JAMA Ophthalmology | Original Investigation

Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India

Varun Gulshan, PhD; Renu P. Rajan, MD; Kasumi Widner, MS; Derek Wu, BS; Peter Wubbels, BA; Tyler Rhodes, BS; Kira Whitehouse, BA; Marc Coram, PhD; Greg Corrado, PhD; Kim Ramasamy, MD; Rajiv Raman, MD; Lily Peng, MD, PhD; Dale R. Webster, PhD

IMPORTANCE More than 60 million people in India have diabetes and are at risk for diabetic retinopathy (DR), a vision-threatening disease. Automated interpretation of retinal fundus photographs can help support and scale a robust screening program to detect DR.

OBJECTIVE To prospectively validate the performance of an automated DR system across 2 sites in India.

DESIGN, SETTING, AND PARTICIPANTS This prospective observational study was conducted at 2 eye care centers in India (Aravind Eye Hospital and Sankara Nethralaya) and included 3049 patients with diabetes. Data collection and patient enrollment took place between April 2016 and July 2016 at Aravind and May 2016 and April 2017 at Sankara Nethralaya. The model was trained and fixed in March 2016.

INTERVENTIONS Automated DR grading system compared with manual grading by 1 trained grader and 1 retina specialist from each site. Adjudication by a panel of 3 retinal specialists served as the reference standard in the cases of disagreement.

MAIN OUTCOMES AND MEASURES Sensitivity and specificity for moderate or worse DR or referable diabetic macula edema.

RESULTS Of 3049 patients, 1091 (35.8%) were women and the mean (SD) age for patients at Aravind and Sankara Nethralaya was 56.6 (9.0) years and 56.0 (10.0) years, respectively. For moderate or worse DR, the sensitivity and specificity for manual grading by individual nonadjudicator graders ranged from 73.4% to 89.8% and from 83.5% to 98.7%, respectively. The automated DR system's performance was equal to or exceeded manual grading, with an 88.9% sensitivity (95% CI, 85.8-91.5), 92.2% specificity (95% CI, 90.3-93.8), and an area under the curve of 0.963 on the data set from Aravind Eye Hospital and 92.1% sensitivity (95% CI, 90.1-93.8), 95.2% specificity (95% CI, 94.2-96.1), and an area under the curve of 0.980 on the data set from Sankara Nethralaya.

CONCLUSIONS AND RELEVANCE This study shows that the automated DR system generalizes to this population of Indian patients in a prospective setting and demonstrates the feasibility of using an automated DR grading system to expand screening programs.

Invited Commentary
page 994

Supplemental content

Author Affiliations: Google Research, Mountain View, California (Gulshan, Widner, Wu, Coram, Corrado, Peng, Webster); Aravind Eye Hospital, Madurai, India (Rajan, Ramasamy); Verily Life Sciences, San Francisco, California (Wubbels, Whitehouse); Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, Tamil Nadu, India (Rhodes, Raman).

Corresponding Author: Lily Peng, MD, PhD, Google Inc, 1600 Amphitheatre Pkwy, Mountain View, CA 94043 (lhpeng@google.com).

JAMA Ophthalmol. 2019;137(9):987-993. doi:10.1001/jamaophthalmol.2019.2004
Published online June 13, 2019.

In India, an estimated 60 million people have diabetes.[1] One serious complication of diabetes is diabetic retinopathy (DR), a major cause of avoidable blindness worldwide. Diabetic retinopathy affects approximately 12% to 18% of patients with diabetes in India.[2-6] In lower-income health care environments, the key challenges to addressing DR include a lack of symptoms until the disease has progressed to vision loss, a large population of patients with diabetes who require screening, and a shortage of eye care specialists. Although most guidelines worldwide recommend yearly screenings,[7] DR screening in India is performed on an ad hoc basis without a cohesive strategy at the national level.[8] Barriers for hospital-based screenings or outreach screening camps include patient awareness, access issues, and the lack of trained ophthalmologists and clinical teams.

Telemedicine[9] is a potential cost-effective solution to the access problems.[10] Patients can have retinal images taken at diabetology offices or primary care clinics and the cases can reviewed by a remote expert. Today, a major impediment in implementing telescreening in tertiary care centers is the lack of trained graders to grade the fundus photography images sent from the remote clinics. Thus, an automated system to assess the severity of DR can help scale screenings.

Recent work has demonstrated highly accurate deep-learning algorithms for various medical image classification tasks,[11-13] including retinal imaging.[14-18] Specifically for DR, multiple groups have shown that deep learning can be leveraged to produce expert-level diagnoses for grading fundus photography images and resulting products have since been validated prospectively to obtain regulatory approval.[19] In this study, we build on this work by studying the use of an automated DR grading software and comparing its performance with that of manual grading in a prospective setting in 2 centers in India.

Key Points

Question What is the performance of a deep-learning model in a cohort of patients with diabetes in India?

Findings In this observational study of moderate or worse diabetic retinopathy and referable diabetic macular edema, the automated diabetic retinopathy system's performance was equal to or exceeded manual grading.

Meaning Deep-learning models performed well within a population of patients with diabetes from India.

Methods

Algorithm Development

Deep neural networks were trained and validated using the methods described by Gulshan et al[14] to produce algorithms that grade retinal fundus photography images according to the International Clinical Diabetic Retinopathy (ICDR) severity scale.[20] The network was trained to make multiple binary classifications: (1) moderate or worse DR (ie, moderate, severe, or proliferative), (2) severe or worse DR, (3) referable diabetic macular edema (DME), (4) fully gradable. In addition to these binary classifications used by Gulshan et al,[14] we also trained the model to make a multiway classification of the 5-point ICDR grade. While the model was trained to make the various predictions described previously, only 2 outputs of the model were used and measured during the trial; one was for referable DR and the other was for referable DME. An image was considered referable for DR if it had moderate or worse DR. Hard exudates within 1 disc diameter of the macula was used as a proxy for referable DME. The algorithm's threshold for the presence of referable DR (also known as the operating point) and DME was optimized for a high sensitivity suitable for a screening use case (eMethods in the Supplement).

Study Population

This prospective study was conducted at 2 tertiary eye care centers in South India, Aravind Eye Hospital and Sankara Nethralaya. The study protocol was approved by the ethics committee of both institutions. Written informed consent was obtained from each patient. Data collection and enrollment took place between April 2016 and July 2016 at Aravind and May 2016 and April 2017 at Sankara. A total of 997 patients were enrolled at Aravind and 2052 at Sankara. At Aravind, approximately 499 patients (50%) were recruited at the general ophthalmology clinics from patients who were known to have diabetes but had not previously received a retinal examination and the remaining 498 patients with diabetes (50%) who presented directly to the vitreoretinal clinic. At Sankara, approximately 841 patients (41%) were recruited from patients with diabetes who were visiting the vitreoretinal clinic and the remaining 1211 patients (59%) with diabetes who presented at the teleophthalmology community screenings during the same period.

The inclusion criteria consisted of patients who were older than 40 years and previously received a diabetes diagnosis. The exclusion criteria consisted of patients with a history of any intraocular surgery other than cataract surgery; ocular laser treatments for any retinal disease; ocular injections for DME or proliferative disease; a history of any other retinal vascular disease, glaucoma, or other diseases that may affect the appearance of the retina or optic disc; medical conditions that would be a contraindication to dilation; overt media opacity; or gestational diabetes.

Study Procedure

Patient eligibility was determined by reviewing their medical records on presentation to the clinic. All eligible patients underwent nonmydriatic fundus photography using a retinal fundus camera (NM TRC; Topcon Medical Systems; 3nethra; Forus Health) to capture a macula-centered 40° to 45° fundus photograph. As per the usual center-specific workflow, all images for Aravind were taken using the Forus 3nethra camera, and for Sankara approximately 94% of images were taken using the Forus 3nethra; the rest were taken using the Topcon NM TRC. Following imaging, patients underwent a routine, dilated fundus examination by a retinal specialist. Patients were advised and provided treatment based on the retinal specialist examination per standard guidelines. The results of additional grading by the software and additional graders were not available to the treating retinal specialist to ensure that standard clini-

Table 1. Baseline Characteristics[a]

Characteristic	No. (%)			
	Development		Clinical Validation	
	Train	Tune	Aravind	Sankara
Images, total No.	103 634	40 790	1983	3779
Patient demographics				
Unique individuals, total No.[b]	54 149	20 860	997	2052
Age, mean (SD), y[c]	55.3 (11.2)	55.2 (11.1)	56.6 (9.0)	56.0 (10.0)
Female/total patients for whom sex was known, %	27 760/ 46 360 (59.9)	10 457/ 17 260 (60.6)	418/997 (41.9)	673/2052 (32.8)
Image quality distribution				
Images for which DR was gradable/total images where gradeability was assessed[d]	41 984/ 55 265 (76.0)	23 176/ 27 951 (82.9)	1905/1983 (96.1)	3747/3779 (99.2)
Images where DME was gradable/total images for which gradeability was assessed, %[d]	41 984/ 55 265 (76.0)	23 176/ 27 951 (82.9)	1946/1983 (98.1)	3737/3779 (98.9)
Disease severity distribution				
Total images for which DR was assessed	98 688 (100.0)	39 190 (100.0)	1905 (100.0)	3747 (100.0)
No DR	49 082 (49.7)	23 045 (58.8)	1213 (63.7)	2518 (67.2)
Mild	20 220 (20.5)	6625 (16.9)	52 (2.7)	76 (2.0)
Moderate	21 417 (21.7)	6844 (17.5)	477 (25.0)	676 (18.0)
Severe	4070 (4.1)	1384 (3.5)	77 (4.0)	150 (4.0)
Proliferative	3899 (4.0)	1292 (3.3)	86 (4.5)	327 (8.7)
Total images for which DME was assessed	96 394 (100.0)	38 776 (100.0)	1946 (100)	3737 (100.0)
Referable DME	14 159 (14.7)	4634 (12.0)	429 (22.0)	780 (20.9)

Abbreviations: DME, diabetic macular edema; DR, diabetic retinopathy.

[a] A summary of image characteristics and available demographic information in the development and clinical validation data sets. The adjudicated reference standard was used for computing the DR and DME distributions on the clinical validation data sets, and the majority reference standard was used for the development data sets.

[b] Unique patient codes (deidentified) were only available for 89 997 images (86.8%) in the training set and 36 976 images (90.6%) in the tuning set.

[c] Age was available only for 46 351 individuals in the training set and 17 254 individuals in the tuning set.

[d] For the training and tuning sets, only a single image quality assessment was done as opposed to separate DR and DME gradeability assessments for the clinical validation sets.

cal care was not affected by the study. All patient-related data were deidentified before transferring for analysis.

Grading and Adjudication

A detailed description of grading and adjudication is described in the eMethods in the Supplement. Nonmydriatic fundus photography images were sent for manual grading by a trained grader (a nonphysician) and retinal specialist at each of the sites using the ICDR scale. At Aravind, the trained grader had 7 months of DR grading experience and the retinal specialist had been practicing for 15 months. At Sankara, the trained grader had 5 years of DR grading experience and the retinal specialist had been practicing for 10 years. Each of the graders was masked to the grading by other graders, algorithm, and the results of the in-person dilated fundus examination.

For Aravind, all images from the study were adjudicated by a panel of 3 senior retinal specialists using the protocol described by Krause et al.[16] The adjudicating retinal specialists first graded each of the images independently. Any disagreements between adjudicating retinal specialists were discussed until a full consensus was achieved. For Sankara, because of the larger number of images, the reference standard was determined using a modified protocol as follows: if all graders, including the algorithm, selected the same grade (ie, a 5-point ICDR grade and referable DME), this grade was accepted as the ground truth. Otherwise, the image was sent for adjudication using the same protocol as Aravind. In addition, 10% of full-agreement images (ie, images for which the algorithm and the clinical site's retinal specialist and trained grader all agreed on the grade) were sent for adjudication by the panel of retinal specialists.

Subsequent Model Development

During the course of the prospective data collection period, we made additional improvements to the model, including tuning the models with adjudicated data as reported by Krause et al.[16] The improvements can be summarized as (1) larger training sets, (2) better hyperparameter exploration (tuning), (3) larger input image resolution, and (4) using the improved Inception-v4[21] neural network architecture. We graded the images using the model from Krause et al[16] retrospectively at the conclusion of the study.

Statistical Analysis

To characterize the sensitivity and specificity of the algorithm with respect to the reference standard, 2 × 2 tables were generated. The 95% confidence intervals for the sensitivity and specificity of the algorithm were calculated to be exact Clopper-Pearson intervals[22] that corresponded to separate 2-sided confidence intervals with individual coverage probabilities of the square root of 0.95 being approximate to 0.975. These simultaneous 2-sided confidence intervals were computed using StatsModels, version 0.6.1 (Python) and statistical significance was set at $P < .05$. Additional details on sample size calculation are in the eMethods in the Supplement.

Results

In total, 3049 patients were enrolled in this study (**Table 1**). The mean (SD) age of enrolled patients was 56.6 (9.0) and 56.0 (10.0) years at Aravind and Sankara, respectively. Women comprised 418 patients (41.9%) at Aravind and 673 patients (32.8%)

Table 2. Sensitivity and Specificity of Various Graders for the Clinical Trial Data From Aravind and Sankara

Hospital	% (95% CI)			
	Moderate DR +		DME	
	Sensitivity	Specificity	Sensitivity	Specificity
Aravind				
Retina specialist (C.O.[a])	89.8 (86.9-92.4)	83.5 (81.0-85.8)	89.5 (85.7-92.6)	93.8 (92.3-95.1)
Trained grader (L.V.[a])	75.7 (71.7-79.5)	94.2 (92.5-95.6)	74.0 (68.9-78.7)	95.6 (94.2-96.7)
Model	88.9 (85.8-91.5)	92.2 (90.3-93.8)	97.4 (95.2-98.8)	90.7 (88.9-92.3)
Sankara				
Retina specialist (R.R.[a])	73.4 (70.4-76.3)	98.7 (98.1-99.2)	57.5 (53.5-61.5)	99.3 (98.9-99.6)
Trained grader (S.S.[b])	84.2 (81.6-86.5)	98.6 (98.0-99.1)	75.1 (71.5-78.5)	97.7 (97.0-98.3)
Model	92.1 (90.1-93.8)	95.2 (94.2-96.1)	93.6 (91.3-95.4)	92.5 (91.3-93.5)

Abbreviations: DME, diabetic macular edema; DR, diabetic retinopathy.

[a] Nonauthor technician.

[b] Dr Raman.

at Sankara. Because this study recruited from both eye clinics and a general screening pool, the study population was enriched for more severe forms of DR.

The performance of each type of grader is shown in **Table 2**. Of all gradable images for moderate or worse DR, the trained grader from Aravind had a 75.7% sensitivity and 94.2% specificity and the retinal specialist had a 89.8% sensitivity and 83.5% specificity. The trained grader from Sankara had a 84.2% sensitivity and 98.6% specificity and the retinal specialist had a 73.4% sensitivity and 98.7% specificity. At the predefined operating point, the algorithm had a sensitivity of 88.9% and specificity of 92.2% on the Aravind data (with an area under the curve [AUC] of 0.963) and a sensitivity of 92.1% and specificity of 95.2% on the Sankara data (with an AUC of 0.980) (**Figure 1**).

For referable DME, the trained grader from Aravind had a 74.0% sensitivity and 95.6% specificity and the retinal specialist had a 89.5% sensitivity and 93.8% specificity. The trained grader from Sankara had a 75.1% sensitivity and 97.7% specificity and the retinal specialist had a 57.5% sensitivity and 99.3% specificity. At the predefined operating point, the algorithm had a sensitivity of 97.4% and specificity of 90.7% on Aravind data (AUC, 0.983) and a sensitivity of 93.6% and specificity of 92.5% on Sankara data (AUC, 0.983).

The performance of the model on the combined data set from Aravind and Sankara is shown in eFigure 1 in the Supplement. The intragrader reliability of manual grading at Aravind was performed on 10% of images. For the trained grader, there was a 64.5% exact concordance of the 5-point DR grade, 79.8% concordance for referable DR, and 84.2% concordance for referable DME. For the retinal specialist, these numbers were 78.5%, 90.8%, and 92.3%, respectively (eFigures 2-5 and eTables 1-6 in the Supplement).

Examples of difficult cases are shown in **Figure 2**. Figure 2, A shows a case in which all graders and the algorithm were discordant with the adjudicated ground truth because of a subtle neovascularization of the disc and neovascularization elsewhere. Figure 2, B depicts cases in which the adjudicated grade was not referable. The graders were correct but the algorithm was not. Figure 2 C, illustrates a case in which the algorithm was correct but the graders were not. The fibrous proliferation at the disc was picked by the algorithm but missed by the graders.

Validation of an Improved Automated DR Grading System

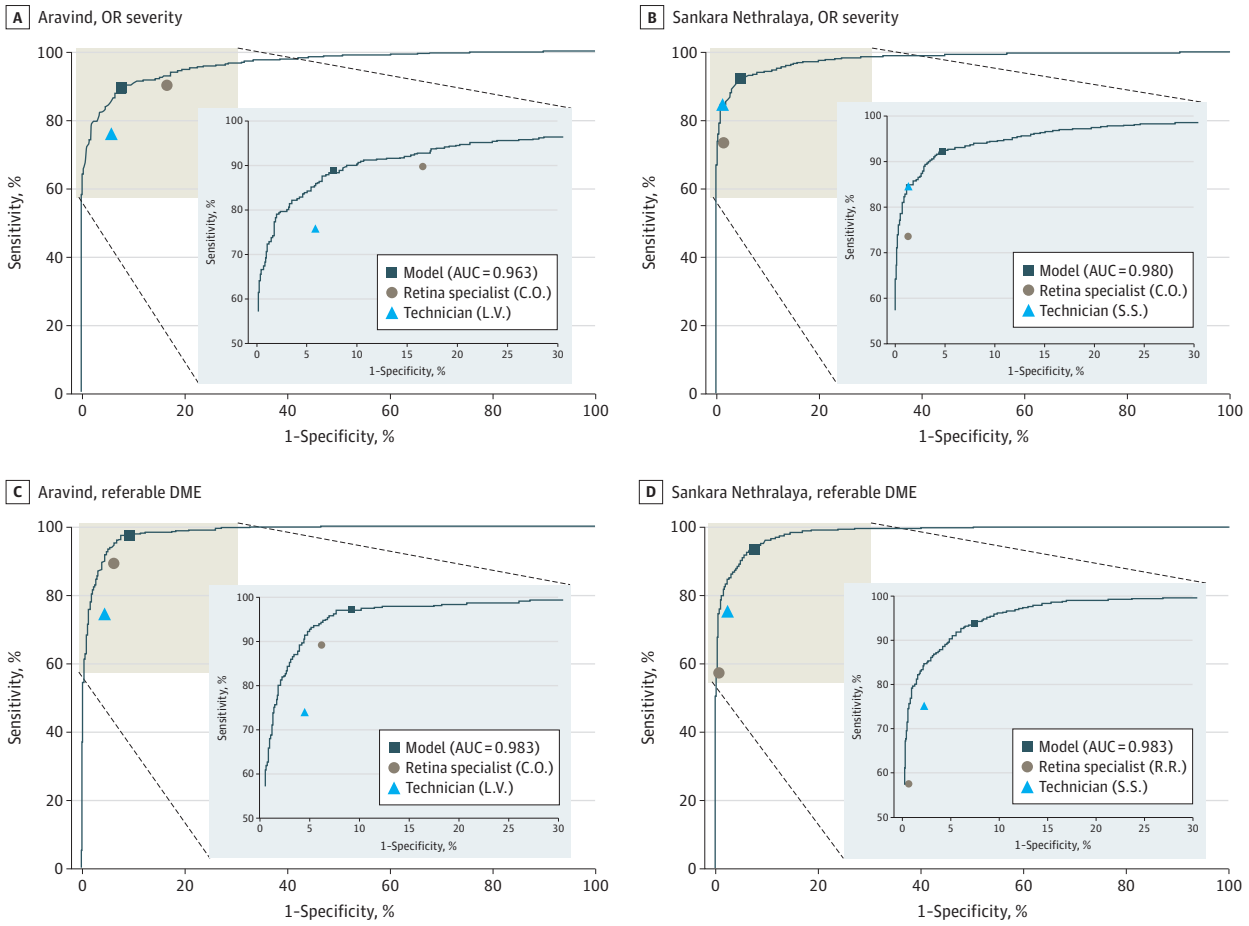
During the course of this study, an improved model was published.[16] This model was optimized for the 5-point ICDR grading, which allowed us to study model performance at each level of severity. On the combined data set, this corresponded to an AUC of 0.986 for the new model vs 0.974 on the old model for detecting moderate or worse DR and an AUC of 0.984 for the new model vs 0.983 on the old model for detecting referable DME. This corresponds to a sensitivity of 92.2% (95% CI, 90.7-93.6) and specificity of 96.9% (95% CI, 96.2-97.5) for detecting moderate or worse DR. To measure the level of agreement across the 5 classes, we used a quadratic weighted κ. Compared with the reference standard, on Aravind data, quadratic κ scores were 0.74 (95% CI, 0.71-0.76), 0.75 (95% CI, 0.72-0.79), and 0.85 (95% CI, 0.83-0.87) for the retinal specialist, trained grader, and new model, respectively. On Sankara data, quadratic weighted κ scores were 0.82 (95% CI, 0.80-0.84), 0.88 (95% CI, 0.86-0.89), and 0.91 (95% CI, 0.90-0.93) for the retinal specialist, trained grader, and new model, respectively (**Table 3**). Overall, the new agreement between the model and the adjudicated reference standard was higher than that of the individual graders.

Discussion

Our results demonstrate that an automated algorithm identified referable DR with performance equal to or exceeding the retinal specialists and trained graders in a prospective clinical setting. These results were consistent across 2 hospitals and suggests good model generalization. This was encouraging because the cameras used in the training data sets and prospective studies were different (the prospective data from Aravind and 94% of the data from Sankara were from Forus 3Nethra, and only 320 images [0.2%] in the development set were from this camera).

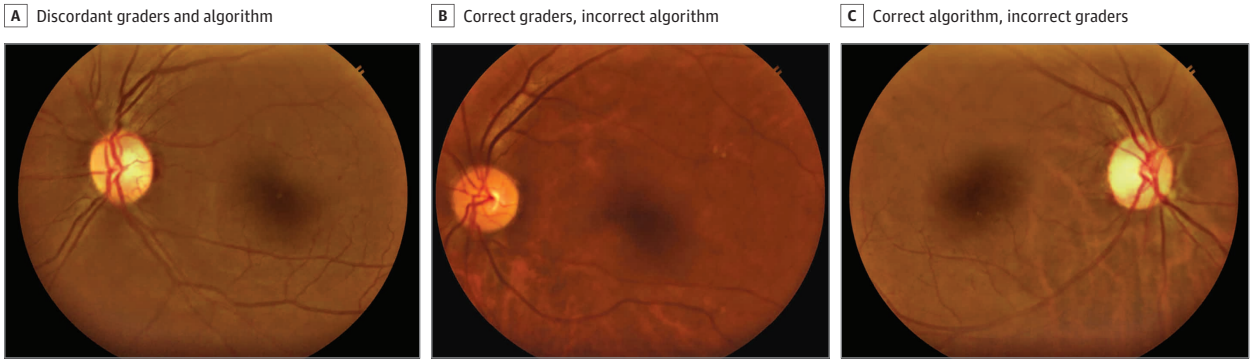
Because the reference standard, cameras, and clinical setting vary between previous studies, comparing this study with

Figure 1. Comparison of the Algorithm, Graders, and Retinal Specialists Using the Adjudicated Reference Standard for Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME)



Diabetic retinopathy severity and referable DME for Aravind (A and C) and Sankara Nethralaya (B and D). Diabetic retinopathy severity is greater than or equal to moderate nonproliferative DR. AUC indicates area under the curve. C.O., L.V., and S.S. refer to nonauthor technicians. R.R. refers to Dr Raman.

Figure 2. Examples of Disagreements Between Different Graders and the Adjudicated Reference Standard



A, Cases in which all graders and the algorithm were discordant with the adjudicated ground truth. B, Cases in which the graders were correct but the algorithm was not. C, Cases in which the algorithm was correct but the graders were not.

others is not straightforward. For example, a previous study by our team used a majority decision as the reference standard,[14] whereas this study used adjudication by a panel of retinal specialists. In addition, using real-world screening data and nonmydriatic low-cost cameras probably led to more images with lower image quality, which can contribute to trickier cases and lower performance numbers for human graders and the algorithm. While the absolute numbers are diffi-

Table 3. Quadratic Weighted κ Scores for 5-Point Diabetic Retinopathy Grading[a]

Hospital	Quadratic Weighted κ (95% CI)
Aravind	
Retina specialist (C.O.[b])	0.74 (0.71-0.76)
Trained grader (L.V.[b])	0.75 (0.72-0.79)
New model	0.85 (0.83-0.87)
Sankara	
Retina specialist (R.R.[c])	0.82 (0.80-0.84)
Trained grader (S.S.[b])	0.88 (0.86-0.89)
New model	0.91 (0.90-0.93)

[a] Only the κ score from the new model is reported as the original model was not trained to predict 5 class grades.

[b] Nonauthor technician.

[c] Dr Raman.

cult to compare, the relative performance of manual grading and the algorithm shown in this study is consistent with previous studies, specifically that algorithm performance meets or exceeds that of manual grading by trained graders and retinal specialists. IDx recently received US Food and Drug Administration approval[19] for a hybrid feature engineered and deep learning–based algorithm by demonstrating its performance in a prospective study in the United States. Similarly, this study represents a critical milestone in using this technology in clinical settings in India.

Strengths and Limitations

While these results are encouraging, there are a few limitations to this study. First, although we used the 5-point ICDR grade for training, the original and main algorithm used in this study only made a binary call in terms of DR. This is consistent with other deep-learning systems, including the IDx algorithm, that define referable DR as moderate or worse DR and/or DME.[14,15,23] In addition to the binary call, we also validated another model that returns a 5-point grade with a slightly better performance than the algorithm used in this study (eFigure 1 in the Supplement), but this validation was performed retrospectively. The more granular 5-point grading would be especially helpful for screening programs in which patient treatment varies at each level of severity. In particular, the threshold for and timing of referrals in DR screening programs often depend on the resources of the program. Currently, in most programs, sight-threatening cases will be referred urgently while moderate cases without DME will be followed clinically. Identifying mild cases may also be of clinical value for health care systems that have a different screening interval for patients with no disease and mild disease.[24] A more granular grading output, such as the 5-point ICDR scale, would also be more robust to guideline changes.

In addition, hard exudates within 1 disc diameter was used as a proxy for DME. In future studies, using optical coherence tomography imaging could provide a better reference standard. Previous studies have shown that hard exudates may not coincide well with actual retinal thickening, resulting in numerous false-negative and false-positive results.[25] Using a wider field of view equal to or greater than the 7-field Early Treatment DR Study standard would also establish a more robust ground truth. In particular, lesions outside the 45° field of view that would have upgraded a case from mild to moderate would not be detected from the reference standard used in this study. For Sankara, because of the larger volume of images, adjudication was performed for all images with disagreement and 10% with agreement; this might affect the specificity and sensitivity of the graders and model compared with Aravind, where the adjudication process was performed for all images regardless of agreement. We extrapolate that the sensitivity for each grader would be approximately 4% to 5% lower when full adjudication is performed. In addition, image quality was not one of the predictions that was returned by this version of the algorithm, so the results of this study included only images deemed gradable by the adjudication panel. Additional improvements could include an image quality output, distinguishing patients with stable disease from postlaser treatment and those who have disease progression and the detection of other common retinal disease, such as age-related macular degeneration and glaucoma.

The algorithm was used in a process that was parallel to standard clinical care. More work must be done to study the integration of this system into the clinical care workflow. Given the high sensitivity of the system and specificity that is equal to human graders, the automated system holds promise as a point-of-care initial screening solution that does a first pass to rule out patients at lower risk of vision-threatening disease and flagging images that are categorized as abnormal for timely follow-up by a clinician. This will decrease the proportion of patients that might be lost to follow-up because of a failure to return the test results to the patient asynchronously (eg, the patient moved in the interim or does not have accurate contact information) or the need for repeated visits. This would be especially advantageous in low-resource settings. In higher-resourced settings, the algorithm could serve as a concurrent read with manual grading and the discrepant calls could be reviewed by an adjudicator. This could decrease the number of false-positive and false-negative results. The various implementation methods of the algorithm should be evaluated in future studies in the clinic. Finally, cost-effectiveness studies in high-resource and low-resource settings are critical in understanding the economic effects that such deep-learning algorithms will have on health care systems. These studies could inform not only the operating points (ie, referral thresholds) of the algorithms themselves but also subsequent care pathways downstream of the screening visit.

Conclusions

While there are many avenues for future work, this study demonstrates the feasibility of using an automated DR grading system in health care systems and shows that the trained algorithm generalizes to this prospective population of Indian patients.

ARTICLE INFORMATION

Accepted for Publication: February 14, 2019.

Published Online: June 13, 2019. doi:10.1001/jamaophthalmol.2019.2004

Author Contributions: Drs Gulshan and Peng had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Ramasamy, Raman, Peng, and Webster contributed equally to this article.
Concept and design: Gulshan, Wu, Kim, Raman, Corrado, Peng, Webster.
Acquisition, analysis, or interpretation of data: Gulshan, Rajan, Widner, Wu, Wubbels, Rhodes, Whitehouse, Kim, Raman, Coram, Peng.
Drafting of the manuscript: Gulshan, Wu, Rhodes, Whitehouse, Peng.
Critical revision of the manuscript for important intellectual content: Gulshan, Rajan, Widner, Wubbels, Kim, Raman, Coram, Corrado, Peng, Webster.
Statistical analysis: Gulshan, Wu, Coram.
Obtained funding: Widner, Corrado, Webster.
Administrative, technical, or material support: Rajan, Widner, Wu, Wubbels, Rhodes, Whitehouse, Kim, Raman, Peng, Webster.
Supervision: Rajan, Kim, Raman, Corrado, Peng, Webster.

Conflict of Interest Disclosures: Drs Gulshan, Coram, Corrado, Peng, and Webster hold a patent for a mechanism through which to process fundus photography images using machine learning models pending. Drs Gulshan, Peng, Webster, Coram, and Widner; Messrs Wu, Wubbels, and Rhodes; and Mses Widner and Whitehouse reported being an employee of Google and owning Google stock. No other disclosures were reported.

Funding/Support: This study received funding from Google LLC.

Role of the Funder/Sponsor: Google LLC was involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We thank Jonathan Krause, PhD, Yun Liu, Katy Blumer, BS, Bilson Campana, PhD, Anita Misra, BS, Philip Nelson, BS, Chris Stelton, MD, Ehsan Rahimy, MD, Anthony Joseph, MD, Oscar Kuruvilla, MD (Google Research), Florence Thng, MS, Sunny Virmani, MS, Shawn Xu, MS, Dave Watson, Eli Romanova, BS, Tom Stanis, BS, Linus Upson (Verily Life Sciences), A. L. Sivaram, MPT, Sangeetha Srinivasan, PhD, Gayathri Swaminathan, MSc, and Durgasri Jaisankar, BS (Aravind and Sankara) for the thoughtful advice, assistance in reviewing the manuscript, building tools to help with data gathering and analysis, and the execution of the study. They were compensated for their contributions.

REFERENCES

1. Kaveeshwar SA, Cornwall J. The current state of diabetes mellitus in India. *Australas Med J*. 2014;7(1):45-48. doi:10.4066/AMJ.2014.1979

2. Rani PK, Raman R, Sharma V, et al. Analysis of a comprehensive diabetic retinopathy screening model for rural and urban diabetics in developing countries. *Br J Ophthalmol*. 2007;91(11):1425-1429. doi:10.1136/bjo.2007.120659

3. World Health Organization. Prevention of blindness from diabetes mellitus: report of a WHO consultation in Geneva, Switzerland, 9-11 November 2005. https://www.who.int/blindness/Prevention%20of%20Blindness%20from%20Diabetes%20Mellitus-with-cover-small.pdf. Accessed May 12, 2019.

4. Narendran V, John RK, Raghuram A, Ravindran RD, Nirmalan PK, Thulasiraj RD. Diabetic retinopathy among self reported diabetics in southern India: a population based assessment. *Br J Ophthalmol*. 2002;86(9):1014-1018. doi:10.1136/bjo.86.9.1014

5. Namperumalsamy P, Kim R, Vignesh TP, et al. Prevalence and risk factors for diabetic retinopathy: a population-based assessment from Theni District, south India. *Br J Ophthalmol*. 2009;93(4):429-434.

6. Rema M, Premkumar S, Anitha B, Deepa R, Pradeepa R, Mohan V. Prevalence of diabetic retinopathy in urban India: the Chennai Urban Rural Epidemiology Study (CURES) eye study, I. *Invest Ophthalmol Vis Sci*. 2005;46(7):2328-2333. doi:10.1167/iovs.05-0019

7. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. *Expert Rev Ophthalmol*. 2012;7(5):417-439. doi:10.1586/eop.12.52

8. Ramasamy K, Raman R, Tandon M. Current state of care for diabetic retinopathy in India. *Curr Diab Rep*. 2013;13(4):460-468. doi:10.1007/s11892-013-0388-6

9. Liesenfeld B, Kohner E, Piehlmeier W, et al. A telemedical approach to the screening of diabetic retinopathy: digital fundus photography. *Diabetes Care*. 2000;23(3):345-348. doi:10.2337/diacare.23.3.345

10. Rachapelle S, Legood R, Alavi Y, et al. The cost-utility of telemedicine to screen for diabetic retinopathy in India. *Ophthalmology*. 2013;120(3):566-573. doi:10.1016/j.ophtha.2012.09.002

11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118. doi:10.1038/nature21056

12. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. http://arxiv.org/abs/1703.02442. Accessed May 12, 2019.

13. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al; the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

14. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

15. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152

16. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125(8):1264-1272. doi:10.1016/j.ophtha.2018.01.034

17. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124(7):962-969. doi:10.1016/j.ophtha.2017.02.008

18. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135(11):1170-1176. doi:10.1001/jamaophthalmol.2017.3782

19. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Med*. 2018;1:39. doi:10.1038/s41746-018-0040-6

20. American Academy of Ophthalmology. International clinical diabetic retinopathy disease severity scale, detailed table. http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf. Accessed 14 Oct, 2016.

21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. https://ieeexplore.ieee.org/document/7780677. Accessed May 12, 2019.

22. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404-413. doi:10.1093/biomet/26.4.404

23. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351-357. doi:10.1001/jamaophthalmol.2013.1743

24. Scanlon PH. Screening intervals for diabetic retinopathy and implications for care. *Curr Diab Rep*. 2017;17(10):96. doi:10.1007/s11892-017-0928-6

25. Wang YT, Tadarati M, Wolfson Y, Bressler SB, Bressler NM. Comparison of prevalence of diabetic macular edema based on monocular fundus photography vs optical coherence tomography. *JAMA Ophthalmol*. 2016;134(2):222-228. doi:10.1001/jamaophthalmol.2015.5332