Eye (2020) 34:451–460
https://doi.org/10.1038/s41433-019-0566-0

The ROYAL COLLEGE of OPHTHALMOLOGISTS

REVIEW ARTICLE

Artificial intelligence for diabetic retinopathy screening: a review

Andrzej Grzybowski[1,2] · Piotr Brona[1] · Gilbert Lim[3,4] · Paisan Ruamviboonsuk[5] · Gavin S. W. Tan[4,6] · Michael Abramoff[7] · Daniel S. W. Ting[4,6]

Received: 7 April 2019 / Accepted: 19 June 2019 / Published online: 5 September 2019

Abstract

Diabetes is a global eye health issue. Given the rising in diabetes prevalence and ageing population, this poses significant challenge to perform diabetic retinopathy (DR) screening for these patients. Artificial intelligence (AI) using machine learning and deep learning have been adopted by various groups to develop automated DR detection algorithms. This article aims to describe the state-of-art AI DR screening technologies that have been described in the literature, some of which are already commercially available. All these technologies were designed using different training datasets and technical methodologies. Although many groups have published robust diagnostic performance of the AI algorithms for DR screening, future research is required to address several challenges, for examples medicolegal implications, ethics, and clinical deployment model in order to expedite the translation of these novel technologies into the healthcare setting.

Introduction

Diabetes is becoming a global epidemic with the number of people affected worldwide rising from 108 million in 1980 to an estimated 425 million in 2017, and an estimated 629 million in 2045 [1]. The prevalence of diabetic retinopathy (DR), a primary cause of blindness and vision loss worldwide, was estimated at 93 million in 2012—out of which 28 million people had vision-threatening DR—and this is also expected to further increase [2, 3]. There is substantial scientific evidence that early diagnosis and timely treatment can prevent most visual loss from DR [4], and developed countries have therefore established DR screening programmes aimed at early diagnosis, surveillance and timely treatment of DR [5]. Such programmes are mostly based on the analysis of fundus photographs by specially trained graders, often through telemedicine. However, the diagnostic accuracy achieved may not be optimal, and scaling and sustaining such systems has been found to be challenging [6]. In addition, in developing countries, the cost of such systems can put substantial strain on the healthcare systems when both financial and human resources are often in short supply.

Deep learning, a state-of-art machine learning (ML) technique, has shown promising diagnostic performance in image recognition, speech recognition, and natural language processing [7]. It has been widely adopted in many domains including social media, tele-communications, cybersecurity, and medicine. For medical imaging analysis in general, it has achieved robust results in various medical specialities such as radiology [8] and dermatology [9]; for ophthalmology in particular [10–14], deep learning (DL) continues the long tradition of autonomous and assisted analysis of retinal photographs, which has existed since the 1990s and before [15]. Such artificial intelligence (AI) systems have been demonstrated to lower cost, improve diagnostic accuracy, and increase patient access to DR screening. Recent works on DL in ophthalmology showcase its potential to at least partially replace human graders, while

✉ Daniel S. W. Ting
daniel.ting45@gmail.com

1 Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland

2 Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland

3 School of Computing, National University of Singapore, Singapore, Singapore

4 Singapore National Eye Center, Singapore Eye Research Institute, Singapore, Singapore

5 Department of Ophthalmology, Rajavithi Hospital, Bangkok, Thailand

6 Duke-NUS Medical School, National University of Singapore, Singapore, Singapore

7 Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa, USA

providing a similar level of accuracy. This review aims to summarise state-of-art DR screening technologies that have been published in the literature thus far.

Methodology

We evaluated the current state of DR screening programs by searching Google Scholar, Scopus, Web of Science, PubMed, Medline and Embase for studies published in English up to 5 January 2019, using these keywords: "diabetes", "DR screening", "fundus photographs", "automated DR system", "ML", "AI", and "DL". We also examined reference lists and publicly available websites on well-known, commercially available DR screening algorithms.

IDx-DR

The IDx-DR system combines results from multiple, partially dependent biomarker detectors, some of which utilise convolutional neural networks [16, 17]. Earlier versions of IDx-DR have been studied as part of the Iowa Detection Programme (IDP), and included separate algorithms for quantifying image quality and the detection of haemorrhages, exudates, cotton wool spots, neovascularisation, and irregular lesions [18].

IDP has shown good results in Caucasian, North African, and Sub-Saharan populations [18]. In an analysis done on images from the Nakuru Eye Study in Kenya, images from 3640 participants were analysed by the IDP algorithm as well as human graders. In 334 cases, the images were deemed to be of insufficient quality by human graders and IDP. In 20 cases, human graders decided that the image quality was insufficient but IDP analysed the images as gradable with no diabetic eye disease (DED). Discounting ungradable patients, the IDP sensitivity of 86.7% and the specificity of 70.0% were comparable to that of the human graders. Notably, none of the false negatives were assessed as having sight-threatening DED by human graders [19].

Similarly, the IDP has been verified against the Messidor-2 dataset, a publicly available, de-identified set of digital fundus colour images of 1748 eyes in 874 patients with diabetes [18]. The sensitivity for detecting referable diabetic retinopathy (rDR) against the consensus of three retinal expert graders was 96.8%, and the specificity was 59.4$. There were a total of six false negatives, none of which met treatment criteria [18].

The IDX-DR system improved on the IDP by the addition of DL features. It was verified against the same public dataset (Messidor-2) to determine whether the addition of DL algorithms provided an advantage [16]. While the already high sensitivity of IDP (96.8%) remained unchanged, there was a marked improvement in specificity—IDX-DR was able to achieve 87% specificity for rDR, greatly reducing the number of false-positive exams [16].

IDX-DR has also recently been verified in a real-life scenario within a Dutch diabetic care system [20]. Out of 1410 patients, 80.4% were judged to be of sufficient quality by three independent human graders, compared with the 66.3% accepted by the IDX-DR system, though issues with using the systems in-built re-imaging prompt have been highlighted [20]. The study's experts rated the images according to both the EURODIAB and ICDR grading standards. IDX-DR's had a sensitivity/specificity of 91%/84% under EURODIAB grading, and 68%/86% under ICDR. This large discrepancy between the EURODIAB and ICDR performance owes partly due to the judging of a single haemorrhage as at least MDR, according to ICDR scale [20]. The authors note that should this be reconsidered, IDX-DR would have a sensitivity/specificity of 96%/86%. This study was conducted on a population with low DR prevalence, which was attributed to good diabetic control and regular screening [20].

Recently Abramoff et al. published the results of the first preregistered clinical trial of an AI system [17]. An independent contract research organisation enrolled 900 patients in primary care clinics and considered IDx-DR performance in detecting more than mild DR (defined as Early Treatment for Diabetic Retinopathy Study (ETDRS) level 35 or higher, and/or diabetic macular oedema (DMO), in at least one eye). A patient outcome based reference standard using four-widefield stereoscopic fundus photographs equivalent to the area of the retina covered by the older, modified seven-field stereo film protocol and macular OCT imaging, performed by Wisconsin Reading Center certified retinal photographers, assessed for the ETDRS Severity Scale for DR as well as clinically significant and centre-involved macular oedema, was used. Out of the 900 patients, a reference standard was available for 852 patients, and out of those 819 patients (96.1%) had a diagnostic result provided by the IDx-DR system. Overall, the sensitivity for detecting more than mild DR was found to be 87.2%, with a specificity of 90.7%. The measured accuracy of IDx-DR was seemingly lower than the results of the same system on other datasets [16]. The two field 45° fundus images that IDx-DR relied on were taken by staff with minimal training and represented less than half the retinal area available from the multiple stereoscopic images judged by the human expert graders. In addition, the reading centre had additional information in OCT images available for diagnosing centre involved DMO [17]. Nevertheless, this study led to a landmark FDA approval as the first allowed fully-autonomous AI diagnostic system [21].

According to its approved use by the FDA, the IDX-DR system is designed to work in conjunction with the Topcon NW400 non-mydriatic fundus camera [17]. It uses a

macula-centred and a disc-centred image from each eye for analysis, and requires all four pictures to return a result [17]. Notably, the system had to be modified to analyse the previously mentioned Messidor-2 dataset, as it contains no disc-centred images. The system is able to deal with some quality issues thanks to partial overlap of those images.

RetmarkerDR

The RetmarkerDR software is a CE-marked Class IIa medical device developed in Portugal, and has been used in local DR screening for some years [22]. It has been implemented into an already existing, human-grader-based DR screening programme conducted in central Portugal in 2011 [22]. In this case, Retmarker is used in preliminary "disease" or "no disease" sorting, which then specifies the need for human grader assessment of the "disease" sub-group. The exact protocols and other details are explained in a study by Riberio et al. [22]. RetmarkerDR uses feature-based ML algorithms.

The main distinction of RetmarkerDR is its ability to compare current images to those screened at an earlier date, thus establishing whether disease progression occurred. Extending this further, the software is able to detect the rate at which new microaneurysms form and old microaneurysms disappear, called "microaneurysm turnover rate" [23]. This in turn appears to be a promising marker for future progression to diabetic maculopathy and worsening of DR [23–25], with decreasing microaneursym turnover rate noted after intravitreal ranibizumab therapy for DM [26, 27]. Similarly, DM patients treated with a dexamethasone intravitreal implant have shown a decrease in microaneurysm turnover rate [27]. Nevertheless, this is a topic that still requires a lot of further study.

RetmarkerDR was analysed in a large study looking at the potential use of AI. DR screening tools in the United Kingdom's national DR screening [28]. Tufail et al.'s study included two images per eye, one macula-centred and one disc-centred, from over twenty thousand patients screened at one London centre. This study considered three ARIAS—RetmarkerDR, EyeArt, and iGradingM, and provided a detailed screening performance and economic analysis of the outcomes. Compared with the arbitrated results of human grading, the measured sensitivities for RetmarkerDR were 73.0% for any retinopathy, 85.0% for referable retinopathy, and 97.9% for proliferative retinopathy. The false-positive rate was 47%.

EyeArt

EyeArt, similar to other automatic DR detection devices, has been categorized as a Class IIa medical device in the EU and is also commercially available in Canada. It is currently limited to investigational use in the US. It was developed by Eyenuk Inc., based in Los Angeles, USA, which offers another product—Eyemark for tracking DR progression which offers MA turnover measurements, as with RetmarkerDR.

EyeArt is able to take a variable number of retinal images from a patient encounter, automatically excluding images of insufficient quality or images of the outer eye. It can analyse images from the patient's previous encounters to estimate MA turnover. The system is cloud-based and offers an application programming interface for easier implementation into existing imaging and telescreening software.

EyeArt has been verified retrospectively on a database of 78,685 patient encounters with a refer/no refer result and a final screening sensitivity of 91.7% (95%CI: 91.3–92.1%) and specificity of 91.5% (95%CI:91.2–91.7%), as compared with the Eye Picture Archive Communication System graders, however only the abstract for the study is available online. Nevertheless, EyeArt achieved similar results in the aforementioned UK study looking into AI. DR screening viability, sensitivities of 94.7% for any retinopathy, 93.8% for referable retinopathy, and 99.6% for proliferative retinopathy [28].

EyeArt was also measured against the Messidor-2 dataset. The referable DR screening sensitivity was found to be 93.8%, with a specificity of 72.2%. It should be noted that the Messidor-2 dataset does not have a defined grading attached to each image, and a separate set of experts judged the images to produce a gold standard for each study [16, 29].

EyeArt was also implemented in a first-of-its-kind study of DR screening, relying on smartphone app-based fundus images combined with an automated AI screening system. Retinal images of 296 patients taken with a Remidio Fundus on Phone device were analysed [30]. Even though the EyeArt algorithms have not been trained on the use of smartphone-based fundus photography, it achieved a sensitivity of 95.8% for any DR, 99.3% for referable DR, and 99.1% for sight-threatening DR, with specificities of 80,2%, 68,8%, and 80.4% respectively [30]. However, these results might be considered in view of the limited scope of the study, with all images collected from a single source.

Google

A Google Inc. sponsored study validating a new, convolutional neural network based, DR detection algorithm was published in 2016 [10]. Similar to IDX-DR, the system outputs a number between 0 and 1, corresponding to the likelihood of referable DR being present in the analysed image. Therefore, the system can be tweaked for higher

specificity or sensitivity by adjusting the threshold at which referable DR is predicted.

The authors compared their DR detection system against the Messidor-2 dataset, as well as a retrospective dataset of 9963 images taken as part of routine DR screening in US and India. Approximately 40% of images from the second set were obtained after pharmacological mydriasis. US-board certified ophthalmologists were invited to grade the sets, seven graders were chosen for the Messidor-2 images and eight graders for the other data set, majority decision being set as a reference standard for referable retinopathy. Against the Messidor-2 images, depending on the operating point chosen, the DL algorithm achieved a sensitivity of 96.1% and specificity of 93.9% (tuned for high sensitivity) and sensitivity of 87.0%, specificity of 98.5% (tuned for specificity). The respective numbers for the second data-set analysed were 97.5%/93.4% (high sensitivity) and 90.3%/98.1% (high specificity).

This algorithm was validated in another community-based, nationwide screening program of DR in Thailand [11]. A total of 25,326 gradable retinal images from 7517 patients with diabetes were analysed for different DR severity levels and DMO. This is an improvement of the algorithm from detecting binary parameters of referable and non-referable DR into detecting the five severity levels of DR. In addition, this study used adjudication gradings among international retinal specialists from Thailand, India, and the United States as the reference standard. Compared with human graders in the screening program, the algorithm had significant higher sensitivity across all severity levels of DR and DMO ($p < 0.001$ for each category). For detecting different severity levels for referrals (moderate NPDR, severe NPDR, PDR, and DMO) the algorithm also had significantly higher sensitivity (0.97 vs. 0.74, $p < 0.001$) with slightly lower specificity (0.96 vs. 0.98, $p < 0.001$). These results can be translated into reducing the false negatives by 23% at the cost of slightly increasing false positives by 2%.

Singapore SERI-NUS

Another high impact study presenting a DL system for analysis of DR in fundus images was published by researchers from Singapore [12]. In this study, Ting et al. describe the development and validation of their algorithm using approximately half a million retinal images. The system demonstrated a sensitivity of 90.5% for detecting referable DR, comparable to professional graders on the same dataset at 91.5%, with a specificity of 91.6% (lower than professional graders at 99.3%). Impressively, it achieved a higher sensitivity for detecting sight-threatening DR at 100%, with human graders achieving 88.6%. However, this was again at a cost of lower specificity: 91.1% compared with 99.6% by trained graders. It showed no racial or other biases with comparable performance in different subgroups of patients; age, sex, and glycaemic control did not affect the accuracy of the algorithm.

The aforementioned study also described the algorithm's performance in the detection of referable glaucoma suspect and referable AMD using retinal fundus photographs. For referable glaucoma suspect, the definition used was a vertical cup-to-disc ratio of 0.8 or more and/or glaucomatous disc changes (e.g. disc haemorrhages, notching, etc.); referable AMD was defined as intermediate AMD or worse based on the (AREDS). The AUC, sensitivity, and specificity were 0.942, 96.4%, 87.2%, respectively, for the glaucoma algorithm, and 0.931, 93.2%, and 88.7%, respectively, for the AMD algorithm. This is one of the few AI systems described that could also detect non-DR pathologies. It may therefore be used in a DR screening setting to detect non-DR related, but potentially sight-threatening conditions (e.g. glaucoma suspect and AMD) that may require intervention in tertiary settings.

Bosch DR algorithm

The recently described automatic DR screening solution from Bosch, employs the use of a Bosch Mobile Eye Care fundus camera [31]. These images were later analysed by a convolutional neural network based AI software to deliver a disease/no-disease or insufficient quality output. Out of 1128 eyes studied, 44 (3.9%) were deemed inconclusive by the algorithm, with just 4 out of 568 patients having images from both eyes of insufficient quality. Interestingly the study compares the AI output, based on a non-mydriatic, single-field, colour image with grading assessment based on seven-field stereoscopic, mydriatic, ETDRS imaging done on the same eye. The Bosch DR Algorithm achieved impressive results with sensitivity, specificity, PPV, and NPV rates of 91%, 96%, 94%, and 95%, respectively. However, little is known about the graders or grading criteria employed in this study, no further reports of this algorithm's effectiveness are available as of writing of this article.

Retinalyze

Retinalyze is a cloud-based, fundus image analysing software, offering automated screening for DR, AMD, and more recently glaucoma. It is CE-marked as a Class I device. The submission of images is done through a website-based system, offering end-to-end encryption.

The first scientific reports of its efficacy were reported back in 2003, with multiple studies on cohort sizes of 137, 100, and

83 patients, respectively [32–34]. These early reports were performed with images taken on 35 mm film and utilised lesion-based detection methods. Good sensitivities of 93.1%, 71.4%, and 89.9%, and specificities of 71.6%, 96.7%, and 85.7%, were reported [32–34].

Since the above results were published, Retinalyze went through a long period of being commercially unavailable until it was reintroduced in 2013 in its current web-based form, with modern-era DL improvements. However, there are no recent studies regarding its efficacy in this current form.

Other systems

Gargeya and Leng published the results of their deep learning algorithm based on a training set of 75 137 fundus images [35]. These images were first annotated by medical professionals as having DR present or absent. Their final best-performing algorithm was tested against the Messidor-2 and the E-Ophtha datasets, the latter of which is a set of 463 fundus images provided by the French Research Agency. For Messidor-2, the system achieved 93% sensitivity and 87% specificity, for a disease or no disease classification. The authors analysed the software's ability to detect mild DR specifically against no DR and mild DR images from both aforementioned datasets, resulting in 74% sensitivity and 80% specificity for Messidor-2, and 90% sensitivity, 94% specificity for E-Ophtha. The authors ran the algorithm on a standard desktop computer and an iPhone 5, for an average processing time of 6 and 8 s, respectively. This shows great potential for internet-independent local analysis.

Li et al., a member of an initiative titled Healgoo, based in China, published a well-powered study detailing the development and validation of their DR detection DL algorithm with very promising result [36]. The authors used 71,043 images graded by selected ophthalmologists, with each image judged by three different graders. The images were graded according to the NHS diabetic eye screening guidelines. The image set used for development and internal validation had a relatively high prevalence of vision-threatening DR (18.5%) and DMO (27.5%).

For internal validation, the system achieved a sensitivity and specificity for vision-threatening DR of 97.0%, and 91.4%, respectively, with DMO performance of 95.0%/92.9%. In external validation, data from three real-world population studies were used—the National Indigenous Eye Health Survey (NIEHS), the Singapore Malay Eye Study (SiMES), and phases 2 and 3 of the Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). Overall, 13657 graded images were analysed by the algorithm achieving a sensitivity of 92.5%, and specificity of 98.5% for referable DR.

Discussion

Early detection and treatment of DR remains a priority in preventing diabetes-associated sight loss worldwide. Although effective screening systems have already been established in many developed countries, these rely on human graders, a resource both costly and in limited supply. In the last decade, large strides have been made in developing automated retinal analysis systems, some designed specifically to detect DR. These rely on different image analysis methods, generally based either on human-designed lesion detection algorithms, or ML, or a mixture of these two broad approaches.

Automated DR detection algorithms have several advantages over human-based screening. Firstly, they do not get tired and can grade thousands of images a day and are often able to provide results within seconds to minutes of taking the photos. Scaling automated DR screening programs are furthermore largely just a matter of acquiring more hardware. Nevertheless, it is unlikely that we will see wholly-automated DR screening anytime soon, and that human graders will still be needed to judge atypical or low-quality images, as well as for quality monitoring. Some however think that the first ever FDA approval of autonomous AI shows that such systems can be made sufficiently safe, efficient, and equitable, and therefore are acceptable for use in the US and elsewhere [21]. However, even with the introduction of semi-automated DR screening, the barrier-to-entry for establishing screening programmes in resource starved regions remains high, even on a local scale. Disregarding the cost of the automated systems themselves, such screening still requires significant expenditure in terms of equipment, trained staff to operate the equipment, secretarial staff, and much more. The performance of currently commercially available systems also suggests that at least some level of specially trained human grader verification is required, especially with the relatively low specificity of most systems.

An adequate balance between high sensitivity and specificity is the key to establishing cost effective screening programmes. With lower sensitivity, more cases of DR are missed, which goes against the main aim of a DR screening programme—the early detection of DR. With lower specificity, a relatively large number of false positives that warrant further examination are returned, which wastes the resources that automated DR screening is trying to spare. The newer DL systems offer the promise of impressively high sensitivities and specificities, but their performance in actual screening conditions remains to be seen.

Table 1 The summary of the major automated diabetic retinopathy (DR) detection tools, using feature-based and deep learning with regards to the training dataset, testing dataset, and diagnostic performance

DR classification system	Year	Development dataset	Ground truth	Clinical validation	Mydriatic or non-mydriatic	*n* (gradable)	% ungradable	Total n (including ungradable)	Referable DR AUC	Referable DR sensitivity	Referable DR specificity
EyeArt	2015	Proprietary algorithm	EyePACS human graders	EyePACS	Mixed Mydriatic (~46%), Non-Mydriatic (~54%)	101710	4.94%	107001	0.965	91.30%	91.10%
			Full-time and part-time optometrist and non-optometrist graders, with arbitration	Homerton	Mydriatic	19963	1.46%	20258	–	86.00%	54.00%
Abramoff et al.	2016	10,000 to 1,250,000 unique samples of each lesion type graded by one or more experts	Adjudication by 3 retinal specialists until full consensus for all cases using a single 45 degree FOV image	Messidor-2	Mydriatic	840	4.0%	874	0.980	96.8%	87.0%
Gulshan et al.	2016	128,175 images graded 3–7 times	Majority decision of 7 or 8 ophthalmologists for all cases using single macula-centred image with 45 degree FOV	EyePACS-1[a]	Mostly Non-Mydriatic	8788	11.60%	9963	0.991	97.5%	93.4%
				Messidor-2	Mydriatic	1745	0.17%	1748	0.990	96.1%	93.9%
Gargeya and Leng	2017	75,137 images from Kaggle competition graded by "a panel of retinal specialists" (with no additonal detail)	Not clearly described, likely the lesion-based grading that came with the public datasets using a single 45 degree FOV image	Messidor-2	Mydriatic	1748	–	–	0.940 (Any DR)	–	–
				E-Ophtha	Likely Non-Mydriatic	463	–	–	0.960 (Any DR)	–	–
Ting et al.	2017	72,610 images from multiple screening program and clinical studies graded by a	2 trained graders for all cases, using 45 degree FOV a single image. If there is a	SiDRP 14–15[a]	Non-mydriatic	35055	1.1%	35,948	0.936	90.5%	91.6%

Table 1 (continued)

DR classification system	Year	Development dataset	Ground truth	Clinical validation	Mydriatic or non-mydriatic	*n* (gradable)	% ungradable	Total n (including ungradable)	Referable DR AUC	Referable DR sensitivity	Referable DR specificity
		minimum of 2 graders, often with a retinal specialist for arbitration	disagreement, a retinal specialist generated final grade								
			2 graders; arbitration by 1 retinal specialist	Guangdong	Non-mydriatic	–	1.4%	15,798	0.949	98.7	81.6
			1 grader; 1 retinal specialist	SIMES	Mydriatic	–	1.8%	3052	0.889	97.1	82
			1 grader; 1 retinal specialist	SINDI	Mydriatic	–	2.1%	4512	0.917	99.3	73.3
			1 grader; 1 retinal specialist	SCES	Mydriatic	–	1.0%	1936	0.919	100	76.3
			2 ophthalmologists	BES	Mydriatic	–	0.4%	1052	0.929	94.4	88.5
			2 retinal specialists	AFEDS	Mydriatic	–	4.2%	1968	0.98	98.8	86.5
			2 graders	RVEEH	Mydriatic	–	10.9%	2302	0.983	98.9	92.2
			2 retinal specialists	Mexican	Mydriatic	–	0.5%	1172	0.95	91.8	84.8
			2 retinal specialists	CUHK	Mydriatic	–	0.0%	1254	0.948	99.3	83.1
			2 optometrists	HKU	Mydriatic	–	0.0%	7706	0.964	100	81.3)
Bosch	2017	Nearly 80,000 images, 5000 from India verified by 3 ophthalmologists, the remainder from EyePACS-1	Investigators following American Academy of Ophthalmology guidelines	Indian	Non-mydriatic	560	3.9%[c]	564	–	91.18% (Any DR)	96.9% (Any DR)
Krause et al.	2018	1.67 M images with clinical grades for train set 3737 fully adjudicated images for tune set	Adjudication by 3 retinal specialists until full consensus for all cases using a single 45 degree FOV image	EyePACS-2[a]	Mostly Non-Mydriatic	1813	0%	–	0.986	97.1%	92.3%
Abramoff et al.	2018	10,000 to 1,250,000 unique	Patient outcome based ETDRS and	Independent trial by contract	23.6% Mydriatic	819	4.0%	892	–	87.2%	90.7%

Table 1 (continued)

DR classification system	Year	Development dataset	Ground truth	Clinical validation	Mydriatic or non-mydriatic	*n* (gradable)	% ungradable	Total n (including ungradable)	Referable DR AUC	Referable DR sensitivity	Referable DR specificity
		samples of each lesion type graded by one or more experts	Clinically Significant DME grading from stereoscopic, 4-widefield field stereo images, and center-involved DME grading from macular OCT by three independent FPRC readers	research organisation Emmes, through a preregistered, prospective intention to screen trial protocol							
Healgoo	2018	71,043 images	Consensus grading of at least 3 out of the 21 ophthalmologists selected for the study	Internal		4588	6.4%	4900	0.989	97.00%	91.40%
			3 certified senior graders (>2 years experience) supervised by a retinal specialist	AusDiab	Non-mydriatic	–	–	4349	0.9688	94.59%	99.17%
			1 grader; 1 retinal specialist	SIMES	Mydriatic	–	–	6431	0.9621	93.94%	98.48%
			5 certified senior graders (>2 years experience) supervised by a retinal specialist	NIEHS	Mostly Non-Mydriatic	–	–	2877	0.9367	89.76%	97.57%
				External validation total		13394	1.9%	13,657	0.955	92.50%	98.50%
Ruamviboonsuk et al.	2019	As of Gulshan et al. (2016)	A panel of international retinal specialists	Nationwide screening program of DR in Thailand	Both mydriatic and non-mydriatic	25,326	15%	29,800	0.987	0.968	0.956

[a]Estimated from given number of nonclassifiable eyes

[b]Primary validation (e.g. validation set is drawn from population that overlaps with development set, but not the same patients). Lack of * means secondary validation (e.g. validation set from different population than development)

[c]Image-level

A number of systems for automatic detection of DR are already available commercially, with others in the pipeline and others still in early development. Nevertheless, regardless of development phase, most of those systems are still being actively developed with changes and improvements to detection algorithms, user interface, scalability, better detection of DMO, etc. in progress.

A head-on comparison of the available systems has so far proven very difficult, for multiple reasons. Depending on the system, the output may be tuned for a different outcome —DR present/absent, referable DR present/absent, no DR/referable DR/Sight-threatening DR outcome, or others. Sensitivity and specificity data between detecting any DR and referable DR are not directly comparable for multiple reasons, such as differences in reference standards and grader capabilities; the cut-off for referable DR for example, may not always correspond with the same ETDRS grading level between two studies. In real-world situations, the cut-off for referable DR may be different between regions, depending on available resources [37]. A comparison of major modern automatic DR detection tools, their training and validation datasets, and performance is presented in Table 1.

One of the biggest hurdles to overcome in the development of such systems is the acquisition of a sufficiently large set of retinal images on which to train and validate those algorithms. Confidentiality, data protection and other regulations are just some of the difficulties in obtaining a sufficiently large dataset. In addition, such images need to be human-graded and labelled as a reference standard, which is a significant time and cost sink. This further introduces additional uncertainty regarding the graders' level of accuracy. The publicly available Messidor-2 dataset, containing over 1700 labelled retinal images, has been used to validate and compare some of the studies. However, this dataset contains images of excellent quality, of a level that is not representative of real-life screening applications. Finally, developing DL systems against human grading as a gold standard limits the potential effectiveness of the system to that of human graders. It is not unimaginable that the possible performance of automatic retinal image analysis systems might surpass that of human graders in the future; this is however not quantifiable if human grading is taken as the objective truth. Some of us have proposed using patient outcome based truth rather than clinician agreement to validate such AI systems, and shown that it performs better than human graders [16–18].

Currently available systems offer various methods of implementation into new or already existing screening initiatives, such as dedicated computer programmes, browser-based solutions, or even mobile applications. It remains unclear how these different platforms affect the performance or workflows of screening programmes, especially with more such systems entering practical use.

Conclusions

The application of AI to ophthalmology is rapidly evolving. Several novel screening technologies have been described over the past years, and have reported robust performance in detecting DR. However, only a few of these are currently commercially available. Screening for DR using AI might yet play a large role in the prevention of blindness from diabetes. Future research is crucial in tackling some of the potential challenges (e.g. patients' acceptability, patients' confidentiality, medico-legal challenges, and unravelling 'black box' nature) in order to improve the adoption of these technologies within the healthcare setting.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest. GL and DT are the co-inventors of a deep learning system for retinal diseases. MA is the inventor on patents and patent applications of artificial intelligence and machine learning algorithms for diagnosis and treatment. He is a Founder CEO, employee, of and investor in IDx, Coralville, Iowa, USA.

References

1. World Health Organization. Global report on diabetes. World Health Organization; 2016.
2. Zheng Y, He M, Congdon N. The worldwide epidemic of diabetic retinopathy. Indian J Ophthalmol. 2012;60:428.
3. Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35:556–64.
4. 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.
5. Pieczynski J, Grzybowski A. Review of diabetic retinopathy screening methods and programmes adopted in different parts of the world. European Ophthalmic Review. 2015;9:49–55.
6. Mohammadpour M, Heidari Z, Mirghorbani M, Hashemi H. Smartphones, tele-ophthalmology, and VISION 2020. Int J Ophthalmol. 2017;10:1909.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436.
8. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284:574–82.
9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115.
10. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep

learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
11. Raumviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. npj Digit Med. 2019;2:25.
12. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, 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; 318:2211–23.
13. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31. e9.
14. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24:1342.
15. Abramoff MD, Staal J, Suttorp MSA, Polak BC, Viergever MA. Low level screening of exudates and hemorrhages in background diabetic retinopathy. Comp. Assist. Fundus Image Anal. 2000;15.
16. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci. 2016;57:5200–6.
17. 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 Digit Med. 2018;1:39.
18. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131:351–7.
19. Hansen MB, Abràmoff MD, Folk JC, Mathenge W, Bastawrous A, Peto T. Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru Study, Kenya. PLoS ONE. 2015;10:e0139148.
20. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol (Copenh). 2018;96:63–8.
21. US Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. News Release, US Food and Drug Administration; April 2018.
22. Ribeiro L, Oliveira CM, Neves C, Ramos JD, Ferreira H, Cunha-Vaz J. Screening for diabetic retinopathy in the central region of Portugal. Added value of automated 'disease/no disease' grading. Ophthalmologica. 2015;233:96–103.
23. Ribeiro ML, Nunes SG, Cunha-Vaz JG. Microaneurysm turnover at the macula predicts risk of development of clinically significant macular edema in persons with mild nonproliferative diabetic retinopathy. Diabetes Care. 2012;36:1254–9.
24. Pappuru RK, Ribeiro L, Lobo C, Alves D, Cunha-Vaz J. Microaneurysm turnover is a predictor of diabetic retinopathy progression. Br J Ophthalmol. 2018;103:222–6.
25. Haritoglou C, Kernt M, Neubauer A, Gerss J, Oliveira CM, Kampik A, et al. Microaneurysm formation rate as a predictive marker for progression to clinically significant macular edema in nonproliferative diabetic retinopathy. Retina. 2014;34: 157–64.
26. Leicht SF, Kernt M, Neubauer A, Wolf A, Oliveira CM, Ulbig M, et al. Microaneurysm turnover in diabetic retinopathy assessed by automated RetmarkerDR image analysis-potential role as biomarker of response to ranibizumab treatment. Ophthalmologica. 2014;231:198–203.
27. Kim ST, Jeong WJ. Microaneurysm turnover after the use of dexamethasone and bevacizumab to treat diabetic macular edema. J Korean Ophthalmol Soc. 2018;59:332–7.
28. Tufail A, Kapetanakis VV, Salas-Vega S, Egan C, Rudisill C, Owen CG, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. Health Technol Assess (Rockv). 2016;20:1–72. xxviii
29. Solanki K, Ramachandra C, Bhat S, Bhaskaranand M, Nittala MG, Sadda SR. EyeArt: automated, high-throughput, image analysis for diabetic retinopathy screening. Invest Ophthalmol Vis Sci. 2015;56:1429.
30. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye. 2018;32:1138.
31. Bawankar P, Shanbhag N, Dhawan B, Palsule A, Kumar D, Chandel S, et al. Sensitivity and specificity of automated analysis of single-field non-mydriatic fundus photographs by Bosch DR Algorithm—Comparison with mydriatic fundus photography (ETDRS) for screening in undiagnosed diabetic retinopathy. PLoS ONE. 2017;12:e0189854.
32. Larsen N, Godt J, Grunkin M, Lund-Andersen H, Larsen M. Automated detection of diabetic retinopathy in a fundus photographic screening population. Invest Ophthalmol Vis Sci. 2003;44:767–71.
33. Hansen AB, Hartvig NV, Jensen MS, Borch-Johnsen K, Lund-Andersen H, Larsen M. Diabetic retinopathy screening using digital non-mydriatic fundus photography and automated image analysis. Acta Ophthalmol Scand. 2004;82:666–72.
34. Larsen M, Godt J, Larsen N, Lund-Andersen H, Sjølie AK, Agardh E, et al. Automated detection of fundus photographic red lesions in diabetic retinopathy. Invest Ophthalmol Vis Sci. 2003;44:761–6.
35. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124:962–9.
36. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care. 2018;41:2509–16.
37. Wong TY, Sun J, Kawasaki R, Ruamviboonsuk P, Gupta N, Lansingh VC, et al. Guidelines on diabetic eye care: the International Council of Ophthalmology Recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018;125:1608–22.