Automated Screening for Diabetic Retinopathy – A Systematic Review

Mads Fonager Nørgaard^{a, b} Jakob Grauslund^{a, b}

^aDepartment of Ophthalmology, Odense University Hospital, Odense, Denmark; ^bResearch Unit of Ophthalmology, Department of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark

Keywords

Diabetic retinopathy · Screening · Automated retinal image analysis · Systematic review

Abstract

Purpose: Worldwide ophthalmologists are challenged by the rapid rise in the prevalence of diabetes. Diabetic retinopathy (DR) is the most common complication in diabetes, and possible consequences range from mild visual impairment to blindness. Repetitive screening for DR is cost-effective, but it is also a costly and strenuous affair. Several studies have examined the application of automated image analysis to solve this problem. Large populations are needed to assess the efficacy of such programs, and a standardized and rigorous methodology is important to give an indication of system performance in actual clinical settings. **Methods:** In a systematic review, we aimed to identify studies with methodology and design that are similar or replicate actual screening scenarios. A total of 1,231 publications were identified through PubMed, Cochrane Library, and Embase searches. Three manual search strategies were carried out to identify publications missed in the primary search. Four levels of screening identified 7 studies applicable for inclusion.

Results: Seven studies were included. The detection of DR had high sensitivities (87.0–95.2%) but lower specificities (49.6–68.8%). False-negative results were related to mild DR with a low risk of progression within 1 year. Several studies reported missed cases of diabetic macular edema. A meta-analysis was not conducted as studies were not suitable for direct comparison or statistical analysis. **Conclusion:** The study demonstrates that despite limited specificity, automated retinal image analysis may potentially be valuable in different DR screening scenarios with a relatively high sensitivity and a substantial workload reduction.

© 2018 S. Karger AG, Basel

Introduction

The threat posed by a boom in the prevalence of diabetes is well known to health care providers worldwide. According to the World Health Organization, diabetes prevalence has quadrupled since 1980, and 422 million adults are now living with some form of diabetes, and even this number is projected to rise rapidly [1, 2]. As per the Liverpool declaration, many European countries have implemented systematic screening programs for diabetic

retinopathy (DR) to minimize the risk of visual impairment [3]. The fact that DR no longer represents the leading cause in blindness among working-age adults in the UK confirms the importance of timely detection and treatment of the disease [4]. The task to deliver a quality screening service is quite demanding and requires specially trained image graders. Annual screening produces large quantities of digital images that need to be analyzed. It is a costly affair, and the repetitious nature puts a strain on manual graders to maintain a high quality standard. A possible solution comes from automated retinal image analysis (ARIA), based on algorithms capable of detecting lesions associated with DR. As a research field for nearly 3 decades, the systems have followed the evolution of computational performance with a continuous increase in sophistication. Portugal has already automated a part of the screening process, and other screening services around the world might be close to implement an automated system [5]. Even though there have been a few reviews on the subject of automated image analysis for DR [6–10], only 1 study was conducted as a systematic review [8]. This study was published in 2010, and, given the technological improvements, an update is needed. Therefore, the aim of this study was to systematically identify and review studies that evaluate automated DR detection performance in realistic screening scenarios.

DR and Screening

DR is the most common complication in diabetes [11]. At worst, it can result in blindness from either proliferative DR (PDR) or diabetic macular edema (DME). A recent large meta-analysis of 35 studies carried out between 1982 and 2008 reported an estimated worldwide prevalence for any DR at 35.4% [12]. Detailed examination of the retina is required for accurate disease detection. The gold standard examination is 7-field 30° stereographic retinal photography as proposed by the Diabetic Retinopathy Study [13] and further established by the Early Treatment of Diabetic Retinopathy Study (ETDRS) [14]. An extension of the modified Airlie House Classification of DR was developed based on the importance of the different lesions for the overall progression of disease. It differentiates 13 complex levels of disease severity from 10 (no DR) to 85 (severe vitreous hemorrhage or retinal detachment involving the macula). However, the complexity of the protocol makes it a strenuous procedure for both photographer and patient, in particular regarding the 14 images needed per eye for stereoscopic evaluation of the

retina. To address this, the International Clinical Diabetic Retinopathy Disease Severity Scale (ICDR) was developed as a simplified version that was more appropriate for clinical use [15]. This was a validated scale based upon the ETDRS classification, with disease severity classified into 5 different stages: stage 0: no apparent DR; stage 1: mild non-PDR (NPDR, microaneurysms [MA] only); stage 2: moderate NPDR (more than MA – i.e., dot and blot hemorrhages and cotton wool spots – but less than stage 3); stage 3: severe NPDR (>20 hemorrhages in each of all 4 retinal quadrants or definite venous beading in 2 or more retinal quadrants or prominent intraretinal microvascular abnormality in at least 1 retinal quadrant); and stage 4: PDR. For DME, a 2-tier classification was made. The first level was a binary decision (DME apparent or absent) based on the presence of lipid exudates, hemorrhages, or retinal thickening in the posterior pole. The binary system made it possible to apply the ICDR in settings when the mode of examination did not provide stereopsis. The second level characterized the severity of DME in 3 levels based on the distance of retinal thickening and/or hard exudates from the fovea. Traditional grading of DR relies on surrogate signs of macular edema such as hard exudates or hemorrhages, but often 2-dimensional images are inadequate to detect DME. However, the definite diagnosis of DME relies on 3-dimensional retinal scans by optical coherence tomography [16].

Overall, screening of DR is needed to identify patients with visually threatening DR (PDR and/or DME), as they need timely treatment before potentially irreversible vision loss ensues. In a screening population, a large percentage will have no apparent DR, and most established screening services are based on trained graders as a first-level assessment to reduce the number of patients that will need specialized ophthalmological assessment [17].

ARIA and Screening for DR

ARIA has been proposed as a possible solution to alleviate the increasing demands of the established screening services in DR. Recent systems are built from sophisticated algorithms utilizing advanced mathematical modeling. In general, there are 2 available systems. Those with a binary output of disease/no disease and those based on disease severity and with the ability to classify patients in need of referral to ophthalmologists. Several image databases have been established to test and compare algorithms, such as the MESSIDOR (Méthodes d’Evaluation

Fig. 1. Flowchart demonstrating the selection process. ARIA, automated retinal image analysis; DR, diabetic retinopathy.

de Systèmes de Segmentation et d'Indexation Dédiées à l'Ophtalmologie Rétinienne) and MESSIDOR-2 [18]. In 2009, the Retina Online Challenge presented the first competition of its kind with a set of previous unseen images, with 23 participating research groups [18, 19]. The competition focused on MA detection. Recently, the Kaggle competition was held, with the task to create a system that could perform image grading and assign a grade based on a 5-level scheme [20]. These initiatives help propel the research field forward, but barriers still have to be surpassed to move from algorithm testing to implementation to clinical practice [7]. Some of these issues have been addressed, such as performance with different image modalities and lesion detection in eyes with different levels of retinal pigmentation. Recently, Hansen et al. [21] assessed an automated system on 3 different European populations and image sets with different settings, with

sensitivities >93% and specificities >80% for disease/no-disease grading. As for retinal pigmentation, different studies have presented results that showed no difference between groups of ethnicity [22, 23]. Further challenges still remain, but the strain already imposed on screening services might require action before all issues have been solved.

Method

A search strategy was created to cover all peer-reviewed literature that involved automated retinal fundus image analysis in diabetes mellitus, and to minimize the influence of heterogeneous populations on the analysis. The final criteria for inclusion were constructed to identify studies with a realistic screening scenario: (1) The image analysis system must be fully automated and include an image quality assessment and a lesion detection module and

have some form of patient-based output in terms of disease/no-disease or disease level. (2) Data must be based on digital (mydriatic or nonmydriatic) images from consecutively recruited patients with any form of diabetes mellitus who have never been diagnosed with referable DR. (3) Studies must be based on patients from the same cohort and not a selection of different trials.

Searches were performed in PubMed, Embase (October 21, 2016), and the Cochrane Library (October 21, 2016), even though no results were found in the latter. Search terms used in PubMed (as of October 10, 2016) were:

“diabetic retinopathy” [MeSH Terms] OR (“diabetic” [All Fields] AND “retinopathy” [All Fields]) OR “diabetic retinopathy” [All Fields] AND “ARIA” [All Fields]) OR “automated” [All Fields] AND “retinal” [All Fields] OR “retina” [MeSH Terms] OR “retina” [All Fields] AND “Image (IN)” [Journal] OR “image” [All Fields] AND (“analysis” [Subheading] OR “analysis” [All Fields] OR “automated” [All Fields] AND “grading” [All Fields]).

The search yielded 1,042 results that were immediately filtered to present only studies that involved “humans,” which left 839 hits for revision. These were added to the 189 studies identified through the Embase database search, yielding 1,028 headlines to be screened for eligibility. For reasons such as other primary diseases, other photo modalities, or focus on manual grading etc., 742 studies were excluded. Abstracts for the remaining 286 studies (1 study could not be identified) were retrieved and screened for eligibility. One of these could not be identified, and, in the end, 6 studies met the criteria for inclusion and data extraction in this systematic review. Three additional manual search strategies were conducted to identify studies missed by database searching: (1) Review studies identified in database searching were screened for references. (2) Authors with several publications that involved automated DR screening were identified, and a search was performed to reveal additional publications. (3) A random google search (“automated retinal image analysis”) yielded 4 results, but none of these met the criteria for inclusion. Two studies were discovered through the second step, one of which was included in the final study [24]. The study was not identified in the original search since it used the term “computer aided” as opposed to “automated” detection. The 5 excluded studies were based on a selected population. Figure 1 demonstrates a flowchart of the search process.

The selection process revealed a number of studies that applied automated detection to large populations, but did not fulfill all criteria for inclusion [21, 23, 25–33].

No meta-analysis was performed since none of the included studies were comparable in regard to their reference standards (system performance quantification); most of them were based on retrograde assessment of large image sets and involved algorithms screening for different lesions and different grading schemes.

Out of the 7 included publications, only 3 were previously included in a systematic review [8].

Results

Results are presented in Table 1.

Philip et al. [34] evaluated the efficacy of an ARIA developed for a Scottish screening program. Trained manual graders referred patients with images containing any

level of retinopathy or bad quality to “full grading” by ophthalmologists. Images from 6,722 consecutive diabetic patients screened between 2003 and 2004 formed the test set with single-field 45° images. All images had a reference standard grade appointed by a clinical research fellow, trained in the Scottish Diabetic Retinopathy Grading Scheme. The automated program assessed the images for quality and dot hemorrhage/MA. As for manual grading, any detected level of retinopathy or bad image quality prompted a referral for “full grading.” To sum up, both systems worked on a disease/no-disease referral level, though still appointed a disease level grade. The study reported efficacy per patient, per eye, and per image. For the purpose of comparison, we only extracted results as given per patient. For disease/no-disease (plus ungradable images) grading, the automated system obtained a sensitivity of 90.5% (95% CI 89.3–91.6%) and a specificity of 67.4% (95% CI 66.0–68.8%). The automated system missed 232 cases of mild NPDR retinopathy, 5 cases of referable maculopathy, 2 cases of observable maculopathy, and 1 case of ungradable images. The authors reported a 60% workload reduction for the automated system, as defined as patients with good-quality images without DR analyzed by the program. This number is, however, based on a combination of manual grading for disease/no disease and patients referred to full grading afterwards, totaling 9,267 episodes for 6,722 patients. Implementation of the automated system would provide a workload reduction at disease/no-disease grading level, and, therefore, the value necessary for comparison of workload reduction in our study should be based on this. In this scenario, 3,652 patients would need full image grading, which means the automated program detected good image quality and no disease in 3,070 patients, leading to a workload reduction of 45.7% (3,070/6,722).

Bouhaimed et al. [35] presented a retrospective study comparing Retinalyze, an ARIA developed in Denmark, with a reference standard, performed by a group of senior diabetologist and ophthalmologist. The program assessed the image for quality and detected red (MA and/or hemorrhages) and bright lesions (hard exudates and/or cotton wool spots). Patients were selected sequentially in day clusters from a database consisting of patients screened between 2002 and 2004 from the Bro Taf diabetic retinal screening programme in Wales, UK. The grading reference standard was performed in accordance with the Bro Taf Protocol for DR Screening. The photographic protocol comprised 2-field 45° images for each eye. The study comprised 458 images of 100 patients attending the screening program. Four patients were excluded due to

Table 1. Characteristics and results of the 7 studies that meet the criteria for inclusion

Source	Year	Grading details	Patient episodes	Image protocol	Reference standard	Sensitivity, %	Specificity, %	Missed cases ¹	Workload reduction, %
Philip et al. [34]	2007	Trained manual graders: any DR or PIQ = referral Aberdeen system: MA/DH: any DR or PIQ = referral	6,722	Nonmydriatic Disk/macula centered 45° Single field	Research fellow trained to use the Scottish DR Grading Scale	90.5 (95% CI 89.3–91.6)	67.4 (95% CI 66.0–68.8)	M2: 5 cases M1: 2 cases PIQ: 1 case R1: 232 cases	60.0
Bouhaimed et al. [35]	2008	Retinalyze: red lesions (MA/HEM) and bright lesions (CWS/HE) Disease/no disease	96	Mydriatic 45° Two-field image Optic disk and macula centered	Manual grading by a group of senior clinicians Grading scale: the Bro Taf protocol for DR screening	88 (95% CI 64–99)	52 (95% CI 40–63)	2 cases of mild NPDR	
Niemeijer et al. [24]	2009	IOWA/Utrecht system: MA, HEM, HE, CWS, DRU Disease/no disease	15,000	Staged mydriasis if necessary 30/45° Two-field image Optic disk and macula	One of 3 retinal experts (2nd reader in discrepancies) See text for grading scale	92.9 (no CI)	60.0 (no CI)	PIQ: 11 cases (10 – 2nd reader) 25 false-negative (22 – 2nd reader) See text for further classification	
Fleming et al. [27]	2010	Aberdeen system: MA/DH + artifact detection Disease/no disease	33,535	Staged mydriasis, if necessary Single field 45° Macula centered	Manual graders trained in the Scottish DR Grading Scale (expert arbitration for discrepancies)	87.0 (no CI)	49.6 (95% CI 48.9–50.3)	R1: 1,286 cases M1: 3 cases M2: 31 cases PIQ: 3 cases	36.3
Goatman et al. [37]	2011	Aberdeen system (iGrading): MA, blot HEM, HE Disease/no disease	8,271	Mydriatic 45° Two-field image Optic disk and macula centered ²	Full disease grading by manual graders trained in the English DR Screening Grading Scale Discrepancies settled by experts	95.2 (95% CI 94.4–95.9) ²	60.2 (95% CI 58.8–61.5) ²	R1: 128 cases M1: 5 cases ²	26.4
Ribeiro et al. [5]	2015	RetmarkerSR system: MA + MA turnover Disease/no-disease grading	45,148	Nonmydriatic 45° Two-field image Optic disk and macula centered	Sample of no-disease results (3,287) were manually graded + Comparison of incidence with previous studies MA turnover: none	Incidence per eye: R0 – 71.5, RL – 22.7, M – 2.2, RP – 0.1, NC – 3.5		Sample. 11 (0.3%) missed cases of referable DR None of them were proliferative DR	22.4–48.4 (progressive workload reduction from 2012 to 2014)
Soto-Pedre et al. [38]	2015	Aberdeen system (iGrading): MA + artifact detection Disease/no disease	5,253	Mydriatic 45° Single field Macular centered	Manual grading performed by a retinal specialist Grading scale: ICDR	94.5 (95% CI 92.6–96.5)	68.8 (95% CI 67.2–70.4)	31 cases None of them manually graded as referable DR	44

Comparison between the presented results should not, under any circumstances, be undertaken given different reference standards, populations (ethnicity, type of diabetes, duration of diabetes, level of diabetes control, prevalence of hypertension, control of high blood pressure, or age), photography protocols, grading protocols, and training and quality assurance of graders and screeners. DR, diabetic retinopathy; PIQ, poor image quality; MA, microaneurysms; DH, dot hemorrhages; HEM, hemorrhages; CWS, cotton wool spot; HE, hard exudates; NPDR, non-proliferative DR; DRU, drusen; ICDR, International Clinical Diabetic Retinopathy Disease Severity Scale. Philip et al. and Fleming et al.: R1, mild background retinopathy; M1, observable maculopathy; M2, referable maculopathy. Goatman et al.: R1, background retinopathy; M1, maculopathy. Ribeiro et al.; R0, no retinopathy; RL, nonproliferative retinopathy; M, maculopathy; RP, proliferative retinopathy; NC, not classifiable.

¹ See individual grading scales for level classification. ² See text for further results in different settings.

previous retinal photocoagulation or incomplete sets of images. A crude categorization was made to compare between manual and automated grading. Manual per-patient grade was divided into 2 groups (nonmanifest and manifest DR). All results below grade 2a, which corresponds to moderate NPDR on the ICDR, were defined as nonmanifest DR, and grade 2a and above was manifest DR. Automatically detected DR was assigned when the algorithm detected 1 or more red or white lesions in ei-

ther of the 2 fields, or if image quality was too poor for analysis. Automatic screening performance reached a sensitivity of 88% (95% CI 64–99%) at a specificity of 52% (95% CI 40–63%). The automated system missed 2 cases in total, both graded manually as mild NPDR.

Niemeijer et al. [24] assessed an ARIA system applied to images from a screening cohort in the Netherlands with 15,000 patients screened between 2006 and 2008. The patients attended a project for DR screening named

EyeCheck. The program is a collection of algorithms developed in Utrecht (The Netherlands) and Iowa (USA), and this version detected red lesions, i.e., hemorrhages and MA, and bright lesions, i.e. exudates, drusen, and cotton wool spots. The photographic protocol consisted of two 30 or 45° images per eye centered on the optic disk and on the macula. Every exam was graded by 1 of 3 ophthalmologists who assigned the exam to 1 of 3 classes in accordance with the protocol for the project [36] as not suspect (no signs of referable DR), suspect (signs of referable DR), or ungradable. It is unclear which grading scale was used to determine referable DR, but presumably any detectable signs led to referral, which indicates that this was a disease/no-disease detection model. Sensitivity was 92.9% and specificity was 60.0%. Missed cases included 11 cases (a second reader agreed with the reference standard on 10 of these) of inadequate image quality and 25 cases (22 with second reader agreement) of abnormal exams. Of the latter, 13 cases (indicated in the text as 50%) were 1 or 2 relatively large hemorrhages connected with the vasculature and no other abnormalities, 8 cases (indicated in the text as 32%) were exams with up to 4 small, isolated exudates close to the fovea and no other abnormalities present, and the remaining 4 cases contained either a single MA near the fovea, exams with laser scars, or another abnormality not associated with DR.

Fleming et al. [22] applied the same set of algorithms as Philip et al. [34] to a larger population from different screening centers between January 1, 2007, and January 31, 2008. A new algorithm was added to account for dirt on the camera lens resembling MA and dot hemorrhages. As with Philip et al. [34], the program produced a binary disease/no-disease result based on detection of MA and/or dot hemorrhages. Consensus grading of 7 senior ophthalmologists were used to settle discrepancies between the automated and manual grading. Macula-centered 45° photographs from 33,535 consecutive patient screening episodes were retrospectively obtained. Automatic screening performance was presented with a sensitivity (any DR) of 87.0% (no CI presented) and a specificity (no DR) of 49.6% (95% CI 48.9–50.3%). Missed cases included 1,286 patients with mild NPDR (R1), 3 patients with observable maculopathy (M1), 31 patients with referable maculopathy (M2), and 3 patients with ungradable images. All patients with some form of DR that required more urgent attention than rescreening after 12 months (observable mild DR and worse) were caught by the program. Implementation of the automated system as presented in this study resulted in a workload reduction of

36.3%, which was defined the same way as in the study by Philip et al. [34].

Goatman et al. [37] assessed the performance of an automated program on 8,271 consecutive unique patient episodes from a South London DR screening service in 2009. The system, iGrading, resembled the software used by Fleming et al. [22] and Philip et al. [34] and performed disease/no-disease grading, this time involving 2-field photos for each eye. All images were manually graded by a local screening service, which served as the quality measure for the automated system. All discrepancies between the 2 were settled by internal and external arbitration. The screening service followed a grading scheme similar to the ICDR. The automatic system detected MA, blot hemorrhages, and hard exudates. This study presents 2 versions of the program. One of these detected MA solely (a), and the other detected all lesions (b). Both versions were tested on an image set consisting of either macular field images or 2-field images. All 4 strategies were assessed for workload reduction, defined as the percentage of all cases, graded as having acceptable image quality and no DR. In the 2-field test of version “a,” sensitivity was 95.8% (95% CI 95.0–96.5%) and specificity 54.6% (95% CI 53.2–55.9%). In version “b,” the system obtained a sensitivity of 95.2% (95% CI 94.4–95.9%) and a specificity of 60.2% (95% CI 58.8–61.5%). For macular-centered single-field images, sensitivity and specificity of version “a” was 91.9% (95% CI 90.8–92.8%) and specificity 43.4% (95% CI 42.0–44.7%). In version “b,” sensitivity was 89.9% (95% CI 88.7–90.9%) and specificity was 50.7% (95% CI 49.3–52.1%). Discrepancies between automated and manual grading were settled by arbitration. All cases concerned maculopathy, except for 1 case graded as PDR by the screening service that was downgraded by ophthalmologists to no disease (R0). The systems with the best sensitivity were based on 2-field detection (a and b). In these, 4 and 5 cases of maculopathy were missed. Single-field detection resulted in a bigger workload reduction (34.1–38.1%) compared to 2-field detection (26.4–29.7%), but an additional 5 cases of maculopathy were missed. It seems that the workload reduction is underestimated for the 2-field strategies. In Scenario a, the system found 3,094 of R0 cases, but the numerator used when calculating the workload reduction reads 2,183, consistent with the number of cases misclassified by the system. Irrespective of whether you choose to calculate the workload reduction from the number of no-disease gradings by the system or the number of true negatives, the result is inconsistent. Most of the workload reduction difference between the 1- and 2-field strategies seems to come from a

decrease in specificity, as well as a 5% decrease in sensitivity. As mentioned above, the clinical impact is minimal since missed cases primarily consisted of no-disease or mild retinopathy. This presents the classic struggle when implementing a screening service: cost-effectiveness versus safety. Theoretically, missed cases would be caught by the system the following year, especially in cases with advancing retinopathy since the system is very sensitive to referable DR.

Ribeiro et al. [5] assessed the ARIA program, RetMarkerSR, applied to DR screening in central Portugal. In July 2011, the reading center introduced RetmarkerSR as a first assessment of all images sent to the reading center. In this study, there was no reference grade, as this system was already implemented. Two-field 45° non-mydratic photos were obtained for each eye. Photographers assessed the image quality and referred patients for ophthalmological specialist grading if needed (i.e., due to cataract, corneal problems, or collaboration issues, etc.). All remaining images were transferred to a central reading center for disease/no-disease grading by RetmarkerSR. For first-time patients, no DR led to automated annual rescreening. On the other hand, MA in fields 1 or 2 (ETDRS level 20 and above) resulted in a second assessment by human graders. At subsequent visits, a 2-step approach of automated screening was used. No DR again led to automated annual rescreening. However, if any MA were present, the program carried out a comparison with images from previous screening visits to test for differences in the number or position of red dots. Red dots in field 1 were reported as “disease” and warranted a second assessment by a human grader for severity level. Red dots limited to field 2 with no differences from previous screening visits were reported as NPDR and led to annual automated rescreening without manual assessment. Trained nonophthalmologic graders (under ophthalmologic supervision) performed manual grading. They assessed the images based on a 5-level grading scale with images assessed as not classifiable (NC), no DR (R0), NPDR (RL), maculopathy (M), and PDR (RP). The study was based on all screening visits between July 2011 and June 2014. Subjects included 45,148 visits, which corresponded to 89,626 eyes. Eyes were graded as NC – 3,132 (3.5%), R0 – 64,045 (71.5%), RL – 20,352 (22.7%), M – 1,964 (2.2%), and RP – 133 (0.1%). Based on a comparison with a previous study, an increased workload reduction from 22.4 to 48.4% was presented from 2012 to 2014 [31]. As the program is already implemented, the workload reduction is sampled at 2 different intervals of 14 weeks and compared to the burden of grading before im-

plementation. The first reduction, recorded in 2012, of 22.4% contained the first application of the program to images. The second reduction, recorded in 2014, of 48.4% saw the application of the 2-step approach which involved previous images as well. R0 images were scheduled for a rescreening the following year along with images with “static” MA in field 2 compared to previous screenings. It is worth mentioning that this approach to workload reduction is unique, which makes comparison to other systems difficult. An accuracy assessment was made for a sample of 3,287 patients, automatically graded as R0, who were randomly selected and sent to human grading. Eleven of these cases (0.3%) were graded manually as referable DR pathology, but none of them had PDR.

Soto-Pedre et al. [38] assessed the performance of iGrading compared with manual grading by the ICDR scale as performed by a retinal specialist. Single-field 45° macula-centered photos were obtained from consecutive patients attending a DR screening program in Valencia, Spain, in 2011–2012. The ARIA assessed image quality and the presence of MA in a disease/no-disease grading setup. There were 5,278 patients enrolled in the study, of which 5,253 had images of sufficient quality for analysis. The obtained sensitivity was 94.5% (95% CI 92.6–96.5%) and specificity was 68.8% (95% CI 67.2–70.4%). The automated system misclassified 31 cases as no DR, but none of them were graded as referable DR by the retinal expert. Referable DR was defined as moderate NPDR or worse and/or suspected macular edema. The study found a workload reduction of 44%, defined as the ability of the system to correctly determine good image quality and classify patients without apparent DR – i.e., the proportion of patients who would not need manual grading or referral to an ophthalmologist.

Discussion

This systematic review demonstrates that ARIA can partake in different DR screening scenarios with a relatively high sensitivity and a substantial workload reduction. In the screening of DR, sensitivity is a matter of safety and must have a high priority. Reported missed cases in the included studies were generally in patients with mild NPDR, although there also seemed to be a trend towards missed referable cases with DME.

To our knowledge, the study by Ribeiro et al. [5] represents the first real-life implementation of independent automated detection in a DR screening service. According to the authors, the distribution of the grading scale

levels was similar to results from the previous 8 years of manual screening. This current study revealed 11 cases of manually graded referable DR misclassified as having no DR by the automated system. None of them had PDR. Prior to implementation, the performance of the system as compared to manual grading was assessed in a study by Oliveira et al. [31] that was excluded from the present study because the cohort was selected for adequate image quality. In this study, missed cases were also related to diabetic maculopathy (18 cases), which in the Portuguese screening system warrants a referral to ophthalmology as soon as possible. No cases of PDR were missed. As an added safety measure, all urgent cases from the screening service were assessed for the original automated grading result. Out of 116 cases, only 1 was graded as no DR by the automated system. For the 2-step process, the study reported a sensitivity of 95.8%, which has been criticized [39] to be flawed, as the sample was taken from “false-positive” results manually graded as no DR, and as such it was an assessment of the specificity of the new techniques.

The reported specificities between the 7 studies range from 49.6 to 68.8%. In disease/no-disease grading, the specificity is the ability of the system to correctly identify people without any apparent DR. From an ethical perspective, a low specificity is problematic since healthy individuals will be falsely diagnosed. In the current proposed role of ARIA, however, a positive result will trigger manual grading, and, as such, the final diagnosis does not rely on automated detection. The primary aspect of specificity is the possible relief provided by the automated systems to established screening services. In this study, workload reduction ranged from 26.4 to 60.0%. Added to the reduced number of images which needed to be manually graded, a workload reduction poses financial benefits with lower costs of screening. Two of the included studies

were subsequently used in cost-benefit analyses [22, 34]. Scotland et al. [40, 41] presented 2 studies and concluded that implementation of an automated system was cost-effective as compared to manual grading. Even though the economic modeling is based on the architecture of the Scottish screening service, it does serve as an indication of a possible gain to other screening services in the consideration of implementation of automated grading.

This review is the first in recent years to apply a systematic approach. The aim was to provide a thorough review of studies that assessed ARIAs applied to a realistic screening scenario. Our results are validated by a documented process of method and selection of studies. On the other hand, we had a limited number of eligible studies to review, of which the study by Bouhaimed et al. [35] was based on a limited number of screening episodes. In addition, other methodological concerns should be taken for such studies as the one by Niemeijer et al. [24], which presented results without a confidence interval and did not present a clear reference standard grading scale. For the systems labeled as Aberdeen and Iowa/Utrecht, the involved studies are based on different (Aberdeen) and older generations [22, 24, 34, 37, 38].

In an observational, retrospective measurement comparison study of 102,856 images from 20,258 patients, Tufail et al. [10] concluded that Retmarker and EyeArt were both acceptable with respect to sensitivity for referable DR and numbers of false-positive results (compared to human graders). In addition, they were considered cost-effective alternatives to purely manual grading.

In conclusion, ARIAs have reached the level of maturity to safely partake in DR screening. Future years will show if they can advance to cover a bigger part of the screening process, but for now they will be able to provide the much needed relief in the preliminary screening worldwide.

References

- 1 World Health Organization – Global Report on Diabetes. http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1 (accessed December 14, 2016).
- 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;103:137–149.
- 3 Metelitsina TI, Grunwald JE, DuPont JC, Ying GS, Brucker AJ, Dunaief JL: Foveolar choroidal circulation and choroidal neovascularization in age-related macular degeneration. *Invest Ophthalmol Vis Sci* 2008;49:358–363.
- 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 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.
- 6 Valverde C, Garcia M, Hornero R, Lopez-Galvez M: Automated detection of diabetic retinopathy in retinal images. *Indian J Ophthalmol* 2016;64:26–32.

- 7 Abramoff MD, Niemeijer M, Russell SR: Automated detection of diabetic retinopathy: barriers to translation into clinical practice. *Expert Rev Med Devices* 2010;7:287–296.
- 8 Fleming AD, Philip S, Goatman KA, Prescott GJ, Sharp PF, Olson JA: The evidence for automated grading in diabetic retinopathy screening. *Curr Diabetes Rev* 2011;7:246–252.
- 9 Sim DA, Keane PA, Tufail A, Egan CA, Aiello LP, Silva PS: Automated retinal image analysis for diabetic retinopathy in telemedicine. *Curr Diabetes Rep* 2015;15:14.
- 10 Tufail A, Kapetanakis VV, Salas-Vega S, Egan C, Rudisill C, Owen CG, Lee A, Louw V, Anderson J, Liew G, Bolter L, Bailey C, Sadda S, Taylor P, Rudnicka AR: 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;20:1–72.
- 11 Grauslund J: Long-term mortality and retinopathy in type 1 diabetes. *Acta Ophthalmol* 2010;88(thesis 1):1–14.
- 12 Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, Chen SJ, Dekker JM, Fletcher A, Grauslund J, Haffner S, Hamman RF, Ikram MK, Kayama T, Klein BE, Klein R, Krishnaiah S, Mayurasakorn K, O'Hare JP, Orchard TJ, Porta M, Rema M, Roy MS, Sharma T, Shaw J, Taylor H, Tielsch JM, Varma R, Wang JJ, Wang N, West S, Xu L, Yasuda M, Zhang X, Mitchell P, Wong TY: Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012;35:556–564.
- 13 Diabetic retinopathy study. Report Number 6. Design, methods, and baseline results. Report Number 7. A modification of the Airlie House classification of diabetic retinopathy. Prepared by the Diabetic Retinopathy. *Invest Ophthalmol Vis Sci* 1981;21:1–226.
- 14 Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology* 1991;98:823–833.
- 15 Wilkinson CP, Ferris FL III, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 2003;110:1677–1682.
- 16 Mackenzie S, Schmermer C, Charnley A, Sim D, Vikas T, Dumskyj M, Nussey S, Egan C: SDOCT imaging to identify macular pathology in patients diagnosed with diabetic maculopathy by a digital photographic retinal screening programme. *PLoS One* 2011;6:e14811.
- 17 Andersen N, Hjortdal JO, Schielke KC, Bek T, Grauslund J, Laugesen CS, Lund-Andersen H, Cerqueira C, Andresen J: The Danish Registry of Diabetic Retinopathy. *Clin Epidemiol* 2016;8:613–619.
- 18 Laboratoire de Traitement de l'Information Médicale (LaTIM – INSERM U650). Mesidor-2 dataset (Méthodes d'Evaluation de Systèmes de Segmentation et d'Indexation Dédiées à l'Ophthalmologie Rétinienne). <http://latim.univ-brest.fr/indexfce0.html> 2016 (accessed December 18, 2016).
- 19 The ROC website – University of Iowa. <http://webeye.ophth.uiowa.edu/ROC/> (accessed on December 18, 2016).
- 20 Kaggle, Inc: Diabetic Retinopathy Detection. 2015, <http://www.kaggle.com/c/diabetic-retinopathy-detection> 2015 (accessed December 18, 2016).
- 21 Hansen MB, Tang HL, Wang S, Al Turk L, Piermarocchi R, Speckauskas M, Hense H-W, Leung I, Peto T: Automated detection of diabetic retinopathy in three European populations. *J Clin Exp Ophthalmol* 2016;7:582.
- 22 Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA: Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *BR J Ophthalmol* 2010;94:1606–1610.
- 23 Hansen MB, Abramoff MD, Folk JC, Matthege 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.
- 24 Niemeijer M, Abramoff MD, van Ginneken B: Information fusion for diabetic retinopathy CAD in digital color fundus photographs. *IEEE Trans Med Imaging* 2009;28:775–785.
- 25 Abramoff MD, Reinhardt JM, Russell SR, Folk JC, Mahajan VB, Niemeijer M, Quellec G: Automated Early Detection of Diabetic Retinopathy. *Ophthalmology* 2010;117:1147–1154.
- 26 Walton OB 4th, Garoon RB, Weng CY, Gross J, Young AK, Camero KA, Jin H, Carvounis PE, Coffee RE, Chu YI: Evaluation of automated teleretinal screening program for diabetic retinopathy. *JAMA Ophthalmol* 2016;134:204–209.
- 27 Fleming AD, Goatman KA, Philip S, Williams GJ, Prescott GJ, Scotland GS, McNamee P, Leese GP, Wykes WN, Sharp PF, Olson JA: The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol* 2010;94:706–711.
- 28 Figueiredo IN, Kumar S, Oliveira CM, Ramos JD, Engquist B: Automated lesion detectors in retinal fundus images. *Comput Biol Med* 2015;66:47–65.
- 29 Decencière E, Cazuguel G, Zhang X, Thibault G, Klein JC, Meyer F, Marcotegui B, Quellec G, Lamard M, Danno R, Elie D, Massin P, Viktor Z, Erginay A, Laÿ B, Chabouis A: TeleOphta: machine learning and image processing methods for teleophthalmology. *IRBM* 2013;34:196–203.
- 30 Prescott G, Sharp P, Goatman K, Scotland G, Fleming A, Philip S, Staff R, Santiago C, Borooah S, Broadbent D, Chong V, Dodson P, Harding S, Leese G, Megaw R, Styles C, Swa K, Wharton H, Olson J: Improving the cost-effectiveness of photographic screening for diabetic macular oedema: a prospective, multi-centre, UK study. *Br J Ophthalmol* 2014;98:1042–1049.
- 31 Oliveira CM, Cristovao LM, Ribeiro ML, Abreu JRF: Improved automated screening of diabetic retinopathy. *Ophthalmologica* 2011;226:191–197.
- 32 Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda S, Solanki K: Automated diabetic retinopathy screening and monitoring using retinal fundus image analysis. *J Diabetes Sci Technol* 2016;10:254–261.
- 33 Tang HL, Goh J, Peto T, Ling BW, Al Turk LI, Hu Y, Wang S, Saleh GM: The reading of components of diabetic retinopathy: an evolutionary approach for filtering normal digital fundus imaging in screening and population based studies. *PLoS One* 2013;8:e66730.
- 34 Philip S, Fleming AD, Goatman KA, Fonseca S, McNamee P, Scotland GS, Prescott GJ, Sharp PF, Olson JA: The efficacy of automated “disease/no disease” grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol* 2007;91:1512–1517.
- 35 Bouhaimed M, Gibbins R, Owens D: Automated detection of diabetic retinopathy: results of a screening study. *Diabetes Technol Ther* 2008;10:142–148.
- 36 Abramoff MD, Suttorp-Schulten MS: Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. *Telemed J E Health* 2005;11:668–674.
- 37 Goatman K, Charnley A, Webster L, Nussey S: Assessment of automated disease detection in diabetic retinopathy screening using two-field photography. *PLoS One* 2011;6:e27524.
- 38 Soto-Pedre E, Navea A, Millan S, Hernaez-Ortega MC, Morales J, Desco MC, Perez P: Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload. *Acta Ophthalmol* 2015;93:e52–e56.
- 39 Fleming AD, Olson JA, Sharp PF, Goatman KA, Philip S: Response to “Improved automated screening of diabetic retinopathy” by Carlos M. Oliveira et al. *Ophthalmologica* 2012;227:173; author reply 174.
- 40 Scotland GS, McNamee P, Philip S, Fleming AD, Goatman KA, Prescott GJ, Fonseca S, Sharp PF, Olson JA: Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *Br J Ophthalmol* 2007;91:1518–1523.
- 41 Scotland GS, McNamee P, Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Williams GJ, Wykes W, Leese GP, Olson JA: Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol* 2010;94:712–719.