Acta Ophthalmologica

Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System

Amber A van der Heijden,[1,2] Michael D Abramoff,[3,4,5] Frank Verbraak,[6] Manon V van Hecke,[7] Albert Liem[8] and Giel Nijpels[1,2]

[1]Department of General Practice and Elderly Care Medicine, VU University Medical Centre, Amsterdam, the Netherlands
[2]Amsterdam Public Health Research Institute, VU University Medical Centre, Amsterdam, the Netherlands
[3]Department of Ophthalmology and Visual Sciences, University of Iowa Hospital and Clinics, Iowa City, IA, USA
[4]VA Medical Center, Iowa City, IA, USA
[5]IDx LLC, Iowa City, IA, USA
[6]Department of Ophthalmology, VU University Medical Centre, Amsterdam, the Netherlands
[7]Department of Ophthalmology, Elisabeth-Tweestedenziekenhuis, Tilburg, the Netherlands
[8]Department of Ophthalmology, University Medical Centre Utrecht, Utrecht, the Netherlands

ABSTRACT.

Purpose: **To increase the efficiency of retinal image grading, algorithms for automated grading have been developed, such as the IDx-DR 2.0 device. We aimed to determine the ability of this device, incorporated in clinical work flow, to detect retinopathy in persons with type 2 diabetes.**

Methods: **Retinal images of persons treated by the Hoorn Diabetes Care System (DCS) were graded by the IDx-DR device and independently by three retinal specialists using the International Clinical Diabetic Retinopathy severity scale (ICDR) and EURODIAB criteria. Agreement between specialists was calculated. Results of the IDx-DR device and experts were compared using sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), distinguishing between referable diabetic retinopathy (RDR) and vision-threatening retinopathy (VTDR). Area under the receiver operating characteristic curve (AUC) was calculated.**

Results: **Of the included 1415 persons, 898 (63.5%) had images of sufficient quality according to the experts and the IDx-DR device. Referable diabetic retinopathy (RDR) was diagnosed in 22 persons (2.4%) using EURODIAB and 73 persons (8.1%) using ICDR classification. Specific intergrader agreement ranged from 40% to 61%. Sensitivity, specificity, PPV and NPV of IDx-DR to detect RDR were 91% (95% CI: 0.69–0.98), 84% (95% CI: 0.81–0.86), 12% (95% CI: 0.08–0.18) and 100% (95% CI: 0.99–1.00; EURODIAB) and 68% (95% CI: 0.56–0.79), 86% (95% CI: 0.84–0.88), 30% (95% CI: 0.24–0.38) and 97% (95% CI: 0.95–0.98; ICDR). The AUC was 0.94 (95% CI: 0.88–1.00; EURODIAB) and 0.87 (95% CI: 0.83–0.92; ICDR). For detection of VTDR, sensitivity was lower and specificity was higher compared to RDR. AUC's were comparable.**

Conclusion: **Automated grading using the IDx-DR device for RDR detection is a valid method and can be used in primary care, decreasing the demand on ophthalmologists.**

Key words: automated grading – diabetic retinopathy – type 2 diabetes – validation

Acta Ophthalmol.

doi: 10.1111/aos.13613

Introduction

Diabetes is a global epidemic. It was estimated that in 2013, at least 382 million people were affected by diabetes, and this number was estimated to increase to 592 million in 2035 (Guariguata et al. 2014). In 2012, the overall worldwide prevalence of any form of diabetic retinopathy (DR) was 35% (Yau et al. 2012). Screening for DR has proven to be effective in the prevention of blindness, and some national health authorities and most professional organizations have adopted an annual or biennial screening programme which is usually integrated within regular diabetes care (Olafsdottir & Stefansson 2007; Rodbard et al. 2007; Scanlon 2008; Chalk et al. 2012). However, the proportion of patients that actually undergoes this screening is less than 60% or unknown (Hazin et al. 2011).

Screening for DR is designed to detect sight-threatening retinopathy prior to vision-loss. Retinal microaneurysms have been defined as the first fundoscopically visible sign of DR (Friedenwald 1950). Screening for DR, including the grading of retinal images is currently usually performed by trained retinal specialists or trained readers, which has low efficiency and leads to intra- and interobserver variability (Helmchen et al. 2014; Oke et al. 2016). Due to the prolific increase in diabetes, based on current estimates, almost 3 million eyes will need to be evaluated each working day by 2030 (35 exams per second, assuming annual screening). Despite a 54% increase in the diabetes population, there will be less than a 2% growth in the number of ophthalmologists by 2030. The limited availability of a trained workforce at all levels limits service quality and reach (https://www.iapb.org/knowledge/what-is-avoidable-blindness/diabetic-retinopathy; 2016).

To increase the productivity of retinopathy screening (Helmchen et al. 2014), algorithms for automated grading of retinal images have been developed capable of recognizing signs of DR, including microaneurysms (Valverde et al. 2016). One such commercially algorithm is the IDx-DR 2.0 (IDx LLC, Iowa City, IA, USA) device, which analyses retinal images for the signs of DR (Abramoff et al. 2010, 2013, 2016; Hansen et al. 2015). This algorithm has recently been extended to separately detect vision-threatening diabetic retinopathy (VTDR), which included severe nonproliferative diabetic retinopathy and/or diabetic macular oedema (Abramoff et al. 2016). Before an algorithm can be implemented in the care process, its accuracy and safety have to be assured in real world populations. When validated against the International Clinical Diabetic Retinopathy Severity Scale (ICDR) classification score, IDx-DR showed a high sensitivity in the detection of referable diabetic retinopathy (RDR) (Abramoff et al. 2013; Hansen et al. 2015). The underlying algorithms have been validated (Abramoff et al. 2010, 2013, 2016; Hansen et al. 2015), but not all have been tested after implementation in a clinical setting. The aim of this study was to determine the performance of the IDx-DR device to detect RDR and VTDR compared to retinal specialist reading based on the ICDR (Wilkinson et al. 2003) as well as EURODIAB (Aldington et al. 1995) classification systems in persons with type 2 diabetes, after incorporation of the IDx-DR device in daily clinical work flow.

Materials and Methods

Study population

Persons with type 2 diabetes consecutively presenting at the Hoorn DCS centre (Zavrelova et al. 2011) for the annual visit were eligible for this study. Persons with a history of laser treatment (EURODIAB grade 4) were excluded for the current study. Anonymized computer records were used for this study, and the participants were informed about the use of these records for research purposes.

Sample size calculation

Using an α error of 0.05, a precision rate of 10% (two sided), an estimated sensitivity rate of 87% and an estimated incidence of RDR (EURODIAB grade $\geq$2 and/or presence of macular oedema) of 3.7%, led to a sample size of 1174 participants. Given these assumptions, and expecting that 30% of the photos will be qualified as insufficient quality by the IDx-DR device, 1526 participants will be needed for this analysis.

Retinal images

During the annual visit at the Hoorn DCS center, fundus photography was performed with a nonmydriatic Topcon TRC NW 100 camera (Topcon, Tokyo, Japan) using a standardized protocol (Abramoff & Suttorp-Schulten 2005). All retinal images were 45° of two fields: one field centred on the macula and one nasal field with the optic disc positioned on a disc-diameter from the temporal edge of the field, according to the EURODIAB protocol. The IDx-DR 2.0 device was installed and implemented in the DCS. Research assistants of the DCS received instructions for the use of the IDx-DR device. Retinal images were taken by the research assistants according to the regular procedure which did not include dilation as routine. Image quality analysis was performed by the IDx-DR device. When the quality of an image was too low, immediate feedback was provided by the software and a new photograph was to be taken. However, the scheduled time per patient most often did not allow this. Subject images were analysed by the IDx-DR device and were separately provided to the experts, who were masked to the IDx-DR device output, via the Truthmarker app on provided tablets (Christopher et al. 2012). The experts transmitted the results of their grading to AAvdH via the truthmarker app, who was the only researcher not to be masked to the IDx-DR device output.

Human grading

All retinal images were graded independently by three retinal specialists in masked fashion to each other and IDx-DR device outputs. At first, the experts indicated whether the retinal image was of sufficient quality for grading. In case of sufficient quality, the image was graded according to the EURODIAB (Aldington et al. 1995) as well as the ICDR classification score (Wilkinson et al. 2003) and is described in detail in the Table S1. Moderate diabetic retinopathy (MDR) was present in case of EURODIAB grading of 2 or ICDR grading of 2. Vision-threatening retinopathy (VTDR) was defined as EURODIAB of 3 or higher or ICDR of 3 or higher, including severe nonproliferative diabetic retinopathy and/or presence of macular oedema. Referable diabetic retinopathy (RDR) included MDR and VTDR. EURODIAB does allow classification of macular oedema separately, as hard exudates are contained in the DR levels. The experts were masked to IDx-DR device outputs. After the independent masked gradings were complete, the experts met for a consensus meeting to discuss cases without initial agreement until they achieved consensus.

Automated grading

The IDx-DR device identifies signs of RDR and VTDR by applying highly advanced image filters to retinal images (Abramoff et al. 2016). Image quality analysis was performed by the IDx-DR automated screening system. When the quality of the photograph was too low to rule out RDR, insufficient image quality feedback was provided immediately, allowing the research assistant to reimage, if the time schedule allowed this. Based on the detection of lesions, the IDx-DR device provides an index, a numerical output varying between 0 and 1, where 0 represents the absence of retinopathy and an index closer to 1 indicates a high chance of retinopathy. Using predefined thresholds points, a categorical outcome of the automated grading was provided: no RDR, MDR or VTDR. IDx employees were masked to the experts results.

Other variables

During the annual visit, weight and height were measured, while participants were barefoot and wearing light clothes. Body mas index (BMI) was calculated [weight (in kg) divided by the square of height (in m)]. Systolic and diastolic blood pressure were measured twice (3 min apart) after 5 min of rest in a seated position on the right arm using a random-zero sphygmomanometer. Information on smoking status was obtained by self-report. Using fasting blood and specified by standard operating procedures, HbA1c determination was based on the turbidimetric inhibition immunoassay for haemolysed whole EDTA blood. Blood glucose level was assessed in fluorinated plasma with the UV test using hexokinase. Levels of triglycerides, total cholesterol and high-density lipoprotein (HDL) cholesterol were determined enzymatically (Cobas c501; Roche Diagnostics, Mannheim, Germany). Low-density lipoprotein (LDL) cholesterol concentration was calculated using the Friedewald formula.

Statistical analysis

Characteristics of the people with type 2 diabetes included in this study were presented as means [standard deviation (SD)] or proportions. The number of cases of MDR and VTDR was calculated using the results of the adjudicated human grading according to the EURODIAB as well as the ICDR classification score.

The agreement between the three experts was estimated using methods described by de Vet et al. (2013). A 3 by 3 table, representing the agreement in the three outcome categories (no DR, MDR and VTDR), was constructed for each combination of experts, resulting in three 3 by 3 tables, demonstrating the agreement between experts 1 and 2, between experts 1 and 3 and between experts 2 and 3. Next, these three 3 by 3 tables were summed, meaning that all images that had the same classification by the three experts were summed for no retinopathy, MDR and VTDR separately. For example, the number of images that were classified as MDR by both experts in Table 1 was combined with the number of images that were classified as MDR by both experts of Table 2 and Table 3. All cells of the tables representing disagreement were summed per category. For example, the number of images that were graded as MDR by one expert and as no retinopathy by the other expert was combined with this specific disagreement by the other combinations of experts. This resulted in one 3 by 3 table, representing probabilities that the three experts would provide the same grading. Overall observed agreement was estimated dividing the number of cases with three similar gradings by the total number of cases. Specific agreement between the experts was estimated and expressed as the probability that the experts would provide the same grading, for each outcome category separately. Agreement for the three outcome categories was calculated for the EURODIAB classification score as well as the ICDR score.

The results of the IDx-DR device were compared to the results of human grading, which is considered the gold standard, using measures as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Sensitivity was expressed as the probability that the result of the IDx-DR device was positive when retinopathy was present according to human grading and specificity as the probability that the result of the IDx-DR device was negative when retinopathy was not present according to human grading. Positive predictive value (PPV) was considered the probability that persons with a positive screening test based on the IDx-DR device truly had retinopathy. Negative predictive value (NPV) was considered the probability that persons with a negative test result based on IDx-DR truly do not have retinopathy. To evaluate the discriminatory ability of the IDx-DR device, the area under the receiver operating characteristic curve (AUC) was calculated. For the accuracy measures, we distinguished between RDR (MDR and VTDR) and VTDR only. To allow comparison to the standard grading process in many screening programs where each exam is evaluated by only a single retinal specialist, each individual expert was also compared to the same gold standard human grading, by calculating sensitivity and specificity.

Results

Of the 1415 persons included in this study, 1138 persons (80.4%) had retinal images of sufficient quality according to the experts. The IDx-DR device rated images of 938 persons (66.3%) of sufficient quality, leaving 898 persons (65.5%) available for the analysis on accuracy measures. This was because the image quality feedback was underutilized in practice. Of these 898

Table 1. Characteristics of the population with retinal images of sufficient quality.

Variable	N = 708
Age (years)	65.0 (11.9)
Men (%)	56.1
Diabetes duration (years)	7.9 (3.2–12.9)
BMI (kg/m^2)	30.4 (5.7)
Systolic blood pressure (mmHg)	143.0 (22.0)
Diastolic blood pressure (mmHg)	79.6 (8.4)
Total cholesterol (mmol/l)	4.5 (1.1)
LDL cholesterol (mmol/l)	2.4 (1.0)
HDL cholesterol (mmol/l)	1.3 (0.4)
Triglycerides (mmol/l)	1.5 (1.1–2.1)
Fasting glucose (mmol/l)	8.2 (7.1–9.6)
HbA1c (%)	6.9 (6.3–7.8)
HbA1c (mmol/mol)	52 (45.0–61.3
Smoking status (%)	
Current	16.2
Never	42.3
Former	41.5

Data are presented as means (SD), median (interquartile range) or proportions.

BMI = body mass index, HbA1c = glycated haemoglobin, LDL = low-density lipoprotein, HDL = high-density lipoprotein, SD = standard deviation.

Table 2. Specific interobserver agreement for the three experts according to the EURODIAB and ICDR.

Grade	Total gradings	No RDR	MDR	VTDR
EURODIAB				
No RDR	3148 (97.5%)	$\frac{3116}{3148}$ = 0.99 (0.98–0.99)	$\frac{12}{3148}$ = 0.004 (0.00–0.01)	$\frac{20}{3148}$ = 0.006 (0.00–0.01)
MDR	30 (1.0%)	$\frac{12}{30}$ = 0.40 (0.39–0.41)	$\frac{12}{30}$ = 0.40 (0.39–0.41)	$\frac{6}{30}$ = 0.20 (0.19–0.21)
VTDR	50 (1.5%)	$\frac{20}{50}$ = 0.40 (0.39–0.41)	$\frac{6}{50}$ = 0.12 (0.11–0.13)	$\frac{24}{50}$ = 0.48 (0.47–0.49)
ICDR				
No RDR	3025 (93.7%)	$\frac{2962}{3025}$ = 0.98 (0.98–0.98)	$\frac{55}{3025}$ = 0.02 (0.02–0.02)	$\frac{8}{3025}$ = 0.003 (0.00–0.01)
MDR	165 (5.1%)	$\frac{55}{165}$ = 0.33 (0.32–0.34)	$\frac{101}{165}$ = 0.61 (0.60–0.62)	$\frac{9}{165}$ = 0.06 (0.05–0.07)
VTDR	38 (1.2)	$\frac{8}{38}$ = 0.21 (0.20–0.22)	$\frac{9}{38}$ = 0.24 (0.23–0.25)	$\frac{21}{38}$ = 0.55 (0.54–0.56)

ICDR = International Clinical Diabetic Retinopathy severity scale, MDR = moderate diabetic retinopathy, RDR = referable diabetic retinopathy, VTDR = vision-threatening diabetic retinopathy.

The table represents the agreement between the three experts, as described by de Vet et al. (2013), representing probabilities that the three experts would provide the same grading. All images that had the same classification by the three experts were summed, stratified by outcome (no RDR, MDR and VTDR). All images with differing classifications between experts were summed per outcome. For example, when one expert scored MDR based on the EURODIAB classification scale, the probability that the other experts also scored MDR was 0.40 while there was a probability of 0.40 that the other experts scored no DR and 0.20 for VTDR.

persons, the ID number was available for 708 persons (79%) so that clinical information could be linked. Characteristics of the subset of the entire population are shown in Table 1. Mean age of the population was 65 (SD: 11.9) years with a median diabetes duration of 7.9 years.

In the 898 persons with information on retinopathy provided by the IDx-DR device as well as human grading, RDR was diagnosed in 22 persons (2.4%) according to human grading based on the EURODIAB of which 14 cases (1.6%) were considered VTDR. Using the ICDR classification scale, the number of persons with RDR was much higher 73 (8.1%), with a comparable number of VTDR cases [*n* = 13 (1.4%)], a logical consequence of the different definition of RDR in EURODIAB versus ICDR (see Table S1).

There were 162 disagreements (11.8%) between the experts on the EURODIAB classification score, 86 led to differences after categorization of the retinopathy outcome into no RDR, MDR and VTDR. For the gradings on the ICDR scale, the experts disagreed on 167 cases (12.2%), leading to 146 disagreements in the categorized outcome of retinopathy. The specific inter-rater agreement was estimated for the two classification scores as shown in Table 2. For example, when one expert scored VTDR, the probability that the other two experts also scored VTDR was approximately 50% for both grading scales.

Table 3 shows the sensitivity, which ranged from 58% to 100%, and specificity, which ranged from 99% to 100%, for each of the experts using both EURODIAB and ICDR for RDR and VTDR.

Table 4 displays the comparison between the IDx-DR device grading results and the results of human grading as well as the calculated accuracy measures. Comparing the IDx-DR device to human grading based on EURODIAB, the IDx-DR device had only few false-negative test results and few false-positive test results in the detection of RDR resulting in a high sensitivity [0.91 (95% CI: 0.69–0.98)] and high specificity [0.84 (95% CI: 0.81–0.86)]. Accordingly, the PPV was 0.12 (95% CI: 0.08–0.18) and NPV was 1.00 (95% CI: 0.99–1.00). Using the ICDR classification, the IDx-DR device showed a sensitivity of 0.68 (95% CI: 0.56–0.79), and a specificity of 0.86 (95% CI: 0.84–0.88) in the detection of RDR, resulting in a PPV of 0.30 (95% CI: 0.24–0.38) and NPV of 0.97 (95% CI: 0.95–0.98). Thus IDx-DR showed an overall similar sensitivity and lower specificity compared to the three human experts.

In the detection of VTDR only, few cases were misclassified as MDR resulting in a sensitivity of 0.64 (95% CI: 0.36–0.86; EURODIAB) and 0.62 (95% CI: 0.32–0.85; ICDR), with a wide 95% confidence interval due to the small number of cases. The number of false-positive test results was low, leading to high specificity for both classification scales (0.95 (95% CI: 0.93–0.96). The IDx-DR device had

Table 3. Sensitivity and specificity (95% CI) of each of the three experts against the adjudicated reference standard (which includes each of them) for EURODIAB and ICDR.

	RDR		VTDR	
	Sensitivity	Specificity	Sensitivity	Specificity
EURODIAB				
Expert 1	0.86 (0.64–0.97)	0.99 (0.98–1.00)	0.62 (0.32–0.86)	1.00 (0.99–1.00)
Expert 2	0.59 (0.36–0.79)	0.99 (0.99–1.00)	0.64 (0.35–0.87)	0.99 (0.98–1.00)
Expert 3	0.86 (0.65–0.97)	0.99 (0.98–1.00)	0.93 (0.66–1.00)	0.99 (0.98–1.00)
ICDR				
Expert 1	0.69 (0.57–0.80)	1.00 (0.99–1.00)	0.58 (0.28–0.84)	1.00 (0.99–1.00)
Expert 2	0.92 (0.83–0.97)	0.99 (0.97–0.99)	0.62 (0.32–0.86)	1.00 (0.99–1.00)
Expert 3	0.63 (0.51–0.75)	1.00 (0.99–1.00)	1.00 (0.75–1.00)	0.99 (0.99–1.00)

CI = confidence interval, ICDR = International Clinical Diabetic Retinopathy severity scale, RDR = referable diabetic retinopathy, VTDR = vision-threatening diabetic retinopathy.

Table 4. Classification of diagnosis according to the IDx-DR device compared to the gold standard and accuracy measures [95% confidence intervals (CIs)] of the IDx-DR device using the EURODIAB and ICDR classification score.

	EURODIAB			ICDR			
	Human grading			Human grading			
IDx-DR	No RDR	MDR	VTDR	No RDR	MDR	VTDR	Total
No RDR	732	1	1	711	22	1	734
MDR	101	3	4	76	28	4	108
VTDR	43	4	9	38	10	8	56
Total	876	8	14	825	60	13	898

	RDR	VTDR	RDR	VTDR
Se	0.91 (0.69–0.98)	0.64 (0.36–0.86)	0.68 (0.56–0.79)	0.62 (0.32–0.85)
Sp	0.84 (0.81–0.86)	0.95 (0.93–0.96)	0.86 (0.84–0.88)	0.95 (0.93–0.96)
PPV	0.12 (0.08–0.18)	0.16 (0.08–0.29)	0.30 (0.24–0.38)	0.14 (0.07–0.27)
NPV	1.00 (0.99–1.00)	0.99 (0.99–1.00)	0.97 (0.95–0.98)	0.99 (0.99–1.00)

Accuracy measures are presented with 95% CI.

ICDR = International Clinical Diabetic Retinopathy severity scale, MDR = moderate diabetic retinopathy, NPV = negative predictive value, PPV = positive predictive value, RDR = referable diabetic retinopathy (moderate and vision-threatening diabetic retinopathy), Se = sensitivity, Sp = specificity, VTDR = vision-threatening diabetic retinopathy.

an estimated PPV of 0.16 (0.08–0.29) testing against EURODIAB classification scale and a PPV of 0.14 (0.07–0.27) testing against ICDR classification scale. The NPV was high for EURODIAB and ICDR [both 0.99 (95% CI: 0.99–1.00)].

The AUC was 0.94 (95% CI: 0.88–0.93) when testing the ability of the IDx-DR device to detect RDR based on the EURODIAB score and 0.87 (95% CI: 0.83–0.92) to detect RDR based on the ICDR score. For the detection of VTDR by the IDx-DR device, the AUC was 0.91 (95% CI: 0.83–0.98) and 0.90 (95% CI: 0.82–0.98).

Discussion

The IDx-DR device has a high sensitivity and specificity in the detection of RDR based on the EURODIAB grading score. In the detection of VTDR only, few cases were misclassified as MDR resulting in a lower sensitivity detecting VTDR based on EURODIAB. Testing the IDx-DR device against ICDR grading, sensitivity was lower while specificity was high in the detection of RDR as well as VTDR only. A high discrepancy was observed between the human grading based on EURODIAB and ICDR, in the diagnosis of MDR while the diagnosis of VTDR was about the same between the two grading scores.

The IDx-DR device was earlier validated against human grading based on the ICDR classification score (Abramoff et al. 2013; Hansen et al. 2015). In these previous studies, the sensitivity of the IDx-DR device in the detection of RDR was higher compared to our results based on the ICDR classification score while specificity was higher in our study. Our results on the validation of the IDx-DR device against the EURODIAB classification showed similar or slightly better results on sensitivity and specificity as previously studies.

The difference in sensitivity of the IDx-DR device between the two classification systems (EURODIAB and ICDR) is remarkable. About 70% of the retinal images that were classified as RDR according to the grading based on the ICDR score were classified as no RDR by the experts when using the EURODIAB score, which was in line with the grading by the IDx-DR device. Reason for this discrepancy between the two classification scores might be that the experts scored some aspects of DR in a different way. In a post hoc analysis, it was found that strict adhering to the definition of ICDR, the experts judged any single haemorrhage as 'more than MA's alone', and as grade 2, meaning MDR. This was the case in 21 of the 22 MDR cases scored as MDR by the experts according to the ICDR. Should one reconsider this decision, and change grade 2 to grade 1 in these cases, one would find no differences between the IDx-DR device, EURODIAB or ICDR. Sensitivity of IDx-DR becomes 0.96, and specificity 0.86.

In both classification scores, there was a large disagreement between the three experts. The most common cause of disagreement was the misclassification of the presence of one single haemorrhage, which was categorized as grade 1 (no RDR) by one grader and grade 2 (= RDR) by another grader. The prevalence of RDR was low in this population. When looking at specific agreement in prevalent RDR cases, chances that all experts scored MDR or VTDR were low, approximately 50%. This is typical for retinal specialist grading, and other studies found low agreements as well (Helmchen et al. 2014; Oke et al. 2016). This low agreement is relevant because a larger disagreement between the experts decreases the measurable performance of algorithms (Quellec & Abramoff 2014). The three experts were very experienced ophthalmologists, and it was expected that they were very capable of detecting referable retinopathy. Training of even experienced ophthalmologists would be of added value. Previous research showed that the grading of retinal images performed by ophthalmologists is systematically different from grading by trained technicians (Sallam et al. 2011), which could have influenced the results of our study.

Sensitivities of the experts calculated against a reference standard which includes each of them (thus giving a slight advantage compared to an expert that would not have been part of the gold standard) showed that these ranged between 59% and 93% for EURODIAB and 58–100% for ICDR, which is in line with other studies (Hutchinson et al. 2000; Lin et al. 2002), and most of them were similar to IDx-DR. The experts were not specifically trained before they made their classifications, and perhaps such a training would have improved the consensus between the experts, and probably would have prevented the interpretation differences described above. Other automated grading algorithms that were validated in people with diabetes distinguished only between no retinopathy and referable retinopathy. The sensitivity of the IDx-DR device based on EURODIAB was comparable to the other algorithms while specificity of the IDx-DR device was similar or higher compared to these algorithms

(Hansen et al. 2004; Philip et al. 2007; Oliveira et al. 2011).

There are several limitations to this study. The prevalence of referable retinopathy in this population is small, which limits comparison to other populations with higher disease prevalence (Sabanayagam et al. 2016). The likely reason for the low prevalence is that this population is well controlled and have been undergoing retinopathy screening for several years. Another limitation is that the number of persons included in our study was lower than planned. Due to technical problems at the start of the study, some of the gradings of the images were lost. Another limitation is the high number of retinal images that were considered of insufficient quality by the IDx-DR device which was caused by under-utilization of the immediate quality feedback feature incorporated into the automated detection software and of the opportunity to dilate. Use of this immediate quality feedback would likely have resulted in a larger number of sufficient quality exams to be included in the analyses. Finally, the version of the IDx-DR device used in the current study does not have an output able to distinguish between no DR and mild DR. It is designed for detection of referable DR, limiting its use for identification of people at risk for developing referable DR. The current version of the IDx-DR device, version 2.1, is able to differentiate between no DR and any form of DR (Abramoff et al. 2016). Once this distinction between no and any DR is validated, this version has the potential to be suitable for use in patient education and the estimation of retinopathy risk and a personalized screening interval.

Strength of this study is that the IDx-DR device was implemented in clinical practice and retinal images of consecutive persons presenting at the DCS Center were included in this study. People with all possible presentations of DR were included in the study: persons with early signs of retinopathy as well as persons with more severe forms of retinopathy.

The automated grading method using the IDx-DR device for detection of RDR is a valid method and can be used in primary care. Despite the high number of retinal images that were rejected by the IDx-DR device due to insufficient image quality, use of this automated grading method results in a large reduction in retinal images that need human grading, resulting in increased productivity and decreased demand on ophthalmologists.

References

Abramoff MD & Suttorp-Schulten MS (2005): Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemed J E Health **11**: 668–674.

Abramoff MD, Reinhardt JM, Russell SR, Folk JC, Mahajan VB, Niemeijer M & Quellec G (2010): Automated early detection of diabetic retinopathy. Ophthalmology **117**: 1147–1154.

Abramoff MD, Folk JC, Han DP et al. (2013): Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol **131**: 351–357.

Abramoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC & Niemeijer M (2016): Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci **57**: 5200–5206.

Aldington SJ, Kohner EM, Meuer S, Klein R & Sjolie AK (1995): Methodology for retinal photography and assessment of diabetic retinopathy: the EURODIAB IDDM complications study. Diabetologia **38**: 437–444.

Chalk D, Pitt M, Vaidya B & Stein K (2012): Can the retinal screening interval be safely increased to 2 years for type 2 diabetic patients without retinopathy? Diabetes Care **35**: 1663–1668.

Christopher M, Moga DC, Russell SR, Folk JC, Scheetz T & Abramoff MD (2012): Validation of tablet-based evaluation of color fundus images. Retina **32**: 1629–1635.

Friedenwald JS (1950): Diabetic retinopathy. Am J Ophthalmol **33**: 1187–1199.

Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U & Shaw JE (2014): Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract **103**: 137–149.

Hansen AB, Hartvig NV, Jensen MS, Borch-Johnsen K, Lund-Andersen H & Larsen M (2004): Diabetic retinopathy screening using digital non-mydriatic fundus photography and automated image analysis. Acta Ophthalmol Scand **82**: 666–672.

Hansen MB, Abramoff MD, Folk JC, Mathenge W, Bastawrous A & Peto T (2015): Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru Study, Kenya. PLoS ONE **10**: e0139148.

Hazin R, Colyer M, Lum F & Barazi MK (2011): Revisiting Diabetes 2000: challenges in establishing nationwide diabetic retinopathy prevention programs. Am J Ophthalmol **152**: 723–729.

Helmchen LA, Lehmann HP & Abramoff MD (2014): Automated detection of retinal disease. Am J Manag Care **20**: eSP48–eSP52.

Hutchinson A, McIntosh A, Peters J, O'Keeffe C, Khunti K, Baker R & Booth A (2000): Effectiveness of screening and monitoring tests for diabetic retinopathy–a systematic review. Diabet Med **17**: 495–506.

Lin DY, Blumenkranz MS, Brothers RJ & Grosvenor DM (2002): The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. Am J Ophthalmol **134**: 204–213.

Oke JL, Stratton IM, Aldington SJ, Stevens RJ & Scanlon PH (2016): The use of statistical methodology to determine the accuracy of grading within a diabetic retinopathy screening programme. Diabet Med **33**: 896–903.

Olafsdottir E & Stefansson E (2007): Biennial eye screening in patients with diabetes without retinopathy: 10-year experience. Br J Ophthalmol **91**: 1599–1601.

Oliveira CM, Cristovao LM, Ribeiro ML & Abreu JR (2011): Improved automated screening of diabetic retinopathy. Ophthalmologica **226**: 191–197.

Philip S, Fleming AD, Goatman KA et al. (2007): The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. Br J Ophthalmol **91**: 1512–1517.

Quellec G & Abramoff MD (2014): Estimating maximal measurable performance for automated decision systems from the characteristics of the reference standard. Application to diabetic retinopathy screening. Conf Proc IEEE Eng Med Biol Soc **2014**: 154–157.

Rodbard HW, Blonde L, Braithwaite SS et al. (2007): American Association of Clinical Endocrinologists medical guidelines for clinical practice for the management of diabetes mellitus. Endocr Pract **13**(Suppl 1): 1–68.

Sabanayagam C, Yip W, Ting DS, Tan G & Wong TY (2016): Ten emerging trends in the epidemiology of diabetic retinopathy. Ophthalmic Epidemiol **23**: 209–222.

Sallam A, Scanlon PH, Stratton IM, Jones V, Martin CN, Brelen M & Johnston RL (2011): Agreement and reasons for disagreement between photographic and hospital biomicroscopy grading of diabetic retinopathy. Diabet Med **28**: 741–746.

Scanlon PH (2008): The English national screening programme for sight-threatening diabetic retinopathy. J Med Screen **15**: 1–4.

Valverde C, Garcia M, Hornero R & Lopez-Galvez MI (2016): Automated detection of diabetic retinopathy in retinal images. Indian J Ophthalmol **64**: 26–32.

de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS & Knol DL (2013): Clinicians are right not to like Cohen's kappa. BMJ **346**: f2125.

Wilkinson CP, Ferris FL III, Klein RE et al. (2003): Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**: 1677–1682.

Yau JW, Rogers SL, Kawasaki R et al. (2012): Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care **35**: 556–564.

Zavrelova H, Hoekstra T, Alssema M et al. (2011): Progression and regression: distinct developmental patterns of diabetic retinopathy in patients with type 2 diabetes treated in the diabetes care system west-friesland, the Netherlands. Diabetes Care **34**: 867–872.

Received on December 19th, 2016.
Accepted on September 5th, 2017.

Correspondence:
Amber A van der Heijden, PhD
Department of General Practice & Elderly
Care Medicine
VU University Medical Centre
Van der Boechorststraat 7
1081 BT Amsterdam
the Netherlands
Tel: +31 20 4448409
Fax: +31 20 4448361
Email: a.vanderheijden@vumc.nl

This study was been made possible by collaboration with the Diabetes Care System West-Friesland. The authors thank participants of this study and research of the Diabetes Care System West-Friesland, and Arlette Goddijn and Eric Talmage from IDx LLC, Iowa City, IA, USA, for provision of administrative, technical and material support. This study was presented as an oral abstract at the 52nd annual meeting of the EASD, and an abstract of this study has been published in Diabetologia (2016) 59 (Suppl 1): S64.

This study was supported by IDx LLC, Iowa City, IA, USA.

M.D.A. has direct equity ownership of and receives income from IDx LLC, a company that has licensed inventions from the University of Iowa, on which the Iowa Reference Algorithms used in this study are based. The sponsor did not have any role in the data collection, management and analysis of the data. No other potential conflict of interest relevant to this article was reported.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Criteria of the EURODIAB and ICDR grading scales.