for fundus images quality grading. *Comput Methods Programs Biomed*. 2023;239:107522.
18. Sample PA, Girkin CA, Zangwill LM, et al. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127:1136–1145.
19. Gordon MO, Beiser JA, Brandt JD, et al. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol*. 2002;120:714–720; discussion 829–830.
20. Gordon MO, Higginbotham EJ, Heuer DK, et al. Assessment of the Impact of an Endpoint Committee in the Ocular Hypertension Treatment Study. *Am J Ophthalmol*. 2019;199:193–199.
21. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vision*. 2015;115:211–252.
22. Chollet F. Xception: Deep Learning with Depthwise separable convolutions, 2016:arXiv:1610. 02357.
23. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning, 2016:arXiv:1602. 07261.
24. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics*. 1997;53:567–578.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D.Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2016:arXiv:1610.02391.
26. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*. 2008;50:419–430.
27. You Q, Wan C, Sun J, Shen J, Ye H, Yu Q. Fundus image enhancement method based on CycleGAN. *Annu Int Conf IEEE Eng Med Biol Soc*. 2019;2019:4500–4503.
28. 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.
29. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI conference on human factors in computing systems, 2020:1–12.
30. Stagg BC, Stein JD, Medeiros FA, et al. Special commentary: using clinical decision support systems to bring predictive models to the glaucoma clinic. *Ophthalmol Glaucoma*. 2021;4:5–9.
31. Zago GT, Andreao RV, Dorizzi B, Teatini Salles EO. Retinal image quality assessment using deep learning. *Comput Biol Med*. 2018;103:64–70.
32. Dai L, Wu L, Li H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nat Commun*. 2021;12: 3242.
33. Yuen V, Ran A, Shi J, et al. Deep-learning-based pre-diagnosis assessment module for retinal photographs: a multicenter study. *Transl Vis Sci Technol*. 2021;10:16.
34. Jonas JB, Budde WM, Panda-Jonas S. Ophthalmoscopic evaluation of the optic nerve head. *Surv Ophthalmol*. 1999;43:293–320.