An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs

Zhixi Li,[1] Stuart Keel,[2] Chi Liu,[3] Yifan He,[3] Wei Meng,[3] Jane Scheetz,[2] Pei Ying Lee,[2] Jonathan Shaw,[4] Daniel Ting,[5] Tien Yin Wong,[5] Hugh Taylor,[6] Robert Chang,[7] and Mingguang He[1,2]

Diabetes Care 2018;41:2509–2516 | https://doi.org/10.2337/dc18-0147

EPIDEMIOLOGY/HEALTH SERVICES RESEARCH

OBJECTIVE

The goal of this study was to describe the development and validation of an artificial intelligence–based, deep learning algorithm (DLA) for the detection of referable diabetic retinopathy (DR).

RESEARCH DESIGN AND METHODS

A DLA using a convolutional neural network was developed for automated detection of vision-threatening referable DR (preproliferative DR or worse, diabetic macular edema, or both). The DLA was tested by using a set of 106,244 nonstereoscopic retinal images. A panel of ophthalmologists graded DR severity in retinal photographs included in the development and internal validation data sets (*n* = 71,043); a reference standard grading was assigned once three graders achieved consistent grading outcomes. For external validation, we tested our DLA using 35,201 images of 14,520 eyes (904 eyes with any DR; 401 eyes with vision-threatening referable DR) from population-based cohorts of Malays, Caucasian Australians, and Indigenous Australians.

RESULTS

Among the 71,043 retinal images in the training and validation data sets, 12,329 showed vision-threatening referable DR. In the internal validation data set, the area under the curve (AUC), sensitivity, and specificity of the DLA for vision-threatening referable DR were 0.989, 97.0%, and 91.4%, respectively. Testing against the independent, multiethnic data set achieved an AUC, sensitivity, and specificity of 0.955, 92.5%, and 98.5%, respectively. Among false-positive cases, 85.6% were due to a misclassification of mild or moderate DR. Undetected intraretinal microvascular abnormalities accounted for 77.3% of all false-negative cases.

CONCLUSIONS

This artificial intelligence–based DLA can be used with high accuracy in the detection of vision-threatening referable DR in retinal images. This technology offers potential to increase the efficiency and accessibility of DR screening programs.

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
[2]Centre for Eye Research Australia, University of Melbourne, Melbourne, Australia
[3]Guangzhou Healgoo Interactive Medical Technology Co. Ltd., Guangzhou, China
[4]Baker Heart and Diabetes Institute, Melbourne, Australia
[5]Singapore Eye Research Institute, Singapore National Eye Centre, Duke-NUS Medical School, National University of Singapore, Singapore
[6]Indigenous Eye Health Unit, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia
[7]Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA

Corresponding author: Mingguang He, mingguang_he@yahoo.com.

Received 18 January 2018 and accepted 2 September 2018.

This article contains Supplementary Data online at http://care.diabetesjournals.org/lookup/suppl/doi:10.2337/dc18-0147/-/DC1.

Z.L., S.K., and C.L. contributed equally to this work.

Diabetic retinopathy (DR), the most common microvascular complication of diabetes, is a leading cause of irreversible vision loss in adults of working age (1). Recent estimates suggest the global prevalence of DR is 34.6%, corresponding to nearly 100 million people worldwide (1). With the prevalence of diabetes expected to rise by at least 25% by 2030 (2,3), a significant increase is expected in the burden of DR (4).

Given that the majority of vision loss from DR is avoidable through early detection and effective treatment strategies (5,6), many national and international societies have long endorsed screening for DR (7). This comes most commonly in the form of point-of-care ophthalmoscopy by trained eye care personnel (e.g., ophthalmologists or optometrists) or retinal photography with either local interpretation or telemedicine-based screening programs with centralized grading (8). However, despite growing evidence of the effectiveness of routine assessments and early intervention (9,10), comprehensive DR screening strategies are not widely implemented (11). This is largely because of the inadequate availability of resources, including trained eye care personnel and financing, to cope with the rapidly growing burden of diabetes. This is particularly important in major developing countries such as China and Indonesia (11–13).

Deep learning, a branch of machine learning under the broad category of artificial intelligence (AI), represents a recent advancement of artificial neural networks that permits improved classification predictions from raw image data (14). These deep learning techniques have been applied in highly image-driven medical specialties, including dermatology and radiology, with promising results: the techniques have identified diseases as accurately as or more accurately than board-certified specialists (15,16). Recent studies suggest that these deep learning algorithms (DLAs) can achieve excellent sensitivities and specificities in detecting DR and referable DR (moderate DR or worse) (17–20), thereby offering significant potential benefits to DR screening programs, including increased efficiency, accessibility, and affordability. Despite this, most of the systems were validated using retinal photographs from publicly available databases (EyePACS, MESSIDOR 2, e-ophtha), comprising mainly high-quality photographs from individuals of a single ethnicity. Given this, such systems should be evaluated under real-world screening conditions, where the quality of retinal images varies considerably and retinal pigmentation differs among ethnicities; it is important that DLAs for DR be validated through the use of retinal photographs captured with different imaging protocols and across multiple ethnicities in order to demonstrate their pertinence. We are aware of only one study to date that has evaluated a DLA in a multiethnic cohort (21).

Herein we describe the development and validation of a DLA for the detection of vision-threatening referable DR (preproliferative DR or worse, diabetic macular edema [DME], or both) through the use of a data set of over 70,000 retinal photographs collected during routine assessments from a variety of clinical settings in China. In addition, we validate the DLA using three population-based data sets comprising photographs from individuals of distinct ethnicities.

RESEARCH DESIGN AND METHODS

This study was approved by the institutional review boards of the Zhongshan Ophthalmic Center, Guangzhou, China (2017KYPJ049), and the Royal Victorian Eye and Ear Hospital, East Melbourne, Australia (16/1268H). The study was conducted in accordance with the Declaration of Helsinki.

Development of the DLA

The DLA was developed using 71,043 retinal photographs acquired from a web-based, crowdsourcing platform (LabelMe, Guangzhou, China; http://www.labelme.org) that contains more than 200,000 color fundus photographs. A total of 36 hospital ophthalmology departments, optometry clinics, and screening settings in China contributed deidentified original fundus photographs to this data set. In all cases retinal photographs were captured through the use of common conventional desktop retinal cameras, including from Topcon Corp., Canon, Centervue, and Heidelberg Engineering, and with a variety of imaging protocols. We recruited 27 ophthalmologists as candidates to grade images. Each candidate graded three test sets of 60 images (20 images of normal fundi, 20 images of background DR, and 20 images of preproliferative or proliferative DR), and their results were compared with those of an experienced ophthalmologist (Z.L.) who held an English National Health Screening (NHS) DR grading license. Only those who achieved an unweighted $\kappa \geq 0.70$ (substantial) for vision-threatening referable DR (preproliferative DR or worse, DME, or both) were included as graders in this study. In total, 21 ophthalmologists met this criterion.

Retinal photographs were graded between March and June 2017. Images from the total data set (n = 71,043) were randomly assigned to a single ophthalmologist for grading when their LabelMe account was activated. After a given image was graded, it was returned to the pooled data set; this process continued until three consistent grading outcomes were achieved for an image. At that time, a conclusive annotation was added to the image and it was removed from the original image pool and made available for download by the research team. Graders were blind to the previous grading outcomes in this process, and a given image could be assigned to a grader only once. The consensus grading outcome was assigned as the final, conclusive grade of each image. A retinopathy severity score was assigned according to the NHS diabetic eye screening guidelines (22,23). These guidelines categorize patients as R0 (no DR), R1 (background DR), R2 (preproliferative DR), R3 (proliferative DR), M0 (no visible maculopathy), M1 (maculopathy), or U (unclassifiable). We defined DME as any hard exudates within a one-disc diameter of the fovea or an area of hard exudates in the macular area that encompassed at least 50% of the disc area. Vision-threatening referable DR was defined as preproliferative DR or worse, DME, or both. Images were graded as "poor quality" if 50% or more of the image was obscured or if vessels in the macular region could not be distinguished as DME. Figure 1 and Supplementary Table 1 describe the image grading process, how models were built, and how images were randomized to the training or validation set.

Several preprocessing steps were performed for normalization to control for variation in the data set. The resolution of original images was any of the following: 2,480 × 3,280; 576 × 768; 1,900 × 2,285; 1,958 × 2,588; 1,900 × 2,265; 1,956 × 2,448; and 1,944 × 2,464 pixels. Image pixel values were scaled within a range of 0–1, and local-space average color was

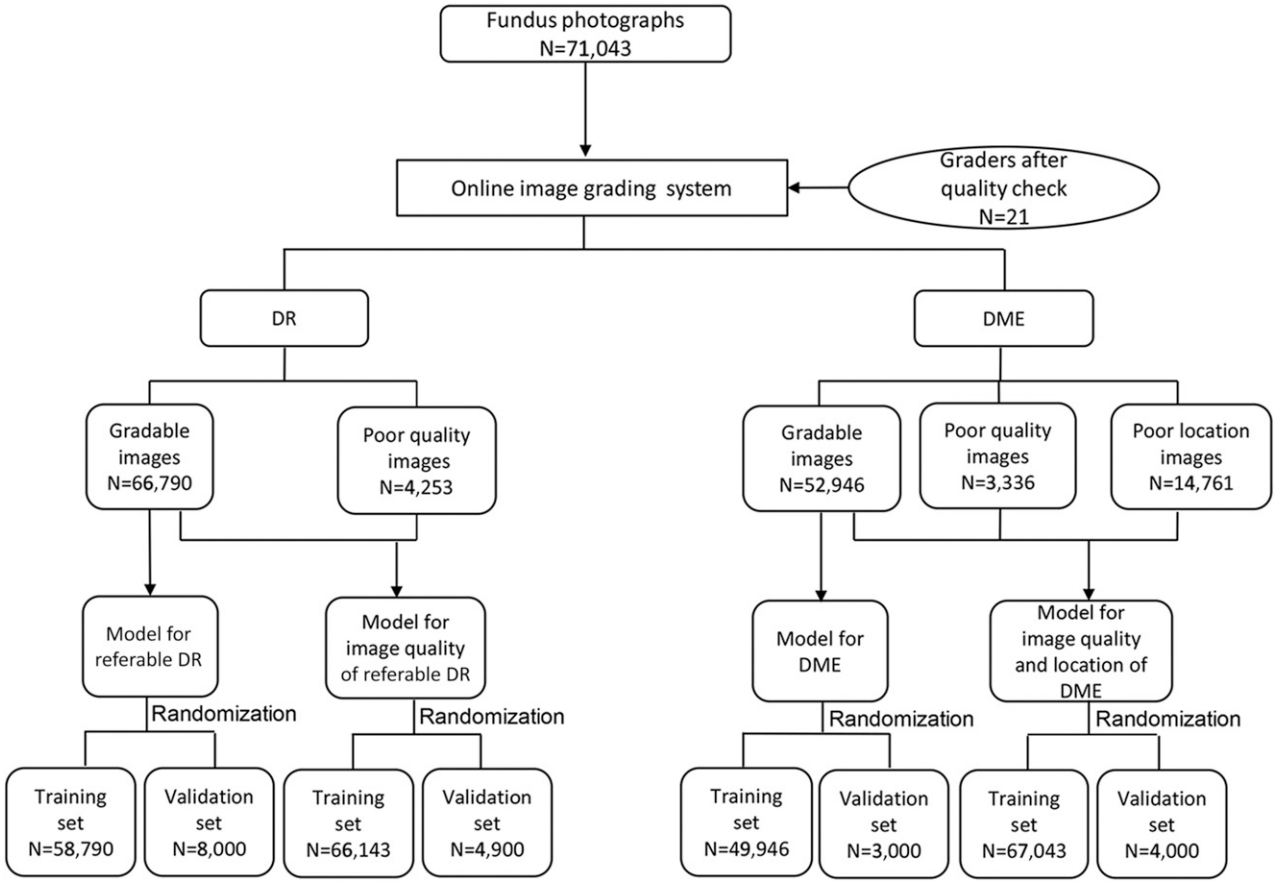

Figure 1—Development workflow for image grading and automated detection.

applied for color constancy (24). Images were then downsized to a resolution of 299 × 299 pixels. Online data augmentation was performed to transform the images by a zero- to three-pixel random horizontal shift and 90°, 180°, or 270° random rotation in order to expand heterogeneity but keep the prognostic features in the image. The study included four deep learning models, all of which use Inception v3 architecture (25) (Supplementary Fig. 1). This included networks for the *1*) classification for vision-threatening referable DR, *2*) classification of DME, *3*) evaluation of image quality for DR, and *4*) assessment of image quality and of the availability of the macular region for DME. The model used in this study was version 20171024.

Validation Data Sets

We initially tested our DLA in our local LabelMe data set by preserving a percentage of images from the original data set of 71,043 images. This included a total of 19,900 images from the four deep learning models; each image was labeled with a consensus standard. An experienced ophthalmologist (Z.L.) classified false-positive and false-negative images into subgroups.

Using identical preprocessing procedures, we also evaluated the performance of our DLA on an independent data set of 35,201 images (14,520 eyes of 7,643 participants) derived from three population-based studies. Population-based samples offer an ideal platform with which to validate this system because they closely mirror the screening setting, capturing a population across the full spectrum of DR severity. These studies included the Indigenous Australians from the National Indigenous Eye Health Survey (NIEHS) (26), Malays from the Singapore Malay Eye Study (SiMES) (27), and Caucasians (95%) from phases 2 and 3 of the Australian Diabetes, Obesity and Lifestyle Study (AusDiab) (28) (Table 1). In each study, two standard, nonmydriatic, 45°, color retinal photographs were taken of each eye, one of the optic disc and the other of the macula, using a Canon retinal camera. Protocols for pupillary dilation differed between studies. In the SiMES, all participants underwent pupillary dilation as part of the retinal imaging protocol, whereas in the NIEHS mydriatic photography was used only when nonmydriatic photography failed; no participants underwent pupillary dilation in the AusDiab. Trained professional graders from each study were used as the reference standard against which the DLA was evaluated. In the case of the NIEHS, images classified as mild and moderate nonproliferative DR (NPDR) were grouped, and as such, a single ophthalmologist (Z.L.) regraded these images to conform to the NHS scheme. In all studies a single manual grading outcome was provided by eye. Given that in the majority of cases multiple images were available for each eye, we adopted the following logic to consolidate a single automated grading result for the right and left eyes: *1*) positive referable DR = any image for a given eye was found upon automated grading to be positive; *2*) negative referable DR = all images for a given eye were found upon automated grading to be negative, or both negative

Table 1—Summary of population-based studies used for external validation

	Data sets		
	NIEHS	SiMES	AusDiab
Year	2008	2004	2004–2012
Country	Australia	Singapore	Australia
Demographics			
Age, years (range)	40–90	40–80	25–90
Ethnicity	Indigenous Australians	Malays	Caucasian Australians
Male sex (%)	41.1	51.9	53
Diabetes prevalence (%)	9.4	35	24.5
Retinal imaging			
Protocol	Two-field, 45° images taken with a Canon CR-DGi camera Mydriasis performed when nonmydriatic photography failed	Two-field, 45° images taken with a Canon CR-DGi camera Mydriasis performed for all participants	Two-field, 45° images taken with a Canon CR6-45NM camera No pupillary dilation
Grader experience	Five certified professional senior graders (>2 years' experience) supervised by a retinal specialist from Australia	One certified professional senior grader (>7 years' experience)	Three certified professional senior graders (>2 years' experience) supervised by a retinal specialist from Australia
Images (*n*)	7,181	15,679	12,341

and ungradable; and *3*) ungradable = all images for a given eye were ungradable upon automated grading.

Statistical Analyses

The accuracy of each grader was calculated as the proportion of grading results that matched the conclusive grading outcome divided by the total number of images graded by that individual. The sensitivity, specificity, accuracy, and area under the curve (AUC) of the DLA in detecting vision-threatening referable DR were calculated and compared to the reference standard (local validation by a retinal specialist, external validation by professional graders) at the level of the individual eye. The 95% CIs were also calculated. Stata version 14.0 (StataCorp, College Station, TX) was used for all statistical analyses in this study.

RESULTS

Each image in the local data set was graded between three and eight times, with a mean agreement of 87.3% (95% CI 85.6–95.4) and a range of 78.6–97.2% among the 21 ophthalmologists. Each ophthalmologist graded between 137 and 21,024 (median 3,501) fundus photographs. Eleven ophthalmologists individually graded more than 3,500 fundus photographs. Of the total 71,043 images, 4,253 (6.1%) were labeled as poor quality, leaving 66,790 images with a conclusive DR severity grading. Using a simple random sampling method, a total of 58,790 images were assigned to the training data set and the remaining 8,000 images were held for internal validation. Among the 58,790 images in the training data set, 10,861 (18.5%) had vision-threatening referable DR and 13,736 (27.5%) had DME (Supplementary Table 1). Training and internal validation of the DLA were completed in October 2017.

Internal Hold-Out Validation

In the internal hold-out validation data set, with reference to the ophthalmologist standard, the AUC, sensitivity, and specificity of the DLA for vision-threatening referable DR were 0.989, 97.0%, and 91.4%, respectively (Fig. 2). For DME, the AUC of the DLA was 0.986, the sensitivity was 95.0%, and the specificity was 92.9%. The AUC, sensitivity, and specificity were 0.950, 90.0%, and 86.2%, respectively, for image quality (to identify referable DR) and 0.989, 96.7%, 90.2%, respectively, for image quality (to identify DR) and availability of the macular region (to identify DME).

The most common clinical features of false-negative cases (*n* = 44) included intraretinal microvascular abnormalities (*n* = 34 [77.3%]) and peripheral retinal photocoagulation laser scars without active proliferative DR lesions (*n* = 4 [9.1%]) (Table 2). An analysis of false-positive cases (*n* = 563) revealed that most images displayed signs of retinal pathology (*n* = 517 [91.8%]), of which the majority (*n* = 482 [93.2%]) were due to background DR being misclassified as vision-threatening DR; the remaining cases were other eye pathologies, including age-related macular degeneration (*n* = 10 [1.9%]) and myopic maculopathy (*n* = 7 [1.4%]). Thus, only few of all false-positive cases (*n* = 46 [8.2%]) had no abnormal ocular findings, and nearly half of these images contained artifacts (*n* = 21 [45.7%]). Examples of typical false-negative and false-positive images can be found in Supplementary Figs. 2 and 3.

External Validation

The external data set contained retinal images from 14,520 eyes from three population-based studies (Supplementary Fig. 4). Of these, 863 eyes (5.9%) had missing or ungradable manual grading outcomes and were subsequently excluded from analysis. Among the total 13,657 eyes (94.1%) included in the external validation data set, 956 (7.0%) had any DR (mild NPDR or worse) and 401 (2.9%) had vision-threatening referable DR.

Also among the 13,657 eyes with images in the external validation data set, 263 images (1.9%) were ungradable for referable DR with the use of the DLA. The DLA showed robust and comparable performance across each of the three population-based studies, achieving accuracy of 97.1%, 98.3%, and 99.1% for the NIEHS, SiMES, and AusDiab, respectively (Supplementary Table 2 details metrics by study). The combined AUC,

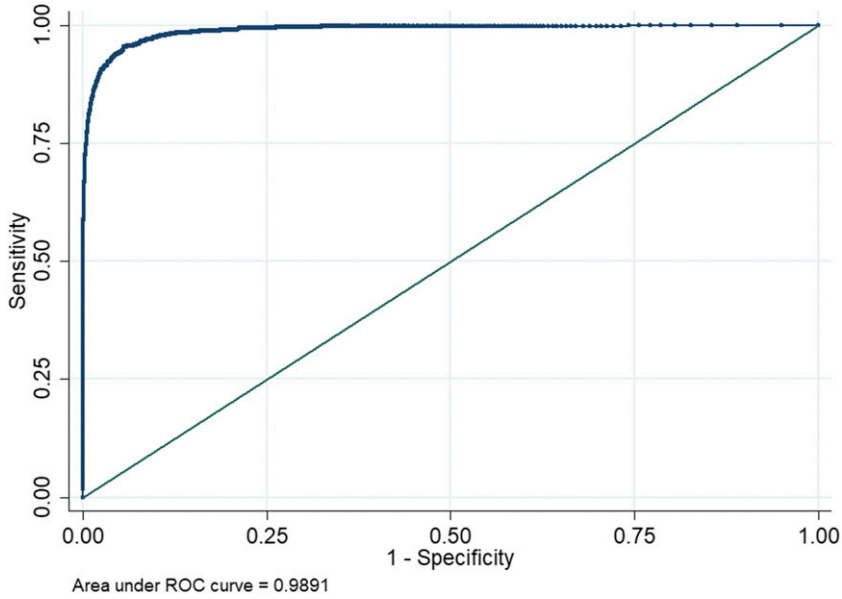

Figure 2—Mean AUC of the model for automated detection of vision-threatening referable DR, derived from internal validation. ROC, receiver operating characteristic.

sensitivity, and specificity of the DLA for referable DR were 0.955, 92.5%, and 98.5%, respectively. This data set consisted of 193 false-positive images (23 [11.9%] of which showed mild or moderate DR), 367 true-positive images, and 30 false-negative images.

To assess repeatability and reliability of the DLA, automated grading was repeated on a sample of 200 eyes from each study data set. We observed 100% consistency in this subset evaluation. An evaluation of real-time runtime of the DLA classification procedure yielded an average of 8.5 s per evaluated image.

CONCLUSIONS

This article describes the development and validation of a novel AI-based DLA for the detection of vision-threatening referable DR. Using a large data set of 71,043 retinal images labeled with a reference standard and 35,201 retinal images from population-based studies, our DLA achieved robust performance in identifying vision-threatening referable DR (AUC = 0.989) and DME (AUC = 0.986) in an independent, local validation data set. Furthermore, the DLA showed excellent diagnostic performance in an external, multiethnic data set (AUC = 0.955). Thus, it offers great potential as an efficient, low-cost solution for DR screening.

Recent reports provide novel data on the accuracy of deep learning systems for the detection of DR (17–19). Gargeya and Leng (18) validated their DLA using 75,137 images, demonstrating high diagnostic accuracy for the detection of any DR (AUC = 0.94–0.95). Abràmoff et al. (17) validated their DLA using 1,748 images and also reported promising results (AUC = 0.980) in the detection of any DR. However, given that early DR (mild NPDR or better) alone does not constitute a change in the routine management of patients with diabetes, identification of referable DR has been the focus of screening programs internationally (i.e., moderate NPDR or worse, DME, or both) (19,29). Adopting this criterion, Gulshan et al. (19) tested their DLA using 9,963 retinal images and achieved excellent performance (AUC = 0.99) for referable DR. Ting et al. (20) validated their DLA using 71,896 images and also reported excellent results for referable DR (AUC 0.936) and vision-threatening DR (AUC 0.958).

Although these studies provide excellent insights, they had some limitations. First, all previously reported DLAs were based on a clinical definition, the International Clinical Diabetic Retinopathy Disease Severity Scale (7), which adopts a criterion for notably earlier referral for DR, that is, any level of retinopathy more severe than mild retinopathy (defined as the presence of microaneurysms only). No specific or effective eye care management is currently available for patients with this milder stage of DR, and therefore these patients should not be referred for specialist eye care; rather, they should be monitored with routine annual screenings (30). This was highlighted by the landmark Early Treatment of Diabetic Retinopathy Study (ETDRS), which

Table 2—Features of false-negative and false-positive images for referable DR in the internal validation data set

Features	No.	Proportion
False-negative images		
IRMA	34	77.3
PRP laser scar	4	9.1
Blurred preretinal hemorrhage	3	6.8
Questionable new vessels	3	6.8
Total	44	100
False-positive images		
Retinal disorders		
Background DR	482	85.6
AMD	10	1.8
Myopic maculopathy	7	1.2
Silicone oil–filled eyes	5	0.9
Retinal vessel occlusion	2	0.4
Retinal detachment	2	0.4
Others	9	1.5
Subtotal	517	91.8
Normal fundus		
With artifacts	21	3.8
Without artifacts	25	4.4
Subtotal	46	8.2
Total	563	100

AMD, age-related macular degeneration; IRMA, intraretinal microvascular abnormality; PRP, peripheral retinal photocoagulation.

reported that none of the patients with mild DR progressed to proliferative DR within 1 year (31). Thus, the adoption of this clinical criterion would lead to over-referral, increasing the strain on eye care resources and adding financial burden to DR screening programs. By comparison, ours is, to our knowledge, the first study to develop a DLA through the use of real-world national screening criteria. The NHS DR classification guideline has been used for many years in national DR screening programs, such as DR screening programs in the U.K. (10). Using this classification, the U.K. DR screening program showed that the number of blindness cases from DR was reduced from 2003 to 2016, to the extent that DR is no longer the leading cause of blindness among working-age adults in the region (32).

Second, three previously reported deep learning systems were validated against public databases comprising mainly high-quality photographs from Caucasian eyes. To demonstrate the generalization of these automated systems, it is important to assess their validity in multiethnic data sets containing images obtained through variable retinal imaging protocols (e.g., nonmydriatic vs. mydriatic) (21).

To address these gaps, in the current study we developed and validated our DLA to detect vision-threatening referable DR (preproliferative DR or worse, DME, or both) based on real-world national DR screening programs (10,33). We tested our DLA in four ethnic groups (Chinese, Malay, Caucasian Australian, and Indigenous Australian), each with distinct retinal pigmentation. Contrast between retinal background and DR lesions (e.g., hemorrhages) varies considerably across ethnicities and is therefore a source of potential error. That is, we hypothesized that a higher level of pigmentation in darker races (e.g., Indigenous Australians) may reduce the contrast between background and DR lesions, making it more difficult for automated methods to detect them. Our finding of robust diagnostic performance across all ethnicities (AUC ranging from 0.94 [Indigenous Australians] to 0.99 [Chinese]) provides evidence that pigmentation does not affect the performance of our DLA and that it is likely generalizable to most populations worldwide.

Poor-quality images are an inevitable consequence of disturbances during the image acquisition process, including poor pupil dilation, media opacities, image contrast or focus issues due to operator problems, or all three. In real-world screening settings, rates of poor-quality or ungradable images have been reported to be as high as 20% (34,35). There is no doubt that sufficient image quality and field definition are key prerequisites for reliable automatic systems for detecting DR and DME. Our DLA successfully offers automated classifiers of image quality (blur detection; AUC = 0.950) and location (macular field vs. other fields; AUC = 0.989), robustly identifying low-quality images in real time and prompting image recapture. Our finding that 98% of retinal images in the external validation data set were gradable using the DLA provides evidence of reliable automated image analysis under different imaging protocols (mydriatic and nonmydriatic) and in older adults who are prone to media opacities. It is worth noting that visual acuity should be assessed in conjunction with retinal analysis in DR screening. Those with reduced vision should be given a referral because other pathology such as cataract, retinal detachment, or vitreous hemorrhage may be the cause of ungradable images. In addition, visual acuity assessment is important to ensure cases that may be undetectable by standard photography (e.g., macular ischemia) are not overlooked.

Indigenous Australians are particularly susceptible to vision loss from DR through a combination of extremely high rates of diabetes (36) and a historical disadvantage in terms of accessing eye care services (37). Our finding that performance indicators (sensitivity, specificity, and AUC) were substantially better among the Indigenous Australian data set than what screening guidelines would typically recommend (>80% sensitivity and specificity) offers great promise to improve the accessibility and efficiency of DR screening among this population.

To our knowledge, this is the first study to explore the characteristics of misclassification (false negatives and false positives) when using a DLA to detect DR. This investigation will help us better understand why this happens in DLA grading and identify strategies to minimize errors in the future. These data are also useful to assist in the clinical acceptance of these systems (21). False-positive referrals result in wasted resources and often cause unnecessary psychological harm to patients (38). Our finding that ~92% of false-positive images displayed abnormal retinal features highlights that most patients in fact may have benefited from referral. More than three-fourths of false-negative cases proved to be undetected intraretinal microvascular abnormality, an often subtle lesion characterized by abnormal branching or dilation of existing retinal vessels. Future optimization of the DLA through the inclusion of more image examples of this lesion in the data set would likely further increase our sensitivity metric.

Key strengths of this study include the use of a large training data set (>70,000 images) of images labeled with a reference standard, development based on real-world screening guidelines, and the robust performance of our DLA when tested on an external validation data set of 35,201 images from three multiethnic, population-based studies. Some limitations must also be considered. First, in the external validation data set (population-based images), the performance of our DLA was assessed against grading by professional nonphysician graders and not by ophthalmologists (i.e., the reference standard). However, given that previous research has consistently reported high interobserver agreement between nonphysician graders and ophthalmologists (39,40), and that similar metrics were observed from internal hold-out validation, it is unlikely that the use of nonphysician graders would have significantly affected the overall diagnostic accuracy. Second, a relatively small representation of referable cases was included in the Caucasian sample (n = 37), which may have resulted in an unstable estimate of diagnostic accuracy in this group. Third, the positive predictive value (PPV) and negative predictive value for the external validation data set were 65.5% ([367 ÷ 560] × 100) and 99.7% ([12,993 ÷ 13,027] × 100), respectively. PPV is highly dependent on the prevalence of disease, and given that our external validation data set contained a relatively low prevalence of referable DR (2.9%) compared with what is expected in real-world screening settings (~10%), the resultant PPV likely

represents an underestimation of the true PPV. Furthermore, the operating threshold for this DLA was purposefully set to be conservative, with an emphasis on not missing vision-threatening referable DR, and thus we focused on achieving a high sensitivity and very high negative predictive value. Fourth, the use of two-field, nonstereoscopic images (as opposed to standard seven-field stereoscopic images) and the exclusion of optical coherence tomography from the study protocol may have resulted in a reduced sensitivity to DR and, in particular, reduced DME detection. Last, although not detailed in this report, it is important to note that our DLA has been developed to detect other common, incidental vision-threatening conditions such as possible cataract (AUC = 0.991), suspected glaucoma (AUC = 0.989), and late age-related macular degeneration (AUC = 0.995) (data not shown). These conditions are typically included within current manual DR screening programs and therefore should not be ignored.

In summary, this AI-based DLA shows robust performance in the detection of vision-threatening referable DR (severe DR or worse, DME, or both) in retinal images from a multiethnic sample. This technology offers great potential to increase the efficiency and accessibility of DR screening programs, particularly in developing countries such as China, Indonesia, and India, and in minority and underserved populations (e.g., Indigenous Australians). Before clinical use, further work is required to investigate where this technology "fits" within the clinical system. For example, whether this software can be successfully incorporated at the point of care to allow non–eye-trained professionals to conduct DR screening or into telemedicine-based screening programs warrants investigation. In addition, public health projects assessing the impact (i.e., adherence to referral, new DR detection rates), end-user acceptance (clinician and patient), and cost-effectiveness of these DLAs would be beneficial.

Acknowledgments. The authors thank the ophthalmologists who volunteered their time to grade the fundus images that were used to train and validate this deep learning algorithm; the EyeGrader team at Guangzhou Healgoo Interactive Medical Technology Co. for their technical assistance during external validation; and the study leads from NIEHS, SiMES, and AusDiab for contributing fundus images for external validation.

Funding. This study was supported in part by the Fundamental Research Funds of the State Key Laboratory of Ophthalmology, National Natural Science Foundation of China (grant no. 81420108008); the Science and Technology Planning Project of Guangdong Province (grant no. 2013B20400003); and the Bupa Health Foundation (Australia grant). J.Sh. is supported by a research fellowship from the National Health and Medical Research Council. M.H. receives support from the Research Accelerator Program at the University of Melbourne and from the CERA Foundation. The Centre for Eye Research Australia receives operational infrastructure support from the Victorian State Government, and Research to Prevent Blindness, Inc., provides support to the Stanford University Department of Ophthalmology.

The sponsor or funding organizations had no role in the design or conduct of this research.

Duality of Interest. W.M. and M.H. report a patent on managing color fundus images using deep learning models (China patent application number ZL201510758675.5; patent filing date 31 May 2017). No other potential conflicts of interest relevant to this article were reported.

Author Contributions. Z.L. and S.K. wrote the manuscript. Z.L., S.K., C.L., Y.H., W.M., J.Sc., P.Y.L., and M.H. acquired, analyzed, and interpreted the data used for internal and external validation. Z.L., C.L., Y.H., W.M., R.C., and M.H. conceived, designed, and developed the deep learning algorithm. J.Sh., D.T., T.Y.W., and H.T. supplied images for external validation. All authors revised and edited the manuscript. M.H. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Yau JW, Rogers SL, Kawasaki R, et al.; Meta-Analysis for Eye Disease (META-EYE) Study Group. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 2012;35:556–564

2. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet 2016;387:1513–1530

3. International Diabetes Federation. IDF Diabetes Atlas, 7th edition [Internet], 2015. Available from http://www.diabetesatlas.org. Accessed 10 November 2017

4. Ting DS, Cheung GC, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clin Experiment Ophthalmol 2016;44:260–277

5. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. Lancet 2010;376:124–136

6. Wong TY, Cheung CM, Larsen M, Sharma S, Simó R. Diabetic retinopathy. Nat Rev Dis Primers 2016;2:16012

7. Wilkinson CP, Ferris FL 3rd, Klein RE, et al.; Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 2003;110:1677–1682

8. Goh JK, Cheung CY, Sim SS, Tan PC, Tan GS, Wong TY. Retinal imaging techniques for diabetic retinopathy screening. J Diabetes Sci Technol 2016;10:282–294

9. Nguyen HV, Tan GS, Tapp RJ, et al. Cost-effectiveness of a national telemedicine diabetic retinopathy screening program in Singapore. Ophthalmology 2016;123:2571–2580

10. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. Acta Diabetol 2017;54:515–525

11. Wang LZ, Cheung CY, Tapp RJ, et al. Availability and variability in guidelines on diabetic retinopathy screening in Asian countries. Br J Ophthalmol 2017;101:1352–1360

12. Sasongko MB, Widyaputri F, Agni AN, et al. Prevalence of diabetic retinopathy and blindness in Indonesian adults with type 2 diabetes. Am J Ophthalmol 2017;181:79–87

13. Wang FH, Liang YB, Zhang F, et al. Prevalence of diabetic retinopathy in rural China: the Handan Eye Study. Ophthalmology 2009;116:461–467

14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444

15. Cheng JZ, Ni D, Chou YH, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454

16. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–118

17. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci 2016;57:5200–5206

18. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 2017;124:962–969

19. 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;316:2402–2410

20. Ting DSW, Cheung CY, 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;318:2211–2223

21. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA 2016;316:2366–2367

22. Peto T, Tadros C. Screening for diabetic retinopathy and diabetic macular edema in the United Kingdom. Curr Diab Rep 2012;12:338–345

23. Public Health England. NHS diabetic eye screening programme: grading definitions for referable disease [Internet], 2016. Available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/582710/Grading_definitions_for_referrable_disease_2017_new_110117.pdf. Accessed 2 May 2015

24. Ebner M. Color constancy based on local space average color. Mach Vis Appl 2009;11:283–301

25. Szegedy C, Vanhouck V, loffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision [article online]. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016;2818–2826. Available from http://arxiv.org/pdf/1512.00567v3.pdf. Accessed 10 May 2017
26. Taylor HR, Xie J, Fox S, Dunn RA, Arnold AL, Keeffe JE. The prevalence and causes of vision loss in Indigenous Australians: the National Indigenous Eye Health Survey. Med J Aust 2010; 192:312–318
27. Wong TY, Cheung N, Tay WT, et al. Prevalence and risk factors for diabetic retinopathy: the Singapore Malay Eye Study. Ophthalmology 2008;115:1869–1875
28. Magliano DJ, Barr EL, Zimmet PZ, et al. Glucose indices, health behaviors, and incidence of diabetes in Australia: the Australian Diabetes, Obesity and Lifestyle Study. Diabetes Care 2008; 31:267–272
29. Ting D, Ng J, Morlet N, et al. Diabetic retinopathy–screening and management by Australian GPs. Aust Fam Physician 2011;40:233–238
30. Scanlon PH. Screening intervals for diabetic retinopathy and implications for care. Curr Diab Rep 2017;17:96
31. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Ophthalmology 1991;98(Suppl.):823–833
32. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open 2014;4:e004015
33. Olafsdóttir E, Stefánsson E. Biennial eye screening in patients with diabetes without retinopathy: 10-year experience. Br J Ophthalmol 2007;91:1599–1601
34. Scanlon PH, Malhotra R, Thomas G, et al. The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy. Diabet Med 2003;20:467–474
35. Scanlon PH, Foy C, Malhotra R, Aldington SJ. The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. Diabetes Care 2005;28:2448–2453
36. Keel S, Foreman J, Xie J, van Wijngaarden P, Taylor HR, Dirani M. The prevalence of self-reported diabetes in the Australian National Eye Health Survey. PLoS One 2017;12: e0169211
37. Foreman J, Xie J, Keel S, Taylor HR, Dirani M. Utilization of eye health-care services in Australia: the National Eye Health Survey. Clin Experiment Ophthalmol 2018;46:213–221
38. Davey CJ, Harley C, Elliott DB. Levels of state and trait anxiety in patients referred to ophthalmology by primary care clinicians: a cross sectional study. PLoS One 2013;8:e65708
39. Bhargava M, Cheung CY, Sabanayagam C, et al. Accuracy of diabetic retinopathy screening by trained non-physician graders using non-mydriatic fundus camera. Singapore Med J 2012; 53:715–719
40. Islam FM, Nguyen TT, Wang JJ, et al. Quantitative retinal vascular calibre changes in diabetes and retinopathy: the Singapore Malay eye study. Eye (Lond) 2009;23:1719–1724