Artificial Intelligence in Diabetic Retinopathy

11

Andrzej Grzybowski and Piotr Brona

Epidemiology of Diabetic Retinopathy

Over the last four decades the number of people living with diabetes has more than quadrupled from 108 million in 1980 to an estimated 422 million in 2014. At the same time diabetes prevalence among adults has almost doubled to 8.5% [1]. Future projections estimate that, by 2035, 592 million people will have diabetes, with the largest rise in low- and middle-income regions [2]. There is no doubt that diabetes constitutes a significant problem for global health and wellbeing. It is a disease that is prevalent all over the world, in the affluent, resource rich countries and much poorer developing countries. Diabetes can cause a number of significant complications, each of them associated with significant morbidity, requiring different, highly qualified medical personnel to diagnose and treat them. This poses a challenge for the local health services which often struggle with either delivering or funding the appropriate care.

A. Grzybowski (✉)
Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland

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

P. Brona
Department of Ophthalmology, Poznan City Hospital, Poznan, Poland

Diabetic retinopathy is one of the major complications of diabetes, estimated to be the leading cause of blindness among working-age adults globally [3].

Prevalence of DR and of proliferative DR (pDR) varies between type 1 and 2 diabetes and among the different regions of the world. Prevalence of DR among type 2 diabetics is reported between 20 and 40% in most studies. In type 1 diabetes, in Europe and the USA, reported prevalence vary widely between reports ranging from 36.5 to 93.6% [3]. Of those with DR an approximate one third may have vision threatening DR with either proliferative DR or diabetic macular edema (DME). Overall DR prevalence is higher among Western communities as compared to Asian regions [3]. Singapore is a notable exception to this, with a much higher prevalence of DR, but comparable to developed Western countries.

It appears incidence of DR among diabetics is also increasing in some regions. A study based in Spain found yearly incidence of DR to increase by almost 1% over an 8-year lifespan, from 8.09% in 2007 to 8.99% in 2014, with incidence of DME also increasing [4]. The increasing worldwide population, coupled with increasing prevalence of diabetes and increased incidence of DR all lead to increasing number of patients with ocular complications of diabetes.

Adding to the global burden of pDR and DME, these appear to be prognostic of other

A. Grzybowski (ed.), *Artificial Intelligence in Ophthalmology*,
https://doi.org/10.1007/978-3-030-78601-4_11

diabetes complications like nephropathy, peripheral neuropathy and cardiovascular events [3].

Conventional Screening Initiatives of DR: Telemedicine

There have been many DR screening initiatives throughout the world with varying degrees of coverage and longevity. Nevertheless, only a few countries were able to successfully establish and continue DR screening on national level, most prominently—UK and Singapore. It appears such programme is also functioning within Denmark, however very little information regarding it is available in English.

United Kingdom

Each country within the UK established their own national screening programme. The specific protocols and grading methods vary, however all are based on digital, colour fundus photography. The programmes cover all diabetics over the age of 12 years old with vision of at least light perception in one eye.

England

The NHS Diabetic Eye Screening Programme (NDESP) is a continuation of an English screening programme set up in 2006. Patients are screened on annual visits, with two fundus images per eye—one macula- and one disc-centred. Images are taken after mydriasis. These images are later digitally sent to one of centralised grading centres. The sheer scale of piloting and implementing such an initiative is impressive, in years 2015–2016 the programme invited more than 2.5 million diabetics to attend screening with an uptake of 82.8% [5]. It also gives us an important insight into the epidemiology of DR in the local population. Between 2015 and 2016 screening resulted in just under 8000 urgent referrals with suspected proliferative disease and over 52,000 referrals for suspected maculopathy or pre-proliferative diabetic retinopathy, with an overall rate of DR of 2.8%.

The aforementioned screening programme was expected to reduce the number of people considered legally blind in England from 4200 to less than 1000. It appears this goal has been accomplished with a 2014 report showing DR is no longer the leading cause of certifiable blindness in England and Wales for the first time in 50 years [5].

Wales

The Diabetic Retinopathy Screening Service for Wales (DRSSW), established in 2002, is a mobile screening service. Similarly to the English programme, two fundus images are taken per eye. Patients with sight-threatening DR are referred to a hospital-based retinal service. 30 screening teams serve 220 locations within Wales, achieving patient over of about 80%.

Scotland

Scotland started its DR screening in 2006. Qualifying patients are identified automatically using the Scottish Care Information-Diabetes Collaboration database. Screening is based on a single macula-centred image per eye, with mydriasis as required. Images are later sent to grading centres. Thanks to automatic patient selection, patient coverage is above 99%.

Northern Ireland

The Northern Ireland Diabetic Retinopathy Screening Programme (NIDRSP) was established in 2002. Its similar in functioning to the welsh programme DRSSW. Patients are referred to the programme by their GPs, with trained readers grading the photographs.

Ireland

DR screening was first identified as a key goal in 2008, but only introduced in 2013 under the name Diabetic RetinaScreen. The Irish programme screens diabetics 12 and older. Diabetic RetinaScreen supervises both annual fundus image-based screening and DR treatment and consists of both stationary and mobile community-based screening centers. The grading follows the English system closely both in terms of the grading matrix and the quality assurance protocols. According to the latest report, screening uptake is around 67% and rose considerably since the screening was first introduced [6].

Singapore

Singapore began widespread DR screening almost three decades ago in 1991. At that time Polaroid non-mydriatic fundus photography was chosen, as images could be taken by trained staff, instead of ophthalmologists. Images were reviewed by the local hospital-based ophthalmologist. At the time it was the first and only nationwide DR screening programme. The Singaporean screening initiative was revived in an updated version reflecting the technological advancements and possibilities and is now known as Singapore Integrated Diabetic Retinopathy Programme (SIDRP). Based on primary care clinics equipped with retinal cameras and specialised reading centres employing trained graders the programme aims for a result within 24 h of screening.

Cost effectiveness analysis has shown that this telemedicine-based model generated $173 of cost savings per patient compared to the previous screening model where family physicians graded the images themselves after special training [7].

Local Screening Initiatives

Other than the established national screening programmes, there have been a large number of smaller-scaled local screening initiatives all other the world. Some of those are similarly long-standing projects that control their population yearly, while most were discontinued or singular screening efforts. Even though so many screening projects were attempted, only the few aforementioned countries were able to implement nationwide screening, further highlighting the difficulty of such undertaking.

Automated Screening for Diabetic Retinopathy

The idea of adopting a computer programme in assessing fundus images for DR is certainly not new. First report, that we were able to find, of such endeavour was published in 1996 by Gardner and colleagues. Almost 25 years ago, the authors, established a neural network trained on 147 diabetic and 32 normal fundus images and aimed to train it to recognise particular features of an image—vessels, haemorrhages and exudates. Due to the many constraints, including computational capacity, each images was divided into small squares 20 or 30 pixels wide and later assessed by an ophthalmologist as containing either vessels, exudates, haemorrhages and micro-aneurysms or normal retina without vessels [8].

Another study done in 2000 describes a mixed technique where algorithms designed to enhance round features in an image were used to select for micro-aneurysms in a fundus image. This was later assessed by a neural network to determine the significance of the round feature extracted. This resulted in a sensible detection rate for images containing microaneurysms (81%) as compared to the opinion of a clinical research fellow [9].

Several studies explored the subject in the early 2000s, without the use of neural networks, relying on various pre-established image-analysis techniques, such as automated detection of anatomical landmarks in fundus images (optic disc, blood vessels, fovea etc.) coupled with specifically designed algorithms for detection of DR features. Among those, first three reports of a system, later known

as Retinalyze and discussed in further sections, were published showing relatively good sensitivities of 71–93% and specificities of 72–86%, these were based on small sample sizes reaching 137 patients in the largest study [10, 11].

All of those studies were done in the pre-digitisation era, meaning images, in the form of slides, taken from a fundus camera had to be scanned by hand. This was done using a slide reader or scanner to achieve a workable digital version of the image. The process was time consuming and required specialized equipment, and additional processing steps introduced potential image artefacts and loss of quality. The lack of centralised databases and digital storage of fundus images meant training and verification images were hard to acquire. As a consequence, most studies suffered from low number of images used, as compared to modern models using tens of thousands of fundus images to establish and validate a system.

Even though at that time automated screening was severely limited from a technical standpoint, a number of people already attempted devising suitable screening methods, recognising the potential of new technology to enhance or substitute human-based grading.

Deep Learning Algorithms

In subsequent years, with increasing digitisation, new ways of approaching the subject of automated image analysis were made possible. Up until 2010s experts designed algorithms for detection of specific features of DR like micro aneurysms or haemorrhages. In deep learning the software is presented with a fundus image as a whole and a pre-specified result for that image. Over the course of analysing many such images, often thousands, it starts being able to distinguish between images with different results. What separates one result from another does not have to be explicitly specified by its designers. The advent of deep learning-based DR detection revealed a significant improvement in the accuracy of newly developed or improved systems. Abramoff and colleagues reported how the introduction of deep learning techniques, allowed a significant improvement to the already established, classically designed, automated DR software—the Iowa Detection Program. Based on the publicly available set of fundus images with/without DR—the Messidor-2 dataset, the sensitivity improved from 94.4% to 96.8% and specificity from a confidence interval of 55.7%–63% to 87% [12]. For the Iowa Detection Program, deep learning features were added on top of already existing algorithms, many other initiatives attempted to establish entirely new deep-based learning DR detection software. Establishing automated or semi-automated screening, with the use of AI, will require striking a careful balance between sensitivity and specificity, imaging modality, gradeability of the images, all of which will need to be weighed against the potential cost. The cost-benefit balance is not universal and will vary depending on the relationship of those parameters with the relevant population characteristics, such as the prevalence of DR and sight-threatening DR, availability of treatment, cost and availability of trained staff etc. A recent paper explores the potential approaches to making a health economic assessment and safety analysis of implementing novel AI DR solutions into widespread screening [13]. Deep learning DR detection has been found to be cost-effective in developed countries, like Singapore and United Kingdom. However, there are no published studies looking into the feasibility of implementing AI DR screening in countries without a robust teleophthalmology screening programme setup beforehand and other resource-limited settings Table 11.1 [13].

Described further are several significant initiatives for AI-based diabetic retinopathy detection.

IDx-DR

IDx-DR is combined DR screening solution that incorporates the aforementioned DR screening algorithm with image quality assessment and feedback system. Submission of images is done using the IDx-DR client, which is a stand-alone piece of software. The IDx-DR client features a system for resubmission of images deemed to be of too low quality. The threshold for a positive result has been set as ‘more than mild’ diabetic retinopathy according to the ICDR grading scale or signs of diabetic macular edema. IDx-DR offers one additional result level of vision threat-

Table 11.1 The list of deep learning - based DR screening algorithms available at the end of 2020

Name of the software	Country of origin	Classification level	Comments
IDx-DR	USA	Per patient rDR/no rDR	First AI autonomous diagnostic device to be FDA approved. Class IIa medical device in EU
Eyeart	USA	Per patient rDR/no rDR	Second AI software to receive FDA approval. Approved by Canadian FDA Class IIa medical device in EU
RetmarkerDR	Portugal	DR/no DR Microaneurysm turnover rate	Previously used in various screening initiatives in Portugal Class IIa medical device in EU
SELENA +	Singapore	Per patient rDR/no rDR	Scheduled to be implemented into national DR screening in Singapore
Google algorithm	USA	Per picture rDR/no rDR	Studies surrounding real-world implementation based in India, Thailand. Currently no official software package available outside of research studies
MediosAI	India	Per patient rDR/no rDR	Integrated into an offline smartphone app to be paired with the Remidio fundus-on-phone device
Verisee	Taiwan	rDR/no rDR	Relatively new algorithm, recently approved by the Taiwanese FDA-equivalent government body
Pegasus	United Kingdom	rDR/no rDR	Operated by the Orbis non-profit organisation
RetCAD	Netherlands	rDR/no rDR	Detects referable AMD as well
Retinalyze	Denmark	Per image, retinal changes/no changes	Detects AMD related changes as well, also offers an automated glaucoma screening module
OphtAI	France	Per patient rDR/no rDR and DR grade	Also detects glaucoma and AMD

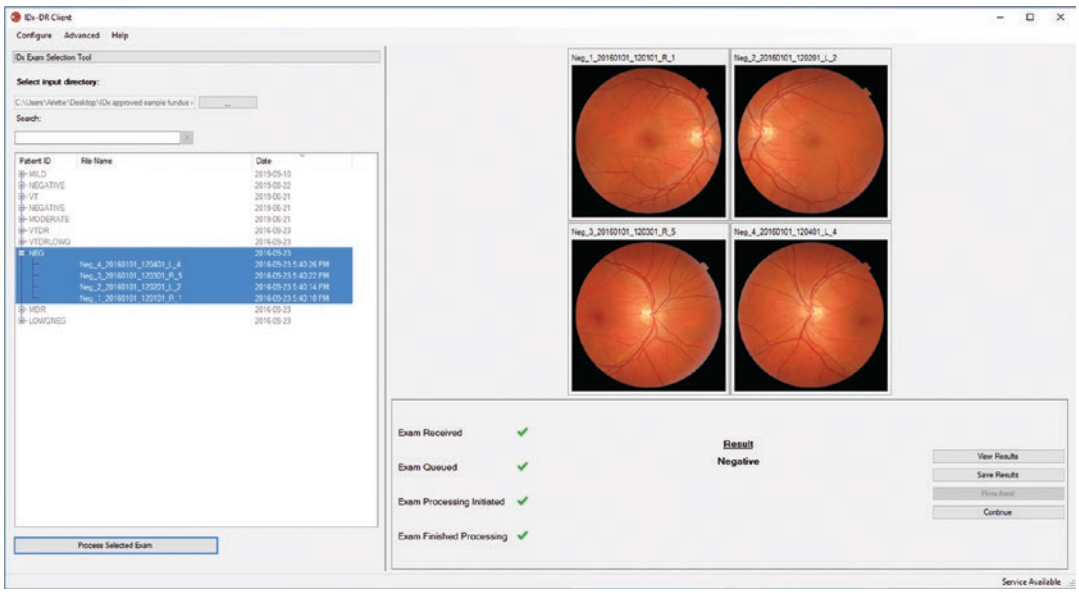

Fig. 11.1 IDx-DR image submission screen. Printed with Permission © IDx Technologies

ening DR, indicative of a suspicion of more advanced, possibly proliferative DR. Screening is based on four fundus images per patient, two from each eye, one macula- and one disc-centred and all images need to be submitted for a result to be produced. The algorithm is able to cope with some quality loss utilizing the overlap of the two image fields (Fig. 11.1).

Although on the front-end, the user is presented with a screening result in one of the four categories—no DR, mtmDR, vision threatening DR and insufficient quality, on the back-end IDx-DR produces a numerical value representing its assessment of likelihood of mtmDR. Currently it uses defined cut-offs to sort the patient into an appropriate category. Theoretically, this means that the IDx-DR output could be adjusted to maximise either sensitivity or specificity depending on the needs of a given screening initiative.

IDx-DR is the first autonomous diagnostic software and one of the very first AI-based software's in medicine to receive Federal Drug and Administration (FDA) approval. In a self-titled pivotal trial, IDx-DR software was studied in a real-world application. A little under 900 patients were screened using IDx-DR coupled with Topcon NW-400 automatic fundus camera in a primary care setting. The staff operating the IDx-DR client and taking the fundus images were not IDx-DR or clinical trial staff, but pre-existing employees of those clinics who underwent standardised training. This is important as in a scenario of large-scale DR screening deployment specialised staff, say in ophthalmology imaging may be harder to produce and acquire that the necessary technical equipment. In previous trials of IDx-DR and other AI algorithms the performance of the AI was compared to human grading with the same information available, which was mostly the fundus images. Sometimes to strengthen the human grading standard against which the AI was compared, several persons graded each image with a consensus grading that followed. This trial took an even more stringent, extreme approach—giving the human graders a lot more available information, while keeping the AI limited to the four fundus images taken by relatively inexperienced staff, albeit with an automatic fundus camera and selective mydriasis. This was compared to grading done on four stereoscopic, widefield fundus images taken by professional technicians and graded by an established, independent reading center—the Wisconsin Fundus Photograph Reading Center. Presence of clinically significant diabetic macular edema (CSME) was additionally established based on macula OCT imaging, which of course the algorithm had no access to. With odds stacked against it, the AI was still able to exceed all endpoints set before the trial began, endpoints at sensitivity of 87.2% (>85%), specificity of 90.7% (>82.5%), and imageability rate of 96.1% (among patients deemed imageable by the reading center). The landmark FDA decision to allow IDx-DR to operate within the United States was largely based on the results of this study [14]. In US, according to the FDA approved use, IDx-DR needs to be coupled with the Topcon NW-400 non-mydriatic fundus camera.

Previously to this study, were a number of studies published on IDx-DR, though none as significant. Notably its performance against the Messidor-2 dataset was significantly higher than in the above described trial, with 96.8% sensitivity and specificity of 87%. In another real-life study, performed in Netherlands, 1410 patients were screened within the Dutch diabetic care system. Three experts graded the resultant images according to ICDR and EURODIAB grading scales, resulting in significantly different algorithm performance depending on the scale used. For EURODIAB IDx-DR sensitivity and specificity was 91% and 84%, whereas for ICDR they were 68% and 86% respectively. The significantly lower performance when compared to ICDR criteria could all be attributed to a single aspect of ICDR—judging a single haemorrhage as at least moderate DR, the authors note that should this be changed the sensitivity changes from 68% to 96.1% [15].

This is a great illustration of how important grading criteria are. A number of differing criteria have been used in different studies so far, Eurodiab, ICDR, ETDRS, some studies use local grading guidelines, with each being one of the most significant parts affecting the outcome and final performance indicators published. The first question and most important question in establishing DR screening is 'what is the screening trying to accomplish?'. In the simplest form the aim of a DR screening initiative should be finding those patients, who will require a specialty ophthalmology visit before the next screening episode. This seems to hold true for established traditional screening programmes in developed countries. However, depending on the region and

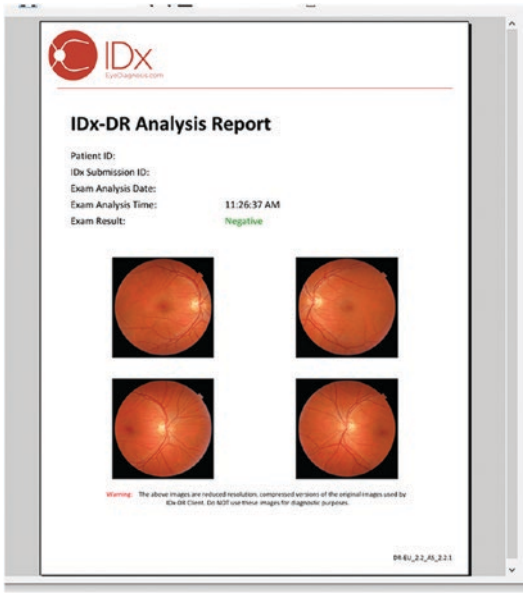

Fig. 11.2 IDX-DR result page. Printed with Permission © IDx Technologies

resources available this can change. In a setting of poorer countries, with many-fold less ophthalmologists and low availability of treatment, one might want to put the bar for referral higher. Nevertheless, the scale used to measure and qualify the retinal changes present, needs to be backed-up by the risk of DR progression and risk to vision at a given level (Fig. 11.2).

Retinalyze

Retinalyze is a DR screening system developed in Denmark. As mentioned, it is one of the very first published automated DR analysis programs, with initial reports of its efficacy starting in 2003, based on scanned 35 mm film fundus images. It features a web-based interface, with a per-image result. Images are submitted through the interface, utilising a secure internet protocol. Results are presented in terms of number/severity of detected retinal changes as either no changes, mild retinal changes, or severe changes. An interesting feature of Retinalyze is being able to see an annotated image with the detected retinal changes highlighted, therefore being able to get a glimpse into what led to the algorithms final result. Since its introduction it went through a period of inactivity and was reintroduced in 2013, with modern era machine learning improvements. It is certified with the Conformité Européenne (CE) level I under the previous regulations. Retinalyze additionally offers screening towards AMD and glaucoma, from the same fundus photos.

RetmarkerDR

Retmarker is a DR detection system originating in Portugal. It is one of the first AI screening tools successfully implemented into real life screening, not just for the purpose of a clinical trial. The central region of Portugal has a longstanding DR screening programme established back in 2001. In 2011 RetmarkerDR has been implemented into the already existing, human-grader based DR screening programme. This screening is based on several teams of photographers equipped with mobile fundus cameras. These screening units rotate between different healthcare providers covering their whole route over the course of 12 months and then repeating this cycle. Patients with diabetes and no history of DR treatment are invited for screening at a local health centre. These images are later collated and sent on a weekly basis to a centralized reading centre (Fig. 11.3).

The Retmarker software forms the first line of analysis for those submitted images. Images in which the algorithm detects signs of DR, or progression of DR in case of repeat screening, are sent for human grading, similarly with images deemed low quality by the algorithm. In this case Retmarker is used in the preliminary ‘disease’ or ‘no disease’ sorting, which then specifies the need for human grader assessment of the ‘disease’ sub-group. For quality assurance a certain number of DR negative results are sent for human analysis as well, with the human graders blinded to the AI decision.

Such implementation of an AI algorithm to detect DR relies on very high sensitivity, as false negatives will rarely be discovered, but can compromise on specificity. As long as it eliminates a significant number of images from human analysis, without missing cases of advanced disease,

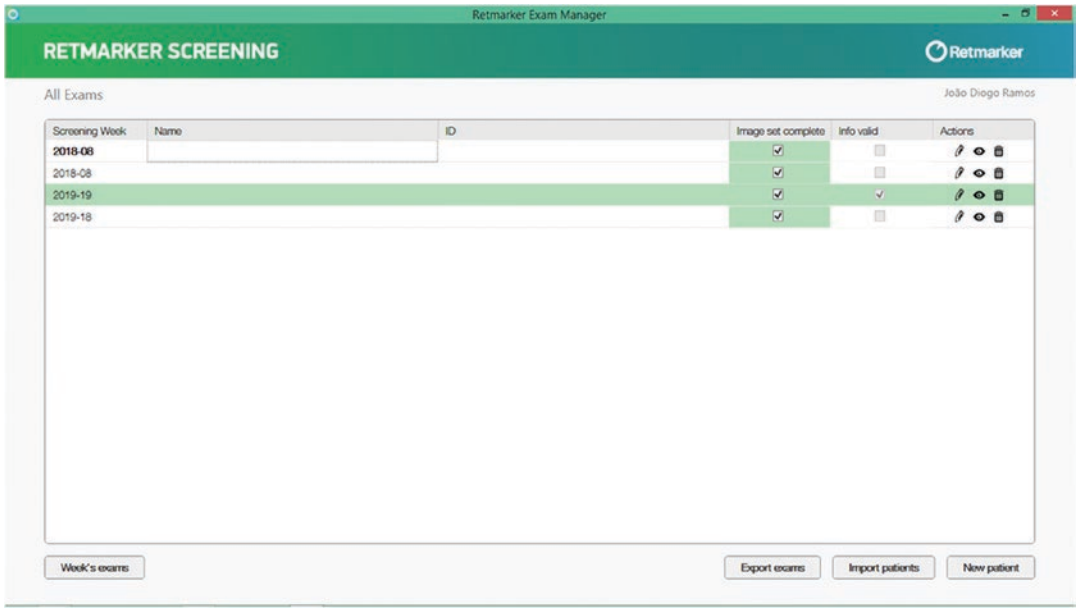

Fig. 11.3 RetmarkerDR exam manager

the process will likely be resource effective, as even a specificity of 50% means almost halving the human grader work.

A noteworthy feature that distinguishes Retmarker from other algorithms is its ability to take previous screenings into account. By comparing the fundus images taken on a previous screening visit, the system is able to track retinal changes and determine if progression occurred. This leads to another interesting avenue—tracking microaneurysms. Microaneurysms disappear over time and new ones form. Tracking those changes using traditional, human-grader based, methods is very labour intensive, but is virtually instantaneous for an AI. The rate of microaneurysms appearing and disappearing was named microaneurysm turn-over rate. A number of studies have been published showing this parameter is a promising predictive factor for future DR progression [9, 16–18]. Although these studies consistently linked increased MA turn-over to increased chance of DR, to establish a clinically significant and actionable link between lesion turn-over and diabetic retinopathy progression would require further work (Fig. 11.4).

In addition to being introduced as a part of screening in Portugal, RetmarkerDR was also studied in one of the only head-to-head comparisons of AI DR systems ever published [19]. This study, done for the purpose of assessing a potential introduction of autonomous DR detection software into the existing English DR screening programme, invited AI DR software makers to submit their algorithm for the testing. Three systems participated in the testing, RetmarkerDR, Eyeart and iGradingM. Because of technical issues iGradingM, a DR detection software born in Scotland, was disqualified from the study and its parent company has since dissolved. The study involved images taken from consecutive, routine screening visits of over 20,000 patients to an English DR screening centre, which were previously graded as per the national screening protocol were processed by the systems, and any discrepancies in grading between the AI and human-graders were sent to an external reading centre. Both the efficiency in detecting DR, referable DR and cost-effectiveness were studied [19]. The study concluded with the following sensitivity levels:

- EyeArt 94.7% for any retinopathy, 93.8% for referable retinopathy (human graded as either ungradable, maculopathy, preproliferative, or proliferative), 99.6% for proliferative retinopathy;

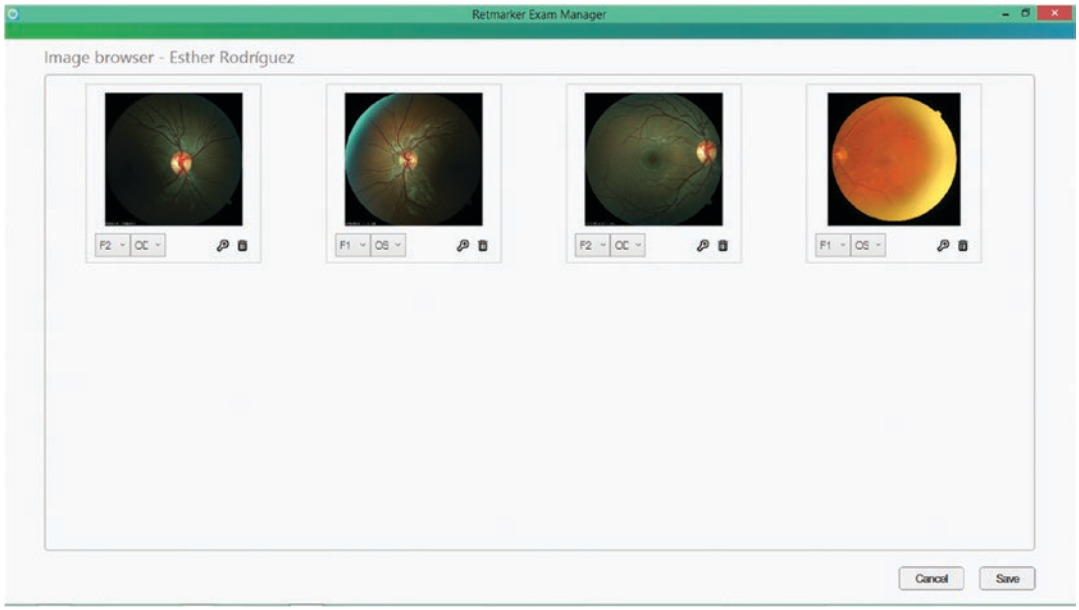

Fig. 11.4 RetmarkerDR image submission

- Retmarker 73.0% for any retinopathy, 85.0% for referable retinopathy, 97.9% for proliferative retinopathy.

Specificity:

- 20% for Eyeart for any DR
- 52.3% for Retmarker for any DR

Although the sensitivity levels are much higher for Eyeart, this is equalised by the reverse situation happening in specificity. Of note are the remarkably low specificity levels for both systems as compared to more recent reports and estimates of those and other software. It is important to realise that although the study was originally published in 2016, it started some years prior, during that period of time machine-learning and image analysis methods were improved dramatically and one can assume the algorithms established for this period of time improved as well.

Eyeart

Eyeart, the second software compared for the purpose of the British screening programme, as described above is being developed by Eyenuk Inc., based in Los Angeles, USA. It additionally offers another product—Eyemark for tracking DR progression which, similarly to Retmarker, offers MA turnover measurements. Eyeart is able to take in a variable number of pictures per patient, making it suitable for various screening scenarios without further adjustments needed, in contrast to some of its competitors. This solves a number of issues, as was illustrated by IDx-DR, which had to be specially modified to accept the single image per eye Messidor-2 dataset, instead of its typical input of two images.

Eyeart had been verified retrospectively on a database of 78,685 patient encounters (total of 627,490 images) with a refer/no refer result and a final screening sensitivity of 91.7% and specificity of 91.5%, as compared to the Eye Picture Archive Communication System (EyePACS) graders, however only the abstract for the study was available on-line. It appears Eyeart has decided to pursue this line of enquiry further with publishing of a full study, done on more than 100,000 consecutive patient visits from the EyePACS database. A total of 850, 908 images were analysed, collected from 404 primary care facilities between 2014 and 2015. Patients generally had eight images taken, four per eye; one

image of external eye, and a single macula-disc-centred image and an image temporal to the disc, though no patient was disqualified because of number of images taken or their resolution. The images almost evenly split between non-mydriatic at 54% and mydriatic at 46%. The final results in terms of detecting referable DR were 91.3% sensitivity and 91.1% specificity, in line with the previous partial results. Sensitivity for detecting higher DR levels that are treatable—either severe or proliferative DR was 98.5% and 97.1% for detecting CSME (as compared to human graders assessing the same fundus pictures). The systems accuracy did not seem to change depending on mydriasis, with 98.0% and 98.8% sensitivities for detecting treatable DR, in non-mydriatic and mydriatic encounters respectively. Only 910 patient encounters, less than 1%, were deemed non-screenable by Eyeart, of those 198 encounters were assigned as insufficient for full human grading previously. Nevertheless, of those 910 screening episodes over one third had severe or proliferative DR, the authors note that the system treats non-screenable patients as positive, for the purpose of patient safety [20].

Eyeart analysed the whole cohort of over 100,000 screening encounters, almost a million images in less than 2 full days [20]. Assuming an average 30 seconds of grading time per image, the same task would take about 7000 work-hours or about 4 full time graders working for a whole year, showing just how much faster computer analysis can be. Of course in the actual screening scenario no one is grading thousands of images at a time, with a quick result available within minutes of the screening being much more satisfactory, but AI can do that too, 24 h a day, every day of the year (Figs. 11.5 and 11.6).

Eyeart achieved similar results in terms of sensitivity, to the aforementioned UK study looking into AI DR screening viability, though there is a very considerable discrepancy in specificity between the two studies [19, 21]. As mentioned

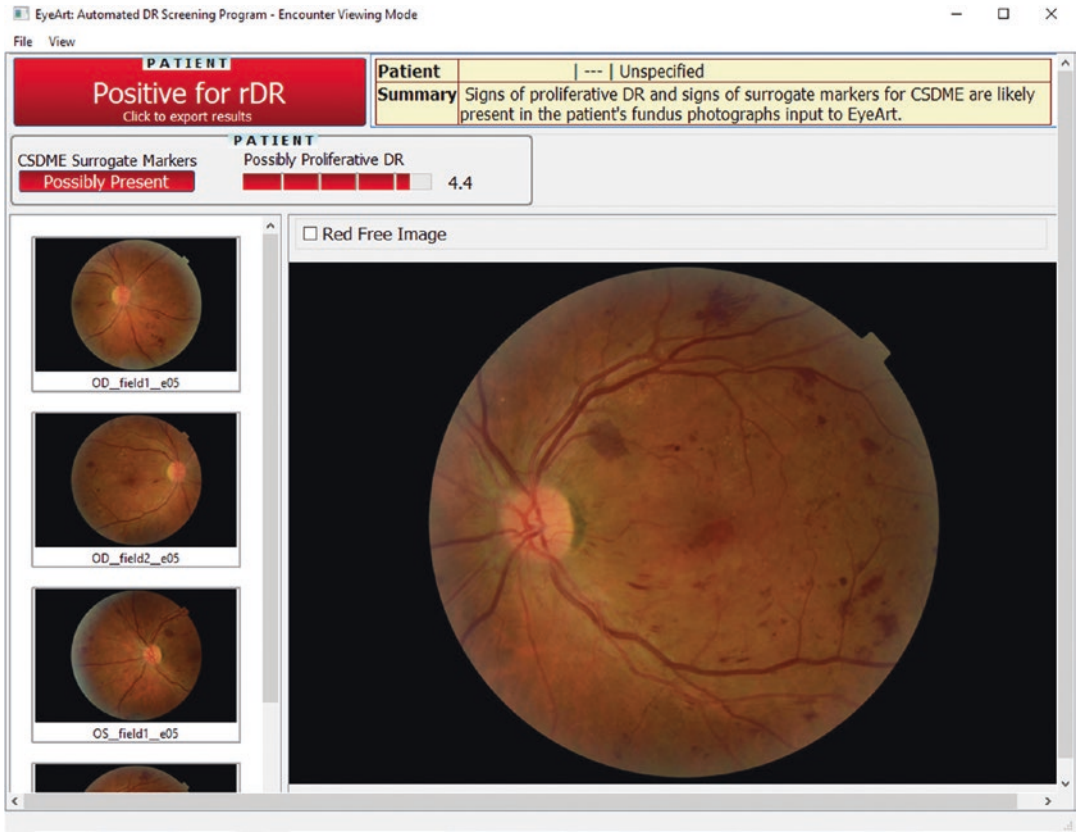

Fig. 11.5 EyeArt result page

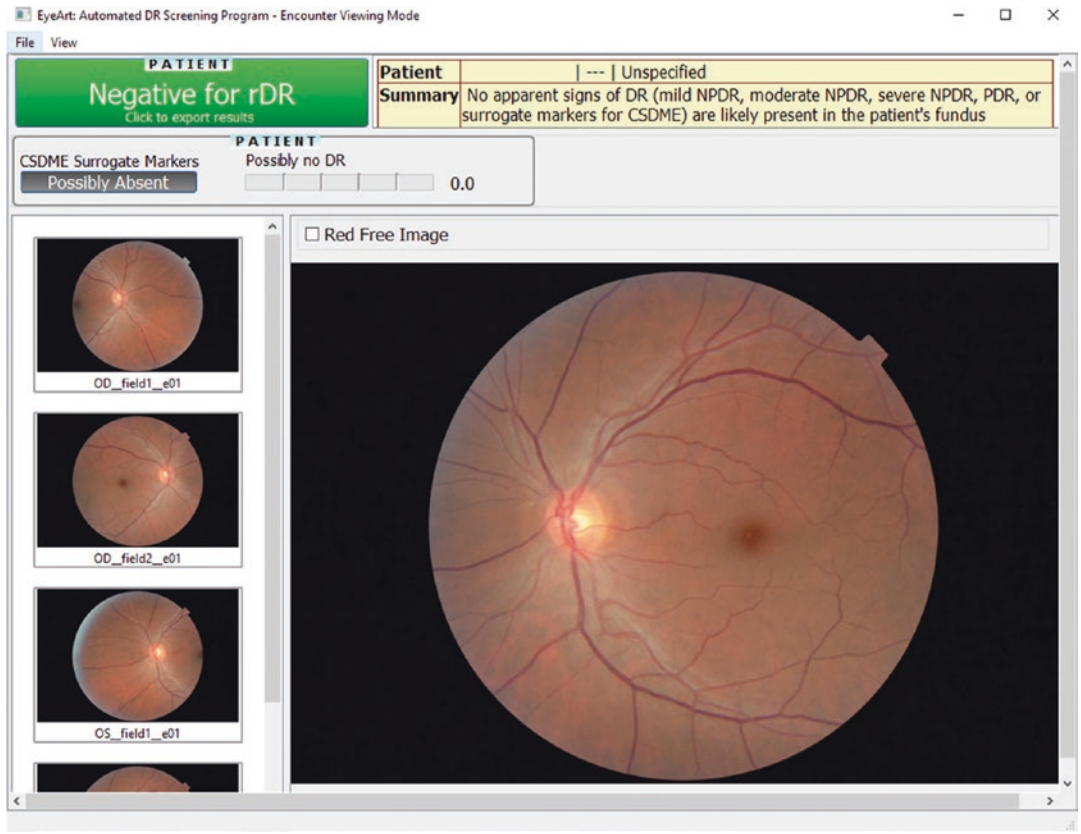

Fig. 11.6 EyeArt result page

before, these studies were not done in the same time-period, and further improvements to the system probably account for the increase in its accuracy. Indeed, the authors themselves describe the improvement that the 1.2 version of Eyeart (still based on traditional image analysis techniques) has undergone with the inclusion of multiple convolutional neural networks.

Eyeart was also measured against the Messidor-2 dataset. Referable DR screening sensitivity was 93.8%, specificity of 72.2%. Importantly this dataset does not have a predefined result or grading attached to it, therefore necessitating a separate set of graders to judge it for the standard that the AI is compared against, this grading is separate for each study, further hampering the ability to directly compare any systems involved.

Eyeart has recently published the results of its most robust clinical trial to date. The study was pre-registered, as with the IDx-DR pivotal trial, and comprised of a similar number of patients—893 patients screened in total. The screening was performed in primary-care clinics with two-field non-mydriatic fundus photography first and 4-field mydriatic imaging second. The study compared the ability of Eyeart to detect clinically significant DME, moderate non-proliferative DR or higher based on the two-field imaging with external reading centre (the Wisconsin Fundus Photograph Reading Center, as was used in the IDx-DR trial) grading decision using the four wide-field stereoscopic images per eye. For non-mydriatic screening EyeArt's was shown to have high sensitivity at 95.5%, good specificity at 86%, and gradeability of 87.5%. When dilating patients from the initially ungradable group, the systems overall gradeability rose to 97.4%, while retaining the same sensitivity and a rise in specificity of 0.5% to 86.5%. Although this trial did not involve OCT imaging for the detection of DME, in all other respects this trial appears similar to the IDx-DR clinical trial, with

similar results in terms of the both systems' accuracy.

Another result, perhaps even more surprising than the stellar performance of the AI, was a comparison based on a subset of the patients in this trial, that have undergone dilated ophthalmoscopy after the fundus imaging. A total of 497 patients were tested across 10 U.S. clinical centres, with some specialty retinal centres and others general ophthalmology clinics. This was compared against the adjudicated decision of the Wisconsin reading center based on the 4 wide-field stereoscopic fundus photography. Although the ophthalmoscope-based examinations had high specificity of 99.5%, this was coupled with an abysmal sensitivity of 28.1% overall. Even among the retina specialty centres the sensitivity rate was only 59.1% [22]. This shows that human-based grading, using ophthalmoscopy, as one of the tools commonly available in primary-care clinics, is very unlikely to be a sensible screening solution, if even ophthalmologists struggle with its accuracy.

The most recent study regarding Eyeart was done on 30,000 images taken from the English DR screening programme and followed a very similar protocol and analysis pattern to the only comparative study on AI in DR screening [19, 23]. Images from three different centers were graded according to the established national screening protocol. Among 30,405 screening episodes, Eyeart flagged all 462 cases of moderate and severe DR. Overall sensitivity for rDR was 95.7% for rDR and 54% specificity. Although the specificity is once again lower than in other studies, it is still a very significant increase from the 20% specificity in the previous study [19, 23]. The authors concluded that with the introduction of such an AI system into the currently established national screening protocol replacing the primary grader, the overall human grading workload could be halved.

'Google' Algorithm

The potential application of new artificial intelligence solutions for analysis of fundus images, DR particularly, caught the attention of not only smaller, independent teams and companies but also industry giants—Google. This is not Google's only foray into medical AI, with teams at Google collaborating to find solutions for automated analysis of histopathology images and other non-image analysis related publications. A Google inc. sponsored study introducing their automated DR screening algorithm was published in 2016 by Gulshan and colleagues. To develop the algorithm the authors gathered over 128,000 macula-centered images from patients presenting for their diabetic screening in India and US. To validate the resultant algorithm a random set of images from the same data source was chosen, those images were not used in creation of the algorithm. The image set for both development and validation consisted of mixed mydriatic and non-mydriatic photos from several different fundus camera models. Additionally, the authors tested the algorithm against the aforementioned French dataset—Messidor-2. The algorithm achieved impressive results at 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 Messidor-2 data-set were 97.5% and 93.4% (high sensitivity) and 90.3% and 98.1% (high specificity) [24]. Although these accuracy results are among the highest published, and the sample size is considerable, this study stood out in that it put a lot of emphasis on selection of human graders and their validation. Initially, for the development of the dataset, the study invited 54 US-licensed ophthalmologists or ophthalmology trainees at last year of residency, with each grading between 20 and 62,508 images. As a result, each image was graded between 3 and 7 times. Final DR status and gradeability of the image were set based on majority-decision. Graders were sometimes shown the images they have previously marked, to measure intra-grader reliability, or how often given the same image, the grader decides on the same result. Sixteen graders went through enough volume of images for this to be feasibly calculated, and the top 7 or 8 ophthalmologists, based on this measure, were chosen to grade all the images from the validation datasets. Inter-

grader reliability was also measured for 26 of the ophthalmologists. The mean intra-grader reliability for the 16 graders for referable DR was 94%, and inter-grader reliability for the 26 graders was 95.5%.

Even when choosing the most self-consistent graders out of several board-certified ophthalmologists, the mean agreement rate for referable DR images was only 77.7% for the EyePacs-1 dataset, with complete agreement among all eight graders achieved in less than 20% of referable DR images. Grader agreement was much better for non-referable DR images, with complete agreement on 85.1% of the nonreferable cases [24]. This highlights just how many caveats; the current universally acceptable grading method and gold standard of certified human grading can have. Out of 16 graders, on average, 4 out of 100 images were marked differently each time they were assessed by the same person. Out of 8 most self-consistent graders, only 20% of referable DR cases were judged as such by all graders.

Issues surrounding human grading were further explored in a subsequent 2018 study [25]. In it, authors build up on the previously described work by Gulshan in terms of developing an improved algorithm, expanding the training dataset and exploring different presently used grading protocols. The authors implemented a solution where the software outputs several numbers ranging from 0 to 1, each indicating its confidence that the image represents a given severity level of DR. This appears to be very similar to the back-end solutions implemented by IDx-DR, which also output's its confidence level in the result being more than moderate DR, although this is not presented to the end-user. This allows relatively easy adjustments to the systems sensitivity-specificity balance, focusing on either of those measures.

This study ended up with three different 'grading pools'—EyePacs graders, Certified Ophthalmologists and Retinal specialists. Additionally, an adjudication protocol was introduced in cases of disagreement by the retinal specialists with both asynchronous and live adjudication sessions until an agreement was reached [25]. This is in contrast to the first work, which relied only on majority decision. The new algorithm was based on well over 1.5 million retinal images, with 3737 images with adjudicated grading used to fine tune the system and 1958 images used for validation. The validation set was graded by three retinal specialists on their own, and was repeated later with face-to-face adjudication of all images between all three specialists. Additionally, three separate ophthalmologists graded the images on their own. The adjudicated grade was set as the gold standard for further comparisons.

All of the graders had high specificity—97.5%, 97.9% and 99.1% for ophthalmologists and 99.1%, 99.3%, 99.3% for retinal specialists. Sensitivities however were much lower with ophthalmologists ranging from 75.2% to 76.4% individually and 83.8% as majority decision as compared to the adjudicated grading [25]. Even the majority decision grading of retinal specialists showed room for improvement at 88.1% and individual sensitivities of 74.6%, 74.6% and 82.1%. Most cases of discrepancy between the majority grading of ophthalmologists and the adjudicated result stemmed from missed MAs—36%; misinterpreted image artefacts that can be construed as MAs or small haemorrhages—20%; and misclassified haemorrhages—16%.

After implementing the adjudication procedure and fine-tuning the autonomous system it achieved accuracy levels comparable to any of the retinal specialists or ophthalmologists involved [25].

A prospective trial was done to assess the real-world viability of the algorithm, utilising many of the lessons learned from the two above-described studies [26]. The trial was done in two hospitals in India on a total of 3049 diabetics attending their appointments in the local general ophthalmology and vitreoretinal clinics, as well as, telescreening initiatives. During their appointments macula-centered 40–45 degree fundus images were taken mainly with a Forus 3nethra camera, a compact, low-cost fundus camera [26]. All images were non-mydriatic and were not included in further therapeutic decisions for the patients, as they carried on with their appointments. All images were later graded by a non-

physician trained grader, a retinal specialist. All images from taken from one of the two centre, 997 patients total, also underwent grading by three retinal specialists with an adjudication process as in the previous study. Additionally, any images from the second centre with any discrepancies between any of the graders or algorithm output (5-point DR grading and DME status) were also adjudicated. The results, in terms of human grading accuracy in detecting rDR, were largely similar to those in the previous study—the four human graders had sensitivities between 73.4% and 88.8%, with specificities between 83.5 to 98.7%. The algorithm had comparable performance, at a sensitivity of 88.9% at the first centre and 92.1% at the second centre respective specificities of 92.2% and 95.2% [26]. The 'Google' DR algorithm was trained on images taken from many different cameras of which only 0.3% were taken by this specific fundus camera, yet it has showed very good performance on images taken using it, suggesting the algorithm is able to deal with different equipment being used to take the images [26]. Although the algorithm and its results appears very promising, with good accuracy, it does require further work in order to be used in a clinical setting, which the authors point out themselves. Firstly, as it currently has no image quality assessment capabilities, only images deemed gradable by the adjudication panel were included in this latest study. Additionally, as with all other algorithms, their place within and the precise protocols of widespread screening and integration into the existing clinical workflow or outside of it remains to be devised and assessed. This latest study was designed specifically for the algorithm not to interfere with established clinical set-up.

SELENA+, Singapore Algorithm

Singapore, one of the very few countries that have an established national DR screening programme, is also at the forefront of testing deep learning for DR detection. Ting and colleagues used images from the on-going Singapore National Diabetic Retinopathy Screening Program (SIDRP), which were additionally graded by two senior non-physician graders and adjudicated by a senior retinal specialist in case of conflicting grading. Overall 72,610 images were included in the training dataset, taken from the years 2010–2013, and further 71,896 from years 2014–2015 were used for the primary validation dataset. The system was additionally validated using images from multi-ethnic populations from Singapore, and using images taken in screening studies from around the world—China, African-American Eye disease study (US based), Royal Victoria Eye Hospital (Australia), Mexico and University of Hong Kong. These studies included between 1052 and 15,798 images for a total validation dataset of 112,618 images, more than 56 thousand patients. Reference standards varied between the different studies, but all included at least two graders, with the largest study by image volume (n = 15,798) also including retinal specialist arbitration.

For the primary validation, that is the data from SIDRP years 2014–2015, the system demonstrated a sensitivity of 90.5% for detecting referable DR, comparable to professional graders on the same dataset at 91.5%, as compared to the final retinal specialist arbitration decision. Specificity of this solution was 91.6%, lower than that of professional graders at 99.3%. Interestingly the system proved better at detecting sight-threatening DR at 100%, with trained graders rated at only 88.6%, again, at a cost of the lower specificity. As the study included multiple ethnic populations, yet was devised only on the basis of SIDRP images, the authors analysed if it showed racial or other biases. This was made possible by the large racial diversity among the validation datasets—Malay, Indian, Chinese, White, African-American and Hispanic. The algorithm achieved comparable performance in different subgroups of patients by race, additionally age, sex, and glycaemic control did not affect the accuracy of the algorithm.

Verisee

Verisee, an algorithm developed in Taiwan, was described in a recent paper. The algorithm was developed based on single-field images taken

previously at the National Taiwan University, with a single fundus camera [27]. The images were graded by two ophthalmologists undergoing fellowship training, with an experienced retinal specialist employed for adjudication. The algorithm was trained on about 37,000 images, with 1875 images used for validation. The validation dataset was not used for training, but was taken with the same camera at the same location. The algorithm achieved 92.2% specificity and 89.5% sensitivity for any DR, and 89.2% and 90.1% for rDR. The algorithm exceeded the sensitivity for detecting rDR achieved by ophthalmologists in this study, which was calculated at 71.1%, and did much better than internal physicians at detecting any DR (64.3% sensitivity, 71.9% specificity, based on diagnosis available in chart records) [27]. Although these results are promising, due to the low volume and homogeneity of validation dataset, the performance of the algorithm in other scenarios remains uncertain. Nevertheless, the algorithm has been approved by the Taiwanese FDA-equivalent body and is scheduled to be implemented into real-world screening in Taiwan in the near future.

RetCAD

A recently published system, developed in the Netherlands, allowing for joint detection of DR and AMD from fundus images [28]. It is the only study to show algorithm's effectiveness at screening for both AMD and rDR at the same time. The validation dataset was rather small, relative to other studies described here, and comprised of 600 images. Nevertheless the software achieved good accuracy and was able to distinguish between rDR and referable AMD rather well with sensitivity of 90.1% and specificity of 90.6% [28]. Unlike the SELENA software, which can also detect both AMD and DR, both diseases were tested at the same time, instead of testing the accuracy against AMD and DR on separate data sets [29]. RetCAD was tested against the publicly available datasets of Messidor-2, for DR detection and Age-Related Eye Disease Study dataset for AMD, achieving favourable results. However, for all of the above datasets, including the development and validation dataset, only images of good quality were chosen.

OphtAI

OphtAI is a relatively new entry to the commercial AI DR detection market. It originates from a joint venture of two French medical IT companies Evolucare and ADCIS, it was developed in France and possesses a class IIa CE certification. The DR algorithm was developed based on a dataset of over 275,000 eyes from a French medical imaging database [30]. It is mostly a cloud-based service accessible through a web interface, my.ophtai.com, which allows between 1 and 6 images per patient to be sent for analysis and offers a DR grading result in a few seconds along with a confidence rating and heatmap of the suspect retinal changes. OphtAI is also available as a locally hosted platform, dependent on local regulations. While software additionally detects referable DR, diabetic macular edema, glaucoma and AMD from fundus images, there are plans for the next version to detect general eye health in addition to detection of over 10 specific pathologies and 27 disease signs to expand the number of detected pathologies to over 30. The DR detection algorithm was compared against the Messidor-2 dataset with very promising results [30, 31]. We would expect further publications related to the verification and efficacy of this algorithm in the coming years.

Other AI DR Solutions

The initiatives described so far focused mostly on the aspect of image analysis. One of the hurdles to go through with their development regarded equipment and technique used to take the fundus images, and how that might affect the system's diagnostic ability or its image quality detection protocols. Use of different fundus cameras by many different technicians can introduce a lot of variability in picture quality, its resolution or sharpness. IDx-DR, for example, is only approved for use in US when coupled, not only with a sin-

gle brand of fundus devices, but with a specific fundus camera—the Topcon NW-400. Other initiatives employ a number of computational techniques to normalize each image to a standard deemed appropriate for the system. Another line of thinking is that using images from multiple fundus cameras in training the algorithm may help it ignore the non-relevant, fundus camera-dependent changes in the images. This strategy appears to be working with most developers reporting their systems as having no significant impact on final accuracy in regards to fundus camera used. This issue is particularly important in case of low-cost or mobile fundus cameras. Introducing DR screening in low-resource regions of the world is costly not just in terms of grading but also in terms of equipment cost and portability, establishing permanent, stationary screening points is unlikely to be viable in settings with low population density and low patient mobility. Even in developed, wealthy countries, wide-spread screening is often done utilising mobile screening units, as exemplified by some of the UK-based screening strategies. The rapid development of AI in diabetic retinopathy did not go unnoticed by companies that already function in the fundus image field, with companies developing dedicated AI DR screening solutions for their existing fundus imaging hardware.

DR Detection with the Use of Mobile Devices

Another widespread invention of the digital era—the smartphone, and its relative cheap cost and ubiquity appears promising in regards to mobile, low-cost screening. In one study, images taken Retinal images of 296 patients taken with a smartphone-based add-on and software—'Remidio Fundus on phone' device were analysed by Eyeart software. Even though the Eyeart algorithms have not been trained on the use of smartphone based fundus photography, it achieved sensitivity 99.3% for referable DR and 99.1% for sight-threatening DR, with specificities of 68.8% and 80.4% respectively [32]. Since that study was done, Remidio have developed their own in-house DR analysis software, embedded into their current generation Fundus on phone devices (Fig. 11.7).

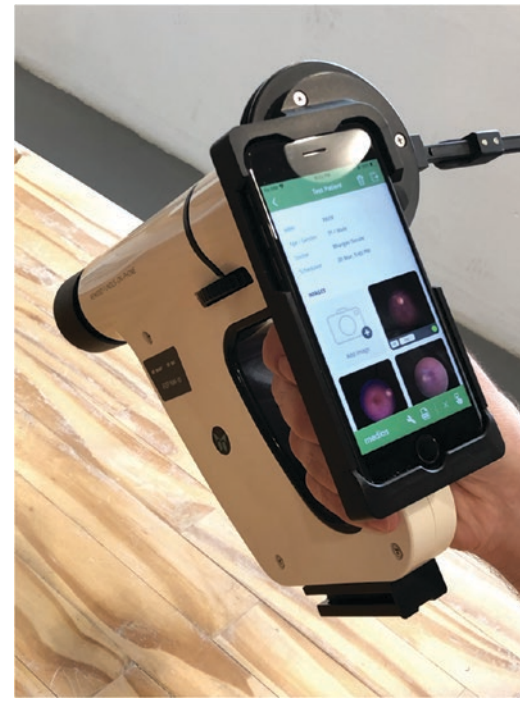

Fig. 11.7 Remidio FOP device. Printed with permission from Remidio

The software side of Remidio's DR detection system was named Medios AI. These results have since been replicated in another similar study by V and colleagues, where 3-field, dilated retinal images taken with the Remidio mobile camera were compared to the diagnosis of a vitreoretinal speciality resident and specialist based on the same pictures. The images were taken by a healthcare professional with no experience in using fundus cameras, with the offline system achieving similarly high accuracy results [33]. In a similar study done on 297 the systems performance was measured against that of 2 vitreoretinal specialists, with final sensitivity and specificity of the AI in detecting referable DR at 98.8% and specificity of 86.7% [34]. This was further corroborated by a study looking into 900 adult subjects with diabetes in India, where five retinal specialists graded images taken with the Remidio mobile camera for any DR or rDR. This was later compared to the Medios AI software running offline on an Iphone 6, a 6-year-old mobile device that currently costs less than 200

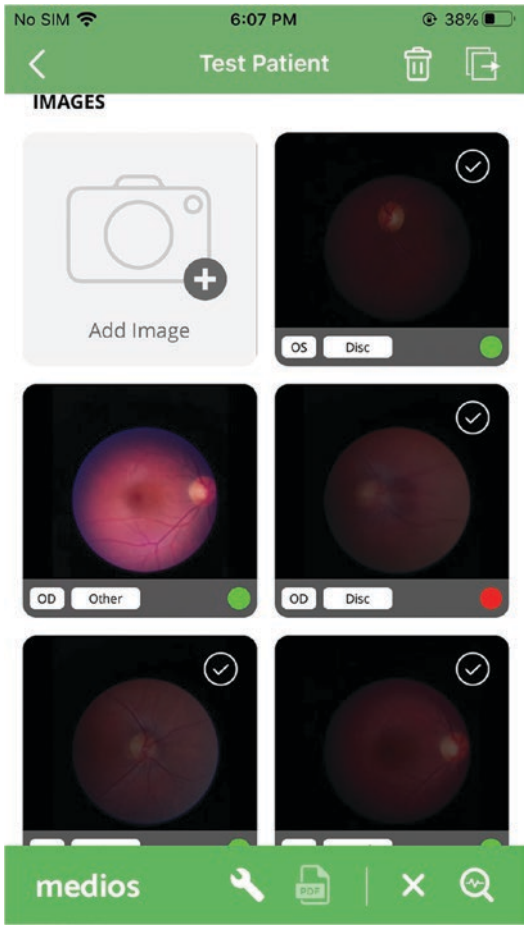

Fig. 11.8 MediosAI Image selection. Printed with permission from MediosAI

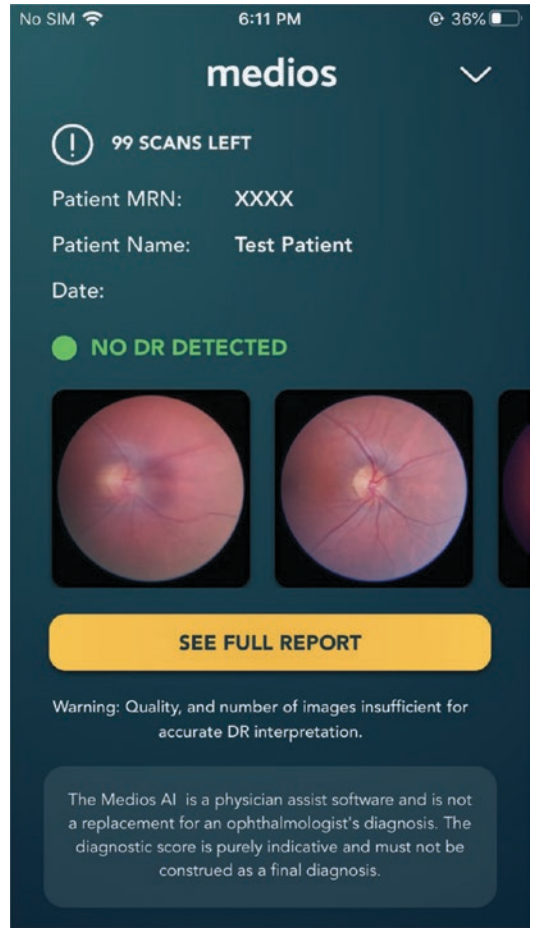

Fig. 11.9 MediosAI report. Printed with permission from MediosAI

USD for a refurbished model. Medios AI achieved good results with sensitivity and specificity pair for any DR of 83.3% and 95.5% and for rDR 93% and 92.5% [35]. For Medios AI, all studies so far compared AI and grader performance on the same source material - pictures taken with the mobile camera. A study similar to that done by IDx-DR and Eyeart, where the chosen system is compared to diagnosis based on professional, multi-field fundus imaging might provide additional insight and comparability of those systems with the mobile approach (Fig. 11.8).

The big difference in Remidio's DR screening system, other than implementing it directly into the fundus imaging device, is performing the analysis entirely offline, without need for internet access. Although the access to wireless internet sources is spreading all over the world, this can be a hugely important factor in screening remote and underprivileged communities, where internet access is sometimes not possible and very often unreliable. This approach is picking up steam with more mobile, smartphone based or smartphone aided fundus imaging solutions being studied and considered for adoption in DR screening. Smartphones, coupled with a compatible mobile fundus camera attachment or device provides a low cost, highly mobile and highly scalable DR screening solution, especially if the analysis is integrated into the smartphone itself. A recent study conducted in India compared effectiveness of four such devices in human based DR grading [36] (Fig. 11.9).

It appears the company Bosch has also taken a similar approach in improving its 'Bosch Mobile Eye Care' fundus camera and developing an in-house DR diagnostic algorithm to be implemented within the fundus camera itself. Single-field images taken with their camera, without pharmacological mydriasis, were analysed by a convolutional neural network-based AI software to deliver a disease/no-disease or insufficient quality output. The system is cloud based and would require internet access. 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. The study compared AI's performance with grading based on 7-field stereoscopic, mydriatic, ETDRS imaging done on the same eye. Bosch DR Algorithm achieved good results with sensitivity, specificity, PPV, and NPV rates of 91%, 96%, 94%, and 95% respectively [37]. However little is known about the grading criteria employed in this study, in contrast to other similar works, it employs purely a disease positive/negative criteria, rather than the more useful rDR, non rDR distinction [37]. Unfortunately no further reports of this algorithms effectiveness are available at this time.

Even though mobile screening does appear very appealing, and as exemplified above the results are very promising, it is conceivable that the lower image quality obtained when using mobile fundus cameras might affect the accuracy of the AI system used to grade it. A recent study compared the performance of a deep learning based DR detection system against a benchmark, curated image set, taken with a desktop camera against its accuracy with images taken with a handheld fundus camera [38]. Although the software, dubbed Pegasus, did exceptionally well on the curated, desktop dataset with 93.4% and 94.2% sensitivity and specificity, this did not translate to equal detection rate in the handheld camera images with a statistically significant decrease in accuracy. The parameters for the handheld camera dataset were 81.6% sensitivity and 81.7% specificity—a drop of more than 10% for each of the parameters [38]. Mobile screening setups and portable cameras are very attractive means for introducing widespread screening. However, testing on curated, high quality data sets will overestimate the real-world testing accuracy. Testing the software should be done in a scenario as close to the desired implementation as possible, to achieve accuracy metrics that will be true to real-life screening.

New Technologies in Retina Imaging and DR Screening

Although most DR screening efforts are directed towards analysis of fundus images, there have been significant advancements in employing AI for analysis of optical coherence tomography (OCT). OCT is commonly used in assessing and monitoring DR and DME on an individual patient basis. Several metrics like central macular thickness help us establish some objective parameters, nevertheless the evaluation of OCT scans is still subjective, user-dependent, similarly to evaluating fundus pictures. A further development of OCT—OCT angiography (OCTA), allows for non-invasive tracing of retinal and choroid vasculature, the role of OCTA in common ophthalmic practice is not firmly defined, and there are few objective quantifications possible. First attempts at using OCTA data for machine-learning and automated analysis of DR patients have already been made. OCTA data from 106 patients with type II diabetes and either no DR (n = 23) or mild non proliferative DR (n = 83) was used to train the algorithm to detect DR features from superficial and deep retinal maps [39]. Using a combined approach of using both layers, the system demonstrated overall accuracy of 94.3%, sensitivity of 97.9%, specificity of 87.0%, and an area under curve (AUC) of 92.4% [39]. Although the relatively high reliability measures are promising, it is important to note that the validation was done on the training subset. Nevertheless, the study has shown that OCTA can be subjected to deep learning and automated analysis and we may very well see more such initiatives in the future. The specific computational techniques for detecting DR from OCTA have been further explored in a recent study comparing different neural network approaches to analysing OCTA

and their results. The best performing algorithm achieved accuracy of 0.90–0.92 [40].

Teaching general practitioners (GPs) to take photos with a mobile fundus camera and subsequently grade them, might be an alternative method of widening access to DR screening, without the use of AI or automated systems. A recent study looked into training GPs in Sri-Lanka to take and grade fundus photos taken with a mobile camera (Zeiss-Visuscout100®). The GPs underwent a training programme delivered by two retinologists, however of the nine doctors that undertook the training only two with the best test grading results were chosen for the study. The GPs took and graded non-dilated and subsequently mydriatic fundus images, their performance was graded against a decision of a retinal specialist after performing a dilated fundus examination using slit lamp biomicroscopy and indirect ophthalmoscopy. Assuming ungradable subjects as referable, the two GPs achieved sensitivities for detecting rDR of 85%, 87% with specificities of 72%, 77% for non-mydriatic screening, rising to 89%, 93% specificity and 95%, 96% sensitivity after mydriasis. Although, this shows that training GPs to screen for rDR is theoretically feasible and can achieve good diagnostic accuracy, both the availability of GPs and their ability to take on additional workload is limited. In the aforementioned study only the two best performing GPs (measured as agreement with the retinal specialist on a test image set) were included, unlike an automated system the accuracy would likely vary between different GP graders [41].

Approaching the issues surrounding DR screening from a different direction is RetinaRisk, a software developed in Iceland. RetinaRisk aims to decrease the overall burden of yearly DR screening by safely extending the time between screening for part of DR population. Although not explicitly derived from machine learning, it is based on analysis of extensive datasets. The algorithm takes in patients' parameters, such as gender, age, HbA1c level, DR status, diabetes type and duration, and blood pressure level. As a result, the algorithm presents a recommended time till next screening interval, which may be longer than the traditionally accepted yearly interval, but may also be shorter, for a subset of patients with high risk for developing DR complications. In a recent study based in one Norwegian ophthalmic practice between 2014 and 2019 average screening interval was extended to 23 months as compared to 14 month average for the control group with fixed screening intervals [42].

Conclusions

Deep learning DR diagnostic software is currently a rapidly developing topic. During the last decade we have seen the concepts surrounding automatic DR screening evolve from few expert-designed algorithms with varying measures of accuracy to a multitude of different approaches employing the newest developments in deep learning and other fields. We have seen progressively more robust studies emerge, proving the diagnostic or decision-support algorithms to be accurate and reliable, some basing on millions of images, others with particularly rigorous setting of their gold-standard. During the last 2 years, a number of software packages have been approved by regulatory bodies around the world and are well on their way to be implemented into widespread screening in the respective countries. Following the general worldwide trend, increasing emphasis is being placed on mobile solutions, which may prove to be a better fit for resource starved regions. Although the body of evidence speaking for the various algorithms is quite large and constantly increasing, there are significant shortcomings in our current study of AI in DR. Virtually all of the current studies looking into and measuring DR algorithms are sponsored or dependent on the respective algorithm's' company. Independent studies are very few and far between. For a long time the only independent and the only robust comparison available, published by Tufail and colleagues in 2016, compares algorithms tested in 2013. Since that time deep learning and related concepts progressed almost beyond recognition, and many of the algorithms described here are being constantly updated. This situation changed only recently with the publishing of a study comparing multiple AI DR detec-

tion algorithms in an anonymised fashion, which made it clear the algorithms' accuracy can vary significantly, but unfortunately not giving readers any insight into the performance of any given algorithm [43]. We recently published a much smaller study comparing two algorithms on a local dataset [44]. Nevertheless independent studies, particularly comparisons or studies establishing objective criteria through which the respective algorithms could be compared are missing, with organisations, end-users or consumers left with a considerable dilemma when trying to choose and algorithm for screening their local population.

References

1. Klein BEK. Overview of epidemiologic studies of diabetic retinopathy. Ophthalmic Epidemiol. 2007;14(4):179–83.
2. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract. 2014;103(2):137–49.
3. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. Eye Vis [Internet]. 2015 Sep 30 [cited 2020 Feb 7];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4657234/
4. Romero-Aroca P, de la Riva-Fernandez S, Valls-Mateu A, Sagarra-Alamo R, Moreno-Ribas A, Soler N. Changes observed in diabetic retinopathy: eight-year follow-up of a Spanish population. Br J Ophthalmol. 2016;100(10):1366–71.
5. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003–2016. Acta Diabetol. 2017;54(6):515–25.
6. Pandey R, Morgan MM, Murphy C, Kavanagh H, Acheson R, Cahill M, et al. Irish National Diabetic RetinaScreen Programme: report on five rounds of retinopathy screening and screen-positive referrals. (INDEAR study report no. 1). Br J Ophthalmol. 2020;Published Online First: 17 December 2020.
7. Nguyen HV, GSW T, Tapp RJ, Mital S, DSW T, Wong HT, et al. Cost-effectiveness of a national telemedicine diabetic retinopathy screening program in Singapore. Ophthalmology. 2016;123(12):2571–80.
8. Gardner GG, Keating D, Williamson TH, Elliott AT. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. Br J Ophthalmol. 1996;80(11):940–4.
9. Hipwell JH, Strachan F, Olson JA, KC MH, Sharp PF, Forrester JV. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. Diabet Med. 2000;17(8):588–94.
10. 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(6):666–72.
11. 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(2):761–6.
12. 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(13):5200–6.
13. Xie Y, Gunasekeran DV, Balaskas K, Keane PA, Sim DA, Bachmann LM, et al. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. Transl Vis Sci Technol. 2020;9(2):22.
14. 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(1):1–8.
15. 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(1):63–8.
16. 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(1):157–64.
17. Nunes S, Pires I, Rosa A, Duarte L, Bernardes R, Cunha-Vaz J. Microaneurysm turnover is a biomarker for diabetic retinopathy progression to clinically significant macular edema: findings for type 2 diabetics with nonproliferative retinopathy. Ophthalmologica. 2009;223(5):292–7.
18. Pappuru RK, Ribeiro L, Lobo C, Alves D, Cunha-Vaz J. Microaneurysm turnover is a predictor of diabetic retinopathy progression. Br J Ophthalmol. 2019;103(2):222–6.
19. 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. 2016;20(92):1–72.
20. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda SR, et al. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. Diabetes Technol Ther. 2019;21(11):635–43.
21. Solanki K, Bhaskaranand M, Bhat S, Ramachandra C, Cuadros J, Nittala MG, et al. Automated diabetic retinopathy screening: large-scale study on con-

secutive patient visits in a primary care setting. In: Diabetologia. Springer 233 SPRING ST, New York; 2016. p. S64.
22. Ipp E, Shah VN, Bode BW, Sadda SR. 599-P: diabetic retinopathy (DR) screening performance of general ophthalmologists, retina specialists, and artificial intelligence (AI): analysis from a pivotal multicenter prospective clinical trial. Diabetes [Internet]. 2019 [cited 2020 Feb 26];68(Supplement 1). Available from: https://diabetes.diabetesjournals.org/content/68/Supplement_1/599-P
23. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br J Ophthalmol. 2020;bjophthalmol-2020-316594.
24. 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(22):2402–10.
25. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology. 2018;125(8):1264–72.
26. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. 2019;137(9):987–93.
27. Hsieh Y-T, Chuang L-M, Jiang Y-D, Chang T-J, Yang C-M, Yang C-H, et al. Application of deep learning image assessment software VeriSee™ for diabetic retinopathy screening. J Formos Med Assoc. 2021;120(1, Part 1):165–71.
28. González-Gonzalo C, Sánchez-Gutiérrez V, Hernández-Martínez P, Contreras I, Lechanteur YT, Domanian A, et al. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. Acta Ophthalmol (Copenh). 2020;98(4):368–77.
29. DSW T, Cheung CY-L, Lim G, GSW T, 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(22):2211–23.
30. Quellec G, et al. Instant automatic diagnosis of diabetic retinopathy. arXiv e-prints: arXiv-1906. 2019. https://arxiv.org/abs/1906.11875.
31. Quellec G, et al. Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med Image Anal. 2020;61:101660. https://doi.org/10.1016/j.media.2020.101660. https://arxiv.org/abs/1907.09449.
32. Rajalakshmi R, Subashini R, Anjana RM, Mohan V. Automated diabetic retinopathy detection in smartphone-based fundus photography using artificial intelligence. Eye. 2018;32(6):1138–44.
33. Natarajan S, Jain A, Krishnan R, Rogye A, Sivaprasad S. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmol. 2019;137(10):1182–8.
34. Sosale B, Sosale AR, Murthy H, Sengupta S, Naveenam M. Medios–An offline, smartphone-based artificial intelligence algorithm for the diagnosis of diabetic retinopathy. Indian J Ophthalmol. 2020;68(2):391–5.
35. Sosale B, Aravind SR, Murthy H, Narayana S, Sharma U, SGV G, et al. Simple, mobile-based artificial intelligence algorithm in the detection of diabetic retinopathy (SMART) study. BMJ Open Diabetes Res Amp Care. 2020;8(1):e000892.
36. MWM W, Mishra DK, Hartmann L, Shah P, Konana VK, Sagar P, et al. Diabetic retinopathy screening using smartphone-based fundus imaging in India. Ophthalmology. 2020;127(11):1529–38.
37. Bawankar P, Shanbhag N, SS K, Dhawan B, Palsule A, Kumar D, 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(12):e0189854.
38. Rogers TW, Gonzalez-Bueno J, Franco RG, Star EL, Marín DM, Vassallo J, et al. Evaluation of an AI system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the MAILOR AI study. Eye. 2020:1–7.
39. Sandhu HS, Eladawi N, Elmogy M, Keynton R, Helmy O, Schaal S, et al. Automated diabetic retinopathy detection using optical coherence tomography angiography: a pilot study. Br J Ophthalmol. 2018;102(11):1564–9.
40. Heisler M, Karst S, Lo J, Mammo Z, Yu T, Warner S, et al. Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography. Transl Vis Sci Technol. 2020;9(2):20.
41. Piyasena MMPN, Yip JL, MacLeod D, Kim M, Gudlavalleti VSM. Diagnostic test accuracy of diabetic retinopathy screening by physician graders using a hand-held non-mydriatic retinal camera at a tertiary level medical clinic. BMC Ophthalmol. 2019;19(1):89.
42. Estil S, Steinarsson ÆÞ, Einarsson S, Aspelund T, Stefánsson E. Diabetic eye screening with variable screening intervals based on individual risk factors is safe and effective in ophthalmic practice. Acta Ophthalmol (Copenh). 2020;98(4):343–6.
43. Lee, A. Y., Yanagihara, R. T., Lee, C. S., Blazes, M., Jung, H. C., Chee, Y. E., ... & Boyko, E. J. (2021). Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes care. 2021;44(5), 1168–1175.
44. Grzybowski, A., & Brona, P. (2021). Analysis and Comparison of Two Artificial Intelligence Diabetic Retinopathy Screening Algorithms in a Pilot Study: IDx-DR and Retinalyze. J Clin Med. 2021;10(11), 2352.