



**UK**  
National  
Screening  
Committee

# Automated grading in the Diabetic Eye Screening Programme

## External review against programme appraisal criteria for the UK National Screening Committee

Version: 3.7.1

Author: Zhivko Zhelev, Jaime Peters, Morwenna Rogers, Michael Allen, Jenny Lowe, Goda Kijauskaite, Elizabeth Wilkinson, Farah Seedat, Christopher Hyde

Date: 24 November 2021

**The UK National Screening Committee secretariat is hosted by the Department of Health and Social Care**

# About the UK National Screening Committee (UK NSC)

The UK NSC advises ministers and the NHS in the 4 UK countries about all aspects of [population screening](#) and supports implementation of screening programmes.

Conditions are reviewed against [evidence review criteria](#) according to the UK NSC's [evidence review process](#).

Read a [complete list of UK NSC recommendations](#).

UK National Screening Committee, Southside, 39 Victoria Street, London, SW1H0EU

[www.gov.uk/uknsc](http://www.gov.uk/uknsc)

Blog: <https://nationalscreening.blog.gov.uk/>

For queries relating to this document, please contact: <https://view-health-screeningrecommendations.service.gov.uk/helpdesk/>

© Crown copyright 2016

You may re-use this information (excluding logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0. To view this licence, visit [OGL](#) or email [psi@nationalarchives.gsi.gov.uk](mailto:psi@nationalarchives.gsi.gov.uk). Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned.

Published November 2021

# Contents

About the UK National Screening Committee (UK NSC)	2
List of tables	5
List of figures	6
List of abbreviations	7
Plain English summary	8
Executive summary	10
Purpose of the review	10
Background	10
Focus of the review	12
Findings and gaps in the evidence of this review	13
Recommendations	20
Introduction and approach	22
Background	22
Objectives	33
Methods	34
Databases/sources searched	37
Results	39
Question level synthesis	40
Criterion 4 and 5. Accuracy of ARIASs	40
Eligibility for inclusion in the review	40
Description of the evidence	41
Discussion of findings	66
Summary of Findings Relevant to Criteria 4 and 5: Criterion met.	68
Criterion 11. Effectiveness of DESPs using ARIASs at level 1 grading	69
Eligibility for inclusion in the review	69
Description of the evidence	69
Discussion of findings	70
Summary of Findings Relevant to Criterion 11: Criterion not met.	71
Criterion 14. Cost-effectiveness of ARIASs	72
Eligibility for inclusion in the review	72
Summary of findings	72
Criterion 12. Social and ethical implications of AI in screening programmes	75
Eligibility for inclusion in the review	75
Summary of findings	75
Review summary	77

Conclusions and implications for policy	77
Limitations	80
Appendix 1 — Search strategy	82
Electronic databases	82
Search Terms	83
Appendix 2 — Included and excluded studies	91
PRISMA flowchart	91
Appendix 3 — Summary and appraisal of individual studies	126
Appendix 4 – Additional information	148
Links to the manufacturer/developer’s website for each of the ARIASs included in the accuracy analysis (or a brief description if no website was found)	148
Systematic reviews investigating the accuracy of AI algorithms for detection of diabetic retinopathy	149
Factors reported to affect the performance of ARIAS	153
Appendix 5 – UK NSC reporting checklist for evidence summaries	157
References	160

## List of tables

Table 1 Differences between DESPs across the 4 nations.....	24
Table 2 Key questions for the evidence summary, and relationship to UK NSC screening criteria.....	33
Table 3 Studies evaluating DL-based ARIASs.....	48
Table 4 Results from the methodological quality assessment using QUADAS-2 checklist of studies evaluating the accuracy of DL-based ARIASs .....	50
Table 5 Diagnostic accuracy of the included DL-based ARIASs.....	51
Table 6 Studies evaluating traditional ML-based ARIASs.....	58
Table 7 Results from the methodological quality assessment using QUADAS-2 checklist of studies evaluating the accuracy of traditional ML-based ARIASs.....	59
Table 8 Diagnostic accuracy of the included traditional ML-based ARIASs.....	59
Table 9 Summary of the results from Tufail 2016 which compared directly 3 ARIASs ('ungradable' images included).....	62
Table 10 Studies comparing directly the accuracy of the ARIAS and human graders.....	64
Table 11 Studies evaluating the impact of implementing ARIAS in a DESP .....	70
Table 12 Summary of electronic database searches and dates.....	82
Table 13 Search strategy for Ovid MEDLINE(R) ALL <1946 to June 25, 2020>.	83
Table 14 Search strategy for Embase <1974 to 2020 June 25> .....	84
Table 15 Search strategy for the Cochrane Library .....	85
Table 16 Search strategy for the Clinical trials database .....	86
Table 17 Search strategy for Ovid MEDLINE(R) ALL <1946 to June 30, 2020>.	87
Table 18 Search strategy for Embase <1974 to 2020 June 30>.....	88
Table 19 Search strategy for PsycINFO <1806 to June Week 5 2020>.....	88
Table 20 Search strategy for CINAHL.....	89
Table 21 Publications included after review of full-text articles and the questions they address .....	92
Table 22 Publications excluded after review of full text articles .....	95
Table 23 Studies meeting the inclusion criteria for question 1 but excluded after prioritisation.....	117
Table 24 Methodological quality assessment of the included studies reporting on the clinical effectiveness and impact of ARIASs in DESPs using the Downs and Black checklist .....	126
Table 25 Question 3: Evidence map of UK-based health economic evaluations of AI in DESPs .....	129
Table 26 Question 4: Evidence map of studies investigating the social and ethical implications of implementing AI in screening programmes: Primary studies.....	132
Table 27 Question 4: Evidence map of studies investigating the social and ethical implications of implementing AI in screening programmes: Review and opinion papers.....	143
Table 28 QUADAS-2 checklist with definitions of the signalling questions.....	145
Table 29 QUADAS-C checklist: Comparative accuracy .....	146

Table 30 Systematic reviews investigating the accuracy of AI algorithms for detection and grading of diabetic retinopathy.....	149
Table 31 Quality assessment of the identified systematic reviews.....	151
Table 32 UK NSC reporting checklist for evidence summaries.....	157

## List of figures

Figure 1 Photographic fields. ....	25
Figure 2 Patient pathway in the EDESP (PHE 2021) .....	26
Figure 3 Grading pathway in the EDESP (Tufail 2016 (1)) .....	27
Figure 4 Proposed EDESP modification, using an automated system for level 1 grading (Tufail 2016 (1)).....	31
Figure 5 ARIAS used as a filter prior to level 1 manual grading, as investigated in the cost-effectiveness analysis conducted by Tufail 2016 (1) .....	32
Figure 6 Summary of publications included and excluded at each stage of the review.....	91

## List of abbreviations

AI	Artificial Intelligence
ARIAS	Automated Retinal Image Analysis System
AUC	Area Under The [ROC] Curve
CE	Conformité Européenne
CI	Confidence Interval
CEA	Cost Effectiveness Analysis
CNN	Convolutional Neural Networks
CUA	Cost Utility Analysis
DESP	Diabetic Eye Screening Programme
DL	Deep Learning
DM	Diabetes Mellitus
DMO	Diabetic Macular Oedema
DR	Diabetic Retinopathy
DSS	Decision Support System
EDESP	English Diabetic Eye Screening Programme
ETDRS	Early Treatment of Diabetic Retinopathy Study [severity scale]
EURODIAB	EURODIAB IDDM Complications Study [grading standard]
F&T	Flow And Timing
FDA	Food and Drug Administration
ICDR	International Clinical Diabetic Retinopathy [severity scale]
ICER	Incremental Cost Effectiveness Ratio
IT	Index Test
ML	Machine Learning
mtmDR	More-than-mild Diabetic Retinopathy
NN	Neural Networks
NPV	Negative Predictive Value
OCT	Optical Coherence Tomography
PPV	Positive Predictive Value
PDR	Proliferative Diabetic Retinopathy
QALY	Quality Adjusted Life Year
rDR	Referable Diabetic Retinopathy
ROC	Receiver Operating Characteristics [curve]
RS	Reference Standard
RoB	Risk Of Bias
SE	Sensitivity
SDESP	Scottish Diabetic Eye Screening Programme
SiDRP	Singapore national Diabetic Retinopathy Screening Program
SP	Specificity
SpDESP	Spanish Diabetic Eye Screening Programme
vtDR	Vision-threatening Diabetic Retinopathy
WHO	World Health Organization

## Plain English summary

Diabetes can cause eye problems called diabetic retinopathy. If left untreated, it can get worse and lead to sight loss. Diabetic retinopathy is one of the leading causes of blindness in the working-age population in the UK.

Eye screening tests can detect eye problems before they affect sight. In the UK, all patients with diabetes aged 12 and over are invited to attend the UK Diabetic Eye Screening Programme (DESP). The frequency of screening is once a year in England, Wales and Northern Ireland. The frequency in Scotland is every two years following two consecutive negative screens.

During the visit, health professionals take images of each eye. Then, trained professionals called level 1 graders study the images. All images of patients with suspected diabetic retinopathy are sent for further assessment by level 2 graders. They also assess some of the images with no disease, to make sure that the programme is working well. If there is disagreement between level 1 and level 2 graders, level 3 graders make a final decision. Patients without diabetic retinopathy and most patients with mild disease can return for another screen after 12 months. Some patients with diabetic retinopathy need to have further assessment and treatment.

Currently, the DESP is under a lot of pressure. The number of people with diabetes continues to increase leading to greater demand for trained professionals. There are artificial intelligence (AI) systems that can study the images. So, it has been suggested that AI systems could be used for reading images of eyes. Such systems have already been implemented in Scotland and Portugal.

The UK National Screening Committee has not looked at the evidence on the use of AI in the English DESP before. Therefore, it is not currently recommended.

This review looked for evidence to see whether:

- AI systems are accurate enough to identify cases of diabetic retinopathy



- AI systems provide better health and value for money compared to the current DESP
- there is any evidence on social and ethical aspects of using AI systems in screening programmes.

The UK NSC review concluded that further research is needed before implementing AI in the English DESP. The results from the review indicate that:

- some AI systems are accurate enough to do the initial reading of images; 3 systems have been evaluated in good quality studies conducted in the UK
- a system called iGradingM is implemented in Scotland and there is evidence that its performance is good; however, we could not find much evidence on the experience with its implementation and use
- there is some evidence that using the above AI systems for initial screening of images provides better value for money, but the analyses didn't take into account some important factors
- there is sufficient evidence from England that a system called EyeArt v2.1 may be safe and provide better value for money; its implementation should be investigated further
- the evidence on social and ethical aspects of using AI systems in screening programmes should be assessed further.

# Executive summary

## Purpose of the review

The purpose of this review is to provide a reference point for a discussion on the major modification proposal to the English Diabetic Eye Screening Programme (DESP). The proposal involves the automated grading using Automated Retinal Image Analysis Systems (ARIASs) to triage patients prior to manual grading. This is achieved by identifying and synthesising research evidence on the accuracy, clinical and cost-effectiveness of ARIASs, and studies and review/opinion papers addressing the social and ethical implications of the implementation of Artificial Intelligence (AI) technology in screening programmes.

## Background

Diabetic eye disease is the most common microvascular complication of diabetes. The microvasculopathy can lead to visual dysfunction by causing diabetic retinopathy and maculopathy, and increases the risk of developing cataracts and glaucoma. Sight-threatening diabetic retinopathy is one of the leading causes of blindness in working-age population in the UK. It is caused by damage to the blood vessels that nourish the retina. This may result in areas of ischaemia and dysfunction, and may trigger the aberrant response of new vessels growing into the vitreous cavity (proliferative retinopathy) that can cause haemorrhage and scarring of the retina. Patients with severe levels of diabetic retinopathy are reported to have poorer quality of life and reduced levels of physical, emotional, and social well-being, and they utilize more health care resources. Diabetic maculopathy is caused by the leakage of fluid in the macula, which is the part of the eye responsible for detailed vision such as reading and face recognition.

In the UK all people with type 1 and type 2 diabetes aged 12 or over (excluding women who have gestational diabetes) are invited to the national screening programme for diabetic eye disease every year. In Scotland the interval is extended to 2 years following 2 consecutive years of negative screens. The aim of the screening programme is to reduce the risk of vision loss for people with diabetes mellitus through early detection of retinopathy or maculopathy during its common asymptomatic

stage. This is achieved by providing appropriate monitoring and early treatment, as patients often do not present until advanced complications when treatment outcomes are less favourable and costs are higher.

There are some variations in the screening protocols and grading schemes used in the DESPs across the UK nations. The English DESP (EDESCP) involves taking two-field 45° colour fundus photographs using pupil dilation (mydriasis) that are manually assessed by human graders. These photographs are graded as follows, in order of progression:

- no retinopathy (R0)
- background retinopathy (R1)
- pre-proliferative retinopathy (R2)
- proliferative retinopathy (R3; R3A (active); R3S (treated and stable))
- no maculopathy (M0)
- maculopathy (M1) - early maculopathy (does not require treatment) or clinically significant macular oedema (requires treatment).

Patients that are graded R0M0, or R1M0 in the more severely graded eye are invited to return for rescreen after 12 months. Patients graded R2 or those who have early maculopathy (M1) are referred to either the Hospital Eye Service (HES) or Digital Surveillance (DS) clinic where they are kept under surveillance and screened more frequently (every 3-6 months). DS clinics or HES may interface with Optical Coherence Tomography (OCT) assessment to detect maculopathy which could not be diagnosed using non-stereoscopic fundus photography. Patients graded R3A or those who have clinically significant macular oedema (M1) are referred to HES where they receive treatment which involves laser photocoagulation or injections of anti-vascular endothelial growth factor drugs. If digital images are not clear enough to allow grading, then a second test using a method called slit lamp biomicroscopy is required.

The EDESCP grading pathway is a multi-level manual grading systems. All fundus photographs are reviewed by level 1 graders. Level 2 graders review fundus photographs of patients who were graded as R1, R2, R3, and M1 by level 1 graders. Level 2 graders also review all ungradable photographs, as determined by level 1 graders, and 10% of the photographs graded as R0 (for quality control). Discrepancies between level 1 and level 2 graders are reviewed by level 3 graders.

DESPs are labour intensive and for the EDESP to cope with the increasing burden of diabetes, the UK NSC received a proposal to modify the programme. The modification involves using an ARIAS for level 1 grading to triage patients into low and high-risk cases. The proposal suggested that the detection of diabetic eye disease through ARIASs would reduce the need for human graders, thus reducing the cost of screening.

ARIASs have been used for level 1 grading in the DESPs of Scotland and Portugal, and are now considered for clinical use in other countries, such as the USA, the Netherlands, Singapore, Korea, Spain, Thailand and India. The systems used in Scotland (iGradingM) and Portugal (RetmarkerSR) are based on traditional Machine Learning (ML) algorithms. They extract from the images pre-specified 'hand-crafted' features, such as microaneurysms, and use the information to classify patients into those who have 'any disease' (images sent for human grading) and 'no disease' (routine recall). Both programmes have quality assurance protocols that monitor the performance of the ARIAS.

Most of the recently developed systems use Deep Learning (DL) algorithms that do not depend on pre-specified features. Instead, they use raw data to build up high level features that allow better classification of images and higher accuracy.

## Focus of the review

The aim of the current review is to identify and synthesise the evidence on the major modification proposal to the EDESP which involves the automated grading with an ARIAS to triage patients prior to manual grading. However, the UK NSC will consider the evidence reviewed here in the context of the whole UK. The review aimed to answer the following questions:

**Question 1 (rapid review, criteria 4 & 5):** What is the diagnostic accuracy of ARIASs at detecting diabetic eye disease in patients with diabetes mellitus? (The term 'diagnostic accuracy' does not imply that the system is used to diagnose diabetic retinopathy or any other condition. The only role of ARIASs investigated in the current review is to identify patients with low risk defined as 'no disease' or 'non-referable disease' and high risk defined as 'any disease' or 'referable disease', prior to

manual grading and as part of a multi-level screening programme, such as the EDESP).

**Question 2 (rapid review, criteria 11 & 12):** What is the clinical impact of DESPs using an ARIAS for level 1 grading compared with DESPs with fully manual grading?

**Question 3 (evidence map, criterion 14):** What is the cost-effectiveness of DESPs using ARIASs compared to DESPs with manual grading?

**Question 4 (evidence map, criterion 12):** What are the social and ethical implications of implementing artificial intelligence (AI)-based tools in screening programmes and would it be acceptable to health professionals and the public?

Two methodologies were used: Question 1 and 2 were addressed with a rapid review whereas questions 3 and 4 with an evidence map. For questions 1 to 3 we prioritised UK-based studies and studies conducted in similar populations and settings; studies evaluating commercially available and CE-marked/FDA-approved ARIASs; we also prioritised RCTs and prospective cohort studies. For question 4, all primary studies evaluating the social and ethical impact of AI-based technology in medical screening programmes or similar settings were included; we also included, but reported separately, all review and opinion papers that appeared to be relevant to this question (based on the information provided in their abstracts).

Two separate searches were carried out from January 2000 to June 2020. The first one addressed questions 1-3 (accuracy, effectiveness and cost-effectiveness). The second search addressed question 4 (social and ethical aspects of AI in screening programmes). We also searched the reference lists of included studies and other publications and, if necessary, contacted authors and manufacturers to request further information.

## Findings and gaps in the evidence of this review

Ninety two studies were judged to be relevant to at least one question and included in the review; of those 3 studies addressed more than one question. We identified:

- 52 studies relevant to question 1 (accuracy of ARIASs), of which 28 studies met one or more of the prioritisation criteria detailed above and were included in the narrative synthesis

- 2 studies relevant to question 2 (effectiveness of ARIASs)
- 5 studies (9 publications) relevant to question 3 (health economic evaluation of ARIASs)
- 57 studies relevant to question 4 (social and ethical impact of ARIASs), of which 19 were primary studies and 38 were review and opinion papers.

### **Question 1 (accuracy of ARIASs; rapid review):**

Twenty eight studies evaluating 10 ARIASs were prioritised for inclusion in the narrative synthesis. Sixteen studies evaluated 7 DL-based systems: EyeArt v2.1 (n=5 studies), EyeGrader (n=1), IDx-DR v2 (n=4), Google AI (n=2), RedCAD (n=1), SELENA (n= 2) and VUNO (n=1). Further 12 studies evaluated ARIASs based on traditional ML algorithms: iGradingM (n=7), RetmarkerSR (n=4), RetinaLyze (n=2) and EyeArt v1 (two versions of EyeArt were included: an earlier ML-based version and the current DL-based version). One study compared the performance of 3 systems: EyeArt v1, iGradingM and RetmarkerSR).

#### *DL-based ARIASs*

Only 2 of the DL studies, both evaluating EyeArt v2.1, were conducted in the UK: Heydon 2020 which was a large prospective multi-centre study and Olivera-Barrios 2020 which was a retrospective study comparing the performance of EyeArt v2.1 when used with the EDESP photographic protocol and an alternative wide-field platform. Only EyeArt v2.1 and IDx-DR v2 were evaluated in prospective clinical studies: Heydon 2020 (UK), Lim 2019 (USA) and Liu 2020 (USA) evaluated EyeArt v2.1, and Abramoff 2018 (USA) and van der Heijden 2018 (the Netherlands) evaluated IDx-DR v2. By ‘prospective study’ we mean that data collection was planned before the index test and reference standard were performed (see the definition in the STARD checklist<sup>1</sup>). None of the prospective clinical studies compared the performance of alternative ARIASs in the same cohort of patients.

The number of participants ranged from 96 to >30 000. There was considerable variation across studies in terms of selection and characteristics of the included participants; prevalence and spectrum of diabetic retinopathy; setting; screening pathway and protocol; reference standard; grading scheme and handling of ungradable images. All these sources of variation are likely to affect the performance of ARIASs,

---

<sup>1</sup> <https://www.equator-network.org/reporting-guidelines/stard/>

especially when the system is evaluated away from the setting in which it was developed and initially evaluated. Some of these factors were investigated in the included studies or other publications and the results are summarised and reported in the review.

Most of the studies were considered to be at high or unclear risk of bias in at least one of the QUADAS-2 domains. The main issues concerned the selection of patients (only 3 studies were considered at low risk of bias); the reference standard (the reference grading did not involve a panel of ophthalmologists or retinal specialists, or the final grade from a multi-level screening programme with established training and quality assurance protocols) and the exclusion of ungradable images. Since we could not determine whether the mix of patients in non-UK studies was similar to that in the UK DESPs *in all significant ways*, we graded all such studies 'unclear' for applicability concerns unless the population was clearly different from that in the UK (in which case the study was considered at high risk). In addition, 7 studies were graded 'high' or 'unclear' for applicability concerns in the Index Test domain as the photographic protocol differed from those used in the UK DESPs.

Across all studies, sensitivity for referable diabetic retinopathy was consistently >90%. There were 3 exceptions, all relating to the IDx-DR v2 system: Abramoff 2018 (USA) reported 87.2% (95% CI 81.8% to 91.2%) sensitivity when stereoscopic wide-field fundus photography was used as a reference standard and 85.9% (95% CI 82.5% to 88.7%) when the latter was combined with OCT; Verbraak 2019 (the Netherlands) reported sensitivity of 79.4% (95% CI 66.5% to 87.9%) due to the presence of a single isolated haemorrhage or cotton wool spot (but no microaneurysms) in the 13 false negative results; and van der Heijden 2018 (the Netherlands) reported considerable difference in sensitivities when ICDR (3) and EURODIAB (4) criteria were applied.

The picture was similar when considering the results for sight-threatening or vision-threatening diabetic retinopathy. Most studies reported sensitivity of around or more than 95% including Abramoff 2018 (IDx-DR) against both reference standards ('4W-D' and '4W-D & OCT'). The prospective pivotal study evaluating EyeArt v2.1 against a '4W-D' photographic protocol reported variable sensitivities ranging from 78.6% to 100% (reported separately for different cohorts and settings, with wide CIs due to the small samples sizes). Another exception was van der

Heijden 2018 (IDx-DR v2, the Netherlands) who reported sensitivity of 64% (36% to 86%) using the EURODIAB criteria (4) and 62% (32% to 85%) using the ICDR criteria (3). The reason for these very different results was unclear from the paper.

The specificity ranged considerably, from 54.0% (95% CI 53.4% to 54.5%) for R0M0 & R1M0 in Heydon 2020 (UK; EyeArt v2.1) to >95% in some studies, which reflects the inclusion/exclusion of ungradable images and other sources of variation. The relatively low specificity in Heydon 2020 is most likely due to the fact that all images that would normally be sent for manual grading were graded by the ARIAS and included in the analysis. The two studies that used 4W-D protocol in the reference standard reported much higher specificities. Lim 2019 (USA; EyeArt v2.1) reported specificity of 86.5% (95% CI 84.3% to 88.7%) when 'dilation-if-ungradable' protocol was used and 86.0% (95% CI 83.7% to 88.4%) for 'no dilation'. Abramoff 2018 (USA; IDx-DR v2) reported specificity of 90.7% (95% CI 88.3% to 92.7%) against the reference standard of 4W-D and 90.7% (95% CI 86.8% to 93.5%) against '4W-D & OCT', which remained unaffected (89.8%) in the sensitivity analysis when ungradable images were included. Finally, van der Heijden 2018 (IDx-DR) reported that specificity was comparable (86% and 84%, respectively) when using the ICDR (3) and EURODIAB criteria (4).

Of all included studies evaluating DL-based ARIASs, we considered only Heydon 2020 to be of sufficient quality and to allow direct generalisation to the EDESP. The study included a cohort of 30 405 consecutive patient episodes from 3 current EDESP centres. For referable disease (M1, R2, R3) sensitivity was 95.7% (95% CI 94.8% to 96.5%); specificity (for the combination of R0M0 and R1M0) was 54.0% (95% CI 53.4% to 54.5%); the detection rate for R2 was 100% (95% CI 98.7% to 100%) and for R3 100% (95% CI 97.9% to 100%); and the system was able to identify 89.4% (95% CI 87.0% to 91.5%) of the images classified as 'ungradable' by human graders. The accuracy was similar across the 3 centres. The study was rated 'low' for risk of bias and applicability concerns in all QUADAS-2 domains. No subgroup analysis was reported in the paper, but the authors provided additional information which is included in the report (personal communication).

#### *ML-based ARIASs*



RetinaLyze was evaluated in 2 small studies at high risk of bias. The most recent study was published in 2008 and was conducted in the UK. The manufacturer confirmed that this was the most recent evaluation and that the system has not been upgraded since then.

Four studies evaluated RetmarkerSR, of which only the most recent one, Tufail 2016, was conducted in the UK and was judged to be at low risk of bias. The sensitivity for any retinopathy was 73.0% (95% CI 72.0% to 74.0%); for referable retinopathy was 85.0% (95% CI 83.6% to 86.2%) and for proliferative retinopathy was 97.9% (95% CI 94.9% to 99.1%). The specificity for R0M0 & R1M0 was 47.7% (95% CI 47% to 48.5%). The other 3 studies were conducted in Portugal. Of those Ribeiro 2015 reported that in the Portuguese DESP the false negative rate of the system was 0.3% in quality control cases and 0.02% in all screened patients

iGradingM/Aberdeen system has been in use in the SDESP since 2011. Prior to its implementation it had been evaluated in 3 large SDESP cohorts one of which was prospectively recruited. The reported sensitivities were consistently >90% for any or referable diabetic retinopathy and approached 100% for sight-threatening disease. The specificity was around 67% in two studies while the rest reported detection rate for individual grades that will translate into similar results. One study published only as a conference abstract reported the results from an internal quality assessment from the SDESP. Sensitivity was comparable to that in previous evaluations but specificity was lower (sensitivity 97%, specificity 38% and false negative rate ranging from 0 to 0.6%) (13). It was also reported that the number of episodes handled by the programme had increased by 20.3% in the period from 2010 to 2015 and in the observed 6-month period in 2015, 58.1% of all episodes were passed on to the autograder.

iGradingM failed to read disc-centred images when evaluated in the EDESP by Tufail et al (1). However, the results reported by Goatman 2011 who compared the EDESP and SDESP photographic protocols, were in line with those reported in the Scottish evaluations (5). According to Tufail et al the discrepant results could be explained with pre-processing of the images in Goatman 2011 (personal communication). In the comparative HTA assessment conducted by Tufail et al, EyeArt v1 achieved higher sensitivity but lower specificity compared to RetmarkerSR. The accuracy of EyeArt v1 was not affected by ethnicity, sex or camera type, but sensitivity was marginally lower with increasing

patient age. The accuracy of RetmarkerSR appeared to vary with patient age, ethnicity and camera type.

#### *ARIAS compared to human graders*

We included 4 studies that compared ARIASs to human graders not involved in the reference grading: 3 evaluated DL-based ARIASs (Google AI, SELENA and RetCAD) and one evaluated iGradingM. The study evaluating iGradingM was the only prospective evaluation and the only study conducted in the UK (Scotland). All studies were judged to be at high risk of bias in at least one QUADAS-2 domain, but none of these issues concerned comparative accuracy. On the whole, DL-based ARIASs had higher sensitivities and lower specificities compared to human graders. The study evaluating iGradingM in the SDESP reported that the system had sensitivity of 90.5% (95% CI 89.3% to 91.6%) and specificity of 67.4% (95% CI 66.0% to 68.8%) for referable diabetic retinopathy. In comparison, the sensitivity of manual grading was 86.5% (95% CI 85.1% to 87.8%) and the specificity 95.3% (95% CI 94.6% to 95.9%). iGradingM and human graders misclassified as normal 240 and 341 patients with diabetic retinopathy, respectively, and the difference was statistically significant ( $p < 0.001$ ). Of the patients with M1, R2, M2, R3 or R4, iGradingM and human graders classified 7/330 and 3/330 as 'no retinopathy', respectively, but the difference was not statistically significant ( $p = 0.125$ ).

**Question 2 (effectiveness of ARIAS; rapid review):** We did not identify any RCTs comparing DESP with level 1 human graders to DESP with level 1 ARIAS grading in terms of clinical outcomes and other measures of impact. Some accuracy studies reported projected impact, such as workload reduction, but since these outcomes were just a different way of expressing the accuracy estimates, we do not include the results here.

We included 2 prospective cohort studies. Keel 2018 (Australia) investigated the acceptability of ARIAS-based DR screening ( $n = 96$ ) in which the patient is immediately provided with the test result, relative to manual grading in which the patient can access the result after 2 weeks. The authors reported that the average time for ARIAS-based screening was 6.9 min; 96% of participants were very satisfied or satisfied with the screening and 78% said they preferred ARIAS. However, patients served as their own controls (no proper control group) and the study was judged to be at high risk of bias.

Liu 2020 (USA) investigated whether using ARIAS to make the result from the examination immediately available to patients improved their adherence to follow-up. They reported that in the study cohort (n=180) the adherence rate was 55.4% at 1 year versus historical adherence rate of 18.7% ( $P < 0.0001$ ). However, the study was judged to be at high risk of bias due to the following issues: historical controls; lack of clarity as to whether the patients in the historical cohort received the same level of encouragement to attend a follow up examination; and considerable delay in making the results available to some patients.

**Question 3 (cost-effectiveness of ARIAS; evidence map):** After excluding all non-UK studies, 5 studies evaluating 3 ML-based systems (EyeArt v1, iGradingM and RetmarkerSR) were included in the evidence map. The studies generally found automated grading to be less effective than manual grading, but also less costly. Therefore, many of the results are reported in terms of the additional costs associated with manual grading to gain additional health benefits when compared to automated grading. The HTA conducted by Tufail et al investigated the cost-effectiveness of EyeArt v1 and RetmarkerSR in the EDESP comparing two alternative strategies: 1) ARIAS replacing level 1 manual grading, and 2) ARIAS acting as a filter prior to level 1 manual grading. They found that both ARIASs are cost-effective with either strategy, but strategy 1) was preferred. Although these studies provide good starting point for further evaluations, they need updating to include new data (e.g. costs and accuracy of the new versions of the software) and to capture the cost-effectiveness of the ARIAS over time.

**Question 4 (social and ethical implications; evidence map):** Nineteen primary studies and 38 reviews and opinion papers were included in this evidence map after reviewing their abstracts. Five of the primary studies investigated the impact of AI in the context of screening, while the rest had a more general focus (e.g. survey of radiologists) or were conducted in a different healthcare setting (e.g. hospital), but were judged to address relevant questions.

All primary studies were surveys of clinicians (n=9), clinicians and the general public (n=2), the general public (n=2) and patients (n=5). The surveys were conducted in the USA (n=4), UK (n=3), France (n=1), and one each in Australia, China, Europe, Germany, India, Italy, Singapore

and Sweden. They investigated a broad range of questions including participants' knowledge, training needs, perceptions, attitudes and satisfaction in relation to AI-based technology. Most of the participants had positive attitudes towards the implementation of AI in healthcare. They acknowledged the benefits that such technologies are likely to bring, both in terms of improved patient outcomes and benefits to the healthcare system as a whole. The studies also reported a range of concerns regarding the impact of AI-based technology on clinicians' professional role and identity, clinician-patient relationship, dealing with uncertainty, the impact on clinical decision making, and the need of training and better understanding of AI by healthcare professionals, patients and the general public.

Only one of the papers was a systematic review. It looked at the characteristics and usability features of tele-ophthalmology for the elderly population, including AI-based screening (Fatehi 2020).

## Recommendations

There is sufficient high quality evidence from the EDESP that EyeArt v2.1 is safe and cost-effective either as level 1 grading or as a filter prior to level 1 manual grading. An implementation study should be carried out to investigate other aspects of its performance, such as interoperability, cyber security, impact on human behaviour, acceptability by staff and patients, and to inform an update of the cost-effectiveness analysis. As a precaution, the system should be implemented initially to screen out patients with no disease while its performance at a higher threshold (referable/non-referable disease) and longer screening intervals (similar to those in the SDESP) is investigated as part of the study.

We also recommend an evidence review on the social and ethical aspects of implementing AI-based technology in population screening. Such review could identify the range of relevant topics, summarise the existing evidence and identify gaps that need further investigation. However, the traditional rapid review approach with narrow focus on screening may not be appropriate. Instead, methods such as realist synthesis could be employed to make use of evidence from related settings that deal with similar issues and generate evidence transferable to screening.

*Future research:*

- Should be done independently from the software developer, in the clinical setting in which the system is meant to be used, under conditions that reflect everyday clinical practice; if possible, they should compare the performance of alternative ARIASs that may have different advantages and disadvantages.
- Should look at outcomes beyond accuracy, such as the actual consequences of false negative and false positive results, and the consequences of accidental findings (e.g. missed by ARIAS but referred by human graders).
- They should include a cost-effectiveness evaluation of the system.
- They should investigate the experience and perceptions of healthcare professionals who interact with and/or are directly affected by the ARIAS; the expectations of those who have not yet had this experience (e.g. those in the control arm); the experience and perceptions of relevant patient groups; and the overall impact on the NHS.

We identified a considerable number of papers looking at the social and ethical aspects of AI implementation in screening programmes. An evidence review of this growing literature will help identify all relevant aspects of the above question, to summarise the existing evidence and identify any gaps that need to be addressed in future research.

## Limitations

The following methodological limitations of the review should be acknowledged: only the main electronic databases were searched; searches were limited to records published since 2000, and included only peer-reviewed, English-language journal articles; only 20% of the titles and abstracts were double-screened; papers were excluded after assessment of the volume of published evidence (although prioritisation was based on pre-specified criteria); the definition of one of the signalling questions in QUADAS-2 checklist was changed following a discussion with experts in the field and after the initial grading of studies; as a result some studies were later regraded using the new definition.

# Introduction and approach

## Background

### Target condition

#### **Diabetes**

Diabetes Mellitus is a chronic metabolic condition in which the body doesn't make enough insulin (the hormone that regulates blood sugar) or is unable to use it effectively. This results in hyperglycaemia (raised blood sugar) which can lead to serious damage of multiple body systems, especially the nerves and the blood vessels. According to the World Health Organisations (WHO) the global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014 and continues to rise, especially in low- and middle-income countries. WHO estimates that diabetes was the 7<sup>th</sup> leading cause of death in 2016 and a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation (6).

In the UK, there were 3.9 million people with diabetes in 2019 and this number is expected to rise to 5.3 million by 2025. Of those, 90% have type 2 diabetes, 8% type 1 diabetes and about 2% have rarer forms of diabetes (7).

#### **Diabetic eye disease**

Diabetic eye disease is the most common microvascular complication of diabetes mellitus. Microvasculopathy can lead to visual dysfunction by causing diabetic retinopathy and maculopathy, and increases the risk of developing cataracts and glaucoma.

Diabetic retinopathy is caused by damage to the blood vessels that nourish the retina (the tissue that lines the inner surface of the eye surrounding the vitreous cavity). Damage to the capillaries causes closure of the vessels and/or leakage of fluid. These vascular changes may result in areas of ischaemia and dysfunction of the retina, which may trigger the aberrant response of new vessels growing into the vitreous cavity (proliferative retinopathy). This, in turn, can cause haemorrhages and scarring of the retina.

Diabetic maculopathy is caused by the leakage of fluid in the macula which is the part of the eye responsible for detailed vision, such as reading, counting and face recognition. Clinically significant maculopathy is arbitrarily defined by the presence of retinal thickening or hard exudates within one disc diameter of the fovea (8).

Diabetic retinopathy/maculopathy is the second most common cause of blindness in working-age population in the UK. It accounts for 14.4% of all cases in the period between 1 April 2009 and 31 March 2010 (the leading cause was hereditary retinal disorders with 20.2%). Although this percentage is still quite high, it is a considerable improvement from the previous period used as a comparator in the study, April 1999 and 31 March 2000. During the latter diabetic retinopathy/maculopathy was the leading cause of blindness accounting for 17.7% of all cases. The improvement is attributed to the introduction of a national diabetic retinopathy screening programme and improved glycaemic control (9).

Worldwide, around 35% of people with diabetes mellitus have some form of diabetic retinopathy and 10% are affected by sight-threatening diabetic retinopathy (10). These estimates will soon be revised in an upcoming systematic review (11), but the results might be difficult to interpret given the considerable variation in populations and methods for data collection and analysis.

There are several risk factors for diabetic retinopathy and maculopathy including the duration and type of diabetes, degree of hyperglycaemia, hypertension, hyperlipidaemia and diabetic nephropathy. Patients with severe levels of diabetic retinopathy are reported to have poorer quality of life and reduced levels of physical, emotional and social well-being, and to utilize more health care resources (12).

## Current policy context and previous reviews

### **Diabetic eye screening programmes in the UK**

The UK was the first country to introduce a national screening programme for diabetic eye disease. In England, Scotland and Wales, the Diabetic Eye Screening programme (DESP) has been implemented since 2003 (2), and in Northern Ireland since 2008. The aim of the screening programme is to reduce the risk of vision loss for people with diabetes mellitus through an early detection of retinopathy and maculopathy during

the asymptomatic stage of the disease, and by providing them with appropriate monitoring and treatment. Without screening patients often present with advanced complications when treatment outcomes are less favourable and costs are higher.

The eligible population for these programmes is all people with type 1 and type 2 diabetes aged 12 or over (excluding women who have gestational diabetes). The differences between DESPs across the 4 nations are summarised in Table 1.

**Table 1 Differences between DESPs across the 4 nations.**

<b>Nation</b>	<b>Scotland</b>	<b>England</b>	<b>Wales</b>	<b>Northern Ireland</b>
<b>System</b>	Single commissioned National	Multiple individually commissioned Regional	Single commissioned National	Single commissioned National
<b>Software</b>	Single commissioned National system	No National system, 2 suppliers	Single commissioned National system	Single commissioned National system
<b>Automation</b>	Automated primary grading	None	None	None
<b>Images</b>	1-field	2-field	2-field	2-field
<b>Extended Intervals</b>	Yes	No	No	No
<b>Added OCT</b>	Yes			
<b>OCT - Optical Coherence Tomography</b>				

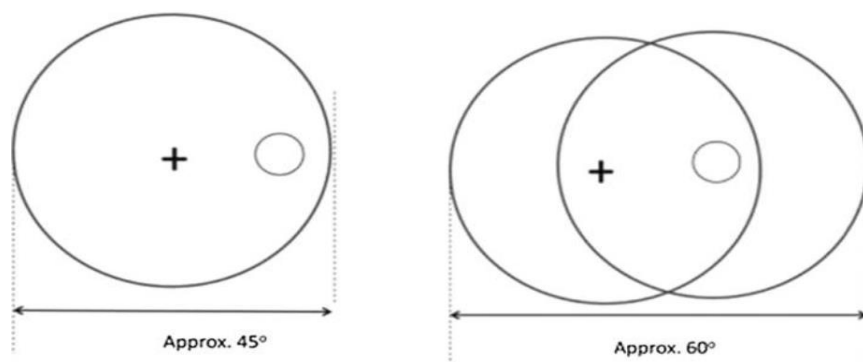
Generally, fundus photography is used as the screening test for detecting retinopathy. Fundus photography is an ophthalmic imaging technique that shows a magnified and subtle view of the surface of the retina. Two or more photographs may be overlapped to create a wider field of view (Figure 1). Diabetic maculopathy is difficult to identify by non-stereo fundus photography as oedema is transparent and therefore surrogate markers are used (i.e. signs of retinopathy close to the macula). Visual acuity decline may also be used to screen for diabetic macula oedema (DMO) that causes a functional decline. The diagnostic test for DMO is Optical Coherence Tomography (OCT) or examination by an ophthalmologist.



The English Diabetic Eye Screening programme (EDESP) involves taking two-field (disc- and macula-centred) 45° colour fundus photographs using pupil dilation (mydriasis). These photographs are assessed by human graders for the presence of both retinopathy and maculopathy, and graded as follows, in order of progression:

- no retinopathy (R0)
- background retinopathy (R1)
- pre-proliferative retinopathy (R2)
- proliferative retinopathy (R3; R3A (active); R3S (treated and stable))
- no maculopathy (M0)
- maculopathy (M1) - early maculopathy (does not require treatment) or clinically significant macular oedema (requires treatment)

The final outcome indicates the presence or absence of both retinopathy and maculopathy (e.g. R1M0).



**Figure 1**

a) One standard 45° fundus image (Scottish Diabetic Eye Screening Programme); b) Two 45° fundus images combined creating 60° field of view (NHS Diabetic Eye Screening programme); adapted from Scanlon 2017 (2)

The patient pathway in the EDESP is depicted in **Error! Reference source not found.** while Figure 3 provides a more detailed representation of the grading part of the pathway. More information about the EDESP can be found in the “*NHS Diabetic Eye Screening Programme Overview of patient pathway, grading pathway, surveillance pathways and referral pathways*” (13) and on the Public Health England’s website.<sup>2</sup>

---

<sup>2</sup> <https://www.gov.uk/topic/population-screening-programmes/diabetic-eye>

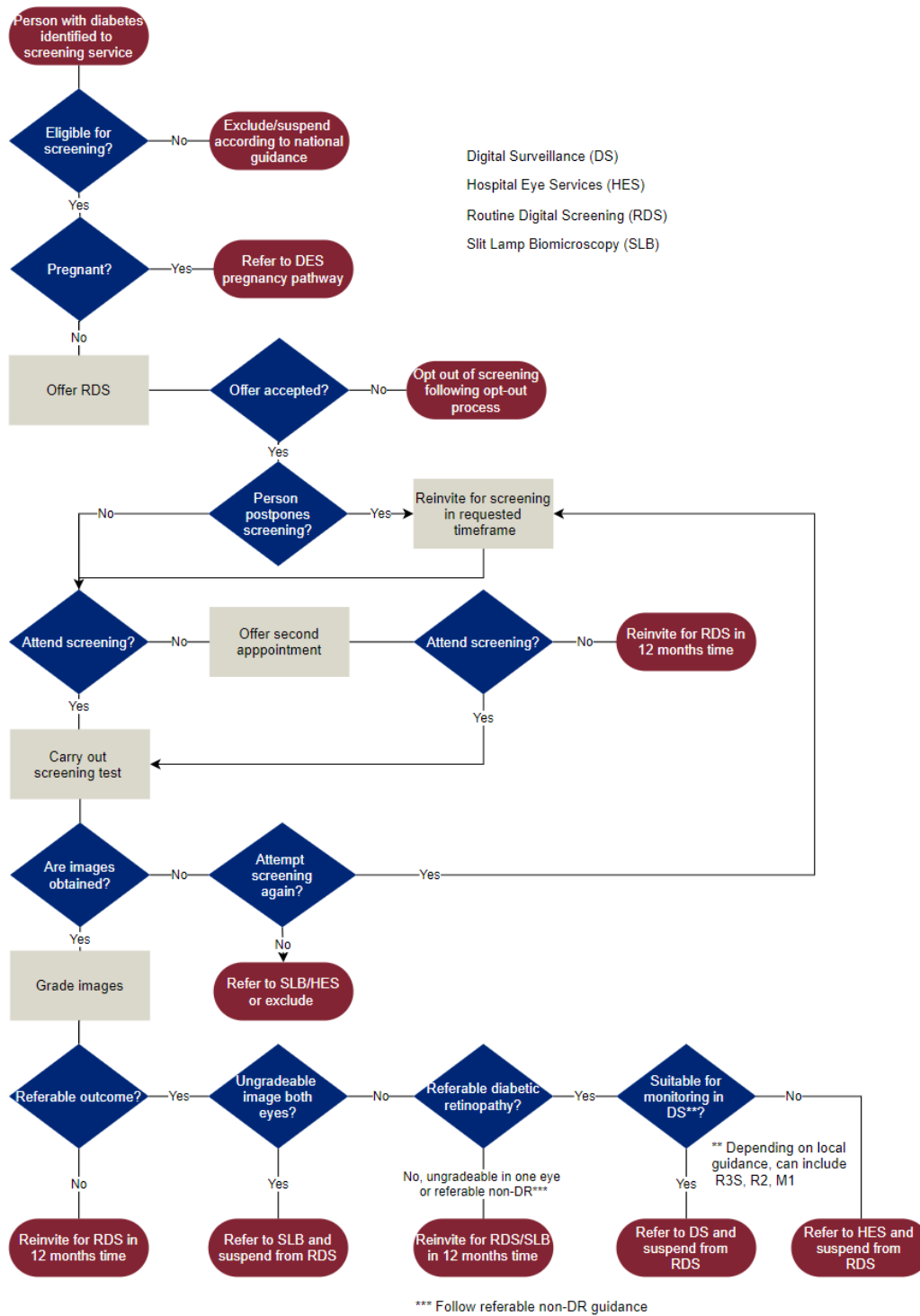
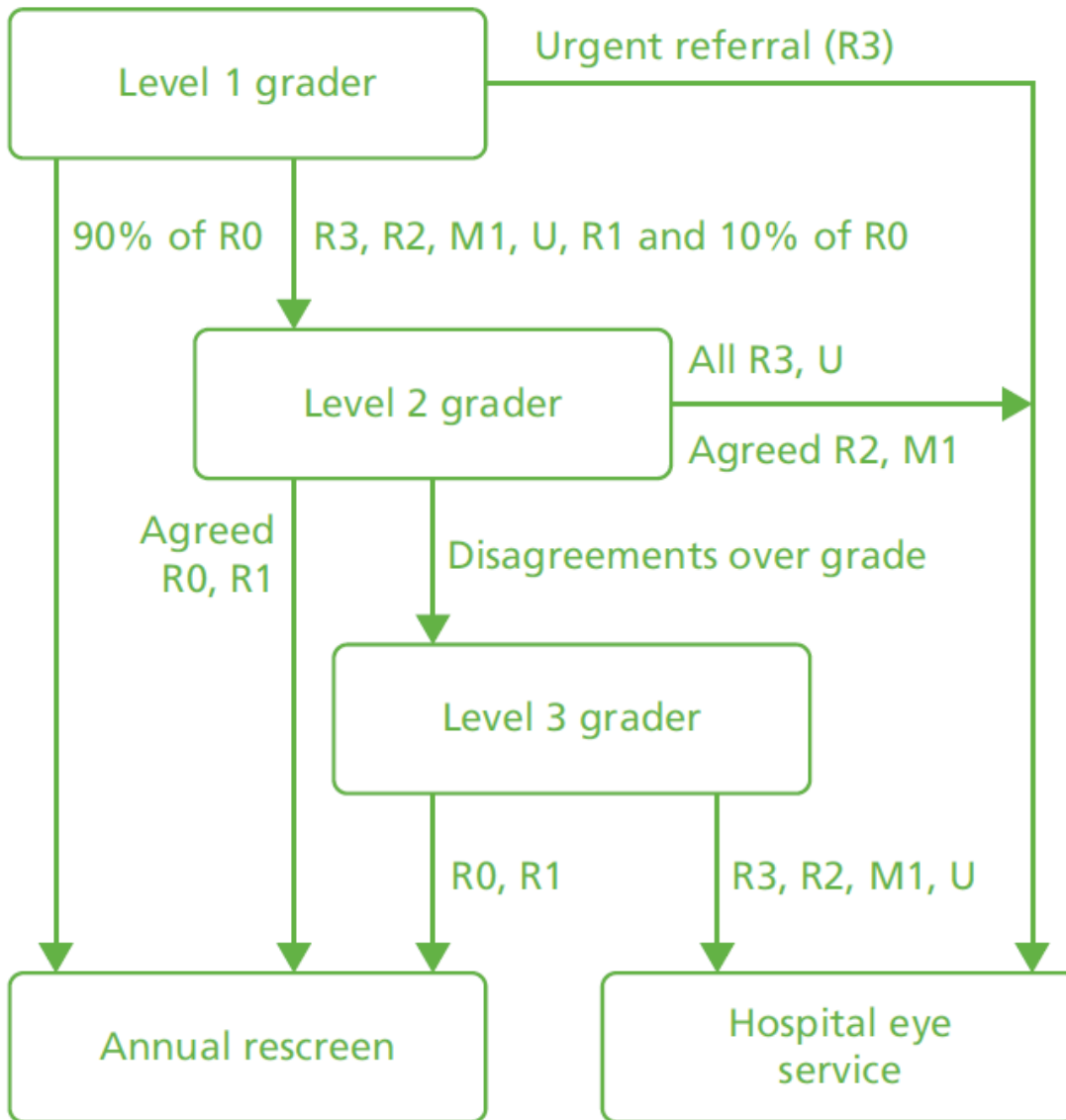


Figure 2 Patient pathway in the EDESP (PHE 2021)



**Figure 3** Grading pathway in the EDESP (Tufail 2016 (1))

Patients who are graded R0M0 or R1M0 in the more severely graded eye are invited to return for rescreen after 12 months. Screening for diabetic eye disease was introduced at annual intervals for pragmatic and administrative reasons. However, the evidence base to support this interval was very limited. In 2016, the UK NSC recommended extending the screening interval to 2 years in patients with low risk of developing sight-threatening diabetic eye disease. However, this recommendation is yet to be implemented (14) and currently 2-year interval is used only in Scotland<sup>3</sup> (Table 1).

Patients graded R2 or those who have early maculopathy (M1) are referred to either the Hospital Eye Service (HES) or Digital Surveillance (DS) clinic where they are kept under surveillance and screened more frequently (every 3-6 months) to monitor the progression of the disease. DS clinics or HES may interface with OCT assessment to detect maculopathy. The use of OCT is not currently included in the EDESP but can be added if commissioned by the local Clinical Commissioning Groups; therefore, practices vary across the country.

Patients with active proliferative retinopathy (R3A) or those who have clinically significant macular oedema (M1) are referred to HES for treatment which involves laser photocoagulation or injections of anti-vascular endothelial growth factor drugs (15). If the digital images are not clear enough to allow grading, then a second test using a method called slit lamp biomicroscopy might be required.

A common multi-level manual grading pathway of fundus photographs is shown in Figure 3. All fundus photographs are reviewed by level 1 graders. Level 2 graders review fundus photographs of patients who were graded as R1, R2, R3, and M1 by level 1 graders. Level 2 graders also review all ungradable photographs and, for quality control, 10% of the photographs which were graded as R0 by level 1 graders. Discrepancies between level 1 and level 2 graders are reviewed by level 3 graders. Based on these grades, patients are referred according to the pathway described above.

---

<sup>3</sup> <https://www.nhsinform.scot/healthy-living/screening/diabetic-eye-screening/diabetic-eye-screening-des#your-screening-appointment>

### **Proposal for a major modification to the EDESP**

DESPs are labour intensive and the number of individuals with diabetes mellitus are projected to escalate in future. The UK NSC received a proposal to modify the EDESP by using an ARIAS for level 1 grading to triage patients into low and high-risk cases. The proposal suggests that incorporating ARIAS in the screening pathway would reduce the need for human graders, thus reducing the cost of screening. An HTA assessment conducted by the research group who made the proposal investigated two different roles of ARIAS in the screening pathway: 1) using an ARIAS for level 1 grading (Figure 4) and 2) using an ARIAS prior to level 1 manual grading (Figure 5). The study found both strategies to be cost-effective, with an ARIAS acting as level 1 grader being slightly more cost-effective (1).

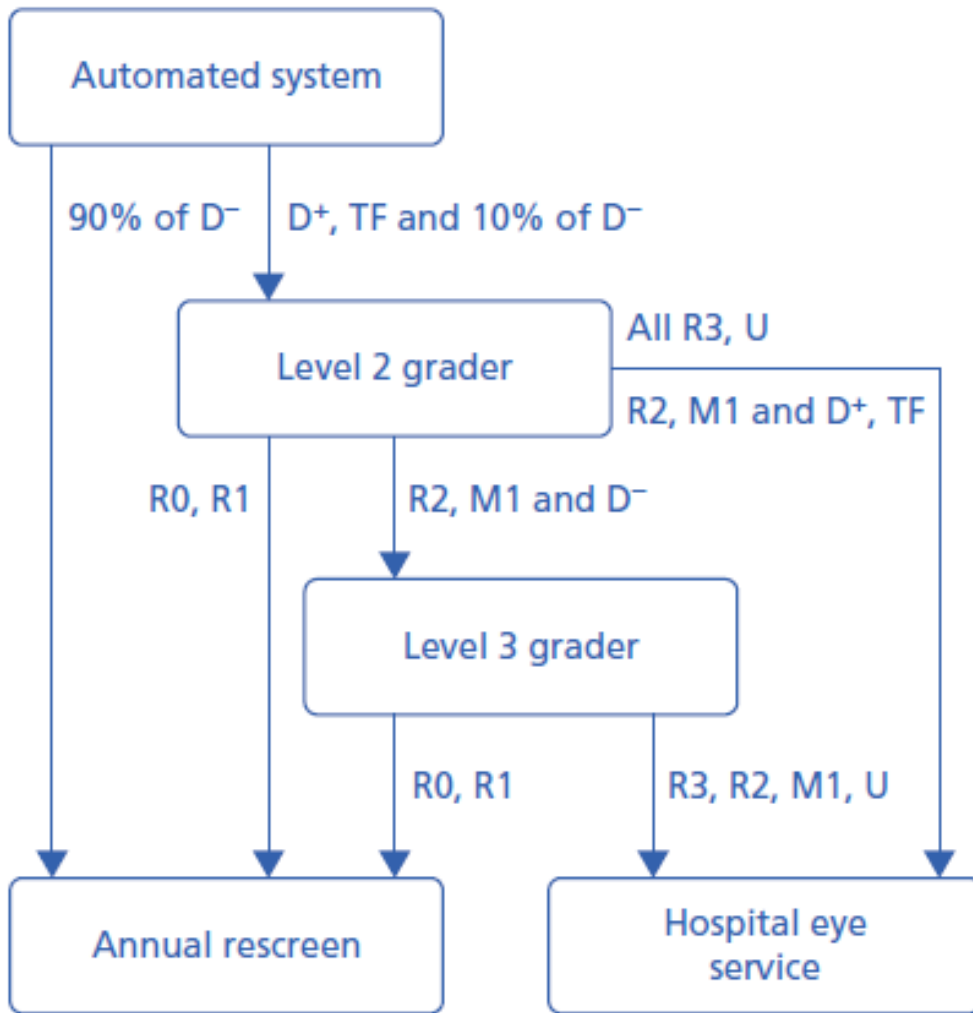
ARIASs are Artificial Intelligence (AI) based systems designed to read digital retinopathy images and classify them according to pre-specified criteria (e.g. images with and without microaneurysms). Although no single definition of AI is available, the *Evidence standards framework for digital health technologies [ECD7]* published by the National Institute for Health and Care Excellence defines AI as "an area of computer science that makes it possible for 'machines' to learn from new experiences, adjust outputs and perform human-like tasks." (16)

ARIASs have been used for level 1 grading in the DESPs of Scotland (17) and Portugal (18) since 2011 and are now considered for clinical use in other countries (19), such as the USA (20), The Netherlands (21, 22), Singapore (23), Korea (24), Spain (25), Thailand (26) and India (27). The systems used in Scotland (iGradingM) and Portugal (RetmarkerSR) are based on Machine Learning (ML) algorithms which extract from the images pre-specified 'hand-crafted' features (e.g. microaneurysms) and use them to classify image sets into 'disease', 'no disease' and 'ungradable images'. Most of the recently developed systems use Deep Learning (DL) algorithms that do not depend on pre-specified features; instead, they use raw data to build up high level features which leads to better accuracy (19).

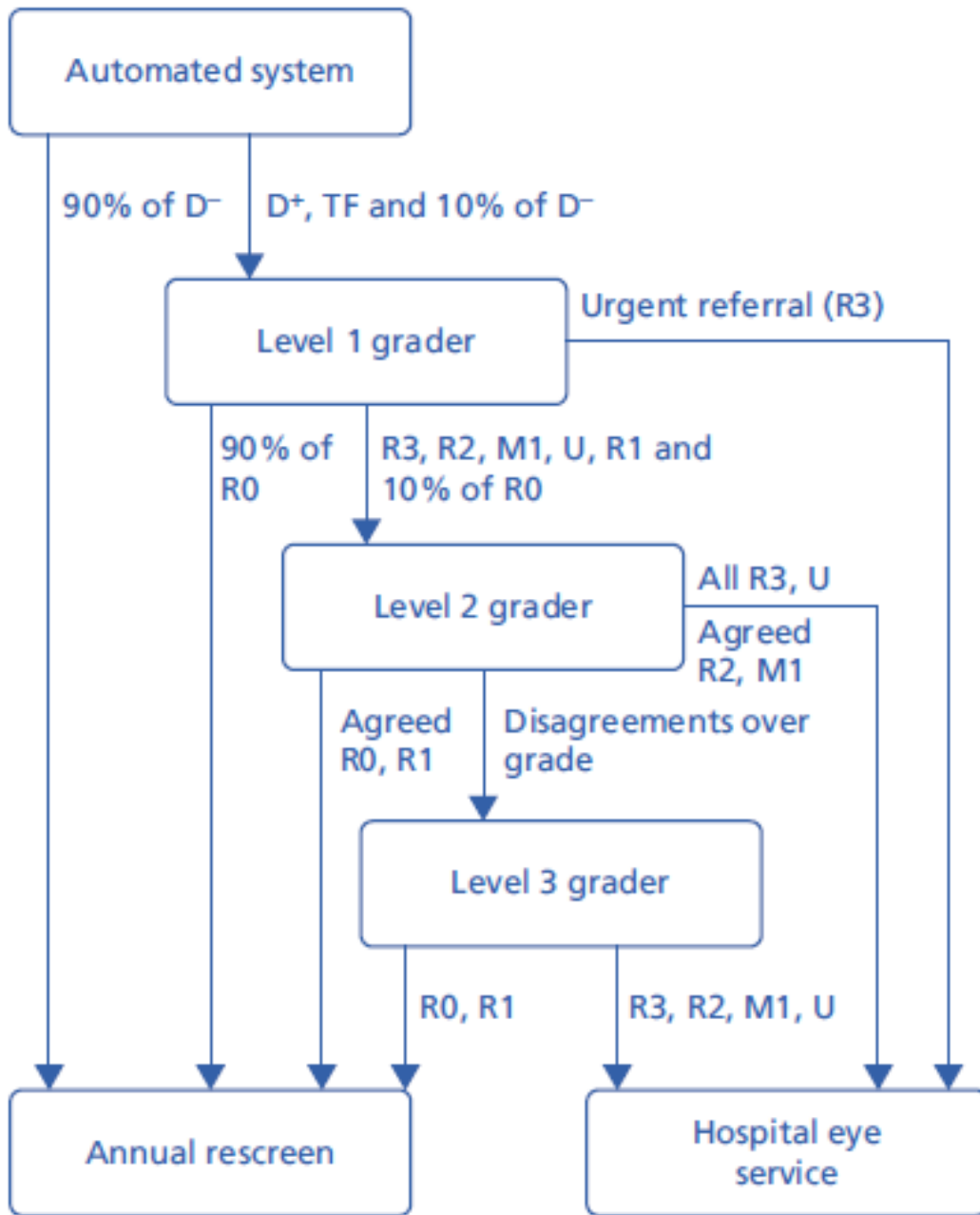
In Scotland, eligible patients undergo acuity testing and non-mydratric (45° image) fundus photography. The stored images go through 3 levels of grading. Most of level 1 grading is done by the ARIAS, which detects microaneurysms and separates the images with disease (plus ungradable

images) from those with no disease. Level 2 graders are either optometrists or nurse practitioners who have the appropriate training. They assign the images with no sight-threatening retinopathy to one of the two outcomes: 1) rescreen in 6 months, or 2) rescreen in 12 months. Images with signs of sight-threatening retinopathy or maculopathy are sent to Level 3 graders. The level 3 graders also undertake internal quality assurance of level 1 and level 2 graders with 500 random images from each grader included yearly for quality assurance assessment. In addition, once a year all Scottish graders must participate in external quality assurance which involves grading 100 images over a period of 4 weeks (28).

In Portugal, eligible patients complete a small questionnaire and undergo visual acuity testing and the acquisition of two 45° non-mydratic fundus images per eye. Images are graded at a central reading centre in two steps: 1) automated analysis using RetmarkerSR, and 2) human grading. The software separates images with and without signs of diabetic retinopathy. It also allows comparison within the same retinal location between different screening visits and detects disease progression. Images with signs of pathology/evolution are sent for human grading. In addition to the quality assurance process for human graders, there is a quality control protocol for the software. A configurable percentage of the images graded by the algorithm as not requiring human grading are randomly selected and sent to human graders who are blinded to this process (18).



**Figure 4** Proposed EDESP modification, using an automated system for level 1 grading (Tufail 2016 (1))



**Figure 5** ARIAS used as a filter prior to level 1 manual grading, as investigated in the cost-effectiveness analysis conducted by Tufail 2016 (1)



## Objectives

The aim of the current review is to identify and synthesise the evidence on the major modification proposal to the EDESP which involves automation of grading using ARIASs to triage patients into low and high-risk cases. However, the UK NSC will consider the evidence reviewed here in the context of the whole UK. The key questions the review attempts to answer are presented in Table 2 below against the UK NSC screening criteria.

**Table 2 Key questions for the evidence summary, and relationship to UK NSC screening criteria**

Criterion	Key questions	Studies Included
<b>THE TEST</b>		
4	There should be a simple, safe, precise and validated screening test.	Question 1: What is the diagnostic accuracy of the automated retinal image analysis systems (ARIASs) at detecting diabetic eye disease in patients with diabetes mellitus? 28
5	The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.	Question 1
<b>THE SCREENING PROGRAMME</b>		
11	There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an “informed choice” (e.g. Down’s syndrome, cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.	Question 2: What is the clinical impact of diabetic eye screening programmes with the use of automated retinal image analysis systems (ARIASs) for level 1 grading compared with diabetic eye screening programmes with fully manual grading? 2
12	There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.	Question 2 Question 4: What are the social and ethical implications of implementing artificial intelligence-based tools in screening programmes and would it be acceptable to health professionals and the public? 19 primary studies, 38 review/opinion papers
14	The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e. value for money). Assessment	Question 3: What is the cost-effectiveness of an ARIAS in diabetic eye screening programmes compared with diabetic eye screening programmes with manual grading? 5 (UK only)

Criterion	Key questions	Studies Included
against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.		

## Methods

The current review was conducted by the Exeter Test Group<sup>4</sup> in keeping with the UK NSC evidence review process. Database searches were conducted on 26th June 2020 (the search covering questions 1-3) and 1st and 2nd July 2020 (the search covering question 4) to identify studies relevant to the questions detailed in Table 2. All searches were limited to the period from the beginning of 2000 to the date of the search.

### Eligibility for inclusion in the review

The following review process was followed:

1. After removing duplicates, the records identified in the two searches were imported in the EndNote X8.2 (Thomson Reuters) and combined. Each abstract was reviewed against the combined inclusion/exclusion criteria by a single reviewer (ZZ). Where the applicability of the inclusion criteria was unclear, the article was included at this stage in order to ensure that all potentially relevant studies were captured. A second independent reviewer (JP) provided input in cases of uncertainty, and validated 20% of the first reviewer's screening decisions. Any disagreements were resolved by discussion until a consensus was reached.
2. Full-text articles required for the full-text review were acquired.
3. Each full-text article was reviewed against the combined inclusion/exclusion criteria by one reviewer (ZZ), who determined whether the article was relevant to one or more of the review questions. A second independent reviewer (JP) provided input in cases of uncertainty, and validated 20% of the first reviewer's decisions. Any disagreements were resolved by discussion until a consensus was reached.

Eligibility criteria for each question are presented in **Error! Reference source not found.** below.

---

<sup>4</sup> <http://medicine.exeter.ac.uk/testgroup/>

**Table 2 Inclusion and exclusion criteria for the key questions**

Key question	Inclusion criteria							Exclusion criteria
	Population	Target condition	Intervention	Reference Standard	Comparator	Outcome	Study type	
1. What is the accuracy of ARIAS at detecting diabetic eye disease in patients with DM?	People with type 1 and type 2 DM ≥12 years of age	Referable DR and/or maculopathy (M1, R2, R3); Any DR (R1M0 or worse)	ARIAS	Any	Manual grading, no comparator, head-to-head comparisons of ARIASs	Accuracy measures, overall and at each grade of DR and maculopathy, where possible	RCTs, prospective or retrospective cohort studies, SRs and meta-analyses	Non-English language, published before 2000, CA, internal evaluations of ARIAS
2. What is the clinical impact of an ARIAS on a DESP?	People with type 1 and type 2 DM ≥12 years of age	N/a	DESP for detection of DR using ARIAS on fundus images prior to manual grading	N/a	DESP for detection of DR that uses human manual grading on fundus images at all levels	Any clinical utility outcomes	RCTs, cohort studies and SR and meta-analyses of those	Non-English language, published before 2000, CA
3. What is the cost-effectiveness of an ARIAS in a DESP?	People with type 1 and type 2 DM ≥12 years of age	N/a	DESP for detection of DR using ARIAS on fundus images prior to manual grading	N/a	DESP for detection of DR that uses manual grading on fundus	Any cost-effectiveness or modelled clinical outcomes	Economic evaluations and reviews of those	Non-English language, published before 2003, CA, non-UK-

			followed by manual grading		images at all levels			based evaluations
4. What are the social and ethical implications of implementing AI-based tools in screening programmes?	Health professionals, providers and users of screening programmes, and the general public	Any	Implementation of AI-based tools in any screening programme	N/a	N/a	Social and ethical implications and acceptability	Qualitative studies and opinion and discussion documents	Non-English, CA, published before 2000

AI - Artificial Intelligence, ARIAS - Automated Retinal Image Analysis Systems, CA – Conference Abstract, DESP – Diabetes Eye Screening Programme, DM – Diabetes Mellitus, DR - Diabetic Retinopathy/Maculopathy, RCT – Randomised Control Trial, RS - Reference Standard, SR – Systematic Review

## Appraisal for quality/risk of bias tool

The following tools were used to assess the quality and risk of bias of each study included in the review:

- systematic reviews and meta-analyses: A Measurement Tool to Assess systematic Reviews (AMSTAR 2)\*\*
- diagnostic accuracy studies: Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Table 28), with additional questions for comparative studies from the pilot version of QUADAS-2C (Table 29); we also categorised studies as prospective or retrospective according to the definition provided in the STARD checklist: “5. Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)”††
- interventional non-RCTs: Downs and Black checklist‡‡
- RCTs: The Cochrane Collaboration’s “Risk of Bias” Tool was listed in the protocol but not used in the review as no relevant RCTs were identified.

## Databases/sources searched

Two separate search strategies were developed by an information specialist (MR) with experience in healthcare research. The first one covered questions 1-3 (accuracy, effectiveness and cost-effectiveness) combining terms for eye pathology with terms for screening and terms for ARIAS. This search was carried out on MEDLINE (via OvidSp), EMBASE (via OvidSp), the Cochrane Library (CDSR and CENTRAL) and the ClinicalTrials.gov provided by the U.S. National Library of Medicine. We were unable to search the WHO ICTRP trial registry (as stated in our protocol) as the database was not available outside WHO due to the Covid-19 pandemic. The search strategy combined free-text and medical subject headings. The search was limited to the period from the beginning of 2000 to the date of the search and restricted to studies published in English.

The second search aimed to identify studies relevant to question 4 (social and ethical aspects of AI implementation in screening programmes). It combined terms for screening programmes, AI-based tools, and terms relevant to qualitative research, such as experience, perceptions, acceptability or interviews. This search was carried out on MEDLINE (via OvidSp), EMBASE (via OvidSp), CINAHL (via EBSCOhost) and PsycINFO (via OvidSp) and used a combination of free-text and medical subject headings. The search

---

\*\* <https://amstar.ca/>

†† <https://www.equator-network.org/reporting-guidelines/stard/>

‡‡ <https://bjsm.bmj.com/content/bjsports/52/6/387/DC3/embed/inline-supplementary-material-3.pdf?download=true>

was not limited by publication date or language, but only studies published in English since 2000 were considered for inclusion.

We also searched the reference lists of all included studies, systematic reviews and other relevant publications, and emailed authors to check for additional titles (e.g. when a potentially relevant study was recently published as a conference abstract). The websites of known ARIAS were also searched and, if necessary, the manufacturer/development team were emailed to request further information.

## Results

The search addressing questions 1 to 3 identified 1921 titles of which 544 were excluded as duplicates. The search addressing question 4 yielded 744 titles of which 211 were excluded as duplicates. As the search addressing question 1 to 3 could identify publications relevant to question 4 and vice versa, we combined all unique titles and screened them together. The total number of studies screened at title and abstract level was 1910. Of those 397 were selected for full text screening or further assessment, plus another 9 identified through hand searching. Ultimately, 58 studies were judged to be relevant to question 1 and 28 of them (including two conference abstracts and one FDA approval letter) were prioritised for inclusion in the evidence synthesis; 2 studies were judged to be relevant to question 2; 5 studies (9 publications) to question 3 and 57 studies to question 4; of the latter 19 reported on primary research and 38 were review or opinion papers.

Details of the selection process are provided in Figure 6; all publications included in the review and the questions they are relevant to are listed in Table 21 (0. Table 23 details the 24 primary studies that met the inclusion criteria for question 1 but were not included in the evidence synthesis (excluded after prioritisation). They were deprioritised because the ARIAS was not commercially available and/or CE-marked/FDA-approved (or we failed to ascertain this); studies were conducted in populations that were very different from the target population in the UK (e.g. higher prevalence of diabetic retinopathy; higher proportion of patients with severe forms of the disease; or, general populations including a large proportion of non-diabetic participants).

We also identified 6 systematic reviews with aims similar to ours. Since their inclusion criteria differed from ours, we decided to review the primary studies addressing question 1 rather than to rely on the results from previous systematic reviews. Nevertheless, we summarised the main characteristics and results of these reviews, assessed their methodological quality using the AMSTAR II checklist and report the results in Appendix 5, Table 30 and Table 31; they are included in the total number of 58 studies judged to be relevant to question 1 (Table 21).

# Question level synthesis

## Criterion 4 and 5. Accuracy of ARIASs

Criterion 4. There should be a simple, safe, precise and validated screening test.

Criterion 5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

*Question 1 (rapid review) – What is the diagnostic accuracy of the automated retinal image analysis systems (ARIASs) at detecting diabetic eye disease in patients with diabetes mellitus?*

## Eligibility for inclusion in the review

Studies were included in the review if they met the following inclusion criteria:

- Population: People with type 1 or type 2 DM aged 12 years and over.
- Index tests: ARIAS, alone or in combination with manual grading.
- Comparator: Manual grading, no comparator, head-to-head comparison of ARIASs.
- Outcome measures: Accuracy measures (e.g. sensitivity and specificity), overall and, if reported in the paper, at each grade of diabetic retinopathy and maculopathy.
- Target condition: Referable diabetic retinopathy and/or maculopathy (M1, R2, R3) or disease present (R1M0 or higher).
- Reference standard: Any, as defined by the authors.
- Study design: RCTs, prospective or retrospective cohort studies, case-control studies, SRs and meta-analyses (studies were prioritised by design, see below).

Studies were excluded if:

- They were published in a language other than English, before 2000 or as conference abstracts (we included 2 conference abstracts, Philip 2017 (17) and Lim 2019 (29), and one FDA approval letter (30) which were considered to be of particular importance).
- The algorithm was evaluated in a proportion of the dataset used for development (internal validation). This was a post hoc decision not included in the initial inclusion/exclusion criteria. Although techniques, such as cross-validation, are usually employed to minimise the risk of overfitting, internal validation is more likely to overestimate the performance of the algorithm



relative to its performance in clinical practice. We included studies in which the development and validation datasets came from the same screening programme, but from different patient cohorts.

We prioritised studies for inclusion in the narrative evidence synthesis if they evaluated:

- Commercially available ARIASs.
- ARIASs that are CE-marked and/or FDA-approved.
- The latest version of the software.
- ARIAS prospectively.
- ARIAS retrospectively, but in a large dataset derived from consecutive patients attending routine screening within a national screening programme.
- ARIAS in the UK.

The rest of the section is structured as follows: 1) we describe the volume and type of evidence relevant to this question; 2) we summarise the findings from studies evaluating DL-based ARIASs; 3) we summarise the findings from studies evaluating ARIASs based on traditional ML algorithms; 4) we discuss studies comparing ARIAS to human graders who were not involved in the reference grading; and, finally, we discuss the evidence reviewed in this section and state our conclusions regarding criteria 4 and 5.

The only role of ARIASs investigated in the current review is to identify patients at low risk defined as ‘no disease’ or ‘non-referable disease’ and high risk defined as ‘any disease’ or ‘referable disease’, prior to manual grading and as part of a multi-level screening programme, such as the EDESP.

## Description of the evidence

Fifty eight studies were judged to be relevant to question 1. Of those 28 studies were prioritised and included in the narrative synthesis. They evaluated 10 ARIASs: 7 DL-based systems and 4 systems based on traditional ML algorithms (including two versions of EyeArt: an earlier ML-based version and the current DL-based version). One of the studies, Lim 2019 (29) was published only as a conference abstract providing limited information on methods and results. However, a full description of the study was included in an FDA approval letter published after our search date. In addition to the design of the study and the test accuracy results, the letter contained detailed description of the system, information from a precision study (repeatability and reproducibility) and human factors validation testing (30).

## DL-based ARIASs

Sixteen studies in total evaluated 7 DL-based ARIAS: EyeArt v2.1 (n=5 studies), EyeGrader (n=1), IDx-DR v2 (n=4), Google AI (n=2), RedCAD (n=1), SELENA (n= 2) and VUNO (n=1). Son 2020 (24) (VUNO) report accuracy estimates only at the level of individual lesions. However, we decided to include this study as the system is commercially available, CE-marked and approved by the Korea Ministry of Food and Drug Safety; and uses a slightly different approach in which a wider range of conditions are targeted. The characteristics of the included studies, the results from their methodological quality assessment and the reported accuracy estimates are summarised in **Table 3**,

Table 4 and Table 5, respectively.

Only 2 studies, both evaluating EyeArt v2.1, were conducted in the UK: Heydon 2020 (31) evaluated the accuracy of the system in a large prospective multi-centre study; Olivera-Barrios 2020 (32) conducted a retrospective analysis comparing the performance of EyeArt v2.1 when used with the EDESP photographic protocol and an alternative wide-field platform. The rest of the studies were conducted in the USA (n=5), the Netherlands (n=3), Singapore (n=2, but in the same cohort), and 1 each in Australia, Korea and Spain. Ting 2017 (Singapore) also reported results from 10 additional cohorts, mainly from counties with no established DESPs, in which various reference standards were used. The latter are not included in this section, but a brief summary of the results is provided in **Error! Reference source not found.**

Only EyeArt v2.1 and IDx-DR v2 were evaluated in prospective clinical studies: Heydon 2020 (31) (UK), Lim 2019 (29) (USA) and Liu 2020 (33) (USA) evaluated EyeArt v2.1, and Abramoff 2018 (20) (USA) and van der Heijden 2018 (21) (the Netherlands) evaluated IDx-DR v2. We defined study design as ‘prospective’ or ‘retrospective’ according to the definition provided in the STARD checklist: “5. *Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)*”.<sup>§§</sup> None of the prospective clinical studies compared the performance of alternative ARIAS in the same cohort of patients. Three studies compared ARIAS to manual graders not involved in the reference grading: Krause 2018 (34) (Google AI), Ting

---

<sup>§§</sup> <https://www.equator-network.org/wp-content/uploads/2015/03/STARD-2015-checklist.pdf>

2017 (23) (SELENA) and Gonzalez-Gonzalo 2020 (35) (RetCAD) (the results are presented in the section *Studies comparing an ARIAS to human graders not involved in the reference grading* below).

The number of participants ranged from 96 (36) to >30 000 (31). There was considerable variation in study characteristics in terms of:

- Participants: selection (e.g. publicly available datasets providing limited background information, such as MESSIDOR; proprietary datasets collected prospectively or retrospectively, from a single or multiple cohorts, with or without prior selection and enrichment of the sample); demographics (e.g. type and duration of diabetes, racial composition), prevalence and spectrum of retinopathy.
- Screening pathway and protocol: setting (the majority were conducted within a national screening programme and/or primary care setting); photographic protocols (e.g. 1-, 2- or 3-field; use of mydriasis; camera; reimaging in case of technical failure); pre-selection of images (e.g. including only some of the available images per eye), handling of the images determined to be ungradable by the ARIAS (included or excluded; reported separately or combined with the referrals), grading scheme (e.g. EDESP, SDESP, ICDR, ETDRS).
- Reference standard: technology (type of fundus photography and photographic protocol, use of OCT to diagnose DMO); background, training and experience of graders; single vs. multiple independent graders; routine manual grading vs. dedicated study graders; method of arriving at the final grade, such as majority voting vs. consensus vs. multi-level grading; handling of ungradable images (as determined by the reference standard).

All these sources of variation are likely to affect the performance of ARIASs, especially when the system is evaluated away from the setting in which it was developed and initially evaluated. Some of these factors have been investigated in the included studies or other publications and their impact on the performance of the systems have been reported. To avoid overwhelming the reader with too much detail, we summarised these sources of variation (and, in some cases, bias) and report them separately in Appendix 5 (*Factors reported to affect the performance of ARIAS*), with references to the respective publications.

The results from the QUADAS-2 assessment are presented in

Table 4 below. Most of the studies were considered to be at high or unclear risk of bias in the Patient Selection domain. The main reason was failure to include unselected consecutive patients or to report the selection process in sufficient detail. Since we could not determine whether the mix of patients in non-UK studies was similar to that in the UK DESP in *all significant ways*, we graded such studies ‘unclear’ for applicability concerns except for studies with clear indication of limited applicability. Most of the studies were judged to be at low risk of bias in the Index Test domain (as the performance of ARIAS was independent from the reference standard or manual graders), but applicability was graded ‘unclear’ or ‘high’ in 7 of the studies as the screening protocol differed significantly from the one used in the UK EDESP or no sufficient information was provided in the paper to make such a judgement.

As mentioned earlier, there was considerable variation in the reference standards across studies (see Appendix 5). We judged studies to be at low risk of bias in the Reference Standard domain if the reference grading involved a panel of experts (ophthalmologists or retinal specialists) blinded to the ARIAS results, who independently graded all images using a pre-specified protocol to resolve disagreements. We also considered studies to be at low risk of bias when the final grade from routine manual grading within a national screening programme (such as EDESP) was used as a reference standard. Initially, the latter criterion also required external arbitration of disagreements between ARIAS and the human graders. This condition was included because a study conducted in the UK, Tufail 2016 (1), reported some disagreements between the final grades from the EDESP and those from the Doheny Image Reading Centre (USA) where the external arbitration was carried out. However, the disagreements had little impact on the reported accuracy estimates and for most of the results the EDESP graders and the external adjudicators agreed. After discussion with experts in the field, it was agreed that if a national screening programme uses multi-level grading and established training and quality assurance procedures, as in the EDESP, the risk of bias is low and an external arbitration is not necessary. According to these final criteria, 5 studies were judged to be at ‘high’ and another 2 at ‘unclear’ risk of bias in the Reference Standard domain.

However, the results pertaining to the risk of bias in the Reference Standard domain should be interpreted with caution. The above criteria were selected for pragmatic reasons and do not take account of the photographic protocol used as part of the reference standard. In all but 2 studies the reference standard involved the same 1- to 3-field non-stereoscopic fundus photography used in screening. Only Abramoff 2018 (20) (IDx-DR v2) and Lim 2019 (29) (EyeArt v2.1) used a superior reference standard: a wide-field stereoscopic retinal imaging protocol (4W-D), that included 4 stereoscopic pairs of digital images per eye. Each

pair covers 45°–60° which is equivalent to the area of the retina covered by the modified 7-field stereo protocol (20). The images in these studies were graded by experienced and certified readers at a Fundus Photograph Reading Centre. In Abramoff 2018, 3 readers graded the images independently and used majority voting to arrive at the final grade. They also tested the performance of IDx-DR v2 against two reference standards: '4W-D fundus imaging only' and '4W-D fundus imaging + OCT' (for diagnosis of DMO). In Lim 2019, 2 readers graded the images independently and a third reader adjudicated any disagreements. The variation in the reference standard across studies is likely to affect the reported accuracy estimates and should be taken into consideration when comparing results from different studies.

With regards to the Flow and Timing domain, the main issue was that in some studies images determined as ungradable, either by the ARIAS or by the reference standard, were excluded from the analysis. This could lead to biased accuracy estimates and overestimate the accuracy of the system.

Across all studies, sensitivity for referable diabetic retinopathy was consistently >90%. There were 3 exceptions, all relating to the IDx-DR v2 system. Abramoff 2018 (USA) reported 87.2% (95% CI 81.8% to 91.2%) sensitivity against the 4W-D reference standard and 85.9% (95% CI 82.5% to 88.7%) against the composite reference standard of '4W-D & OCT'. In both cases the reported sensitivities exceeded the pre-specified goal of >85% sensitivity. Verbraak 2019 (the Netherlands) reported that there were 13 false negative results with a single isolated haemorrhage or cotton wool spot and no microaneurysms, leading to sensitivity of 79.4% (95% CI 66.5% to 87.9%). van der Heijden 2018 (the Netherlands) reported considerable difference in sensitivities when ICDR (3) and EURODIAB criteria (4) were applied, 68% (95% CI 56% to 79%) and 91% (95% CI 69% to 98%), respectively. The authors showed that the main source of disagreement was the interpretation of any single haemorrhage as 'more than microaneurysms alone' which, according to the ICDR criteria (3), results in a diagnosis of moderate diabetic retinopathy (21).

The specificity ranged considerably, from 54.0% (95% CI 53.4% to 54.5%) for R0M0 & R1M0 in Heydon 2020 (UK; EyeArt v2.1) to >95% in some studies, which reflects the inclusion/exclusion of ungradable images and other sources of variation. The relatively low specificity in Heydon 2020 is most likely due to the fact that all images that would normally be sent for manual grading were graded by the ARIAS and included in the analysis. The two studies that used 4W-D protocol in the reference standard reported much higher specificities. Lim 2019 (USA; EyeArt v2.1) reported specificity of 86.5% (95% CI 84.3% to 88.7%) when 'dilation-if-ungradable' protocol was used and 86.0% (95% CI 83.7% to 88.4%) for 'no dilation'. Abramoff 2018 (USA; IDx-DR v2) reported specificity of 90.7%

(95% CI 88.3% to 92.7%) against the reference standard of 4W-D and 90.7% (95% CI 86.8% to 93.5%) against '4W-D & OCT'; including the ungradable images did not change the results (specificity 89.8%). Finally, van der Heijden 2018 (IDx-DR) reported that specificity was comparable (86% and 84%, respectively) when using the ICDR (3) and EURODIAB criteria (4).

The picture was similar when considering the results for vision-threatening diabetic retinopathy. Most studies reported sensitivity of around or more than 95% including Abramoff 2018 (IDx-DR) against both reference standards ('4W-D' and '4W-D & OCT'). Lim 2019 (EyeArt v2.1) did not report results for vtDR, but data from the same study provided in the FDA approval letter (30) showed considerable variability, with sensitivity ranging from 78.6% to 100% (reported separately for different cohorts and settings, with wide CIs due to the small samples sizes). Another exception was van der Heijden 2018 (IDx-DR v2, the Netherlands) who reported sensitivity of 64% (36% to 86%) using the EURODIAB criteria (4) and 62% (32% to 85%) using the ICDR criteria (3). The reason for these very different results was unclear from the paper.

Given the considerable number of factors that may affect the performance of ARIASs (discussed earlier), it would be safer to assume that the accuracy of the systems is likely to vary from setting to setting, as illustrated by the studies evaluating IDx-DR v2. Of all included studies evaluating DL-based ARIASs only Heydon 2020 was considered to be of sufficient quality and to allow direct generalisation of the results to the EDESP. Hence, below we provide further details of this study.

### **Heydon 2020 (EyeArt v2.1, EDESP)**

Heydon 2020 included a cohort of 30 405 consecutive patient episodes from 3 current EDESP centres (North East London [NEL], South East London [SEL] and Gloucestershire) which had slightly different populations in terms of age, sex, ethnicity, uptake of screening and prevalence of diabetic retinopathy. The software was run in parallel to usual care. The reference standard was the final outcome from the manual grading at each centre. The study was considered to have no applicability concerns and to be at low risk of bias in all domains of the QUADAS-2 tool.

Particular strengths of Heydon 2020 were that 3 different EDESP centres were involved, allowing for investigation of between-centre performance. Also, all images captured for each eye in the episode were included in the dataset without any editing or selection prior to being processed with the EyeArt v2.1. None of the poor quality images or those classified as ungradable by human graders were excluded.

For referable disease (M1, R2, R3) sensitivity was 95.7% (95% CI 94.8% to 96.5%) and specificity (for the combination of R0M0 and R1M0) was 54.0% (95% CI 53.4% to 54.5%). The system had sensitivity of 100% (95% CI 98.7% to 100%) for R2 and 100% (95% CI 97.9% to 100%) for R3, and was able to identify 89.4% (95% CI 87.0% to 91.5%) of the images classified as ‘ungradable’ by human graders. The detection rate for non-referable retinopathy (R1M0) was 89.1% (95% CI 88.4% to 89.9%) and 68% (95% CI 67% to 69%) of the images graded as R0M0 by human graders were also graded as ‘no retinopathy’ by the software. The accuracy was similar across the 3 centres. Although no subgroup analysis was reported in the paper, the authors provided additional information included here verbatim:

*“In Heydon 2020 as all R2 and R3 were test positive it was not possible to investigate any differences with sex/age/ethnic differences in ARIAS performance. Only 17 cases of R1M1 were test negative – hence again in this sample of 30,000 with high sensitivity there is no scope to explore sex/age/ethnic differences in ARIAS performance. For R1M0 and R0M0 there was no difference in ARIAS performance by sex but individuals of older age were more likely to test positive and hence would require manual grading of images. In NEL DESP patients of Black African Caribbean (BAC) ethnicity were more likely than South Asians and Whites to have a test positive ARIAS results for R1 and R0 (49.9% vs 45.7% vs 46.3% respectively) but this equates to a maximum difference 4 percentage points meaning that a slightly higher proportion of BAC would require manual grading after ARIAS. Whereas in SEL patients of South Asian ethnicity were more likely than White or BAC to test positive with ARIAS in presence of R1 or R0 (47.9% vs 51.6% vs 47.1% respectively) and hence require manual grading after ARIAS.”* (Heydon et al, personal communication).

The authors reported that approximately 50% of all screening episodes would require further human grading (which ranged from 47% to 51% across the 3 centres). They claim that this would not result in an increased workload for the secondary grader and the workload of the tertiary grader (arbitration) is likely to reduce (31): *“The observation that currently 40-50% of level 1 (primary) grading passes to the level 2 (secondary) grader agrees with observation in other NHS DESP centres (ARR Personal communication with programme leads at North East London and Gloucestershire DESPs)”* (personal communication).

**Table 3 Studies evaluating DL-based ARIASs**

Study	Country	Setting	Prospective study*	Compared to humans	N of patients	Photographic protocol	Reference standard
<b>EyeArt v2</b>							
Heydon 2020	UK	EDESP	Yes	No	30 000** (30 405 episodes)	2 fields (EDESP)	The final grade from the EDESP manual grading
Olvera-Barrios 2020	UK	EDESP	No	No	1257	2 fields (EDESP); 100% mydriasis	The final grade from the EDESP manual grading
Lim 2019 and FDA 2020	USA	Primary care and general ophthalmology	Yes	No	893 (1718 eyes***)	2 fields; results reported for no dilation and dilation-if-ungradable	FPRC using 4W-D protocol and independent grading with adjudication
Liu 2020	USA	Primary Care	Yes	No	180	Unclear; no mydriasis	A single fellowship-trained retina specialist (1 out of 5 specialists involved in the study; relation to setting unclear)
Bhaskaranand 2019	USA	EyePACS	No	No	n/a (101 710 episodes)	3 fields; 45.8% mydriasis	EyePACS certified graders (192 assessed externally)
<b>EyeGrader</b>							
Keel 2018	Australia	2 urban endocrinology dept.	Yes	Yes (but not reported in detail)	96	1-field; no mydriasis	A single ophthalmologist
<b>IDx-DR</b>							
Abramoff 2018	USA	Primary care	Yes	No	892****	2 fields; 23.6% mydriasis	FPRC using 4W-D protocol and independent grading with adjudication; OCT for DMO
van der Heijden 2018	The Netherlands	Primary care	Yes	No	898	2 fields; mydriasis if needed	3 retinal specialists
Verbraak 2019	The Netherlands	Primary care	No	No	1425	2 fields; mydriasis if needed	2 readers, adjudicated by a retinal specialist
Shah 2020a	Spain	Primary care	No	No	2680	2 fields; mydriasis if needed	3 ophthalmologists, adjudication by a retinal specialist
<b>Google AI</b>							
Krause 2018	USA	EyePACS	No	Yes	998	1 field	3 retinal specialists
Raumviboonsuk 2019	Thailand	Diabetes registry	No	Yes	7517	1 field	Regional graders, subsets adjudicated by retinal specialists
<b>RedCAD</b>							
Gonzalez-Gonzalo 2020	The Netherlands	MESSID OR 1&2	No	No	1200 + 874	1 field	Dataset ground truth
<b>SELENA</b>							



UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study	Country	Setting	Prospective study*	Compared to humans	N of patients	Photographic protocol	Reference standard
Ting 2017	Singapore + 10 other	SiDRP + 10 other	No	Yes	8589****	2 fields	SiDRP: Single retinal specialist; other cohorts – routine grading
Yip 2020	Singapore	SiDRP	No	No	n/a (455 491 images)	2 fields	As in Ting 2017
<b>VUNO</b>							
Son 2020	Korea	IDRiD, e-optha	No	No	IDRiD: 143, e-optha: 434	unclear	unclear

\*Prospective and retrospective study designs are defined according the definition provided in the STARD checklist (see p. 36)

\*\*The number of patients (n = 30 000) in Heydon 2020 is given only in the title, but the exact number is not reported in the paper.

\*\*\*After excluding the images classified as ‘ungradable’ by the reference standard; the cohort combined a sequential enrolment cohort and an enrichment permitted cohort; the analysis reported in the FDA approval letter reporting on the same study included only 655 participants; the exclusions were made to “better align the analysis population with the proposed intended use population” (p. 4); excluded from the analysis were patients younger than 22 years, patients enrolled at retinal sites and patients with recorded history of DR; there are some discrepancies in the numbers reported in Lim 2019 (29) and in the FDA letter (30).

\*\*\*\* To recruit sufficient numbers of mtmDR participants Abramoff 2018 used an enrichment strategy actively seeking higher risk participants with elevated HbA1c (>9.0%) levels or elevated fasting plasma glucose; the enrichment was independently activated by the statistician while always remaining masked to the ARIAS outputs and the disease levels.

\*\*\*\*\*The main analysis in Ting 2020 was based on 14 880 patients, of whom 8589 were unique patients (not involved in the development of SELENA); only the results from this sensitivity analysis (of unique patients) are included in results reported here.

Publicly available datasets: e-optha; MESSIDOR – Méthodes d’Évaluation de Systèmes de Segmentation et d’Indexation Dédiées à l’Ophthalmologie Rétinienne; IDRiD - Indian Diabetic Retinopathy Image Dataset

Abbreviations: EDESP – English Diabetic Eye Screening Programme, DM – Diabetes Mellitus, DMO - Diabetic Macular Oedema, FPRC - Fundus Photograph Reading Centre, OCT – Optical Coherence Tomography, SiDRP - Singapore National Diabetic Retinopathy Screening Program

**Table 4 Results from the methodological quality assessment using QUADAS-2 checklist of studies evaluating the accuracy of DL-based ARIASs**

ARIAS	Study	PS: RB	PS: A	IT: RB	IT: A	RS: RB	RS: A	F&T: RB
EyeArt v2	Bhaskaranand 2019	High						
	Heydon 2020	Low						
	Liu 2020	Unclear						
	Lim 2019 & DFA 2020							
	Olvera-Barrios 2020							
EyeGrader	Keel 2018							
Google AI	Krause 2018							
	Raumviboonsuk 2019							
IDx-DR	Abramoff 2018							
	van der Heijden 2018							
	Verbraak 2019							
	Shah 2020a							
RedCAD	Gonzalez-Gonzalo 2020							
SELENA	Ting 2017*							
	Yip 2020							
VUNO	Son 2020							

\* SiDRP dataset only; PS – Patient selection, IT – Index test, RS – Reference standard, F&T – Flow and timing, RB – Risk of bias; A – Applicability concerns;

Comparative accuracy is incorporated into each domain according to the questions listed in Table 29. If either outcome (comparative accuracy or single test accuracy assessment) is 'high RB' then the domain is rated 'high RB'.

**Table 5 Diagnostic accuracy of the included DL-based ARIASs**

Study	Country	N of participants	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
<b>EyeArt v2</b>					
Heydon 2020	UK	30000 <sup>1</sup> (30405 episodes)	M1, R2, R3, ungradable according to human graders (EDESP)	SE 95.7% (94.8% - 96.5%), SP 54.0% (53.4% - 54.5%) for R0M0 & R1M0	68% (67% - 69%) for R0M0 89.1% (88.4% - 89.9%) for R1M0 98.3% (97.3% - 98.9%) for R1M1 100% (98.7% - 100%) for R2 100% (97.9% - 100%) for R3 89.4% (87.0% - 91.5%) for ungradable
Olvera-Barrios 2020 <sup>2</sup>	UK	1257	R1M1, R2, R3, ungradable (EDESP)	SE 96.4% SP 55.9%	Any DR SE 92.26% (88.37% to 94.69%) vtDR SE 100% (99.53% to 100%) PDR SE 100% (61% to 100%)
Lim 2019 and FDA 2020	USA	893 (1718 eyes) <sup>3</sup>	Moderate NPDR or higher (ICDR)	<u>Dilation-if-ungradable:</u> SE 95.5% (92.6% - 98.4%) SP 86.5% (84.3% - 88.7%) Imageability <sup>4</sup> 97.5% (96.4% - 98.5%)  <u>No dilation:</u> SE 95.5% (92.4% - 98.5%) SP 86.0% (83.7% - 88.4%) Imageability 87.5% (85.4% - 89.7%)	<i>Images ungradable by ARIAS are included in the SE &amp; SP estimates!</i> <u>vtDR in the sequentially enrolled cohort</u> <u>Primary care</u> (n=45 subjects): SE 100.0% (51.0% - 100%) SP 93.9% (86.5% - 98.8%) Imageability 96.5% (90.6% - 100.0%) <u>Ophthalmology</u> (n=190 subjects): SE 88.9% (60.0% - 100.0%) SP 92.5% (88.7% - 95.6%) Imageability 98.6% (97.0% - 99.7%)  <u>vtDR in the enrichment permitted cohort</u> <u>Primary care</u> (n=335 subjects): SE 78.6% (58.8% - 95.8%) SP 89.6% (86.6% - 92.4%) Imageability 96.7% (94.8% - 98.5%) <u>Ophthalmology</u> (n=85 subjects): SE 100.0% (51.0% - 100%) SP 87.0% (80.1% - 93.1%) Imageability 97.0% (92.9% - 100.0%)
Liu 2020	USA	180	Moderate or worse DR or inconclusive screening results (ICDR)	SE 100% (92.3% - 100%) SP 65.7% (57.0% - 73.7%) including inconclusive results	n/a (but reported that 29.4% of the results were inconclusive)
Bhaskaran and 2019	USA	n/a	Moderate or severe NPDR,	SE 91.3% (90.9% - 91.7%) SP 91.1% (90.9% - 91.3%)	<u>Treatable DR: Mydriatic vs. non-mydriatic</u> SE 98.8% vs 98.0%

Study	Country	N of participants	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
		(101 710 episodes)	PDR, and/or clinically significant DMO (ERGS, based on the ETDRS)	<u>Mydriatic vs. non-mydriatic:</u> SE 93.0% vs 89.6% (no p-value reported) SP 90.4% vs 91.7%	<u>Other measures:</u> 95.4% of the FNs were moderate NPDR and did not meet the general treatment criteria Severe or proliferative DR (potentially treatable): SE 98.5%, the fraction of FNs in the entire cohort was 0.08%
<b>EyeGrader</b>					
Keel 2018	Australia	96	Moderate NPDR or worse and/or DMO (EDESCP)	SE 92.3% (no CIs reported) SP 93.7% (unclear if ungradable images were included)	93/96 (96.9%) had a retinal photograph in at least one eye that was gradable for rDR; 10/93 (10.8%) had gradable images in only one eye according to ARIAS
<b>IDx-DR</b>					
Abramoff 2018	USA	892 <sup>5</sup>	mtmDR defined as: ETDRS level $\geq$ 35, and/or clinically significant DMO (ETDRS)	<u>SE for mtmDR:</u> 87.2% (81.8% – 91.2%) (pre-specified goal >85%), against 4W-D fundus imaging RS 85.9% (82.5% – 88.7%) against '4W-D + OCT' RS <u>SP for mtmDR (excluding ungradable by the software or the reading centre):</u> 90.7% (88.3% – 92.7%) (pre-specified goal >82.5%) against 4W-D fundus imaging RS 90.7% (86.8% – 93.5%) against '4W-D + OCT' RS	<u>SE for vtDR:</u> 97.4% (86.2% – 99.9%) against 4W-D fundus imaging RS 92.2% (81.1% – 97.8%) against '4W-D + OCT' RS <u>Sensitivity analysis</u> (worst case scenario including all intention-to-screen patients and using mtmDR as a threshold): SE 80.7% (76.7% – 84.2%) SP 89.8% (85.9% – 92.7%) <u>Ungradable by the software</u> (after excluding the ungradable by the RS): 33/852
van der Heijden 2018	The Netherlands	898	Moderate or vtDR (ICDR and EURODIAB)	<u>EURODIAB criteria:</u> SE 91% (69% – 98%) SP 84% (81% – 86%) <u>ICDR criteria:</u> SE 68% (56%–79%) SP 86% (84%–88%)	<u>vtDR, EURODIAB criteria:</u> SE 64% (36% – 86%) SP 95% (93% – 96%) <u>vtDR, ICDR criteria:</u> SE 62% (32%–85%) SP 95% (93%–96%) <u>'Ungradable' by ARIAS:</u> 477/1415 (unclear what proportion of those were also rated 'ungradable' by the RS)
Verbraak 2019	The Netherlands	1425	mtmDR and / or DMO (ICDR)	SE 79.4% (66.5–87.9) SP 93.8% (92.1–94.9)	<u>vtDR:</u> SE 100% (77.1–100) SP 97.8% (96.8–98.5)

## UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study	Country	N of participants	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
				All 13 FNs had a single isolated haemorrhage or cotton wool spot and no microaneurysms	
Shah 2020a	Spain	2680	Moderate or vtDR and/or DMO (ICDR mapped onto the ETDRS)	SE 100% (97%-100%) SP 81.82% (80%-83%) (ungradable images excluded)	<u>vtDR:</u> SE 100% (95%-100%) SP 94.64% (94%-95%) (ungradable images excluded) <u>Ungradable</u> by the ARIAS: 404/3531
<b>Google AI</b>					
Krause 2018	USA	998	Moderate or worse DR or referable DMO (ICDR)	SE 97.1%, SP 92.3% (no CIs reported; ungradable images excluded) <u>Referable DMO:</u> SE 94.9%, SP 94.4%	ARIAS classified 4/16 cases of PDR as severe and 2/50 cases of severe DR as moderate
Raumvibonsuk 2019	Thailand	7517	Moderate or worse DR or referable DMO (ICDR)	SE 96.8% (range: 89.3% – 99.3%) SP 95.6% (range: 98.3% – 98.7%) <u>Referable DMO:</u> SE 95.3% (range: 85.9% – 100.0%) SP 98.2% (range: 94.4% – 99.1%) (ungradable images and cases of other retinal diseases excluded)	<u>Severe or worse NPDR and/or DMO:</u> SE 93.6% (range: 85.2%–98.4%) SP 98.2% (range: 94.8%–99.3%) (similar for PDR and/or DMO) 12.6% of all images were classified by ARIAS as ungradable
<b>RedCAD</b>					
Gonzalez-Gonzalo 2020	The Netherlands	Messidor: 1200; Messidor-2: 874	Stage 2 and 3 (Messidor)	<u>Messidor:</u> SE 92.0% (89.1%-95.9%) SP 92.1% (88.7%-95.2%) <u>Messidor-2:</u> SE 92.6% (88.4%-97.4%) SP 93.4% (89.9%-97.2%) (images classified as 'ungradable' by human graders were excluded)	n/a
<b>SELENA</b>					

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study	Country	N of participants	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
Ting 2017	Singapore (+ other countries)	SiDRP: 8589 <sup>6</sup>	Moderate NPDR or worse and/or DMO and/or ungradable images (ICDR)	SE 89.56% (85.51%-92.58%) SP 83.49% (82.68%-84.27%)	vtDR: SE 100% (90.97%-100.0%) SP 81.4% (80.57%-82.22%)
<b>VUNO</b>					
Son 2020	Korea	IDRiD : 143, e-ophtha: 434	n/a	<u>Haemorrhage:</u> E-ophtha: SE 89.2% (83.0%–93.7%) SP 91.4% (87.1%–94.7%) IDRiD: SE 88.9% (77.4%–96.6%) SP 96.6% (90.5%–99.3%) <u>Hard exudate:</u> E-ophtha: SE 93.6% (82.5%–98.7%) SP 97.1% (85.1%–99.9%) IDRiD: SE 92.6% (82.1%-97.9%) SP 100% (95.9%–100%) <u>Cotton wool patch:</u> IDRiD: SE 92.3% (74.9%–99.1%) SP 94.0% (88.1%–97.6%)	n/a

Yip 2020 is not included here as it was an exploratory analysis looking at the impact of different technical factors on the performance of SELENA in the same SiDRP cohort investigated in Ting 2017; detailed results are reported in **Error! Reference source not found.**

<sup>1</sup>The number of patients (n = 30 000) in Heydon 2020 is given only in the title, but the exact number is not reported in the paper.

<sup>2</sup>Olvera-Barrios 2020 did not report specificity. The study compared the performance of EyeArt v2 using the standard EDESP photographic protocol vs. EIDON wide-field platform.

Study	Country	N of participants	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
<p><sup>3</sup>This is the total number of participants after excluding the images classified as ‘ungradable’ by the reference standard. The study included 2 sequential enrolment cohorts and 2 enrichment permitted cohorts, recruited in primary care and at general ophthalmology clinics, respectively. The accuracy estimates for referable disease are based on Lim 2019 (29) (conference abstract). It is unclear if the sensitivity and specificity estimates included ungradable images. The accuracy estimates for vtDR are based on the FDA approval letter (30) reporting only on 655 patients. In the latter, the sensitivity and specificity estimates take into account ungradable images, but the results are reported separately by cohort and setting, hence the small sample sizes and the wide confidence intervals.</p>					
<p><sup>4</sup> ‘Imageability’ is defined as the proportion of results classified as ‘gradable’ by the software out of all results classified as ‘gradable’ by the reference standard.</p>					
<p><sup>5</sup> In order to recruit sufficient numbers of mtmDR participants, Abramoff 2018 used an enrichment strategy actively seeking higher risk participants with elevated HbA1c (&gt;9.0%) levels or elevated fasting plasma glucose. The enrichment was independently activated by the statistician while always remaining masked to the ARIAS outputs and the disease levels.</p>					
<p><sup>6</sup>The main analysis in Ting 2020 was based on 14 880 patients, of whom 8589 were unique patients not involved in the development of SELENA. Here we report the results from the sensitivity analysis which included only the latter cohort.</p>					
<p><u>Publicly available datasets:</u> e-optha; MESSIDOR – Méthodes d’Évaluation de Systèmes de Segmentation et d’Indexation Dédiées à l’Ophthalmologie Rétinienne; IDRiD - Indian Diabetic Retinopathy Image Dataset</p>					
<p><u>Abbreviations:</u> CI – confidence interval, EDESP – English Diabetic Eye Screening Programme, DM – Diabetes Mellitus, DMO - Diabetic Macular Oedema, DR – Diabetic Retinopathy, ETDRS - Early Treatment of Diabetic Retinopathy Study, ICDR - International Clinical Diabetic Retinopathy severity scale, mtmDR – more than mild DR, OCT – Optical Coherence Tomography, RS – Reference Standard, SE – Sensitivity, SiDRP - Singapore National Diabetic Retinopathy Screening Program, SP – Specificity, vtDR – vision-threatening DR, 4W-D - a wide-field stereoscopic fundus imaging protocol which included 4 stereoscopic pairs of digital images per eye, each pair covering 45–60° visual field.</p>					

## Traditional ML-based ARIASs

Twelve studies evaluating 4 ML-based ARIASs were included in the review: iGradingM and its predecessor the Aberdeen system (n = 7 studies), RetmarkerSR (n = 4), RetinaLyze (n = 2) and EyeArt v1 (n=1). Tufail 2016 evaluated iGradingM, RetmarkerSR and EyeArt v1 in the same cohort of patients (1). The characteristics of the included studies, the results from their methodological quality assessment and the reported accuracy estimates are summarised in Table 6, Table 7 and Table 8, respectively. Table 9 details the comparative accuracies of the 3 systems evaluated in Tufail 2016 (1).

Of the 7 studies evaluating iGradingM, one was conducted in Spain, 4 in Scotland and 2 in England. However, the system failed to read disc-centred images in Tufail 2016 (1) (EDESCP) and no valid accuracy results were reported. Only Philip 2007 (37) had a prospective design and compared the performance of the system against human graders who were not involved in the reference standard grading (discussed in the following section). The largest study, Fleming 2010a (38), included 33 535 consecutive patients from the SDESP and used as a reference standard the programme's final grade with external adjudication of discrepant results. Philip 2017 (17) reported on the analysis of audit data from the SDESP but the results were published only as a conference abstract with no details provided.

All 4 studies evaluating the RetmarkerSR were retrospective. Three were conducted in Portugal and one in England. We contacted the manufacturer to ascertain that no relevant publications have been missed. They had updated the system which means that the results from studies evaluating earlier versions may not necessarily reflect the performance of the current version of the RetmarkerSR. Ribeiro 2014 (18) reported results based on audit data from the Portuguese DESP.

Both studies evaluating the accuracy of RetinaLyze were retrospective: Hansen 2004 (39) was conducted in Denmark and Bouhaimed 2008 in Wales (UK) (40). We found 2 more studies evaluating the same system, Larsen 2003 (41) and Larsen 2007 (42), but they used images taken on a 35-mm colour transparency film and later digitalised, and were eventually excluded from the review. We contacted the manufacturer who confirmed that the most recent study was Bouhaimed 2008 (40) which evaluated RetinaLyze v.1.0.6.1 validated for commercial use.

We also included the accuracy estimates for EyeArt v1 reported by Tufail 2016 (1) (UK), despite the fact that the current version (discussed in the previous section) is v2.1 and is DL-based. Since Tufail 2016 is a large, UK-based study of good methodological quality and the only study comparing directly 3 CE-marked and commercially available ARIASs, we



believe that such data might be informative and will provide additional evidence for the performance of the system.

Overall, the studies evaluating iGradingM were of better methodological quality compared to those evaluating RetmarkerSR and RetinaLyze. Both studies evaluating RetinaLyze were judged to be at high risk of bias in at least one domain (39, 40), while the risk of bias was low for all domains in 3 of the studies evaluating iGradingM (1, 5, 38) and one of the studies evaluating RetmarkerSR (1). Also, since 7/12 studies were conducted in the UK, they were considered to have no applicability concerns. Tufail 2016 was judged to be at low risk of bias and applicability concerns including the risk of bias related to comparative accuracy (Table 7).

The sensitivity of iGradingM was consistently >90% for referable diabetic retinopathy and approached 100% for higher grades of retinopathy across all included studies. Philip 2007 (SDESP) reported specificity of 67.4% (95%CI 66.0–68.8), Soto-Pedre 2015 (Spain) 68.77% (95%CI 67.18–70.36) while the rest of the studies reported detection rate for individual grades that will translate into similar results. The audit data from the SDESP reported by Philip 2017 was in line with the above results but with lower specificity: sensitivity 97%, specificity 38% and false negative rate ranging from 0 to 0.6% (17). As mentioned earlier, iGradingM failed to read disc-centred images in Tufail 2016. However, the results reported by Goatman 2011 (5) who compared the EDESP and SDESP photographic protocols were in line with those reported in the Scottish evaluations. Tufail and colleagues explained the different results with pre-processing of the images in Goatman 2011 which was not done in their study (personal communication).

The sensitivity of RetmarkerSR for referable diabetic retinopathy was lower in Tufail 2016 (UK) compared to that reported by the development team in Oliveira 2011 (43) (85% vs 96%) while specificity was comparable (around 50%). The results from the audit reported by Ribeiro 2011 (18) show high negative predictive value with only 11 false negatives cases out of 3 287 screened cases. The sensitivity of RetinaLyze was variable and depended on the specific setting of the algorithm and the use of mydriasis.

Tufail 2016 reported higher sensitivity and lower specificity for EyeArt v1 compared to RetmarkerSR (Table 9). EyeArt v1 achieved 94.7% sensitivity for any retinopathy, 93.8% for referable retinopathy and 99.6% for proliferative retinopathy. However, the system misclassified 80% of the patients with 'no disease' as 'referrals'. The accuracy of EyeArt v1 was not affected by ethnicity, sex or camera type, but sensitivity was marginally lower with increasing patient age. The accuracy of RetmarkerSR appeared to vary with patient age, ethnicity and camera type (1).

**Table 6 Studies evaluating traditional ML-based ARIASs**

Study	Country	Setting	Prospective study	Compared to humans	N of patients	Photographic protocol	Reference standard
<b>iGradingM</b>							
Philip 2007	UK, Scotland	SDESP	Yes	Yes	6722	1 field; mydriasis if needed	A single grader (clinical research fellow)
Fleming 2010a	UK, Scotland	SDESP	No	No	33 535	1 field; mydriasis if needed	SDESP's manual grading + arbitration by 7 senior ophthalmologists
Fleming 2010b	UK, Scotland	SDESP	No	No	7 586	Not reported (probably SDESP)	SDESP's manual grading + a research fellow's grading, arbitrated by a lead clinician
Goatman 2011	UK, England	EDESP	No	No	8271	1 and 2 fields; 100% mydriasis	EDESP's manual grading + 2 levels of arbitration
Philip 2017 (CA)	UK, Scotland	SDESP	No*	No	Not reported	Not reported but probably 1 field (SDESP)	Not reported
Soto-Pedre 2015	Spain	SpDESP	No	No	5278	1 field; mydriasis if needed	SpDESP's manual grading
<b>RetmarkerSR</b>							
Oliveira 2011	Portugal	PDESP	No	No	5386 (289)	2 fields; mydriasis if needed	A single ophthalmologist
Ribeiro 2011	Portugal	PDESP	No*	No	3287	2 fields; mydriasis if needed	PDESP's manual grading
Figueiredo 2015	Portugal	PDESP	No	No	11 511 (4 datasets)	2 fields; no mydriasis	PDESP's manual grading
<b>RetinaLyze</b>							
Bouhaimed 2008	UK, Wales	WDESP	No	No	100	2 fields; 100% mydriasis	WDESP's manual grading
Hansen 2004	Denmark	hospital	No	No	83	5 fields; mydriasis used in 1 arm	2 graders, disagreements adjudicated by a third grader
<b>EyeArt v1, RetmarkerSR, iGradingM</b>							
Tufail 2016	UK, England	EDESP	No	No	20258	2 fields; mydriasis if needed	EDESP's manual grading with arbitration at a US-based fundus photographic reading centre
*Audit studies of the respective national screening programmes.							
Abbreviations: CA - conference abstract, EDESP – English Diabetic Eye Screening Programme, PDESP – Portuguese Diabetic Eye Screening Programme, SDESP – Scottish Diabetic Eye Screening Programme, SpDESP – Spanish Diabetic Eye Screening Programme, WDESP – Welsh Diabetic Eye Screening Programme							

**Table 7 Results from the methodological quality assessment using QUADAS-2 checklist of studies evaluating the accuracy of traditional ML-based ARIASs**

ARIAS	Study	PS: RB	PS: A	IT: RB	IT: A	RS: RB	RS: A	F&T: RB
iGradingM	Fleming 2010a	Green	Green	Green	Green	Green	Green	Green
	Fleming 2010b	Red	Green	Green	Green	Green	Green	Green
	Goatman 2011	Green	Green	Green	Green	Green	Green	Green
	Soto-Pedre 2015	Green	Yellow	Green	Green	Green	Green	Red
	Philip 2007	Green	Green	Green	Green	Red	Green	Green
	Philip 2017*	Unclear	Green	Yellow	Green	Yellow	Green	Yellow
RetinalLyze	Bouhaimed 2008	Low	Green	Green	Green	Green	Red	Green
	Hansen 2004	High	Yellow	Red	Red	Green	Red	Green
RetmarkerSR	Figueiredo 2015	Yellow	Yellow	Yellow	Yellow	Green	Yellow	Green
	Oliveira 2011	Red	Yellow	Green	Green	Red	Green	Green
	Ribeiro 2011	Green	Yellow	Green	Green	Green	Green	Green
RetmarkerSR, EyeArt v1, iGradingM**	Tufail 2016	Green	Green	Green	Green	Green	Green	Green

\*Conference abstract;

\*\*iGradingM failed the evaluation.

Abbreviations: PS – Patient selection, IT – Index test, RS – Reference standard, F&T – Flow and timing, RB – risk of bias; A – Applicability concerns; Comparative accuracy is incorporated into each domain according to the questions listed in Table 28. If either outcome (comparative accuracy or single test accuracy assessment) is 'high RB' then the domain is rated 'high RB'.

**Table 8 Diagnostic accuracy of the included traditional ML-based ARIASs**

Study	Country	N of patients	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
<b>iGradingM</b>					
Philip 2007	UK, Scotland	6722	Any retinopathy including R1, M1, R2, M2, R3, R4 or technical failure (SDESP)	SE 90.5% (89.3% – 91.6%) SP 67.4% (66.0% – 68.8%)	Detection rate by patient: R1 85.9% (84.1% – 87.5%) M1 97.4% (90.9% – 99.3%) R2 100% (67.6% – 100%) M2 97.2% (93.6% – 98.8%) R3 100% (91.0% – 100%) R4 100% (87.9% – 100%) Technical failure 99.8% (99.0% – 100%)

Study	Country	N of patients	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
Fleming 2010a	UK, Scotland	33 535	M1, R2 (rescreen in 6 months), M2, R3 or R4 (refer to ophthalmology), (SDESP)	The error rate for cases requiring referral to ophthalmology, 6 month recall or slit-lamp examination was 0.11% (37 out of all 33535 patients)	<u>Detection rate by patient:</u> R0 49.6% (48.9% - 50.3%) R1 83.9% (83.0% - 84.6%) M1 99.2% (97.8% - 99.7%) R2 100% (97.9% - 100%) M2 97.3% (96.1% - 98.1%) R3 100% (98.8% - 100%) R4 100% (98.1% - 100%) Ungradable 99.8% (99.5% - 99.9%) No statistically significant differences found between ethnic groups.
Fleming 2010b	UK, Scotland	7 586	R2, R3, R4, M1, M2 (SDESP)	Adding the detection of exudates (EX) and haemorrhages (HM) to the detection of microaneurysms (MA) increased SE for observable retinopathy (R2) or worse, from 94.9% (93.5% - 96.0%) to 96.6% (95.4% - 97.4%) (p=0.001) without affecting manual grading workload	<u>MA+EX+HM (similar for MA alone):</u> R0: 36.8% (35.3% - 38.3%) R1: 79.0% (76.9% - 80.9%) M1: 92.2% (85.3% - 96.0%) M2: 95.9% (94.0% - 97.2%) R2: 100% (82.4% - 100%) R3: 98.9% (96.8% - 99.6%) R4: 97.4% (94.4% - 98.8%) Technical failure: 98.8% (97.6% to 99.4%)
Goatman 2011 <sup>1</sup>	UK, England	8271	M1, R2, R3 or ungradable (EDESP)	SE (range) 98.3% (MA/BH/EX, 1field) to 99.3% (MA only, 2 fields)	SE for <u>any retinopathy</u> ranged from 89.9% (MA/BH/EX, macular field) to 95.8% (MA only, both fields) SE for <u>NPDR and PDR</u> was 100% for all strategies; SE for detecting ungradable images ranged from 97.4% to 99.1% across strategies; <u>'MA+BH+EX' x 2-field:</u> R0: 60.2% (58.8% - 61.5%) R1: 94.2% (93.1% - 95.1%) <u>'MA+BH+EX' x 1-field:</u> R0: 50.7% (49.3% - 52.1%) R1: 87.6% (86.2% - 88.9%) <u>Workload reduction</u> ranged from 26.4% to 38.1%
Philip 2017 (CA)	UK, Scotland	n/a	SDESP (assumed)	SE 97%, SP 38%, FNR of 0 to 0.6%	n/a

## UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study	Country	N of patients	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
Soto-Pedre 2015	Spain	5278	Any DR (ICDR)	SE 94.52% (92.56% – 96.49%) SP 68.77% (67.18% – 70.36%) (ungradable images excluded)	Ungradable patients: 26.16% (n=1374)
Tufail 2016	UK, England	20258	M1, R2, R3 or ungradable (EDESP)	n/a (the system was unable to read disc-centred images and failed the evaluation)	n/a
<b>RetmarkerSR</b>					
Tufail 2016	UK, England	20258	M1, R2, R3 or ungradable (EDESP)	SE 85.0% (83.6% - 86.2%) SP 53% (52% - 54%) for R0M0 SP 47.7% (47% - 48.5%) for R0M0 & R1M0	<u>Any retinopathy:</u> SE 73.0% (72.0% - 74.0%) <u>PDR (R3):</u> SE 97.9% (94.9% - 99.1%)
Oliveira 2011	Portugal	5386 <sup>2</sup>	NPDR with maculopathy or PDR (Portuguese DESP)	SE 96.1% (94.39% – 97.89%) SP 51.7% (50.27% – 53.07%) <u>2-step algorithm (n=289)</u> SE 95.8% (92.8% - 98.4%) SP 63.2% (60.8% -65.7%)	<u>Urgent referrals (n=116)</u> 115 classified as ‘having disease’
Ribeiro 2011	Portugal	3287	NPDR with maculopathy or PDR (Portuguese DESP)	Missed only 11 cases (false negatives) out of 3,287 screened cases (0.3% of quality control cases, 0.02% of total patients)	N/a
Figueiredo 2015	Portugal	11 511 <sup>3</sup>	N/A (most likely Portuguese DESP)	Across the 4 datasets: SE ranged from 89.3% to 100% SP ranged from 57.6% to 73%	N/a
<b>EyeArt v1</b>					
Tufail 2016	UK, England	20258	M1, R2, R3 or ungradable (EDESP)	SE 93.8% (92.9% - 94.6%) SP 20% (19% to 21%)	<u>Any retinopathy:</u> SE 94.7% (94.2% - 95.2%) <u>PDR (R3):</u> SE 99.6% (97.0% - 99.9%)
<b>RetinaLyze</b>					
Bouhaimed 2008	UK, Wales	100	Mild NPDR or worse ( $\geq 2a$ according to the Bro Taf Protocol used in the study)	<u>Red lesions:</u> SE 82%, SP 75%, <u>Red &amp; bright lesions:</u> SE 88%, SP 52%, <u>Red &amp; bright lesions (at elevated thresholds in images of good quality):</u> SE 93%, SP 78%	N/a
Hansen 2004	Denmark	83	Moderate NPDR or worse; DMO not graded (ETDRS)	<u>No mydriasis:</u> SE 89.9%, SP 85.7% <u>Mydriasis:</u> SE 97.0%, SP 75.0%	For moderate NPDR or more severe DR: SE 100% for images captured both with and without mydriasis
<sup>1</sup> Compared 4 screening strategies: 1- or 2-field imaging x ‘MA’ or ‘MA/BH/EX’; the 2-field protocol was the standard EDESP protocol <sup>2</sup> 289 patients included in the evaluation of the second step of the algorithm <sup>3</sup> 4 datasets combined					

Study	Country	N of patients	Definition of referable disease (criteria)	Accuracy for referable DR (95% CI)	Accuracy / detection rate for other grades (95% CI)
<p><u>Abbreviations:</u> BH – blot haemorrhage, DESP – Diabetic Eye Screening Programme, DMO – Diabetic Macular Oedema, DR – Diabetic Retinopathy, EDESP – English Diabetic Eye Screening Programme, EX – exudate, FNR – False Negative Rate, HM – haemorrhage, MA – microaneurysm, NPDR – Non-Proliferative Diabetic Retinopathy, PDR – Proliferative Diabetic Retinopathy, rDR – Referable Diabetic Retinopathy, RS – Reference Standard, SDESP – Scottish Diabetic Eye Screening Programme, SE – Sensitivity, SP – Specificity</p>					

**Table 9 Summary of the results from Tufail 2016 which compared directly 3 ARIASs ('ungradable' images included)**

ARIAS	Sensitivity for rDR (95%CI)	Sensitivity for PDR (95%CI)	Sensitivity for any DR (95% CI)	Specificity (95%CI) [equivalent to detection rate of R0&M0&R1M0]	Detection rate for R0M0
EyeArt v1 (ML)	93.8% (92.9%-94.6%)	99.6% (97.0%-99.9%)	94.7% (94.2%-95.2%)	15.8% (15.3%-16.4%)	20%
RetmarkerSR	85.0% (83.6%-86.2%)	97.9% (94.9%-99.1%)	73.0% (72.0 %-74.0%)	34.7% (34.0%-35.4%)	53%
iGradingM	Failed the evaluation as all disc-centred images were classified as 'ungradable'				
CI – confidence interval, DL- deep learning, rDR – referable diabetic retinopathy, PDR – proliferative diabetic retinopathy					

## Studies comparing an ARIAS to human graders not involved in the reference grading

Table 10 summarises the characteristics and results of studies in which an ARIAS is compared to human graders not involved in the reference standard grading. We included 5 such studies: 3 evaluated DL-based ARIASs (Google AI, SELINA and RetCAD) and one evaluated an ARIAS based on traditional ML (iGradingM). The study evaluating iGradingM was the only prospective evaluation and the only study conducted in the UK (Scotland).

**None of the studies were considered to be at high risk of bias with respect to comparative accuracy while the other limitations, which apply equally to ARIAS and human graders, have already been reported in the previous sections (see**

Table 4 and Table 7). On the whole, DL-based ARIASs had higher sensitivities and lower specificities compared to human graders. However, only Ting 2017 (23) reported the statistical significance of these differences (Table 10). Given the differences between DESPs even in a single country such as the UK (see Table 1), the results from the 3 studies evaluating DL-based ARIASs may not be generalizable beyond the settings in which they were conducted.

Therefore, we considered only Philip 2007 (37), the only prospective study evaluating a traditional ML-based ARIAS, to be directly relevant to the current review. The study included 6722 consecutive patients from the SDESP and compared the performance of iGradingM against that of 3 human graders from the programme who also performed the photography. The comparison was restricted to ‘level 1 grading’ and did not investigate the impact on the SDESP as a whole (i.e. SDESP with manual grading at level 1 vs SDESP with ARIAS at level 1). The study was judged to be at low risk of bias except for the Reference Standard domain which was graded as ‘high risk’ because the reference grading was done by a single grader (a clinical research fellow).

For referable diabetic retinopathy, iGradingM had sensitivity of 90.5% (95%CI 89.3–91.6) and specificity of 67.4% (95%CI 66.0–68.8). In comparison, the sensitivity of manual grading was 86.5% (95%CI 85.1–87.8) and the specificity 95.3% (95%CI 94.6–95.9). iGradingM and human graders misclassified as normal 240 and 341 patients with diabetic retinopathy, respectively, and the difference was statistically significant ( $p < 0.001$ ). Of the

patients with referable retinopathy (M1, R2, M2, R3 or R4), iGradingM and human graders classified 7/330 and 3/330, respectively, as 'no retinopathy' but the difference was not statistically significant ( $p=0.125$ ).



**Table 10 Studies comparing directly the accuracy of the ARIAS and human graders**

ARIAS: study and country	1. Study design 2. Dataset 3. RS	Comparator	Accuracy of human graders (95% CI)	Accuracy of ARIAS (95% CI)
Google AI: Krause 2018, USA	1. Retrospective cohort study, 2. EyePACS-2: 1958 images from 998 unique individuals 3. Consensus by 3 retinal specialists	3 ophthalmologists, individually and as a majority decision	<p><u>Accuracy for rDR:</u> Ophthalmologists' majority decision: SE 83.8% SP 98.1% Individual ophthalmologists (range): SE 74.9% to 76.4%, SP 97.5% to 99.1%</p> <p><u>Accuracy for referable DME:</u> Ophthalmologists' majority decision: SE 83.3% SP 99.0% Individual ophthalmologists (range): SE 62.7% to 86.4%, SP 98.6% to 99.1%</p>	<p><u>Accuracy for rDR:</u> SE 97.1% SP 92.3%</p> <p><u>Accuracy for referable DME:</u> SE 94.9% SP 94.4%</p> <p>No statistical comparison reported</p>
Google AI: Raumviboonsuk 2019, Thailand	1. Retrospective cohort study 2. Diabetes registry 3. Partially adjudicated by independent retinal specialists	Regional graders from Thailand DESP	<p><u>Accuracy for RDR</u> SE 73.4% (range: 40.7% to 91.4% across regional graders)  SP 98.0% (range: 93.9% to 100%)</p> <p><u>Accuracy for severe NPDR/worse</u> AUC 0.993 (range: 0.974 to 0.995)</p>	<p><u>Accuracy for RDR</u> SE 96.8% (range 89.3% to 99.3%) (difference 24%, p&lt;0.001)  SP 95.6% (range 98.3% to 98.7%) (difference -2.5%, p&lt;0.001)</p> <p><u>Accuracy for severe NPDR/worse</u> AUC 0.991 (range: 0.978 to 0.997)</p>
RetCAD: Gonzalez-Gonzalo 2020, Spain, The Netherlands	1. Retrospective cohort study 2. Messidor (n=1200) 3. Dataset's ground truth	2 graders, a general ophthalmologist and a retinal specialist, with 4 and 20 years of DR screening experience, respectively	<p><u>Accuracy for rDR:</u> HG1 SE 79.6% (74.8%-84.8%) SP 97.7% (96.0%-99.2%) HG2 SE 69.0% (62.9%-74.7%) SP 99.1% (97.9%-100%)</p>	<p><u>Accuracy for rDR</u> SE 92.0% (89.1%-95.9%) SP 92.1% (88.7%-95.2%)</p> <p>No statistical comparison reported</p>
SELENA: Ting 2017, Singapore (and other countries, not included here)	1. Retrospective cohort study 2. 8589 unique patients (excluding those used in the development)	2 trained senior nonmedical professional graders with >5 years experience, currently employed in the SiDRP	<p><u>Accuracy for rDR:</u> SE 84.84% (81.28%-88.51%)  SP 98.55% (98.27%-98.79%)</p> <p><u>Accuracy for vtDR:</u></p>	<p><u>Accuracy for rDR:</u> SE 89.56% (85.51%- 92.58%), p=0.04 SP 83.49% (82.68%-84.27%), p&lt;0.001</p> <p><u>Accuracy for vtDR:</u></p>

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

	3. A single retinal specialist with >5 years experience		SE 89.74% (74.77%-96.27%) SP 99.09% (98.86%-99.27%)	SE 100% (90.97%-100%), p=0.04 SP 81.4% (80.57%-82.22%), p<0.001
iGradingM/Aberdeen system: Philip 2007, UK	1. Prospective cohort study 2. 14 406 images from 6722 consecutive patients from the SDESP 3. A single clinical research fellow	3 retinal screeners who also performed the photography	<u>Technical failures:</u> SE 93.7% (91%.3–95.4%) SP 99.0% (98.7%–99.2%) <u>Accuracy for rDR:</u> SE 86.5% (85.1%–87.8%) SP 95.3% (94.6%–95.9%) <u>N of patients misclassified as normal:</u> 341 <u>N of patients with M1, R2, M2, R3 or R4 graded as ‘no retinopathy’:</u> 3/330	<u>Technical failures:</u> SE 99.5% (98.4%–99.8%) SP 84.4% (83.5%–85.3%) <u>Accuracy for rDR:</u> SE 90.5% (89.3%–91.6%) SP 67.4% (66.0%–68.8%) <u>N of patients misclassified as normal:</u> 240, p<0.001 <u>N of patients with M1, R2, M2, R3 or R4 graded as ‘no retinopathy’:</u> 7/330, p=0.125
<u>Abbreviations:</u> CI – confidence interval, DR – Diabetic Retinopathy, HG1 – Human Grader 1, HG2 – Human Grader 2, rDR – Referable Diabetic Retinopathy, RS – Reference Standard, SE – Sensitivity, SDESP – Scottish Diabetic Eye Screening Programme, SiDRP - Singapore National Diabetic Retinopathy Screening Program, SP – Specificity, vtDR – Vision-threatening Diabetic Retinopathy				

## Discussion of findings

There is sufficient evidence from high quality diagnostic accuracy studies that the DL-based system EyeArt v2.1 has sensitivity of around 90%, comparable to that of human graders. It could be safely implemented in the EDESP for level 1 grading or as a filter before level 1 human grading, as suggested in Tufail 2016 (1). The system has been evaluated in multiple studies including a prospective pivotal study that informed the FDA approval (29, 30) and a large UK-based prospective multicentre study, Heydon 2020 (31), that evaluated the system in realistic EDESP conditions (31). In addition, an earlier version of the system, EyeArt v1, was evaluated by Tufail 2016 (1) in an EDESP cohort and showed similar sensitivity (1). Particular strengths of Heydon 2020 are that:

- The study included a large consecutively recruited representative cohort with pre-specified sample size based on Tufail 2016 (1).
- There was no selection or editing of the images prior to their processing with EyeArt v2.1 and all images that would normally be sent for manual grading were included.
- The performance of EyeArt v2.1 was shown to be robust across 3 EDESP sites which varied in a number of ways including mean age and racial composition of the cohorts. Additional subgroup data provided by the authors as well as data from Tufail 2016 (EyeArt v1) show that patient characteristics, such as race and age, and technical factors, such as camera type, have little or no impact on the accuracy of the system.

Depending on whether the system is used to screen out patients with no disease or to differentiate between 'referable' and 'non-referable' cases, the specificity and the respective workload reduction that could be expected will vary, and is likely to have an impact on the cost-effectiveness of the system.

There is high quality evidence for acceptable sensitivity of two ML-based systems, iGradingM in the SDESP and RetmarkerSR in the EDESP. Both ARIASs are currently used in Scotland and Portugal, respectively, and there is some published evidence that their sensitivity remains high after implementation. However, iGradingM may not be able to work with the EDESP photographic protocol without some additional pre-processing of the images; and the evidence for the performance of RetmarkerSR in the EDESP comes from a single study in which the sensitivity of the system was considerably lower than that of EyeArt v1 (1).

We found no high quality studies reporting on the accuracy of the other ARIASs in the UK DESPs. Systems using similar protocols and evaluated in high quality studies in similar settings and populations outside the UK, and especially those evaluated *independently* from the software developer, are expected to have comparable accuracy in the UK DESPs. However, given the large number of contextual factors that could affect the performance of

these systems (see Appendix 5), generalising the results from studies conducted in other countries should be avoided and used only in the initial selection of candidate ARIASs. The selected systems should be evaluated in the DESP in which they are intended to be used.

The evidence on the comparative accuracy of alternative ARIASs is also limited. We identified only one study, Tufail 2016 (1), which compared directly the performance of 3 ML-based ARIASs in a clinical cohort of consecutive participants. Unfortunately, the results from this study may not be applicable to the new versions of the software. Given the considerable clinical heterogeneity across studies, indirect (between-study) comparisons are unlikely to produce valid results and, therefore, the question about the comparative accuracy of ARIASs need to be investigated in future studies.

We found no studies that compared the overall performance of DESP with level 1 human graders versus DESP with level 1 ARIAS grading. The comparisons between ARIAS and human graders were limited to level 1 grading, with manual grading often included in the reference standard. The overall impact of using an ARIAS for level 1 grading was inferred from the accuracy estimates and there was no direct evidence of the actual impact that such change may have (including the impact on human behaviour).

One journal article published in 2014 reported accuracy data from an internal quality assessment of RetmarkerSR in the Portuguese DESP (18) and one conference abstract published in 2017 reported accuracy and usage data for the autograder in the SDESP (which we assumed to be a version of iGradingM, although the system was not explicitly named) (17). Given that both systems have been in use for more than 10 years, data on their impact on the overall performance of the respective DESPs could be very helpful for future research and policy decisions. The audit data from Scotland indicates that in clinical practice the sensitivity of the system remains similar to that reported from previous evaluations, while the specificity is at the lower end of the range (sensitivity of 97%, specificity of 38% and false negative rate of 0 to 0.6%). The number of episodes handled by the programme had increased by 20.3% in the period from 2010 to 2015 and, in the observed 6-month period in 2015, 58.1% of all episodes were passed on to the autograder (17).

### Summary of Findings Relevant to Criteria 4 and 5: Criterion met.

There is sufficient evidence from high quality diagnostic accuracy studies that the DL-based system EyeArt v2.1 has sensitivity of around 90%, comparable to that of human graders, and could be safely implemented in the EDESP as level 1 grading or as a filter before level 1 human grading, as suggested in Tufail 2016. There is also high quality evidence for the acceptable accuracy of iGradingM in the SDESP, but the system may not be able to perform with 2-field photographic protocols. There is also evidence from a single high-quality study of the accuracy of RetmarkerSR in the EDESP (1). The system had higher specificity but lower sensitivity compared to EyeArt v1. iGradingM (or related software) and RetmarkerSR are currently implemented in Scotland and Portugal, respectively. Limited published evidence suggests that their sensitivity remains high after implementation and they are safe to be used at level 1 grading.

There is no high quality evidence for the accuracy of the other ARIASs in the UK DESPs. Given the large number of factors that could affect the performance of these systems and the considerable clinical heterogeneity across studies: 1) the results from studies conducted in other countries should be used only for initial selection of candidate systems; 2) high quality evaluations conducted independently from the software developer / manufacturer should be prioritised; 3) indirect (between-study) comparison of alternative ARIASs is unlikely to lead to valid results and the comparative accuracy of alternative systems should be assessed directly (in the same study) or in the same cohort under similar, pre-specified conditions.

## Criterion 11. Effectiveness of DESPs using ARIASs at level 1 grading

Criterion 11. There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an “informed choice” (such as Down’s syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.

*Question 2 (rapid review) – What is the clinical impact of diabetic eye screening programmes with the use of automated retinal image analysis systems (ARIAS) for level 1 grading compared with diabetic eye screening programmes with fully manual grading?*

## Eligibility for inclusion in the review

Studies were included in the review if they met the following inclusion criteria:

- Population: People with type 1 or type 2 DM aged 12 years and over.
- Intervention: A DESP for the detection of DR and/or maculopathy that uses ARIAS on fundus photographs for level 1 grading followed by manual grading for level 2 and 3.
- Comparator: A DESP for the detection of DR and/or maculopathy that uses human manual grading on fundus photographs at all levels of grading.
- Outcome measures: Any clinical utility outcomes, including clinically significant outcomes and patient management and practical implications outcomes.
- Study design: RCTs, prospective and retrospective cohort studies, and systematic reviews and meta-analyses of these.
- Publication date and language: Studies had to be published in English after 2000.

## Description of the evidence

We did not identify any RCTs that investigated the above question. We included 2 prospective cohort studies (Table 11), both at high risk of bias and of limited applicability (**Error! Reference source not found.**). Some accuracy studies reported projected impact in terms of workload reduction and other outcomes, but these were non-comparative retrospective cohort studies and the reported outcomes (e.g. workload reduction) were inferred from the reported accuracy estimates.

**Table 11 Studies evaluating the impact of implementing ARIAS in a DESP**

Study & country	ARIAS	Study design and PICO	Outcomes	Risk of bias
Keel 2018 (36), Australia	EyeGrader™ (DL-based)	Prospective cohort study, within-patient control: P: 96 patients attending 2 endocrinology departments I: ARIAS-based DR screening providing an immediate result to the patient C: human grading with results available online in 2 weeks O: Satisfaction and preferences (through a questionnaire), mean assessment time of ARIAS-based screening	Mean assessment time for ARIAS-based screening was 6.9 min 96% of patients were very satisfied or satisfied 78% of patients preferred ARIAS	High risk of bias: convenience sample, no control group, within-subjects comparison with ARIAS results always provided to the participant first
Liu 2020 (33), USA	EyeArt 2.0 (DL-based)	Prospective cohort study, historical control: P: 180 patients with type 1 or 2 diabetes (≥18 years old) visiting a primary care centre for low-income patients (from January 1 <sup>st</sup> , 2018 through to August 31 <sup>st</sup> , 2018) I: ARIAS-based DR screening C: Adherence rate of a historical cohort from the same primary care centre. The cohort included consecutive adult patients with diabetes seen over a period of 9 months (July 1, 2016, and March 31, 2017).	Among patients referred for a follow-up exam, the adherence rate was 55.4% at 1 year vs. historical adherence rate of 18.7% (p < 0.0001). No false negative results were found. 17 patients had additional pathologic features that required evaluation earlier than recommended by ARIAS (as identified by human graders: 9 with grade 1 to 2 hypertensive retinopathy; 2 with AMD; 7 were glaucoma suspects, and 1 with nonspecific chorioretinal scarring).	High risk of bias: possible selection bias; historical controls; patients in the ARIAS cohort received 3 telephone calls and a letter to encourage them to attend; it is unclear if the same level of encouragement was used in the historical controls; 42 patients received their results within 2-weeks of the index test (which could serve as another reminder)
AMD – Age-related Macular Degeneration, ARIAS - Automated Retinal Image Analysis System, DESP – Diabetic Eye Screening Programme, DL – Deep Learning, DR – Diabetic Retinopathy, PICO – Population, Intervention, Comparator, Outcome				

## Discussion of findings

The studies are summarised in Table 11. Both studies had key methodological issues identified in the quality assessment and reported in full in **Error! Reference source not found.** Keel 2018 (Australia) compared the time taken and the acceptability of ARIAS-based DR screening where the patients received the result immediately, relative to manual grading where the patients can access their results

after 2 weeks. However, there was no control group, i.e. patients served as their own controls, and the study was judged to be at high risk of bias (36).

Liu 2020 (USA) investigated whether using ARIAS to make the result from the examination immediately available to patients improves adherence to follow-up. However, the study was judged to be at high risk of selection bias; used historical controls and, in some cases, the result was provided to the patient with considerable delay (33).

### Summary of Findings Relevant to Criterion 11: Criterion not met.<sup>9</sup>

We did not identify any relevant RCTs or high quality prospective cohort studies comparing DESP with level 1 manual grading to DESP with level 1 ARIAS grading in terms of clinical outcomes and overall impact. The two prospective cohort studies included here did not include concurrent comparator groups; were judged to be at high risk of bias; and their results may not generalise to the UK DESP. Future studies should start by clarifying the range of relevant clinical outcomes to be investigated by involving all relevant stakeholders. Large prospective trials may not be feasible and alternative study designs should be explored.

---

<sup>9</sup> **Met** -for example, this should be applied in circumstances in which there is a sufficient volume of evidence of sufficient quality to judge an outcome or effect which is unlikely to be changed by further research or systematic review.

**Not Met** - for example, this should be applied in circumstances where there is insufficient evidence to clearly judge an outcome or effect or where there is sufficient evidence of poor performance.

**Uncertain** -for example, this should be applied in circumstances in which the constraints of an evidence summary prevent a reliable answer to the question. An example of this may be when the need for a systematic review and meta-analysis is identified by the rapid review.



## Criterion 14. Cost-effectiveness of ARIASs

Criterion 14. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money).

Assessment against this criterion should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

*Question 3 (evidence map) – What is the cost-effectiveness of an automated retinal image systems (ARIAS) in diabetic eye screening programmes compared with diabetic eye screening programmes with manual grading?*

## Eligibility for inclusion in the review

Studies were included in the review if they met the following inclusion criteria:

- Population: People with type 1 or type 2 DM aged 12 years and over.
- Intervention: A DESP for the detection of DR and/or maculopathy that uses ARIAS on fundus photographs for level 1 grading followed by manual grading for level 2 and 3.
- Comparator: A DESP for the detection of DR and/or maculopathy that uses manual grading on fundus photographs at all levels of grading.
- Outcome measures: Any cost-effectiveness or modelled clinical outcomes, e.g. NNS, NNT, proportion of appropriate screening outcome (true positive or true negative results correctly identified), interval retinopathy or maculopathy, loss of vision, proportion of vision loss prevented, number/proportion of different grades of retinopathy and maculopathy detected (including ungradable images), incidental findings.
- Country in which the study was conducted: UK (non-UK studies were excluded).
- Study design: Economic evaluations (of any type) and reviews of these.
- Publication date and language: Studies have to be published in English after 2002.

## Summary of findings

Eighteen references were identified from title/abstract screening as potentially relevant to the evidence map. Nine of these were subsequently excluded: one conference abstract did not report an evaluation of ARIAS (Harding 2019), 3 were review/opinion papers (Scanlon

2019, Sosale 2019, Dismuke 2020) and 6 were not set in the UK (Ballreich 2016, Australia; Fuller 2019, US; Xie 2019 and 2020, Singapore; Andreasen and Kjellberg 2008, Denmark). The 9 included references report on 5 different evaluations: one reporting separate results for England and Scotland (Olsen 2013, Prescott 2014), one based in England (Tufail 2016 and 2017, Liew 2014, Egan 2016), two based in Scotland (Scotland 2007 and 2010) and one described as based in the UK (Bhaskaranand 2016). The evaluations generally find automated grading to be less costly than manual grading, but less effective. Therefore, many of the results are reported in terms of the additional costs associated with manual grading to gain additional health benefits when compared to automated grading.

Olson 2013 (44) and Prescott 2014 (45) use data from centres in England and Scotland to compare fully automated grading to the English and Scottish manual grading systems (and a further strategy based on the English system). The focus of the study was the presence of maculopathy in a cohort of patients already diagnosed with diabetic retinopathy. The authors use a Markov microsimulation model with a 20 year time horizon, with costs from year 2009/2010 to conduct a cost-effectiveness analysis (CEA) and cost-utility analysis (CUA). In terms of both cost per cases detected and cost per QALYs (after 20 years), Olsen and Prescott estimate that the fully automated system (iGradingM, Medalytix Ltd) dominates the English manual grading system (it is cheaper and more effective). To correct for sampling bias in the Scottish dataset used, the authors adjusted the frequency of different features to reflect expected frequency within a screening programme. They then estimated that the fully adjusted strategy cost an additional £900 per case detected compared to the Scottish manual system. The fully automated system was estimated to cost £113 more than the Scottish manual system and provide incremental QALYs of 0.0005, thus the fully automated system had an ICER of £222,210 compared to the Scottish system.

Tufail 2016 (1) and Tufail 2017 (46) estimate the cost-effectiveness of RetmarkerSR and EyeArt v1 (both ML-based) as a replacement of level 1 manual grading (strategy 1) and as a filter prior to level 1 manual grading (strategy 2). They use a decision tree model, with a 1-year time horizon and costs from 2013/2014. It is assumed that manual grading is implemented as in the Homerton University Hospital, London. For both ARIASs and strategies, automated grading is estimated to be less costly, but less effective, than manual grading. However, since the overall performance is driven down by the relatively high false positive error rate (while the systems are comparable to manual grading in picking up cases with disease), both systems and strategies are cost-effective and could reduce the requirement for manual grading. For strategy 1 and strategy 2, the reduced cost (relative to manual grading) per appropriate outcome missed by the software was £4.51 and £2.80 with EyeArt v1 and £11.81 and £9.71 with RetmarkerSR, respectively. Appropriate outcomes

were defined as either disease present [M1, R2, R3, and U] or absent [R0, R1] which agreed with human graders.

Bhaskaranand 2016 (47) report in a conference abstract the incremental costs for the use of EyeArt v1 (ML-based) compared to fully manual grading strategies (defined to be similar to that in the UK NHS DESP). They conclude that EyeArt v1 could lead to significant cost savings.

Scotland 2007 (48) report a CEA to compare level 1 manual grading in the Scottish DESP with an automated system (most likely an earlier version of iGradingM). They use a decision tree approach to model the number of referable cases detected and appropriate screening outcomes, alongside costs. The automated grading was estimated to identify 50 fewer cases, but save £201,600 per year compared to manual grading. They conclude that automated grading is likely to be cost-effective due to similar effectiveness, but lower costs, compared to manual grading.

Using data from three centres in Scotland, Scotland 2010 (49) compare the automated grading system described in Scotland 2007 to an improved automated grading system (most likely a version of iGradingM), and to manual grading. They take the model used by Scotland 2007, and extrapolate this to allow consideration of the consequences of any missed referable cases. Scotland 2010 estimate that the improved automated grading system performs better than the original automated system, but misses 123 referable cases when compared to manual grading. However, the improved automated grading system is estimated to be less costly than manual grading (by £212 695). They also estimate that given manual grading is more effective, to gain an additional QALY using the manual grading rather than the improved automated system, the additional cost is £25,676 to £267,115 depending on the probability of the improved automated system missing true proliferative cases. The authors, therefore, conclude that the improved automated system is likely to be cost-effective compared to manual grading.

Further details on the studies, as reported in the abstracts, are given in Table 25. Those studies that attempt to capture cost-effectiveness over time are preferred. Those where the outcome is cost per case detected will not capture the longer-term impacts of correct/incorrect identification.

In summary, we identified 5 studies (one published only as an abstract) that evaluated the cost-effectiveness of 3 ARIASs—EyeArt v1, RetmarkerSR and iGradingM—all based on traditional ML algorithms. The studies show that ARIASs are less effective but less costly compared to manual grading and could lead to considerable savings. The performance is driven down by the relatively high false positive error rate while the systems are comparable to human graders in picking up cases with disease. Although these studies provide good starting point for further evaluations, they need updating to incorporate new data and to be able to capture cost-effectiveness over time.

## Criterion 12. Social and ethical implications of AI in screening programmes

Criterion 12. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.

*Question 4 (evidence map) – What are the social and ethical implications of implementing AI-based tools in screening programmes and would it be acceptable to health professionals and the public?*

### Eligibility for inclusion in the review

Studies were included in the review if they met the following inclusion criteria:

- Population: Health professionals, providers and users of screening programmes, and the general public.
- Intervention: Implementation of AI-based tools in any screening programme.
- Outcome measures: Social implications (e.g. public's opinion on the use of AI in their clinical management), ethical implications (e.g. data privacy, involvement of third party i.e. developers of ARIASs) and acceptability.
- Study design: Qualitative studies (e.g. surveys, interviews, clinical audits and service evaluation reports); relevant opinion/discussion papers.
- Publication date and language: Studies have to be published in English since 2000.

### Summary of findings

Eighty three titles were selected at title/abstract level for further inspection. Of those, 19 primary studies (Table 26) and 38 reviews and opinion papers (Table 27) were considered to be potentially relevant and included in the evidence map. Five of the primary studies investigated the impact of AI in the context of screening (36, 50-53) while the rest had a more general focus (e.g. radiologists' attitudes towards AI) or investigated the topic in a different setting (e.g. clinicians perceptions of a ML-based early warning system to predict severe sepsis and septic shock (54)). The reason to include this latter group of studies was that the investigated issues (e.g. clinicians' trust in AI technology) were judged to be relevant to screening, even though the study was conducted in a different setting.

All primary studies were surveys of clinicians (n=9), clinicians and the general public (n=2), the general public (n=2) and patients (n=5). Clinicians were radiologists in 7 of the studies and psychiatrists, physicians and healthcare professionals, respectively, in the other 3. One

of the surveys was aimed at directors of screening programmes, but no further information was given in the abstract. The patients were adults participating in DESPs (n=3), neurosurgery patients and their families (n=1) and women participating in a breast cancer screening programme (n=1). The surveys were conducted in USA (n=4), UK (n=3), France (n=1), and one each in Australia, China, Europe, Germany, India, Italy, Singapore and Sweden.

The surveys investigated a broad range of questions including participants' knowledge, training needs, perceptions, attitudes and satisfaction in relation to AI-based technology. Most of the participants had positive attitudes towards the implementation of AI in healthcare and acknowledged the benefits that such technologies are likely to bring, both in terms of improved patient outcomes and benefits to the healthcare system as a whole. Studies also reported a range of concerns regarding the impact of AI-based technology on clinicians' professional role and identity, clinician-patient relationship, dealing with uncertainty, the impact on clinical decision making, and the need for training and better understanding of AI by healthcare professionals, patients and the general public.

Of the reviews and opinion papers, Fatehi 2020 was a systematic review of the characteristics and usability features of tele-ophthalmology for the elderly population, including AI-based screening (55); Carter 2020 was a non-systematic review of the ethical legal and social implications of AI implementation in breast cancer care (56) and Larson 2020 was a theoretical paper in which the authors proposed an ethical framework for using and sharing clinical data for the development of AI applications (57). The rest of the papers were non-systematic reviews and editorials many of which focused specifically on the implementation of ARIASs in DESPs or ophthalmology in general (Table 27).

In summary, there is a rapidly growing volume of evidence on the social and ethical aspects of AI in screening and healthcare. An evidence review is warranted and would help identify the range of relevant topics, summarise the existing evidence and identify gaps that need further investigation. However, a traditional evidence review limited to screening may not be the most appropriate approach; instead, we suggest that methods, such as realist synthesis, are employed to make use of evidence from related settings (e.g. healthcare and other areas) that deal with similar issues and generate evidence transferable to screening.

# Review summary

## Conclusions and implications for policy

Despite the large number of publications on AI algorithms designed to detect and grade diabetic eye disease, only a few ARIASs have been evaluated in large clinical studies that provide reliable data on their accuracy in clinical practice. Given the large number of contextual factors that could affect the performance of these systems, the results from such studies should not be generalised beyond the setting in which they were conducted. They could be used to inform the initial selection of candidate ARIASs, which then should be evaluated in the setting in which they are to be implemented. Also, indirect (between-study) comparisons of alternative ARIASs are unlikely to produce valid results.

This means that out of the 10 ARIASs included in this review, we have applicable high quality evidence for 3 systems: EyeArt v2.1 (DL-based), iGradingM (ML-based) and RetmarkerSR (ML-based). Evidence from multiple studies show that EyeArt has consistently high sensitivity (~90%), comparable to that of human graders. In the EDESP, it could be used to filter out images with no signs of retinopathy prior to manual grading or as a replacement of level 1 manual grading. Although most of the studies focused on referable disease, the EDESP-based evaluations show that the sensitivity of the system remains high ( $\geq 90\%$ ) for both 'disease/no disease' and 'referable/non-referable disease' and approaches 100% for more severe forms of diabetic retinopathy (1, 31, 46).

The large prospective study conducted by Heydon et al shows that the sensitivity of the most recent version of the software, EyeArt v2.1, remains high when evaluated in realistic conditions and the performance of the system is stable across different EDESP sites. Although there is no direct evidence about the overall impact of EyeArt on the EDESP, Heydon 2020 estimated that, with referable disease as a threshold, approximately 50% of all screening episodes would require further human grading. They report that this would not result in an increased workload for the secondary grader and the workload of the tertiary grader (arbitration) would likely be reduced (31). The HTA conducted by Tufail 2016 also showed that implementing the system either as a replacement of level 1 manual grading or as a filter prior to manual grading is cost-effective. Given that the new version of the system evaluated in Heydon 2020 has comparable sensitivity and higher specificity, its implementation could lead to even greater savings than those reported by Tufail et al (1, 31, 46).

There is also high quality evidence of the performance of iGradingM in the SDESP including an analysis of audit data published in 2017 (17). This analysis shows that the

system is safe to use in clinical practice with sensitivity of 97%, comparable to that in the published evaluations, and a false negative rate of 0 to 0.6%. The study also shows that use of the system had increased over time and, in the observed 6-month period in 2015, 58.1% of all screening episodes in the SDESP were passed on to the autograder. However, iGradingM may not work with the 2-field imaging protocol used in the EDESP (1).

Another ML-based system, RetmarkerSR, has also been evaluated in the EDESP. Tufail 2016 reported that it had lower sensitivity but higher specificity compared to EyeArt v1 and was cost-effective both as a replacement of level 1 manual grading and as a pre-screen before manual grading (1). An analysis of audit data from the Portuguese DESP, where RetmarkerSR is currently in use, shows that the system has very low false negative rate and is safe to be used in clinical practice (18).

Only one study, Tufail 2016 (1) reported on the comparative accuracy of alternative ARIASs, RetmarkerSR and EyeArt v1, and the results may not be applicable to the new versions of the software. Four studies compared the accuracy of ARIASs to that of human graders who were not involved in the reference grading. Overall, ARIASs had higher or similar sensitivity to that of manual grading but lower specificity. However, the results are likely to depend on factors, such as the background and experience of human graders, and may not be generalizable to other programmes.

We did not find any relevant RCTs or high quality prospective cohort studies comparing DESP with level 1 manual grading to DESP with level 1 ARIAS grading in terms of clinical outcomes or other impact measures. We included 2 prospective cohort studies conducted in the USA and Australia, and evaluating EyeArt v2 and EyeGrader, respectively. They were conducted in primary care and looked at the impact of ARIAS screening on patient satisfaction (Australia) and adherence to follow up examination (USA). Both studies were considered to be at high risk of bias and their results are not directly applicable to the UK DESPs.

Three ML-based ARIASs were evaluated in the UK-based health economic studies included in the review: iGradingM (or related software), RetmarkerSR and EyeArt v1. Overall, the studies found that ARIASs are less effective but also less costly compared to manual grading. Their overall performance was driven down by the high false positive rate while their sensitivity was comparable to that of human graders. Given that the current DL-based systems have higher specificity and comparable sensitivity, the implementation of ARIAS as an initial screen could result in even greater savings than those reported in these studies. Tufail 2016 also showed that while using ARIAS as a filter prior to level 1 manual grading is cost-effective, the strategy where an ARIAS is used as a replacement of level 1 manual grading could lead to greater savings. Other factors that are likely to affect the cost-

effectiveness of the systems are the decision threshold ('disease/no-disease' vs 'referable/non-referable disease') and the screening intervals (every year vs every 2 years).

In conclusion, there is sufficient high quality evidence from the EDESP that EyeArt v2.1 may be safe and cost-effective either using it for level 1 grading or as a filter prior to manual grading. An implementation study should be carried out to investigate other aspects of its performance, such as interoperability, cyber security, impact on human behaviour, acceptability by staff and patients, and to inform an update of the cost-effectiveness analysis. As a precaution, the system should be implemented initially to screen out patients with no disease while its performance at a higher threshold (referable/non-referable disease) and longer screening intervals (similar to those in the SDESP) should be investigated as part of the study.

We also recommend an evidence review on the social and ethical aspects of implementing AI-based technology in population screening. Such review could identify the range of relevant topics, summarise the existing evidence and identify gaps that need further investigation. However, the traditional rapid review approach with narrow focus on screening programmes may not be appropriate. Instead, methods, such as realist synthesis, could be employed to make use of evidence from related settings that deal with similar issues and generate evidence transferable to screening.

### *Future research*

Future evaluations of ARIAS should:

- be done independently from the software developer, in the clinical setting in which the system is meant to be used, under conditions that reflect everyday clinical practice. If possible, they should compare the performance of alternative ARIASs that may have different advantages and disadvantages in that setting.
- look at outcomes beyond accuracy, such as the actual consequences (including long term impact) of false negative and false positive results.
- investigate the consequences of accidental findings that lead to referral by human graders but are not captured by the ARIAS.
- include a cost-effectiveness evaluation of the system.
- investigate the experience and perceptions of healthcare professionals who interact with and/or are directly affected by the ARIAS; the expectations of those who have not yet had this experience (e.g. those in the control arm); the experience and perceptions of relevant patient groups; and the overall impact on the NHS.

There is a wide range of methodological questions that require further discussion and investigation:



- Some of these concern study design, stages of evaluation and type of evidence that needs to be available prior to the implementation of an ARIAS in a specific DESP programme. What should be the acceptable reference standard in test accuracy evaluations of ARIASs, both in terms of technology and in terms of protocol to produce the final grade? What other outcomes, in addition to accuracy, should be investigated (e.g. actual consequence of false negative and false positive results)? What study designs should be used at different stages of the evaluation that are both feasible and able to produce reliable evidence? How should the overall impact of ARIAS (relative to DESP with fully manual grading) be investigated?
- AI software is constantly evolving and this is one of its strengths. What is the best way to manage and monitor this, to make sure that the next version is safe, reliable and at least as effective as the previous one? Should the performance of ARIAS, once implemented, be monitored in the same way as DESP with human graders or different process is required? What lessons could be learnt from the programmes in Scotland and Portugal?
- The tasks that ARIASs should be able to perform. Should they be able to identify a broader range of conditions (e.g. glaucoma and age-related degenerative maculopathy, which some ARIASs have been designed to do)? Should they be able to incorporate other information into the final decision, such as information from the previous visit(s) (e.g. RetmarkerSR) and other demographic and clinical data (as some systems not included here have been designed to do)?
- How the impact of ARIAS on healthcare professionals, patients and the NHS as a whole should be investigated? So far, most of the studies are in the form of surveys asking general questions with little regard of respondents' experience and interaction with AI. Can we learn from the experience in other areas of AI research, outside screening, to design more informative and reliable studies of the impact of AI?

## Limitations

The following methodological limitations of the review should be acknowledged: only the main electronic databases were searched; searches were limited to records published since 2000, and included only peer-reviewed, English-language journal articles; only 20% of the titles and abstracts were double-screened; papers were excluded after assessment of the volume of published evidence (although prioritisation was based on pre-specified criteria); the definition of one of the signalling questions in QUADAS-2 checklist was changed following a discussion with topic experts and after the initial grading of studies; as a result some studies were later regraded using the new definition.

High quality evidence on the accuracy of ARIAS in the UK DESPs was found only for 3 systems. Given the large number of contextual factors that may affect the performance of

ARIASs, generalising the results from studies conducted in other countries is not advisable. No RCTs or high quality prospective cohort studies were found that compare directly DESP with level 1 ARIAS grading vs DESP with level 1 human grading and report outcomes beyond accuracy. The identified health economic evaluations show that the systems are cost-effective. However the analyses need updating to include new data and to investigate the long term effects of implementing ARIASs and the potential impact of extended screening intervals.

## Appendix 1 — Search strategy

### Electronic databases

Two separate strategies were developed. The first strategy aimed to identify studies related to questions 1 to 3 (accuracy, effectiveness and cost-effectiveness of ARIASs in DESP) while the second search strategy aimed to identify studies relevant to question 4 (ethical and social aspects of the implementation of AI in screening programmes). The two strategies are detailed in the next section while details of the searches are provided in Table 12. Results were imported into EndNote X8.2 (Thomson Reuters) and de-duplicated.

**Table 12 Summary of electronic database searches and dates**

Database	Platform	Searched on date	Date range of search	Hits
<b>Searches covering question 1-3</b>				
MEDLINE(R) ALL 1946 to June 25, 2020	Ovid SP	26/06/2020	2000 – 25 June 2020	991
EMBASE 1974 to 2020 June 25 Date limited to 2000-current	Ovid SP	26/06/2020	2000 – 25 June 2020	894
Cochrane Database of Systematic Reviews (CDSR)	Wiley Online	26/06/2020	Inception to present	0
Cochrane Central Register of Controlled Trials (CENTRAL)	Wiley Online	26/06/2020	Inception to present	27
ICTRP	Not available at present			N/A
Clinicaltrials.gov		26/06/2020	Inception to present	9
Total 1912 titles plus 9 clinical trials Duplicates: 544 For title and abstract screening: 1377				
<b>Searches covering question 4</b>				
Ovid MEDLINE(R) ALL 1946 to June 30, 2020	Ovid SP	01/07/2020	1946 to present	253
Embase 1974 to 2020 June 30	Ovid SP	01/07/2020	1974 to present	346
APA PsycInfo 1806 to June Week 5 2020	Ovid SP	01/07/2020	1806 to present	42
CINHAL		02/07/2020	Inception to present	103
Total: 744 Duplicates: 211 For title and abstract screening: 533				

## Search Terms

### Search terms included in the strategy covering questions 1 to 3

Search terms included combinations of free text and subject headings grouped into the following categories:

- disease area: eye disease, retinopathy, eye pathology, maculopathy, diabetic eye, diabetic macular, retinal fundus
- index test: diagnostic techniques, ophthalmological; diagnosis, computer assisted, sensitivity and specificity, diagnostic test, diagnostic accuracy, diagnostic performance, screening, imaging, reference standard, artificial intelligence, deep learning, neural network, automated retinal image analysis system, automated grading, automated level, ARIAS, iGradingM, EyeArt, IDx-DR, DR-RACS, RetinaLyze, RetmarkerSR, Singapore Eye Lesion Analyzer, RetinaVue, TRIAD network.

The search terms for Ovid MEDLINE(R) ALL are shown in Table 13; for Embase in Table 14; for the Cochrane Library in Table 15, and for the Clinical trials database in Table 16.

**Table 13 Search strategy for Ovid MEDLINE(R) ALL <1946 to June 25, 2020>**

N	Terms (N of hits)
1	exp eye diseases/ (563274)
2	retinopathy.ti,ab. (43575)
3	eye pathology.ti,ab. (319)
4	maculopathy.ti,ab. (4228)
5	diabetic eye.ti,ab. (732)
6	diabetic macular.ti,ab. (4004)
7	retinal fundus.ti,ab. (315)
8	1 or 2 or 3 or 4 or 5 or 6 or 7 (578122)
9	exp Diagnostic Techniques, Ophthalmological/ (169047)
10	exp Diagnosis, Computer-Assisted/ (83108)
11	"Sensitivity and Specificity"/ (346232)
12	diagnostic test*.ti,ab. (46123)
13	diagnostic accuracy.ti,ab. (43812)
14	diagnostic performance.ti,ab. (16194)
15	screening.ti,ab. (520833)
16	imaging.ti,ab. (798701)
17	(Sensitivity or specificity).ti,ab. (1033830)
18	reference standard.ab. (14139)
19	optical coherence tomography.ti,ab. (35177)
20	or/9-19 (2600413)
21	exp Artificial Intelligence/ (96870)
22	artificial intelligence.ti,ab. (6988)
23	deep learning.ti,ab. (8539)

24	neural network*.ti,ab. (48131)
25	automated retinal image analysis system.ti,ab. (2)
26	automated grading.ti,ab. (72)
27	automated level.ti,ab. (8)
28	(automated adj (tool* or technique* or identification or detection)).ti,ab. (4376)
29	ARIAS.ti,ab. (299)
30	iGradingM.ti,ab. (2)
31	EyeArt.ti,ab. (5)
32	IDx-DR.ti,ab. (6)
33	DR-RACS.ti,ab. (0)
34	RetinaLyze.ti,ab. (1)
35	RetmarkerSR DR.ti,ab. (0)
36	Singapore Eye Lesion Analyzer.ti,ab. (0)
37	RetinaVue.ti,ab. (0)
38	TRIAD network.ti,ab. (0)
39	or/21-38 (135626)
40	8 and 20 and 39 (1048)
41	limit 40 to yr="2000 -Current" (991)

**Table 14 Search strategy for Embase <1974 to 2020 June 25>**

N	Terms (N of hits)
1	exp eye disease/di [Diagnosis] (136263)
2	retinopathy.ti,ab. (59510)
3	eye pathology.ti,ab. (375)
4	maculopathy.ti,ab. (5348)
5	diabetic eye.ti,ab. (1103)
6	diabetic macular.ti,ab. (5986)
7	retinal fundus.ti,ab. (479)
8	1 or 2 or 3 or 4 or 5 or 6 or 7 (194744)
9	exp computer assisted diagnosis/ (1112353)
10	exp "sensitivity and specificity"/ (359564)
11	diagnostic test*.ti,ab. (64444)
12	diagnostic accuracy.ti,ab. (63870)
13	diagnostic performance.ti,ab. (23087)
14	screening.ti,ab. (729183)
15	imaging.ti,ab. (1108924)
16	(Sensitivity or specificity).ti,ab. (1302252)
17	reference standard.ab. (19637)
18	optical coherence tomography.ti,ab. (47622)
19	or/9-18 (3902148)
20	exp Artificial Intelligence/ (39401)
21	artificial intelligence.ti,ab. (9221)
22	deep learning.ti,ab. (10678)
23	neural network*.ti,ab. (59173)

24	automated retinal image analysis system.ti,ab. (4)
25	automated grading.ti,ab. (121)
26	automated level.ti,ab. (7)
27	(automated adj (tool* or technique* or identification or detection)).ti,ab. (5773)
28	ARIAS.ti,ab. (413)
29	iGradingM.ti,ab. (3)
30	EyeArt.ti,ab. (20)
31	IDx-DR.ti,ab. (13)
32	DR-RACS.ti,ab. (1)
33	RetinaLyze.ti,ab. (4)
34	RetmarkerSR DR.ti,ab. (0)
35	Singapore Eye Lesion Analyzer.ti,ab. (0)
36	RetinaVue.ti,ab. (6)
37	TRIAD network.ti,ab. (0)
38	or/20-37 (105880)
39	8 and 19 and 38 (917)
40	limit 39 to yr="2000 -Current" (894)

**Table 15 Search strategy for the Cochrane Library**

ID	Search	Hits
#1	MeSH descriptor: [Eye Diseases] explode all trees	18741
#2	retinopathy:ti,ab	4616
#3	eye pathology:ti,ab	314
#4	maculopathy:ti,ab	316
#5	diabetic eye:ti,ab	1426
#6	diabetic macular:ti,ab	2327
#7	retinal fundus:ti,ab	874
#8	#1 or #2 or #3 or #4 or #5 or #6 or #7	23185
#9	MeSH descriptor: [Diagnostic Techniques, Ophthalmological] explode all trees	7678
#10	MeSH descriptor: [Diagnosis, Computer-Assisted] explode all trees	1871
#11	MeSH descriptor: [Sensitivity and Specificity] explode all trees	15181
#12	"diagnostic test":ti,ab	1124
#13	"diagnostic accuracy":ti,ab	2512
#14	"diagnostic performance":ti,ab	742
#15	screening:ti,ab	48626
#16	imaging:ti,ab	32475
#17	(Sensitivity or specificity):ti,ab	46522
#18	"reference standard":ab	850
#19	"optical coherence tomography":ti,ab	2959
#20	#9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19	139933
#21	MeSH descriptor: [Artificial Intelligence] explode all trees	968

#22	"artificial intelligence":ti,ab	229
#23	"deep learning":ti,ab	229
#24	"neural network*":ti,ab	521
#25	"automated retinal image analysis system":ti,ab	0
#26	"automated grading":ti,ab	1
#27	"automated level":ti,ab	1
#28	(automated NEXT (tool* or technique* or identification or detection)):ti,ab	115
#29	ARIAS:ti,ab	23
#30	iGradingM:ti,ab	0
#31	EyeArt:ti,ab	1
#32	IDx-DR:ti,ab	0
#33	DR-RACS:ti,ab	0
#34	RetinaLyze:ti,ab	0
#35	RetmarkerSR:ti,ab	3
#36	"Singapore Eye Lesion Analyzer":ti,ab	0
#37	RetinaVue:ti,ab	2
#38	"TRIAD network":ti,ab	0
#39	#21 or #22 or #23 or #24 or #25 or #26 or #27 or #28 or #29 or #30 or #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38	1841
#40	#8 and #20 and #39	27

**Table 16 Search strategy for the Clinical trials database**

Terms	Search Results*	Entire Database**
<b>Synonyms</b>		
Eye Diseases	9 studies	9,594 studies
diseases of the eye	--	20 studies
Disorder of eye	--	1 studies
Disorders of the globe	--	1 studies
Eye Disorders	--	47 studies
oculopathy	--	2 studies
ophthalmic disorders	--	1 studies
Ophthalmological disorder	--	9 studies
ophthalmopathy	--	93 studies
Diseases	9 studies	274,158 studies
Disorders	2 studies	98,155 studies
condition	1 studies	39,071 studies
Eye	9 studies	16,012 studies
Ocular	2 studies	4,860 studies
ophthalmic	2 studies	2,545 studies
Oculus	1 studies	207 studies
Optic	1 studies	1,457 studies
Eyeball structure	--	1 studies
Orbital region	--	10 studies
Diabetic Retinopathy	9 studies	547 studies

Retinopathy diabetic	--	1 studies
Retinopathy	9 studies	2,972 studies
Retinal Disease	9 studies	2,933 studies
retina disorder	--	2 studies
Retinal disorder	--	16 studies
Diabetic	9 studies	4,372 studies
artificial intelligence	9 studies	227 studies
Computational Intelligence	--	1 studies
Machine Intelligence	--	2 studies
intelligence	9 studies	355 studies
artificial	9 studies	1,193 studies
Factitious	--	1 studies
Spurious	--	2 studies

### Search terms included in the strategy covering question 4

The search terms included combinations of free text and subject headings grouped into the following categories:

- intervention: diagnosis, computer assisted, diagnostic test, screening, imaging, artificial intelligence, automation, machine learning
- outcomes: attitudes, perception, accept, barriers, appropriate, experience, ethical, social
- study design: qualitative, interview, survey, question.

The search terms for Ovid MEDLINE(R) ALL are shown in Table 17; for Embase in Table 18; for PsycINFO in Table 19, and for CINAHL in Table 20.

**Table 17 Search strategy for Ovid MEDLINE(R) ALL <1946 to June 30, 2020>**

N	Terms (N of hits)
1	exp Diagnosis, Computer-Assisted/ (83147)
2	diagnostic test*.ti,ab. (46187)
3	screening.ti,ab. (521711)
4	imaging.ti,ab. (800097)
5	or/1-4 (1393715)
6	exp Artificial Intelligence/ (97028)
7	Automation/ (18051)
8	artificial intelligence.ti,ab. (7049)
9	(automated adj (tool* or technique* or identification or detection or test* or screening)).ti,ab. (5480)
10	automation.ti,ab. (13393)
11	machine learning.ti,ab. (27261)
12	or/6-11 (147104)



13	(attitude* or perception* or acceptab* or barriers or appropriate* or experience*).ti,ab. (2126676)
14	(ethical* or social*).ti,ab. (600302)
15	13 or 14 (2557367)
16	qualitative research/ (55100)
17	(qualitative or interview* or survey* or question*).ti,ab. (1802109)
18	16 or 17 (1807651)
19	5 and 12 and 15 and 18 (253)

**Table 18 Search strategy for Embase <1974 to 2020 June 30>**

N	Terms (N of hits)
1	exp computer assisted diagnosis/ (1113982)
2	diagnostic test*.ti,ab. (64500)
3	screening.ti,ab. (729884)
4	imaging.ti,ab. (1110040)
5	or/1-4 (2682904)
6	exp artificial intelligence/ (39529)
7	automation/ (57309)
8	artificial intelligence.ti,ab. (9278)
9	(automated adj (tool* or technique* or identification or detection or test* or screening)).ti,ab. (7409)
10	automation.ti,ab. (19034)
11	machine learning.ti,ab. (33905)
12	or/6-11 (141036)
13	(attitude* or perception* or acceptab* or barriers or appropriate* or experience*).ti,ab. (2853491)
14	(ethical* or social*).ti,ab. (751387)
15	13 or 14 (3382481)
16	qualitative research/ (75324)
17	(qualitative or interview* or survey* or question*).ti,ab. (2360248)
18	16 or 17 (2367415)
19	5 and 12 and 15 and 18 (346)

**Table 19 Search strategy for PsycINFO <1806 to June Week 5 2020>**

N	Terms (N of hits)
1	computer assisted diagnosis/ (1578)
2	diagnostic test*.ti,ab. (3143)
3	screening.ti,ab. (63770)
4	imaging.ti,ab. (73432)
5	or/1-4 (140227)
6	exp artificial intelligence/ (21146)

7	automation/ (2264)
8	artificial intelligence.ti,ab. (3787)
9	(automated adj (tool* or technique* or identification or detection or test* or screening)).ti,ab. (608)
10	automation.ti,ab. (2543)
11	machine learning.ti,ab. (5086)
12	or/6-11 (28041)
13	attitudes/ (26982)
14	(attitude* or perception* or acceptab* or barriers or appropriate* or experience* or views).ti,ab. (1209016)
15	(ethical* or social*).ti,ab. (814078)
16	or/13-15 (1752604)
17	exp qualitative methods/ (15326)
18	(qualitative or interview* or survey* or question* or focus group*).ti,ab. (1091184)
19	17 or 18 (1092960)
20	5 and 12 and 16 and 19 (42)

**Table 20 Search strategy for CINAHL**

#	Query	Results
S20	S5 AND S12 AND S15 AND S19	103
S19	S16 OR S17 OR S18	793,966
S18	TI ( qualitative or interview* or "focus group" or survey or questionnaire* ) OR AB ( qualitative or interview* or "focus group" or survey or questionnaire* )	705,051
S17	(MH "Interviews+")	253,066
S16	(MM "Qualitative Studies+")	7,649
S15	S13 OR S14	1,055,374
S14	TI ( ethical* or social* ) OR AB ( ethical* or social* )	332,302
S13	TI ( attitude* or perception* or acceptab* or barriers or appropriate* or experience* ) OR AB ( attitude* or perception* or acceptab* or barriers or appropriate* or experience* )	834,228
S12	S6 OR S7 OR S8 OR S9 OR S10 OR S11	24,350

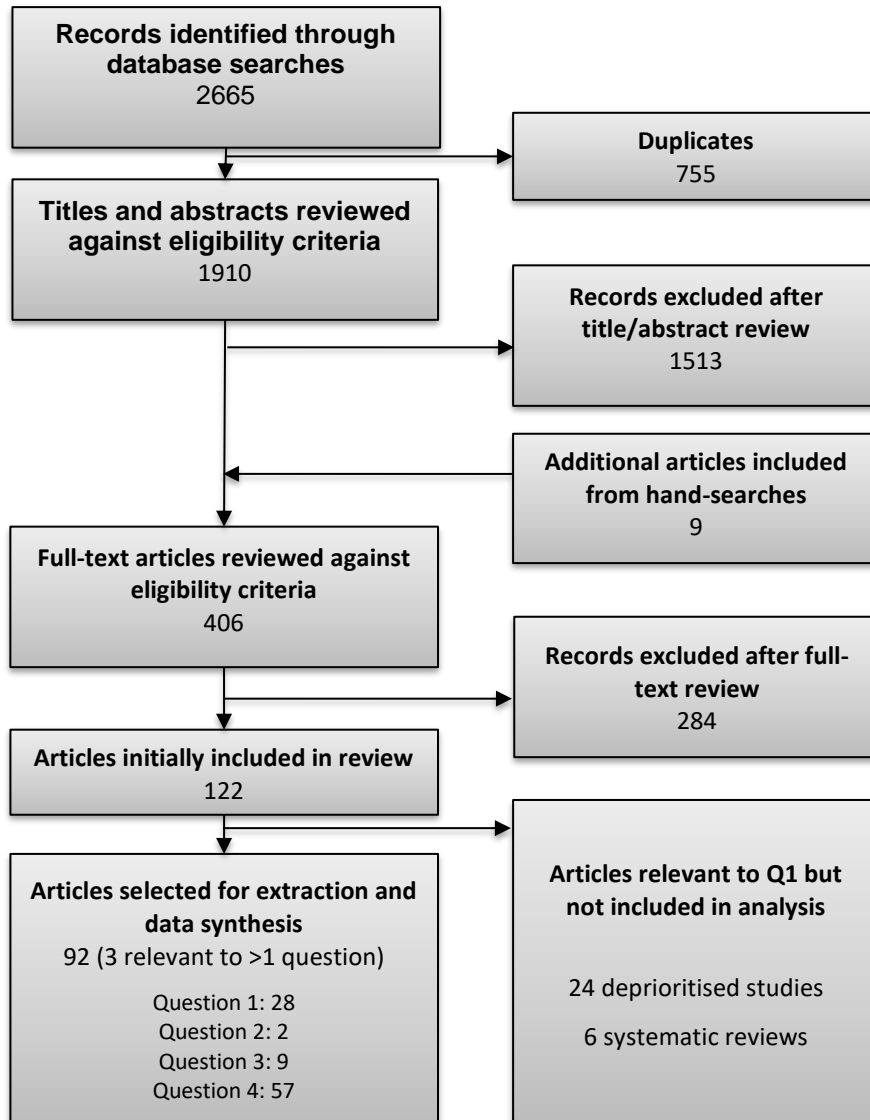
S11	TI "machine learning" OR AB "machine learning"	4,822
S10	TI automation OR AB automation	2,915
S9	TI ( (automated N2 (tool* or technique* or identification or detection or test* or screening)) ) OR AB ( (automated N2 (tool* or technique* or identification or detection or test* or screening)) )	2,603
S8	TI "artificial intelligence" OR AB "artificial intelligence"	2,606
S7	(MM "Automation+")	9,478
S6	(MM "Artificial Intelligence+")	11,742
S5	S1 OR S2 OR S3 OR S4	904,632
S4	TI imaging OR AB imaging	168,567
S3	TI screening OR AB screening	141,479
S2	TI "diagnostic test*" OR AB "diagnostic test*"	11,898
S1	(MM "Diagnosis+")	725,110

## Appendix 2 — Included and excluded studies

### PRISMA flowchart

Figure 6 summarises the volume of publications included and excluded at each stage of the review. Ninety two publications were judged to be relevant to one or more review questions and were considered for extraction. Publications that were included or excluded after the review of full-text articles are detailed below.

**Figure 6 Summary of publications included and excluded at each stage of the review**



## Publications included after review of full-text articles

Studies were prioritised for extraction and data synthesis. It was planned *a priori* that the following approach would be taken to prioritise studies for extraction:

1. For all questions, UK-based studies were prioritised. In the absence or minimal volume of such studies, those from comparable countries were prioritised next.
2. Systematic reviews and meta-analyses, addressing directly the review question, were considered the highest quality of evidence. Following this, study designs were prioritised for questions 1 and 2 in the following (descending) order: RCTs, prospective cohort studies, retrospective cohort studies.
3. In addition, studies reporting on the accuracy of ARIASs (question 1) were prioritised:
  - a. If they evaluated commercially available ARIASs
  - b. If they evaluated ARIASs that are CE-marked and/or FDA approved
  - c. If they evaluated the latest version of the software.

All included studies are listed in Table 21 by relevance to each question.

**Table 21 Publications included after review of full-text articles and the questions they address**

Study	Q1: Accuracy	Q2: Effectiveness	Q3: Cost- effectiveness	Q4: Social and ethical impact: Primary studies	Q4: Social and ethical impact: Review and opinion papers
Abramoff 2010 (58)	No	No	No	No	Yes
Abramoff 2018 (20)	Yes	No	No	No	No
Abramoff 2020 (59)	No	No	No	No	Yes
Alexander 2020 (60)	No	No	No	Yes	No
Anonymous 2018 (61)	No	No	No	No	Yes
Anonymous 2019 (62)	No	No	No	No	Yes
Balyen 2019 (63)	No	No	No	No	Yes
Berens 2020 (64)	No	No	No	No	Yes
Bhaskaranand 2016 (CA) (47)	No	No	Yes	No	No
Bhaskaranand 2019 (65)	Yes	No	No	No	No
Bouhaimed 2008 (40)	Yes	No	No	No	No
Bourla 2018 (66)	No	No	No	Yes	No
Broome 2020 (67)	No	No	No	No	Yes
Carter 2019 (56)	No	No	No	No	Yes
Channa 2020 (68)	No	No	No	No	Yes
Chee 2018 (69)	No	No	No	No	Yes
Coppola 2020 (70)	No	No	No	Yes	No
Egan 2016 (71)	No	No	Yes	No	No
ESR 2019 (72)	No	No	No	Yes	No
Fatehi 2020 (55)	No	No	No	No	Yes

Study	Q1: Accuracy	Q2: Effectiveness	Q3: Cost- effectiveness	Q4: Social and ethical impact: Primary studies	Q4: Social and ethical impact: Review and opinion papers
<b>Figueiredo 2015 (73)</b>	Yes	No	No	No	No
<b>Fleming 2010a (38)</b>	Yes	No	No	No	No
<b>Fleming 2010b (74)</b>	Yes	No	No	No	No
<b>Francolini 2020 (75)</b>	No	No	No	No	Yes
<b>Ginestra 2019 (54)</b>	No	No	No	Yes	No
<b>Goatman 2011 (5)</b>	Yes	No	No	No	No
<b>Gonzalez-Gonzalo 2020 (35)</b>	Yes	No	No	No	No
<b>Gorges 2020 (CA) (76)</b>	No	No	No	Yes	No
<b>Graham 2019 (77)</b>	No	No	No	No	Yes
<b>Halamka 2019 (78)</b>	No	No	No	No	Yes
<b>Hamilton 2002 (50)</b>	No	No	No	Yes	No
<b>Hansen 2004 (39)</b>	Yes	No	No	No	No
<b>Heydon 2020 (31)</b>	Yes	No	No	No	No
<b>Jheng 2020 (79)</b>	No	No	No	No	Yes
<b>Jonmarker 2019 (51)</b>	No	No	No	Yes	No
<b>Jungmann 2020 (80)</b>	No	No	No	Yes	No
<b>Kapoor 2019a (81)</b>	No	No	No	No	Yes
<b>Kapoor 2019b (82)</b>	No	No	No	No	Yes
<b>Keel 2018 (36)</b>	No	Yes	No	Yes	No
<b>Keskinbora 2020 (83)</b>	No	No	No	No	Yes
<b>Koh 2019 (CA) (84)</b>	No	No	No	Yes	No
<b>Krause 2018 (34)</b>	Yes	No	No	No	No
<b>Larson 2020 (57)</b>	No	No	No	No	Yes
<b>Li 2018 (85)</b>	Yes	No	No	No	No
<b>Liew 2014 (86)</b>	No	No	Yes	No	No
<b>Liew 2019 (87)</b>	No	No	No	No	Yes
<b>Lim 2019 (CA) (29)&amp; FDA (30)</b>	Yes	No	No	No	No
<b>Liu 2020 (33)</b>	Yes	Yes	No	No	No
<b>Meyer 2020 (88)</b>	No	No	No	Yes	No
<b>Nadarzynski 2019 (89)</b>	No	No	No	Yes	No
<b>O'Connor 2019 (90)</b>	No	No	No	No	Yes
<b>Oliveira 2011 (43)</b>	Yes	No	No	No	No
<b>Olsen 2013 (44)</b>	No	No	Yes	No	No
<b>Olvera-Barrios 2020 (32)</b>	Yes	No	No	No	No
<b>Ooi 2019 (91)</b>	No	No	No	Yes	No
<b>Ooms 2019 (CA) (52)</b>	No	No	No	Yes	No
<b>Padhy 2019 (92)</b>	No	No	No	No	Yes
<b>Palmisciano 2020 (93)</b>	No	No	No	Yes	No
<b>Patel 2007 (94)</b>	No	No	No	No	Yes

Study	Q1: Accuracy	Q2: Effectiveness	Q3: Cost- effectiveness	Q4: Social and ethical impact: Primary studies	Q4: Social and ethical impact: Review and opinion papers
Paul 2006 (53)	No	No	No	Yes	No
Philip 2007 (37)	Yes	No	No	No	No
Philip 2017 (17)	Yes	No	No	No	No
Prescott 2014 (45)	No	No	Yes	No	No
Rahimy 2018 (95)	No	No	No	No	Yes
Rajalakshmi 2020 (96)	No	No	No	No	Yes
Raumviboonsuk 2019 (26)	Yes	No	No	No	No
Ribeiro 2011 (18)	Yes	No	No	No	No
Ruamviboonsuk 2020 (97)	No	No	No	No	Yes
Scotland 2007 (48)	No	No	Yes	No	No
Scotland 2010 (49)	No	No	Yes	No	No
Scott 2019 (98)	No	No	No	No	Yes
Shaban-Nejad 2018 (99)	No	No	No	No	Yes
Shah 2020 (100)	Yes	No	No	No	No
Sivaprasad 2020 (101)	No	No	No	No	Yes
Son 2020 (24)	Yes	No	No	No	No
Sosale 2019 (102)	No	No	No	No	Yes
Soto-Pedre 2015 (25)	Yes	No	No	No	No
Stolte 2020 (103)	No	No	No	No	Yes
Ting 2017 (23)	Yes	No	No	No	No
Ting 2019a (104)	No	No	No	No	Yes
Ting 2019b (105)	No	No	No	No	Yes
Ting 2020 (106)	No	No	No	No	Yes
Tufail 2016 (1) & 2017 (46)	Yes	No	Yes	No	No
van der Heijden 2018 (21)	Yes	No	No	No	No
Verbraak 2019 (22)	Yes	No	No	No	No
Vollmer 2020 (107)	No	No	No	No	Yes
Wang 2012 (108)	No	No	No	No	Yes
Waymel 2019 (109)	No	No	No	Yes	No
Wong 2019 (110)	No	No	No	No	Yes
Wong 2020 (111)	No	No	No	No	Yes
Xiang 2020 (112)	No	No	No	Yes	No
Yip 2020 (113)	Yes	No	No	No	No

**Publications excluded after review of full-text articles**

Of the 397 publications included after the review of titles and abstracts, 278 were ultimately judged not to be relevant to this review. These publications, along with reasons for exclusion, are listed in Table 22.

**Table 22 Publications excluded after review of full text articles**

Reference	Reason for exclusion
Abbas Q, Fondon I, Sarmiento A, Jimenez S, Alemany P. Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. <i>Med Biol Eng Comput.</i> 2017;55(11):1959-74.	algorithm in development*
Abramoff MD, Niemeijer M, Suttorp-Schulten MS, Viergever MA, Russell SR, van Ginneken B. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. <i>Diabetes Care.</i> 2008;31(2):193-8.	Review**
Abramoff MD, Niemeijer M, Russell SR. Automated detection of diabetic retinopathy: barriers to translation into clinical practice. <i>Expert Rev Med Devices.</i> 2010;7(2):287-96.	Review
Abramoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. <i>JAMA Ophthalmology.</i> 2013;131(3):351-7.	Evaluates older non-DL version of the IDx-DR system
Abramoff MD, Leng T, Ting DSW, Rhee K, Horton MB, Brady CJ, et al. Automated and Computer-Assisted Detection, Classification, and Diagnosis of Diabetic Retinopathy. <i>Telemed J E Health.</i> 2020;26(4):544-50.	position paper
Abramoff MD, Tobey D, Char DS. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. <i>Am J Ophthalmol.</i> 2020;214:134-42.	Review
Acharya UR, Lim CM, Ng EY, Chee C, Tamura T. Computer-based detection of diabetes retinopathy stages using digital fundus images. <i>Proc Inst Mech Eng [H].</i> 2009;223(5):545-53.	algorithm in development
Acharya UR, Ng EY, Tan JH, Sree SV, Ng KH. An integrated index for the identification of diabetic retinopathy stages using texture parameters. <i>Journal of Medical Systems.</i> 2012;36(3):2011-20.	internal validation only
Acharya UR, Mookiah MRK, Koh JEW, Tan JH, Bhandary SV, Rao AK, et al. Automated diabetic macular oedema (DME) grading system using DWT, DCT Features and maculopathy index. <i>Comput Biol Med.</i> 2017;84:59-68.	algorithm in development



Adal KM, Sidibe D, Ali S, Chaum E, Karnowski TP, Meriaudeau F. Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning. <i>Computer Methods and Programs in Biomedicine</i> . 2014;114(1):1-10.	Focus on detection of MA; evaluation at DR level performed on a small subset of the development dataset
Adal KM, van Etten PG, Martinez JP, Rouwen KW, Vermeer KA, van Vliet LJ. An Automated System for the Detection and Classification of Retinal Changes Due to Red Lesions in Longitudinal Fundus Images. <i>IEEE Trans Biomed Eng</i> . 2018;65(6):1382-90.	retinal change detection in images taken at different time points
Akbar S, Akram MU, Sharif M, Tariq A, Yasin UU. Decision Support System for Detection of Papilledema through Fundus Retinal Images. <i>Journal of Medical Systems</i> . 2017;41(4):66.	papilledema
Akram MU, Tariq A, Anjum MA, Javed MY. Automated detection of exudates in colored retinal images for diagnosis of diabetic retinopathy. <i>Appl Opt</i> . 2012;51(20):4858-66.	internal validation only
Akram UM, Khan SA. Automated detection of dark and bright lesions in retinal images for early detection of diabetic retinopathy. <i>Journal of Medical Systems</i> . 2012;36(5):3151-62.	algorithm in development
Akram MU, Tariq A, Khan SA, Javed MY. Automated detection of exudates and macula for grading of diabetic macular edema. <i>Comput Methods Programs Biomed</i> . 2014;114(2):141-52.	algorithm in development
Akyol K, Sen B, Bayir S. Automatic Detection of Optic Disc in Retinal Image by Using Keypoint Detection, Texture Analysis, and Visual Dictionary Techniques. <i>Comput</i> . 2016;2016:6814791.	algorithm in development
Alam M, Le D, Lim JI, Chan RVP, Yao X. Supervised Machine Learning Based Multi-Task Artificial Intelligence Classification of Retinopathies. <i>Journal of Clinical Medicine</i> . 2019;8(6):18.	OCT***
Al-Jarrah MA, Shatnawi H. Non-proliferative diabetic retinopathy symptoms detection and classification using neural network. <i>J Med Eng Technol</i> . 2017;41(6):498-505.	internal validation only
Alqudah AM. AOCT-NET: a convolutional network automated classification of multiclass retinal diseases using spectral-domain optical coherence tomography images. <i>Med Biol Eng Comput</i> . 2020;58(1):41-53.	OCT
Al-Rawi M, Karajeh H. Genetic algorithm matched filter optimization for automated detection of blood vessels from digital retinal images. <i>Computer Methods and Programs in Biomedicine</i> . 2007;87(3):248-53.	algorithm in development

Amin J, Sharif M, Rehman A, Raza M, Mufti MR. Diabetic retinopathy detection and classification using hybrid feature set. <i>Microsc Res Tech.</i> 2018;81(9):990-6.	algorithm in development
Andreasen M, Kjellberg J. A health economic analysis: Retinalyze™ . The Danish Committee for Health Education. December 2008	Non-UK HE study
Anitha J, Vijila CK, Selvakumar AI, Indumathy A, Jude Hemanth D. Automated multi-level pathology identification techniques for abnormal retinal images using artificial neural networks. <i>British Journal of Ophthalmology.</i> 2012;96(2):220-3.	algorithm in development
Anonymous. All eyes are on AI. <i>Nat.</i> 2018;2(3):139.	Review
Anonymous. Ascent of machine learning in medicine. <i>Nat Mater.</i> 2019;18(5):407.	Review
Arcadu F, Benmansour F, Maunz A, Michon J, Haskova Z, McClintock D, et al. Deep Learning Predicts OCT Measures of Diabetic Macular Thickening From Color Fundus Photographs. <i>Invest Ophthalmol Vis Sci.</i> 2019;60(4):852-7.	internal validation only
Armstrong S. The computer will assess you now. <i>Bmj.</i> 2016;355:i5680.	Review
Arsalan M, Owais M, Mahmood T, Cho SW, Park KR. Aiding the Diagnosis of Diabetic and Hypertensive Retinopathy Using Artificial Intelligence-Based Semantic Segmentation. <i>Journal of Clinical Medicine.</i> 2019;8(9):11.	algorithm in development
Arunkumar R, Balakrishnan N. Medical image classification for disease diagnosis by DBN methods. <i>Pakistan Journal of Biotechnology.</i> 2018;15(1):107-10.	theoretical paper
Ayhan MS, Kuhlewein L, Aliyeva G, Inhoffen W, Ziemssen F, Berens P. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. <i>Med Image Anal.</i> 2020;64:101724.	not accuracy/eff study; focus on uncertainty
Badar M, Haris M, Fatima A. Application of deep learning for retinal image analysis: A review. <i>Computer Science Review</i> Volume 35, February 2020	Review
BahadarKhan K, Khaliq AA, Shahid M. A morphological hessian based approach for retinal blood vessels segmentation and denoising using region based otsu thresholding. <i>PLoS ONE.</i> 2016;11 (7) (no pagination)(e0158996).	algorithm in development
Bala MP, Vijayachitra S. Early detection and classification of microaneurysms in retinal fundus images using sequential learning methods. <i>International Journal of Biomedical Engineering and Technology.</i> 2014;15(2):128-43.	internal validation only
Balyen L, Peto T. Promising Artificial Intelligence-Machine Learning-Deep Learning Algorithms in Ophthalmology. <i>Asia Pac J Ophthalmol (Phila).</i> 2019;8(3):264-72.	Review
Banerjee S, Kayal D. Detection of hard exudates using mean shift and normalized cut method. <i>Biocybernetics and Biomedical Engineering.</i> 2016;36(4):679-85.	internal validation only
Banuselvasaraswathy B, Arul Murugan C, Karthigaikumar P. Automatic retinal lesions detection of diabetic retinopathy using curvelet based enhancement. <i>Indian Journal of Public Health Research and Development.</i> 2019;10(2):1029-35.	no info on evaluation methods (looks like an internal evaluation only)

Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. <i>Curr Diab Rep.</i> 2019;19(9):72.	overview paper
Berens P, Waldstein SM, Ayhan MS, Kummerle L, Agostini H, Stahl A, et al. [Potential of methods of artificial intelligence for quality assurance]. <i>Ophthalmologe.</i> 2020;117(4):320-5.	Review
Boucher MC, Qian J, Brent MH, Wong DT, Sheidow T, Duval R, et al. Evidence-based Canadian guidelines for tele-retina screening for diabetic retinopathy: recommendations from the Canadian Retina Research Network (CR2N) Tele-Retina Steering Committee. <i>Can J Ophthalmol.</i> 2020;55(1S1):14-24.	guideline
Broome DT, Hilton CB, Mehta N. Policy Implications of Artificial Intelligence and Machine Learning in Diabetes Management. <i>Curr Diab Rep.</i> 2020;20(2):5.	Review
Buchanan CR, Trucco E. Contextual detection of diabetic pathology in wide-field retinal angiograms. <i>Conf Proc IEEE Eng Med Biol Soc.</i> 2008;2008:5437-40.	internal validation only
Caixinha M, Nunes S. Machine Learning Techniques in Clinical Vision Sciences. <i>Curr Eye Res.</i> 2017;42(1):1-15.	Review
Cao W, Czarnek N, Shan J, Li L. Microaneurysm Detection Using Principal Component Analysis and Machine Learning Methods. <i>IEEE Trans Nanobioscience.</i> 2018;17(3):191-8.	Focus on microaneurysm detection, no DR level evaluation
Cao P, Ren F, Wan C, Yang J, Zaiane O. Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis. <i>Comput Med Imaging Graph.</i> 2018;69:112-24.	internal validation only
Cao K, Xu J, Zhao WQ. Artificial intelligence on diabetic retinopathy diagnosis: an automatic classification method based on grey level co-occurrence matrix and naive Bayesian model. <i>International Journal of Ophthalmology.</i> 2019;12(7):1158-62.	internal validation only
Carter S, Win K, Wang L, Rogers W, Richards B, Houssami N. Ethical, legal and social implications of artificial intelligence systems for screening and diagnosis. <i>BMJ Evidence-Based Medicine.</i> 2019;24 (Supplement 2):A37-A8.	Review
Channa R, Wolf R, Abramoff MD. Autonomous Artificial Intelligence in Diabetic Retinopathy: From Algorithm to Clinical Application. <i>J Diabetes Sci Technol.</i> 2020:1932296820909900.	Review
Chee RI, Darwish D, Fernandez-Vega A, Patel S, Jonas K, Ostmo S, et al. Retinal Telemedicine. <i>Current Ophthalmology Reports.</i> 2018;6(1):36-45.	Review
Chen PC, Liu Y, Peng L. How to develop machine learning models for healthcare. <i>Nat Mater.</i> 2019;18(5):410-4.	Review

Cheung CY, Tang F, Ting DSW, Tan GSW, Wong TY. Artificial Intelligence in Diabetic Eye Disease Screening. <i>Asia Pac J Ophthalmol (Phila)</i> . 2019;24:24.	overview paper
Chi CS. Deep learning based automated detection and grading of diabetic retinopathy for screening programme. <a href="http://wwwwhoint/trialsearch/Trial2.aspx?TrialID=ChiCTR-SON-17010692">http://wwwwhoint/trialsearch/Trial2.aspx?TrialID=ChiCTR-SON-17010692</a> . 2017.	Trial registration
Choi JY, Yoo TK, Seo JG, Kwak J, Um TT, Rim TH. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. <i>PLoS ONE</i> . 2017;12 (11) (no pagination)(e0187336).	internal validation only
Chowriappa P, Dua S, Rajendra Acharya U, Muthu Rama Krishnan M. Ensemble selection for feature-based classification of diabetic maculopathy images. <i>Comput Biol Med</i> . 2013;43(12):2156-62.	internal validation only
Chudzik P, Al-Diri B, Caliva F, Ometto G, Hunter A. Exudates Segmentation using Fully Convolutional Neural Network and Auxiliary Codebook. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2018;2018:770-3.	internal validation only
Chudzik P, Majumdar S, Caliva F, Al-Diri B, Hunter A. Microaneurysm detection using fully convolutional neural networks. <i>Comput Methods Programs Biomed</i> . 2018;158:185-92.	internal validation only
Coyner AS, Campbell JP, Chiang MF. Demystifying the Jargon: The Bridge between Ophthalmology and Artificial Intelligence. <i>Ophthalmol Retina</i> . 2019;3(4):291-3.	Review
Cuadros J. The Real-World Impact of Artificial Intelligence on Diabetic Retinopathy Screening in Primary Care. <i>J Diabetes Sci Technol</i> . 2020:1932296820914287.	comment
Dai L, Fang R, Li H, Hou X, Sheng B, Wu Q, et al. Clinical Report Guided Retinal Microaneurysm Detection With Multi-Sieving Deep Learning. <i>IEEE Trans Med Imaging</i> . 2018;37(5):1149-61.	internal validation only
De Fauw J, Keane P, Tomasev N, Visentin D, van den Driessche G, Johnson M, et al. Automated analysis of retinal imaging using machine learning techniques for computer vision. <i>F1000Res</i> . 2016;5:1573.	study protocol
De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. <i>Nat Med</i> . 2018;24(9):1342-50.	OCT
Dharmawan DA, Boon Poh N. A new two-dimensional matched filter based on the modified Chebyshev type I function for retinal vessels detection. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2017;2017:369-72.	internal validation only
Di X, Shuang Y, Vignarajan J, Dong A, Mei-Ling T-K, Kanagasingam Y. Retinal hemorrhage detection by rule-based and machine learning approach. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2017;2017:660-3.	algorithm in development

Dismuke C. Progress in examining cost-effectiveness of AI in diabetic retinopathy screening. <i>The Lancet Digital Health</i> . 2020;2(5):e212-e3.	comment
Dupas B, Walter T, Erginay A, Ordonez R, Deb-Joardar N, Gain P, et al. Evaluation of automated fundus photograph analysis algorithms for detecting microaneurysms, haemorrhages and exudates, and of a computer-assisted diagnostic system for grading diabetic retinopathy. <i>Diabetes Metab</i> . 2010;36(3):213-20.	internal validation only
Eftekhari N, Pourreza HR, Masoudi M, Ghiasi-Shirazi K, Saeedi E. Microaneurysm detection in fundus images using a two-step convolutional neural network. <i>Biomedical Engineering Online</i> . 2019;18(1):67.	internal validation only
Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. <i>The Lancet Digital Health</i> . 2019;1(5):e232-e42.	internal validation only
FengLi Y, Jing S, Annan L, Jun C, Cheng W, Jiang L. Image quality classification for DR screening using deep learning. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2017;2017:664-7.	internal validation only
Feng S, Zhuo Z, Pan D, Tian Q. CcNet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features. <i>Neurocomputing</i> . 2020;392:268-76.	internal validation only
Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated microaneurysm detection using local contrast normalization and local vessel detection. <i>IEEE Trans Med Imaging</i> . 2006;25(9):1223-32.	algorithm in development
Fleming AD, Philip S, Goatman KA, Williams GJ, Olson JA, Sharp PF. Automated detection of exudates for diabetic retinopathy screening. <i>Phys Med Biol</i> . 2007;52(24):7385-96.	algorithm in development
Fleming AD, Philip S, Goatman KA, Prescott GJ, Sharp PF, Olson JA. The evidence for automated grading in diabetic retinopathy screening. <i>Curr Diabetes Rev</i> . 2011;7(4):246-52.	Review
Francolini G, Desideri I, Stocchi G, Salvestrini V, Ciccone LP, Garlatti P, et al. Artificial Intelligence in radiotherapy: state of the art and future directions. <i>Med Oncol</i> . 2020;37(6):50.	Review
Franklin SW, Rajan SE. An automated retinal imaging method for the early diagnosis of diabetic retinopathy. <i>Technol Health Care</i> . 2013;21(6):557-69.	internal validation only
Ganesan K, Martis RJ, Acharya UR, Chua CK, Min LC, Ng EY, et al. Computer-aided diabetic retinopathy detection using trace transforms on digital fundus images. <i>Med Biol Eng Comput</i> . 2014;52(8):663-72.	internal validation only

Garcia M, Sanchez CI, Lopez MI, Diez A, Hornero R. Automatic detection of red lesions in retinal images using a multilayer perceptron neural network. Conf Proc IEEE Eng Med Biol Soc. 2008;2008:5425-8.	internal validation only
Garcia M, Sanchez CI, Poza J, Lopez MI, Hornero R. Detection of hard exudates in retinal images using a radial basis function classifier. Ann Biomed Eng. 2009;37(7):1448-63.	internal validation only
Garcia M, Lopez MI, Alvarez D, Hornero R. Assessment of four neural network based classifiers to automatically detect red lesions in retinal images. Med Eng Phys. 2010;32(10):1085-93.	internal validation only
Garside K, Henderson R, Makarenko I, Masoller C. Topological data analysis of high resolution diabetic retinopathy images. PLoS ONE. 2019;14(5):e0217413.	algorithm in development
Gonzalez-Gonzalo C, Liefers B, Vaidyanathan A, Van Zeeland H, Klaver CCW, Sanchez CI. Opening the "black box" of deep learning in automated screening of eye diseases. Investigative Ophthalmology and Visual Science Conference. 2019;60(9).	Review
Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. Curr Psychiatry Rep. 2019;21(11):116.	Review
Guo Y, Budak U, Sengur A. A novel retinal vessel detection approach based on multiple deep convolution neural networks. Comput Methods Programs Biomed. 2018;167:43-8.	algorithm in development
Halamka J, Cerrato P. An FP's guide to AI-enabled clinical decision support. Journal of Family Practice. 2019;68(9):486;8;90;92.	Review
Hansen M B, Tang H L, Wang S, Turk L A, Piermarocchi R, Speckauskas M, Hense HW, Leung I, Peto T. Automated detection of Diabetic Retinopathy in Three European Populations. J Clin Exp Ophthalmol 2016, 7:4; DOI: 10.4172/2155-9570.1000582.	Participants criteria not met (non-DM patients included)
Harangi B, Hajdu A. Detection of exudates in fundus images using a Markovian segmentation model. Conf Proc IEEE Eng Med Biol Soc. 2014;2014:130-3.	algorithm in development
Harangi B, Hajdu A. Automatic exudate detection by fusing multiple active contours and regionwise classification. Comput Biol Med. 2014;54:156-71.	algorithm in development
Hassan B, Hassan T, Li B, Ahmed R, Hassan O. Deep Ensemble Learning Based Objective Grading of Macular Edema by Extracting Clinically Significant Findings from Fused Retinal Imaging Modalities. Sensors (Basel). 2019;19(13):05.	algorithm in development
Hatanaka Y, Nakagawa T, Hayashi Y, Hara T, Fujita H. Improvement of automated detection method of hemorrhages in fundus images. Conf Proc IEEE Eng Med Biol Soc. 2008;2008:5429-32.	algorithm in development

Helmchen LA, Lehmann HP, Abramoff MD. Automated detection of retinal disease. Am J Manag Care. 2014;20(11 Spec No. 17):eSP48-52.	overview paper
Hemanth DJ, Anitha J, Son LH, Mittal M. Diabetic Retinopathy Diagnosis from Retinal Images Using Modified Hopfield Neural Network. Journal of Medical Systems. 2018;42(12):247.	internal validation only
Hsieh YT, Chuang LM, Jiang YD, Chang TJ, Yang CM, Yang CH, et al. Application of deep learning image assessment software VeriSee TM for diabetic retinopathy screening. J Formos Med Assoc. 2020;16:16.	internal validation only
Huang H, Ma H, Qian W. Automatic Parallel Detection of Neovascularization from Retinal Images Using Ensemble of Extreme Learning Machine<sup>.</sup>. Conf Proc IEEE Eng Med Biol Soc. 2019;2019:4712-6.	internal validation only
Ibrahim S, Chowriappa P, Dua S, Acharya UR, Noronha K, Bhandary S, et al. Classification of diabetes maculopathy images using data-adaptive neuro-fuzzy inference classifier. Med Biol Eng Comput. 2015;53(12):1345-60.	internal validation only
Imani E, Pourreza HR, Banaee T. Fully automated diabetic retinopathy screening using morphological component analysis. Comput Med Imaging Graph. 2015;43:78-88.	internal validation only
Jaafar HF, Nandi AK, Al-Nuaimy W. Automated detection of red lesions from digital colour fundus photographs. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:6232-5.	algorithm in development
Jaafar HF, Nandi AK, Al-Nuaimy W. Decision support system for the detection and grading of hard exudates from color fundus photographs. J Biomed Opt. 2011;16(11):116001.	algorithm in development
Janaki SD, Geetha K. Enhanced CAE system for detection of exudates and diagnosis of diabetic retinopathy stages in fundus retinal images using soft computing techniques. Polish Journal of Medical Physics and Engineering. 2019;25(2):131-9.	internal validation only
Jaya T, Dheeba J, Singh NA. Detection of Hard Exudates in Colour Fundus Images Using Fuzzy Support Vector Machine-Based Expert System. Journal of Digital Imaging. 2015;28(6):761-8.	algorithm in development
Jeba Derwin D, Tamil Selvi S, Jeba Singh O, Priestly Shan B. A novel automated system of discriminating Microaneurysms in fundus images. Biomedical Signal Processing and Control. 2020;58 (no pagination)(101839).	algorithm in development
Jelinek HF, Cree MJ, Leandro JJ, Soares JV, Cesar RM, Jr., Luckie A. Automated segmentation of retinal blood vessels and identification of proliferative diabetic retinopathy. J Opt Soc Am A Opt Image Sci Vis. 2007;24(5):1448-56.	algorithm in development

Jelinek HF, Rocha A, Carvalho T, Goldenstein S, Wainer J. Machine learning and pattern classification in identification of indigenous retinal pathology. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:5951-4.	algorithm in development
Jheng YC, Chou YB, Kao CL, Yarmishyn AA, Hsu CC, Lin TC, et al. A Novelty Route for Smartphone-based Artificial Intelligence Approach to Ophthalmic Screening. J Chin Med Assoc. 2020;09:09.	Review
Jiang J, Liu X, Zhang K, Long E, Wang L, Li W, et al. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomedical Engineering Online. 2017;16(1):132.	cataracts
Jiang J, Liu X, Liu L, Wang S, Long E, Yang H, et al. Predicting the progression of ophthalmic disease based on slit-lamp images using a deep temporal sequence network. PLoS ONE. 2018;13(7):e0201142.	slit-lamp
Jiang Z, Yu Z, Feng S, Huang Z, Peng Y, Guo J, et al. A super-resolution method-based pipeline for fundus fluorescein angiography imaging. Biomedical Engineering Online. 2018;17(1):125.	fundus fluorescein angiography
Jiayi W, Jingmin X, Lai H, You J, Nanning Z. New hierarchical approach for microaneurysms detection with matched filter and machine learning. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:4322-5.	algorithm in development
John S, Ram K, Sivaprakasam M, Raman R. Assessment of Computer-Assisted Screening Technology for Diabetic Retinopathy Screening in India - Preliminary Results and Recommendations from a Pilot Study. Studies in Health Technology & Informatics. 2016;231:74-81.	algorithm in development
Joonyoung S, Boreom L. Development of automatic retinal vessel segmentation method in fundus images via convolutional neural networks. Conf Proc IEEE Eng Med Biol Soc. 2017;2017:681-4.	algorithm in development
Joshi S, Karule PT. Mathematical morphology for microaneurysm detection in fundus images. European Journal of Ophthalmology. 2019:1120672119843021.	algorithm in development
Kande GB, Subbaiah PV, Savithri TS. Unsupervised fuzzy based vessel segmentation in pathological digital fundus images. Journal of Medical Systems. 2010;34(5):849-58.	algorithm in development
Kandemir M, Hamprecht FA. Computer-aided diagnosis from weak supervision: a benchmarking study. Comput Med Imaging Graph. 2015;42:44-50.	Internal validation only
Kapetanakis VV, Rudnicka AR, Liew G, Owen CG, Lee A, Louw V, et al. A study of whether automated Diabetic Retinopathy Image Assessment could replace manual grading steps in the English National Screening Programme. J Med Screen. 2015;22(3):112-8.	study protocol



Kapoor R, Whigham BT, Al-Aswad LA. Artificial Intelligence and Optical Coherence Tomography Imaging. <i>Asia Pac J Ophthalmol (Phila)</i> . 2019;8(2):187-94.	Review
Kapoor R, Whigham BT, Al-Aswad LA. The Role of Artificial Intelligence in the Diagnosis and Management of Glaucoma. <i>Current Ophthalmology Reports</i> . 2019;7(2):136-42.	Review
Karnowski TP, Aykac D, Giancardo L, Li Y, Nichols T, Tobin KW, Jr., et al. Automatic detection of retina disease: robustness to image quality and localization of anatomy structure. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2011;2011:5959-64.	no external validation
Karperien A, Jelinek HF, Leandro JJ, Soares JV, Cesar RM, Jr., Luckie A. Automated detection of proliferative retinopathy in clinical practice. <i>Clinical Ophthalmology</i> . 2008;2(1):109-22.	algorithm in development
Karthikeyan R, Alli P. Feature Selection and Parameters Optimization of Support Vector Machines Based on Hybrid Glowworm Swarm Optimization for Classification of Diabetic Retinopathy. <i>Journal of Medical Systems</i> . 2018;42(10):195.	algorithm in development
Karthikeyan S, Sanjay Kumar P, Madhusudan RJ, Sundaramoorthy SK, Krishnan Namboori PK. Detection of multi-class retinal diseases using artificial intelligence: An expeditious learning using deep CNN with minimal data. <i>Biomedical and Pharmacology Journal</i> . 2019;12(3):1577-86.	No external validation
Kauppi T, Kamarainen JK, Lensu L, Kalesnykiene V, Sorri I, Uusitalo H, et al. Constructing benchmark databases and protocols for medical image analysis: diabetic retinopathy. <i>Comput</i> . 2013;2013:368514.	not an accuracy study
Kavitha G, Ramakrishnan S. Abnormality detection in retinal images using ant colony optimization and artificial neural networks - biomed 2010. <i>Biomed Sci Instrum</i> . 2010;46:331-6.	algorithm in development
Keane PA. Artificial intelligence: The algorithmic solution to retinal healthcare. <i>Investigative Ophthalmology and Visual Science Conference</i> . 2019;60(9).	Review
Keel S, Wu J, Lee PY, Scheetz J, He M. Visualizing Deep Learning Models for the Detection of Referable Diabetic Retinopathy and Glaucoma. <i>JAMA Ophthalmology</i> . 2019;137(3):288-92.	a single optometrist trying out visualisation system
Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. <i>Cell</i> . 2018;172(5):1122-31.e9.	OCT, age-related maculopathy
Keskinbora K, Guven F. Artificial intelligence and ophthalmology. <i>Turkish Journal of Ophthalmology</i> . 2020;50(1):37-43.	Review
Khojasteh P, Aliahmad B, Kumar DK. Fundus images analysis using deep features for detection of exudates, hemorrhages and microaneurysms. <i>BMC ophthalmol</i> . 2018;18(1):288.	internal validation only

Khojasteh P, Passos Junior LA, Carvalho T, Rezende E, Aliahmad B, Papa JP, et al. Exudate detection in fundus images using deeply-learnable features. <i>Comput Biol Med.</i> 2019;104:62-9.	internal validation only
Khomri B, Christodoulidis A, Djerou L, Babahenini MC, Cheriet F. Particle swarm optimization method for small retinal vessels detection on multiresolution fundus images. <i>J Biomed Opt.</i> 2018;23(5):1-13.	algorithm in development
Koch M. Artificial Intelligence Is Becoming Natural. <i>Cell.</i> 2018;173(3):531-3.	Review
Kose C, Sevik U, Ikibas C, Erdol H. Simple methods for segmentation and measurement of diabetic retinopathy lesions in retinal fundus images. <i>Comput Methods Programs Biomed.</i> 2012;107(2):274-93.	algorithm in development
Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T. Can machines think? Interaction and perspective taking with robots investigated via fMRI. <i>PLoS ONE.</i> 2008;3(7):e2597.	Review
Krishnan MMR, Laude A. An integrated diabetic retinopathy index for the diagnosis of retinopathy using digital fundus image features. <i>Journal of Medical Imaging and Health Informatics.</i> 2013;3(2):306-13.	algorithm in development
Kusakunniran W, Wu Q, Ritthipravat P, Zhang J. Hard exudates segmentation based on learned initial seeds and iterative graph cut. <i>Comput Methods Programs Biomed.</i> 2018;158:173-83.	algorithm in development
Lahmiri S, Boukadoum M. Automated detection of circinate exudates in retina digital images using empirical mode decomposition and the entropy and uniformity of the intrinsic mode functions. <i>Biomed Tech (Berl).</i> 2014;59(4):357-66.	algorithm in development
Lahmiri S. Hybrid deep learning convolutional neural networks and optimal nonlinear support vector machine to detect presence of hemorrhage in retina. <i>Biomedical Signal Processing and Control.</i> 2020;60 (no pagination)(101978).	algorithm in development
Lam BY, Yan H. A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields. <i>IEEE Trans Med Imaging.</i> 2008;27(2):237-46.	internal validation only
Lam C, Yi D, Guo M, Lindsey T. Automated Detection of Diabetic Retinopathy using Deep Learning. <i>AMIA Summits Transl Sci Proc.</i> 2018;2017:147-55.	internal validation only
Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. <i>Radiology.</i> 2020;295(3):675-82.	Theoretical paper
Lemke HU. Machine intelligence and CARS. <i>International Journal of Computer Assisted Radiology and Surgery.</i> 2018;13 (Supplement 1):S125-S6.	Review

Li F, Liu Z, Chen H, Jiang M, Zhang X, Wu Z. Automatic Detection of Diabetic Retinopathy in Retinal Fundus Photographs Based on Deep Learning Algorithm. Transl. 2019;8(6):4.	Algorithm in development
Li Q, Fan S, Chen C. An Intelligent Segmentation and Diagnosis Method for Diabetic Retinopathy Based on Improved U-NET Network. Journal of Medical Systems. 2019;43(9):304.	algorithm in development
Li Z, Guo C, Nie D, Lin D, Yi Z, Chen C, et al. Development and evaluation of a deep learning system for screening retinal hemorrhage based on ultra-wide-field fundus images. Translational Vision Science and Technology. 2020;9 (2) (no pagination)(3).	ultra-wide-field fundus images
Liew CJ, Krishnaswamy P, Cheng LT, Tan CH, Poh AC, Lim TC. Artificial Intelligence and Radiology in Singapore: Championing a New Age of Augmented Imaging for Unsurpassed Patient Care. Ann Acad Med Singapore. 2019;48(1):16-24.	Review
Lin GM, Chen MJ, Yeh CH, Lin YY, Kuo HY, Lin MH, et al. Transforming Retinal Photographs to Entropy Images in Deep Learning to Improve Automated Detection for Diabetic Retinopathy. Journal of ophthalmology. 2018;2018:2159702.	internal validation only
Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine. 2019;9:52-9.	Review
Liu Z, Yao Z, Cao Y, Wu J. Computerized diagnosis of fundus vascular structure based on predictions of diabetic retinopathy grade and risk of macular edema. Journal of Medical Imaging and Health Informatics. 2019;9(5):884-92.	internal validation only
Long S, Huang X, Chen Z, Pardhan S, Zheng D. Automatic Detection of Hard Exudates in Color Retinal Images Using Dynamic Threshold and SVM Classification: Algorithm Development and Evaluation. Biomed Res Int. 2019;2019:3926930.	Focus on detection of exudates, no disease level evaluation
Lyford T, Sheppard J. Diabetic Eye Disease: Advancements in Technology, Detection, and Access to Care. Sr Care Pharm. 2020;35(6):266-72.	Review
Mansour RF. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. Biomedical Engineering Letters. 2018;8(1):41-57.	internal validation only
Mathenge WC. Artificial intelligence for diabetic retinopathy screening in Africa. The Lancet Digital Health. 2019;1(1):e6-e7.	Review
Meza-Kubo V, Morán AL, Carrillo I, Galindo G, García-Canseco E. Assessing the user experience of older adults using a neural network trained to recognize emotions from brain signals. Journal of Biomedical Informatics. 2016;62:202-9.	Review

Mookiah MR, Acharya UR, Chandran V, Martis RJ, Tan JH, Koh JE, et al. Application of higher-order spectra for automated grading of diabetic maculopathy. <i>Med Biol Eng Comput.</i> 2015;53(12):1319-31.	internal validation only
Mumtaz R, Hussain M, Sarwar S, Khan K, Mumtaz S, Mumtaz M. Automatic detection of retinal hemorrhages by exploiting image processing techniques for screening retinal diseases in diabetic patients. <i>International Journal of Diabetes in Developing Countries.</i> 2018;38(1):80-7.	internal validation only
Murugeswari S, Sukanesh R. Examinations on diffuse diabetic macular oedema using neural networks. <i>Journal of Medical Imaging and Health Informatics.</i> 2016;6(8):2019-23.	internal validation only
Murugeswari S, Sukanesh R. Investigations of severity level measurements for diabetic macular oedema using machine learning algorithms. <i>Ir J Med Sci.</i> 2017;186(4):929-38.	internal validation only
Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. <i>The BMJ.</i> 2020;368 (no pagination)(m689).	Review
Naqvi SA, Zafar MF, Haq I. Referral system for hard exudates in eye fundus. <i>Comput Biol Med.</i> 2015;64:217-35.	internal validation only
Naqvi SAG, Zafar HMF, UI Haq I. Automated System for Referral of Cotton-Wool Spots. <i>Curr Diabetes Rev.</i> 2018;14(2):168-74.	algorithm in development
Narasimha-lyer H, Can A, Roysam B, Stewart CV, Tanenbaum HL, Majerovics A, et al. Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy. <i>IEEE Trans Biomed Eng.</i> 2006;53(6):1084-98.	change in DR fundus images
Narasimha-lyer H, Can A, Roysam B, Tanenbaum HL, Majerovics A. Integrated analysis of vascular and nonvascular changes from color retinal fundus image sequences. <i>IEEE Trans Biomed Eng.</i> 2007;54(8):1436-45.	change in DR fundus images
Nayak J, Bhat PS, Acharya R, Lim CM, Kagathi M. Automated identification of diabetic retinopathy stages using digital fundus images. <i>Journal of Medical Systems.</i> 2008;32(2):107-15.	internal validation only
Nayak J, Bhat PS, Acharya UR. Automatic identification of diabetic maculopathy stages using fundus images. <i>J Med Eng Technol.</i> 2009;33(2):119-29.	internal validation only
Nguyen PA, Li YC. Artificial Intelligence in Clinical Implications. <i>Computer Methods and Programs in Biomedicine.</i> 2018;166:A1.	Review

Nidhi MT, Gunaseelan K. Efficient ranking of diabetic retinopathy and glaucoma using echo state neural network and radial basis function (RBF). <i>Journal of Medical Imaging and Health Informatics</i> . 2016;6(3):869-74.	internal validation only
Niemeijer M, van Ginneken B, Russell SR, Suttorp-Schulten MS, Abramoff MD. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. <i>Invest Ophthalmol Vis Sci</i> . 2007;48(5):2260-7.	internal validation only
Noronha K, Acharya UR, Nayak KP, Kamath S, Bhandary SV. Decision support system for diabetic retinopathy using discrete wavelet transform. <i>Proc Inst Mech Eng [H]</i> . 2013;227(3):251-61.	internal validation only
O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? <i>Syst</i> . 2019;8(1):143.	Review
Ogunyemi O, Moran E, Daskivich LP, George S, Teklehaimanot S, Ilapakurthi R, et al. Autonomy versus automation: Perceptions of nonmydriatic camera choice for teleretinal screening in an urban safety net clinic. <i>Telemedicine and e-Health</i> . 2013;19(8):591-6.	different purpose
Ordonez PF, Cepeda CM, Garrido J, Chakravarty S. Classification of images based on small local features: a case applied to microaneurysms in fundus retina images. <i>Journal of Medical Imaging</i> . 2017;4(4):041309.	External evaluation only at the level of lesions
Osareh A, Mirmehdi M, Thomas B, Markham R. Automated identification of diabetic retinal exudates in digital colour images. <i>British Journal of Ophthalmology</i> . 2003;87(10):1220-3.	internal validation only
Osareh A, Shadgar B, Markham R. A computational-intelligence-based approach for detection of exudates in diabetic retinopathy images. <i>IEEE Trans Inf Technol Biomed</i> . 2009;13(4):535-45.	internal validation only
Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: A natural step to the future. <i>Indian J Ophthalmol</i> . 2019;67(7):1004-9.	Review
Pan X, Jin K, Cao J, Liu Z, Wu J, You K, et al. Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. <i>Graefe's Archive for Clinical and Experimental Ophthalmology</i> . 2020;258(4):779-85.	fundus fluorescein angiography
Parekh NU, Bhaskaranand M, Ramachandra C, Bhat S, Solanki K. Explaining an artificial intelligence (AI) system for diabetic retinopathy (DR) screening in primary care. <i>Diabetes Conference: 79th Scientific Sessions of the American Diabetes Association, ADA</i> . 2019;68(Supplement 1).	Review

Patel JL, Goyal RK. Applications of artificial neural networks in medical science. <i>Current Clinical Pharmacology</i> . 2007;2(3):217-26.	Review
Pedrosa M, Silva JM, Matos S, Costa C. SCREEN-DR - Software Architecture for the Diabetic Retinopathy Screening. <i>Studies in Health Technology &amp; Informatics</i> . 2018;247:396-400.	not an accuracy study
Pedrosa M, Silva JM, Silva JF, Matos S, Costa C. SCREEN-DR: Collaborative platform for diabetic retinopathy. <i>International Journal of Medical Informatics</i> . 2018;120:137-46.	not an accuracy study
Pires R, Avila S, Jelinek HF, Wainer J, Valle E, Rocha A. Automatic Diabetic Retinopathy detection using BossaNova representation. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2014;2014:146-9.	internal validation only
Playout C, Duval R, Cheriet F. A Novel Weakly Supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images. <i>IEEE Trans Med Imaging</i> . 2019;38(10):2434-44.	internal validation only
Poly TN, Islam MM, Yang HC, Nguyen PA, Wu CC, Li YJ. Artificial Intelligence in Diabetic Retinopathy: Insights from a Meta-Analysis of Deep Learning. <i>Studies in Health Technology &amp; Informatics</i> . 2019;264:1556-7.	updated systematic review included
Porwal P, Pachade S, Kokare M, Giancardo L, Meriaudeau F. Retinal image analysis for disease screening through local tetra patterns. <i>Comput Biol Med</i> . 2018;102:200-10.	internal validation only
Prakash NB, Hemalakshmi GR, Stella Inba Mary M. Automated grading of diabetic retinopathy stages in fundus images using SVM classifier. <i>Journal of Chemical and Pharmaceutical Research</i> . 2016;8(1):537-41.	internal validation only
Prentasic P, Loncaric S. Weighted ensemble based automatic detection of exudates in fundus photographs. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2014;2014:138-41.	internal validation only
Prentasic P, Loncaric S. Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion. <i>Comput Methods Programs Biomed</i> . 2016;137:281-92.	algorithm in development
Punniyamorthy U, Pushpam I. Remote examination of exudates-impact of macular oedema. <i>Healthc</i> . 2018;5(4):118-23.	internal validation only
Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Detection of lesions in retina photographs based on the wavelet transform. <i>Conf Proc IEEE Eng Med Biol Soc</i> . 2006;2006:2618-21.	algorithm in development
Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Optimal wavelet transform for the detection of microaneurysms in retina photographs. <i>IEEE Trans Med Imaging</i> . 2008;27(9):1230-41.	algorithm in development

Quellec G, Russell SR, Abramoff MD. Optimal filter framework for automated, instantaneous detection of lesions in retinal images. IEEE Trans Med Imaging. 2011;30(2):523-33.	algorithm in development (unclear if independent dataset was used)
Quellec G, Lamard M, Cazuguel G, Bekri L, Daccache W, Roux C, et al. Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs. Invest Ophthalmol Vis Sci. 2011;52(11):8342-8.	internal validation only
Quellec G, Lamard M, Abramoff MD, Decenciere E, Lay B, Erginay A, et al. A multiple-instance learning framework for diabetic retinopathy screening. Med Image Anal. 2012;16(6):1228-40.	Algorithm in development
Quellec G, Lamard M, Erginay A, Chabouis A, Massin P, Cochener B, et al. Automatic detection of referral patients due to retinal pathologies through data mining. Med Image Anal. 2016;29:47-64.	Target condition is any pathology that requires referral to an ophthalmologist; uses data mining (not DL) and combines image analysis with contextual data from patient records
Quellec G, Charriere K, Boudi Y, Cochener B, Lamard M. Deep image mining for diabetic retinopathy screening. Med Image Anal. 2017;39:178-93.	Internal validation only
Quellec G, Lamard M, Conze PH, Massin P, Cochener B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. Med Image Anal. 2020;61:101660.	internal validation only
Rahimy E. Deep learning applications in ophthalmology. Current Opinion in Ophthalmology. 2018;29(3):254-60.	Review
Rajalakshmi R. The impact of artificial intelligence in screening for diabetic retinopathy in India. Eye. 2020;34(3):420-1.	Review
Rajesh IS, Arikerie BM, Reshmi BM. A review on automatic identification of fovea in retinal fundus images. International Journal of Medical Engineering and Informatics. 2020;12(2):169-79.	review
Raju M, Pagidimarri V, Barreto R, Kadam A, Kasivajjala V, Aswath A. Development of a Deep Learning Algorithm for Automatic Diagnosis of Diabetic Retinopathy. Studies in Health Technology & Informatics. 2017;245:559-63.	internal validation only
Randive SN, Senapati RK, Rahulkar AD. A self-adaptive optimisation for diabetic retinopathy detection with neural classification. International Journal of Nano and Biomaterials. 2019;8(3-4):204-27.	algorithm in development, seems like an internal validation only
Reeves A. ES08.07 System Approach to Screening Management. Journal of Thoracic Oncology. 2019;14 (10 Supplement):S33-S4.	Review

Ren F, Cao P, Li W, Zhao D, Zaiane O. Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. <i>Comput Med Imaging Graph.</i> 2017;55:54-67.	internal validation only
Ren F, Cao P, Zhao D, Wan C. Diabetic macular edema grading in retinal images using vector quantization and semi-supervised learning. <i>Technol Health Care.</i> 2018;26(S1):389-97.	internal validation only
Reza AW, Eswaran C. A decision support system for automatic screening of non-proliferative diabetic retinopathy. <i>Journal of Medical Systems.</i> 2011;35(1):17-24.	internal validation only
Riaz H, Park J, Choi H, Kim H, Kim J. Deep and Densely Connected Networks for Classification of Diabetic Retinopathy. <i>Diagnostics.</i> 2020;10(1):02.	internal validation only
Rocha A, Carvalho T, Jelinek HF, Goldenstein S, Wainer J. Points of interest and visual dictionaries for automatic retinal lesion detection. <i>IEEE Trans Biomed Eng.</i> 2012;59(8):2244-53.	Algorithm in development
Roychowdhury S, Koozekanani DD, Parhi KK. DREAM: diabetic retinopathy analysis using machine learning. <i>IEEE j.</i> 2014;18(5):1717-28.	Non-DL algorithm in development (no real life evaluation)
Roychowdhury S, Koozekanani DD, Parhi KK. Automated detection of neovascularization for proliferative diabetic retinopathy screening. <i>Conf Proc IEEE Eng Med Biol Soc.</i> 2016;2016:1300-3.	internal validation only
Ruamviboonsuk P, Cheung CY, Zhang X, Raman R, Park SJ, Ting DSW. Artificial Intelligence in Ophthalmology: Evolutions in Asia. <i>Asia Pac J Ophthalmol (Phila).</i> 2020;9(2):78-84.	Review
S KS, P A. A Machine Learning Ensemble Classifier for Early Prediction of Diabetic Retinopathy. <i>Journal of Medical Systems.</i> 2017;41(12):201.	internal validation only
Saha SK, Fernando B, Cuadros J, Xiao D, Kanagasingam Y. Automated Quality Assessment of Colour Fundus Images for Diabetic Retinopathy Screening in Telemedicine. <i>Journal of Digital Imaging.</i> 2018;31(6):869-78.	ML to determine image quality as 'accept' or 'reject' during acquisition
Saha SK, Xiao D, Kanagasingam Y. A Novel Method for Correcting Non-uniform/Poor Illumination of Color Fundus Photographs. <i>Journal of Digital Imaging.</i> 2018;31(4):553-61.	Method for adjusting for poor illumination
Sahlsten J, Jaskari J, Kivinen J, Turunen L, Jaanio E, Hietala K, et al. Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. <i>Sci.</i> 2019;9(1):10750.	internal validation only
Saleh E, Blaszczyński J, Moreno A, Valls A, Romero-Aroca P, de la Riva-Fernandez S, et al. Learning ensemble classifiers for diabetic retinopathy assessment. <i>Artificial Intelligence in Medicine.</i> 2018;85:50-63.	internal validation only



Sanchez CI, Garcia M, Mayo A, Lopez MI, Hornero R. Retinal image analysis based on mixture models to detect hard exudates. <i>Med Image Anal.</i> 2009;13(4):650-8.	internal validation only
Sanchez CI, Niemeijer M, Abramoff MD, van Ginneken B. Active learning for an efficient training strategy of computer-aided diagnosis systems: application to diabetic retinopathy screening. <i>Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv.</i> 2010;13(Pt 3):603-10.	internal validation only
Sánchez CI, Niemeijer M, Dumitrescu AV, Suttrop-Schulten MS, Abramoff MD, van Ginneken B. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. <i>Invest Ophthalmol Vis Sci.</i> 2011 Jul 1;52(7):4866-71. doi: 10.1167/iops.10-6633. PMID: 21527381.	Older non-DL version of the IDx-DR system
Sangeetha SN, Uma Maheswari P. An Intelligent Model for Blood Vessel Segmentation in Diagnosing DR Using CNN. <i>Journal of Medical Systems.</i> 2018;42(10):175.	internal validation only
Santhi D, Manimegalai D, Parvathi S, Karkuzhali S. Segmentation and classification of bright lesions to diagnose diabetic retinopathy in retinal images. <i>Biomed Tech (Berl).</i> 2016;61(4):443-53.	internal validation only
Savoy M. IDx-DR for Diabetic Retinopathy Screening. <i>Am Fam Physician.</i> 2020;101(5):307-8.	review
Scanlon PH. Update on Screening for Sight-Threatening Diabetic Retinopathy. <i>Ophthalmic Res.</i> 2019;62(4):218-24.	review
Scott IA, Cook D, Coiera EW, Richards B. Machine learning in clinical practice: prospects and pitfalls. <i>Medical Journal of Australia.</i> 2019;211(5):203-5.	Review
Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. <i>npj digit.</i> 2018;1:53.	Review
Shaban M, Ogur Z, Mahmoud A, Switala A, Shalaby A, Abu Khalifeh H, et al. A convolutional neural network for the screening and staging of diabetic retinopathy. <i>PLoS ONE.</i> 2020;15(6):e0233514.	internal validation only
Sharma P, Nirmala SR, Sarma KK. Classification of retinal images using image processing techniques. <i>Journal of Medical Imaging and Health Informatics.</i> 2013;3(3):341-6.	algorithm in development
Sharma S, Maheshwari S, Shukla A. An intelligible deep convolution neural network based approach for classification of diabetic retinopathy. <i>Bio-Algorithms and Med-Systems.</i> 2018;14 (2) (no pagination)(20180011).	internal validation only
Shuang Y, Di X, Kanagasingam Y. Automatic detection of neovascularization on optic disk region with feature extraction and support vector machine. <i>Conf Proc IEEE Eng Med Biol Soc.</i> 2016;2016:1324-7.	algorithm in development

Shuang Y, Di X, Kanagasingam Y. Exudate detection for diabetic retinopathy with convolutional neural networks. Conf Proc IEEE Eng Med Biol Soc. 2017;2017:1744-7.	algorithm in development
Sim D. Historical perspective of diabetic retinopathy screening in the united kingdom - Where do we go from here? West Indian Medical Journal. 2018;67 (Supplement 1):19.	Review
Singh RK, Gorantla R. DMENet: Diabetic Macular Edema diagnosis using Hierarchical Ensemble of CNNs. PLoS ONE. 2020;15(2):e0220677.	internal validation only
Sinthanayothin C, Boyce JF, Williamson TH, Cook HL, Mensah E, Lal S, et al. Automated detection of diabetic retinopathy on digital fundus images. Diabetic Medicine. 2002;19(2):105-12.	only feature-level accuracy in a small number of images
Sivaprasad S, Raman R, Conroy D, Mohan t, Wittenberg R, Rajalakshmi R, et al. The ORNATE India Project: United Kingdom-India Research Collaboration to tackle visual impairment due to diabetic retinopathy. Eye. 2020;34(7):1279-86.	Review
Sosale AR. Screening for diabetic retinopathy-is the use of artificial intelligence and cost-effective fundus imaging the answer? International Journal of Diabetes in Developing Countries. 2019;39(1).	Review
Srivastava R, Wong DW, Lixin D, Jiang L, Tien Yin W. Red lesion detection in retinal fundus images using Frangi-based filters. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:5663-6.	internal validation only
Stevenson CH, Hong SC, Ogbuehi KC. Development of an artificial intelligence system to classify pathology and clinical features on retinal fundus images. Clin Experiment Ophthalmol. 2019;47(4):484-9.	internal validation only
Stolte S, Fang R. A survey on medical image analysis in diabetic retinopathy. Med Image Anal. 2020;64:101742.	Review
Sumathy B, Poornachandra S. Automated dr and prediction of various related diseases of retinal fundus images. Biomedical Research (India). 2018;2018(Special Issue ArtificialIntelligentTechniquesforBioMedicalSignalProcessingEdition-II):S325-S32.	internal validation only
Tang HL, Goh J, Peto T, Ling BW, Al Turk LI, Hu Y, et al. 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(7):e66730.	Validation in a collection of images obtained from screening and non-screening settings
Thomas SA, Titus G. Design of a portable retinal imaging module with automatic abnormality detection. Biomedical Signal Processing and Control. 2020;60 (no pagination)(101962).	internal validation only

Ting DSW, Carin L, Abramoff MD. Observations and Lessons Learned From the Artificial Intelligence Studies for Diabetic Retinopathy Screening. JAMA Ophthalmology. 2019;13:13.	Review
Ting DSW, Peng L, Varadarajan AV, Keane PA, Burlina PM, Chiang MF, et al. Deep learning in ophthalmology: The technical and clinical considerations. Prog Retin Eye Res. 2019;72:100759.	Review
Ting DSW, Cheung CY, Nguyen Q, Sabanayagam C, Lim G, Lim ZW, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study. npj digit. 2019;2:24.	overlapping cohorts with Ting 2017
Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology. 2019;103(2):167.	review
Ting DSJ, Foo VH, Yang LWY, Sia JT, Ang M, Lin H, et al. Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology. British Journal of Ophthalmology. 2020;12:12.	Review
Tobin KW, Chaum E, Govindasamy VP, Karnowski TP. Detection of anatomic structures in human retinal imagery. IEEE Trans Med Imaging. 2007;26(12):1729-39.	algorithm in development
Tobin KW, Abramoff MD, Chaum E, Giancardo L, Govindasamy V, Karnowski TP, et al. Using a patient image archive to diagnose retinopathy. Conf Proc IEEE Eng Med Biol Soc. 2008;2008:5441-4.	algorithm in development
Torok Z, Peto T, Csoz E, Tukacs E, Molnar AM, Berta A, et al. Combined Methods for Diabetic Retinopathy Screening, Using Retina Photographs and Tear Fluid Proteomics Biomarkers. J Diabetes Res. 2015;2015:623619.	internal validation only
Tsai CL, Madore B, Leotta MJ, Sofka M, Yang G, Majerovics A, et al. Automated retinal image analysis over the internet. IEEE Trans Inf Technol Biomed. 2008;12(4):480-7.	not a screening algorithm
Ullah H, Saba T, Islam N, Abbas N, Rehman A, Mehmood Z, et al. An ensemble classification of exudates in color fundus images using an evolutionary algorithm based optimal features selection. Microsc Res Tech. 2019;82(4):361-72.	algorithm in development
Umadevi KS, Jeyapriya J. Cascaded neural network based automated detection of diabetic retinopathy. Indian Journal of Public Health Research and Development. 2017;8(4):1322-8.	algorithm in development
Valverde C, Garcia M, Hornero R, Lopez-Galvez MI. Automated detection of diabetic retinopathy in retinal images. Indian J Ophthalmol. 2016;64(1):26-32.	Review
van Grinsven MJ, van Ginneken B, Hoyng CB, Theelen T, Sanchez CI. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application	Evaluation at the level of lesions only

to Hemorrhage Detection in Color Fundus Images. IEEE Trans Med Imaging. 2016;35(5):1273-84.	
Venkatesan R, Chandakkar P, Li B, Li HK. Classification of diabetic retinopathy images using multi-class multiple-instance learning based on color correlogram features. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:1462-5.	internal validation only
Verbraak FD, Schmidt-Erfurth U, Grzybowski A, Abramoff M, Schlingemann R. Is automated screening for diabetic retinopathy indeed not yet ready as stated by Grauslund et al.? Acta Ophthalmologica. 2020;98(2):e257-e8.	Review
Vidal-Alaball J, Royo Fibla D, Zapata MA, Marin-Gomez FX, Solans Fernandez O. Artificial Intelligence for the Detection of Diabetic Retinopathy in Primary Care: Protocol for Algorithm Development. JMIR Res Protoc. 2019;8(2):e12539.	protocol for algorithm development
Vijayabaskar J, Rajeswari D, Vaithiyanathan V. To detect diabetic retinopathy in fundus enhanced retina images using effective ROI segmentation and kirch's templates. International Journal of Pharmacy and Technology. 2016;8(4):23240-52.	algorithm in development, no info on validation
Vijayalakshmi R, Selvarajan S. A decision support system for detecting the stages of diabetic retinopathy by using fundus images. Journal of Pure and Applied Microbiology. 2015;9(Special Edition):65-70.	unable to obtain full text
Vollmer S, Mateen BA, Bohner G, Kiraly FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. The BMJ. 2020;368 (no pagination)(l6927).	Review
Wang S, Summers RM. Machine learning and radiology. Med Image Anal. 2012;16(5):933-51.	Review
Wang S, Yin Y, Cao G, Wei B, Zheng Y, Yang G. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. Neurocomputing. 2015;149(PB):708-17.	Focus on blood vessel segmentation
Wang S, Tang HL, Al Turk LI, Hu Y, Sanei S, Saleh GM, et al. Localizing Microaneurysms in Fundus Images Through Singular Spectrum Analysis. IEEE Trans Biomed Eng. 2017;64(5):990-1002.	No evaluation at DR level; only at the level of lesions
Wang R, Chen B, Meng D, Wang L. Weakly Supervised Lesion Detection From Fundus Images. IEEE Trans Med Imaging. 2019;38(6):1501-12.	algorithm in development
Wang H, Yuan G, Zhao X, Peng L, Wang Z, He Y, et al. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. Comput Methods Programs Biomed. 2020;191:105398.	Focus on detection of exudates, no DR level evaluation
Welikala RA, Dehmeshki J, Hoppe A, Tah V, Mann S, Williamson TH, et al. Automated detection of proliferative diabetic retinopathy using a modified line	internal validation only

operator and dual classification. <i>Comput Methods Programs Biomed.</i> 2014;114(3):247-61.	
Welikala RA, Fraz MM, Dehmeshki J, Hoppe A, Tah V, Mann S, et al. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. <i>Comput Med Imaging Graph.</i> 2015;43:64-77.	internal validation only
Wong TY, Sabanayagam C. The War on Diabetic Retinopathy: Where Are We Now? <i>Asia Pac J Ophthalmol (Phila).</i> 2019;8(6):448-56.	Review
Wong TY. Artificial intelligence in ophthalmology: Concepts, progress, challenges and myths synopsis. <i>Clinical and Experimental Ophthalmology.</i> 2019;47 (Supplement 1):15-6.	Review
Wong TY, Sabanayagam C. Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence. <i>Ophthalmologica.</i> 2020;243(1):9-20.	Review
Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. <i>The Lancet Digital Health.</i> 2020;2(5):e240-e9.	Non-UK HE study
Xu L, Luo S. Optimal algorithm for automatic detection of microaneurysms based on receiver operating characteristic curve. <i>J Biomed Opt.</i> 2010;15(6):065004.	algorithm in development
Xu K, Feng D, Mi H. Deep Convolutional Neural Network-Based Early Automated Detection of Diabetic Retinopathy Using Fundus Image. <i>Molecules (Basel).</i> 2017;22(12):23.	internal validation only
Yao L, Zhong Y, Wu J, Zhang G, Chen L, Guan P, et al. Multivariable Logistic Regression And Back Propagation Artificial Neural Network To Predict Diabetic Retinopathy. <i>Diabetes Metab Syndr Obes.</i> 2019;12:1943-51.	not image analysis
Yedururi S, Katabathina VS, Jo NH, Rachamalla M, Prasad S, Marcal L. Machine learning and artificial intelligence in oncologic imaging: Potential barriers and solutions, abdominal imagers' perspective. <i>Cancer Imaging Conference: 19th Meeting and Annual of the International Cancer Imaging Society Italy.</i> 2019;19(Supplement 1).	Review
You Z, Hu X, Shi K. Will artificial intelligence replace ophthalmologist in diabetic retinopathy screening? <i>Biomedical Research (India).</i> 2017;28(15):6920.	Review
Yun WL, Mookiah MRK, Koh JEW. Automated detection of proliferative diabetic retinopathy using brownian motion features. <i>Journal of Medical Imaging and Health Informatics.</i> 2014;4(2):250-4.	internal validation only

Zaki WMDW, Zulkifley MA, Hussain A, Halim WHWA, Mustafa NBA, Ting LS. Diabetic retinopathy assessment: Towards an automated system. Biomedical Signal Processing and Control. 2016;24:72-82.	Review
Zapata MA, Royo-Fibla D, Font O, Vela JI, Marcantonio I, Moya-Sanchez EU, et al. Artificial Intelligence to Identify Retinal Fundus Images, Quality Validation, Laterality Evaluation, Macular Degeneration, and Suspected Glaucoma. Clinical Ophthalmology. 2020;14:419-29.	not retinopathy (different target condition)
Zhang L, Feng S, Duan G, Li Y, Liu G. Detection of Microaneurysms in Fundus Images Based on an Attention Mechanism. Genes (Basel). 2019;10(10):17.	internal validation only
Zheng R, Liu L, Zhang S, Zheng C, Bunyak F, Xu R, et al. Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. Biomedical Optics Express. 2018;9(10):4863-78.	Focus on exudate detection, no DR level evaluation
Zhou W, Wu C, Chen D, Wang Z, Yi Y, Du W. Automated Detection of Red Lesions Using Superpixel Multichannel Multifeature. Comput. 2017;2017:9854825.	internal validation only
Zhou K, Gu Z, Liu W, Luo W, Cheng J, Gao S, et al. Multi-Cell Multi-Task Convolutional Neural Networks for Diabetic Retinopathy Grading. Conf Proc IEEE Eng Med Biol Soc. 2018;2018:2724-7.	internal validation only
Zutis K, Trucco E, Hubschman JP, Reed D, Shah S, van Hemert J. Towards automatic detection of abnormal retinal capillaries in ultra-wide-field-of-view retinal angiographic exams. Conf Proc IEEE Eng Med Biol Soc. 2013;2013:7372-5.	ultra-wide-field-of-view exam; target condition: abnormal retinal capillaries
<p>*Algorithm in development – early development of an algorithm or papers in which it appeared that no external validation has been performed but it was not completely clear from the paper to be classified as ‘internal validation only’; due to limited resources, we were unable to investigate this further</p> <p>**Review – non-systematic review of the literature</p> <p>***OCT – the index test was Optical Coherence Tomography</p> <p>DL – Deep Learning, ML – traditional (non-DL) Machine Learning, DR – diabetic retinopathy</p>	

**Table 23 Studies meeting the inclusion criteria for question 1 but excluded after prioritisation**

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
Abramoff 2016, USA, France (114)	IDx-DR X2.1 (DL)	rDR: moderate NPDR or worse and/or macular	Messidor-2	As per dataset	SE 96.8% (95% CI: 93.3%–98.8%), SP 87.0% (95% CI: 84.2%–89.4%),	Comparison with an older version of IDx-DR (the Iowa system) in Messidor-2

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
		oedema (modified ICDR)			NPV 99.0% (95% CI: 97.8%–99.6%) No cases of severe NPDR, PDR, or DMO were missed. SE was not statistically different from published IDP SE which had a CI of 94.4% to 99.3%, SP was significantly better than the published IDP SP CI of 55.7% to 63.0%.	
Araujo 2020, Portugal (115)	DR GRADUATE (DL-based)	R0 - R4	multiple	as per dataset	quadratic-weighted Cohen's kappa ( $\kappa$ ) between 0.71 and 0.84 was achieved in five different datasets	Unclear if commercially available or licensed; validation in multiple public datasets
Bellemo 2019, Zambia (116)	SELENA	Moderate NPDR or worse, DMO and ungradable images (UK National Health Service)	Prospective cohort study	UNCLEAR! if the reference standard was 1) or 2) 1) Nurses and imaging technicians of non-medical background from Kitwe Central Hospital. Images graded separately for DR&DMO 2) Re-graded in Singapore using ICDRSS	rDR: AUC 0.973 (95% CI 0.969–0.978), SE 92.25% (95% CI 90.10%–94.12%), SP 89.04% (95% CI 87.85%–90.28%) <u>Of the referable eyes:</u> SE 99.42% (99.15–99.68) for vtDR SE 97.19% (96.61–97.77) for DMO Comparable performance in patients stratified by age, sex, and HbA1c	Deprioritised country
Bhaskaranand 2016, USA and elsewhere (117)	EyeArt v2	ICDR	EyePACS	Human graders (no further detail)	SE 90.0% (95% CI: 88.0%-92.0%) SP 63.2% (95% CI: 61.7%-64.6%) AUC 0.879 (95% CI: 0.865-0.893).	Older version of EyeArt (v1.2); external validation in a convenience sample from EyePACS (USA)
Gargeya 2017, USA (118)	DL-based	No DR vs any sign of DR	MESSIDOR 2 and E-Ophtha	as per dataset	AUC 0.94 AND 0.95 for MESSIDOR 2 and E-Ophtha	Unclear if commercially available or licensed;

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
						evaluation in public datasets only
Gulshan 2016, USA and India (119)	Google AI (DL)	Moderate or worse DR or referable DMO (ICDR)	EyePACS-1, Messidor-2	A simple majority decision of US board-certified ophthalmologists, EyePACS-1 (n = 8) and Messidor-2 (n = 7)	<p>All-cause referable predictions in EyePACS-1 (including ungradable images)</p> <p>1) At high SP operating point: SE 90.7% (95%CI 89.2%-92.1%), SP 93.8% (95% CI, 93.2%-94.4%)</p> <p>2) At high SE operating point: SE 96.7% (95% CI, 95.7%-97.5%), SP 84.0% (95% CI, 83.1%85.0%)</p> <p>3) Severe or worse DR only: SE 84.0% (75.3, 90.6), SP 98.8% (98.5, 99.0) on EyePACS-1; SE 87.8% (73.4, 96.0), SP 98.2% (97.4, 98.9) on Messidor-2</p> <p>4) Maculopathy only: SE 90.8% (86.1, 94.3), SP 98.7% (98.4, 99.0) on EyePACS-1; SE 90.4% (81.9, 94.8), SP 98.8% (98.1, 99.3) on Messidor-2</p> <p>5) Ungradable images on EyePACS-1: SE 93.9%, SP 90.9%</p> <p>On EyePACS-1 mean agreement for rDR among ophthalmologists was 77.7% (SD, 16.3%), complete agreement 19.6%; for non-referable images, mean agreement 97.4% (SD, 7.3%), complete agreement 85.6%</p> <p>On Messidor-2 the respective values were 82.4% (SD,16.9%), 37.8%,</p>	Older version of the system; Krause 2018 (included) reports on the performance of a new version compared to this one



Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
					rDR, 96.3% (SD, 9.9%) and 85.1%; The accuracy in mydriatic and non-mydriatic images was similar	
Gulshan 2019, India (27)	Google AI	Moderate or worse DR or referable DMO (ICDR)	Prospective cohort study (compared to human graders)	Adjudication by a panel of 3 retinal specialists	Site 1: Trained grader: SE 75.5%, SP 94.2%, Retinal specialist: SE 89.8%, SP 83.5%, Model: SE 88.9%, SP 92.2% Site 2: Trained grader: SE 84.2%, SP 98.6% Retinal specialist: SE 73.4%, SP 98.7% Model: SE 92.1% (90.1 – 93.8), SP 95.2 (94.2 -96.1) the results were similar for referable DMO; the new model had AUC of 0.986 vs 0.974 for the old one	Deprioritised country
Harangi 2019, Hungary (120)	DL-based	5-class DR, 3-class DMO	ISBI 2018 data	as per dataset	Total accuracy 90.07% for the 5-class DR challenge, and 96.85% for the 3-class DMO, respectively	Unclear if commercially available or licensed; validation in a public dataset
He 2020, China (121)	Airdoc, Beijing, China (DL-based)	ICDR	889 diabetic patients at community hospital	2 ophthalmologists, independently	For DR: SE 90.79% (95% CI 86.4–94.1), SP 98.5% (95% CI 97.8–99.0) and AUC 0.946 (95% CI 0.935–0.956), respectively. For rDR, SE 91.18% (95% CI 86.4–94.7), SP 98.79% (95% CI 98.1–99.3) and AUC 0.950 (95% CI 0.939–0.960),	Deprioritised country
Kanagasingam 2018, Australia (122)	DL-based	DR vs no DR; severity of DR based on ICDR; identification of specific pathologies, e.g. microaneuysms,	real-life study, 386 images from 193 patients in primary care	ophthalmologist	This is a pilot study. Of the 193 patients (93 [48%] female; mean [SD] age, 55 [17] years [range, 18-87 years]), the AI system judged 17 as having diabetic retinopathy of	Unclear if commercially available or licensed; pilot study

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
		exudates; image quality			sufficient severity to require referral. The system correctly identified 2 patients with true disease and misclassified 15 as having disease (false-positives). The resulting specificity was 92% (95% CI, 87%-96%), and the positive predictive value was 12% (95% CI, 8%-18%). Many false-positives were driven by inadequate image quality (e.g. dirty lens) and sheen reflections.	
Leibig 2017, Germany (123)	DL-based	R1 or worse; R2 or worse	Messidor	as per dataset	Depending on network capacity and task/dataset difficulty, >85% sensitivity and 80% specificity as recommended by the NHS when referring 0–20% of the most uncertain decisions for further inspection.	Unclear if commercially available or licensed; validation in a single public dataset
Li 2018, China, Australia, Singapore (85)	EyeGrader (DL-based)	EDESP criteria	NIEHS, SiMES, AusDiab	Certified professional senior graders	SE: 89.76% (NIEHS); 93.94% (SiMES), 94.59% (AusDiab); 92.50% (combined) SP: 97.57% (NIEHS); 98.48% (SiMES), 99.17% (AusDiab); 98.52% (combined); no CIs reported	Retrospective study in deprioritised countries and general populations
Liu 2019b, China (124)	WP-CNN (DL-based)	Referral/No referral	STARE	as per dataset	Accuracy of 90.84%, AUC of 0.951 and F1-score of 0.934	Deprioritised country; unclear if commercially available or licensed
Nazir 2019, Pakistan, Korea (125)	Extreme ML vs DL-based and others	both rDR and stage-wise grading	multiple	as per dataset	Median accuracy 0.993 in STARE testing set; 0.993 in Review-DB testing set	Deprioritised country; unclear if commercially available or licensed; validation in public datasets only

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
Pires 2019, Brazil, USA (126)	DL-based	R0-R4	Messidor-2, DR2	as per dataset	The Neural Network-based classifier yields the best result for testing with DR2 dataset, with an AUC of 96.3% (95% CI: 93.8–98.1%).	Deprioritised country; unclear if commercially available or licensed; validation in public datasets only
Ramachandran 2018, New Zealand and elsewhere (127)	Visiona (Hong Kong)	New Zealand MoH guidelines	Otago, Messidor	A single grader, checked by an ophthalmologist	<u>Otago:</u> SE 84.6% SP 79.7% AUC 0.901 (0.807–0.995) <u>Messidor:</u> SE 96.0% SP 90.0% AUC 0.980 (0.973–0.986)	Unclear if commercially available or licensed; external validation in a single study at high risk of bias
Romero-Aroca 2019, Spain (128)	DL-based	The retinographies were placed into four levels: (1) Level 0 = no DR, (2) Level 1 =mild DR (only microaneurysms), (3) Level 2 = moderate DR (microaneurysms with a minimum of 5 and a maximum of 15 and/or retina hemorrhages inferior to 5), and (4) Level 3 = severe DR or proliferative DR (microaneurysms more than 15 and hemorrhages more than 5 or presence of new	38,339 images randomly selected from the Spain's DESP	Four masked senior retina ophthalmologists.	The results of the DLA to detect any-DR were: CWK= 0.886 – 0.004 (95% confidence interval [CI] 0.879–0.894), S = 0.967%, SP = 0.976%, PPV= 0.836%, and NPV = 0.996%. The error type I = 0.024, and the error type II = 0.004. Likewise, the referable-DR results were: CWK= 0.809 (95% CI 0.798–0.819), S = 0.998, SP = 0.968, PPV = 0.701, NPV = 0.928, error type I = 0.032, and error type II = 0.001.	Unclear if commercially available or licensed

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
		vessels elsewhere).				
Roychowdhury 2014, USA (129)	DREAM (non-DL)	DR vs no DR	MESSIDOR	as per dataset	SE 100%, SP 53.16%, AUC 0.904	Unclear if commercially available or licensed; validation in a single public dataset
Sayres 2019, USA (130)	Google AI	5-point scale (based on ICDR): no DR, mild, moderate, severe and proliferative; Referable DR was defined as moderate or worse DR	EyePACS	Adjudication by 3 fellowship-trained retina specialists	Both forms of assistance increased readers' sensitivity for moderate-or-worse DR: unassisted: mean, 79.4% [95%CI 72.3%-86.5%]; grades only: mean, 87.5% [95% CI, 85.1%-89.9%]; grades plus heatmap: mean, 88.7% [95% CI, 84.9%-92.5%] without a corresponding drop in specificity (unassisted: mean, 96.6% [95% CI, 95.9%-97.4%]; grades only: mean, 96.1% [95% CI, 95.5%- 96.7%]; grades plus heatmap: mean, 95.5% [95% CI, 94.8%-96.1%]).	Different focus: comparative evaluation of 3 diagnostic strategies: 1) unassisted human graders (HG); 2) HG provided with ARIAS's grade; 3) HG provided with ARIAS's grade and a heatmap
Shah 2020b, India, Singapore (131)	DL-based	rDR defined as moderate NPDR or worse	MESSIDOR 1	as per dataset	SE 90.4% and SP 91.0% for any DR; SE 94.7% and SP 97.4% for prompt referral	Unclear if commercially available or licensed; validation in a single public dataset
Turk 2020, China, Saudi Arabia, UK, Kenya (132)	DAPHNE (DL-based)	R2, R3, DMO (EDESCP); ICDR also available	China, Saudi Arabia & Kenya	Trained and certified clinicians	SE 95.51% (93.1%–97.50%); SP 91.11% (85.11%–92.63%); No PDR or MDO missed	Unclear if commercially available or licensed; deprioritised countries; external validation in general populations only
Usher 2004, UK (133)	Non-DL	EURODIAB	N/A	clinical research fellow	At a setting with 94.8% sensitivity and 52.8% specificity, no cases of sight-threatening retinopathy were missed (retinopathy warranting immediate ophthalmology	Unclear if commercially available or licensed; only internal validation

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
					referral or re-examination sooner than 1 year by National Institute for Clinical Excellence criteria). If the system was implemented at 94.8% sensitivity setting over half the images with no retinopathy would be correctly identified, reducing the need for a human grader to examine images in 1/3 of patients.	
Wang 2020c, China (134)	DeepDR (DL-based)	ICDR: No DR vs DR Present, Mild or less NPDR vs moderate or more NPDR, moderate or less NPDR vs Severe NPDR or worse, Severe NPDR or less versus PDR	6788 images from local screening programme	Ophthalmologists	For detecting DR, the device had SE 93.50% and SP 77.08%; For moderate or more NPDR output SE 96.57% and SP 78.09%; and SE of the device's severe NPDR or worse output to detect severe NPDR and worse was 98.46% and specificity was 62.15%; and SE of the device's PDR output to detect PDR was 99.16% and SP was 68.75%. The area under the ROC curve were 0.93, 0.96, 0.97 and 0.97 respectively. To further verify the usefulness of the system in the real world, fundus photographs from community screening were evaluated. A total of 20,000 fundus images were selected, and 7593 photos of poor quality were excluded according to quality standards. The	Unclear if commercially available or licensed; deprioritised country

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Study & country	ARIAS	DR grading	External validation dataset	Reference standard	Accuracy results	Reason for not being prioritised
					accuracy of staging of the fundus photos was 0.9179. The sensitivity, specificity and area under the curve (AUC) were 80.58%, 95.77% and 0.9327, respectively	
Zago 2020, Brazil, France (135)	DL-based (patch-based approach)	R0 vs R1-R3; R0&R1 vs R2&R3 (rDR)	Messidor, Messidor-2, IDRiD, DDR, Kaggle	as per dataset	It reached AUC 0.912 (95%CI 0.897 - 0.928) and SE 94% (95%CI 92.1% - 95.9%) for DR screening (SP fixed at 50%)	Unclear if commercially available or licensed; validation in public datasets only
<p>AUC – Area Under the ROC Curve, CI – Confidence Interval, EDESP – English Diabetic Eye Screening Programme, DM – Diabetes Mellitus, DMO - Diabetic Macular Oedema, DL – Deep Learning, DR – Diabetic Retinopathy, FPRC - Fundus Photograph Reading Centre, ICDR - International Clinical Diabetic Retinopathy, NPDR – Non-Proliferative DR, OCT – Optical Coherence Tomography, PDR – Proliferative DR, rDR – referable DR, SE – Sensitivity, SP – Specificity, SiDRP - Singapore National Diabetic Retinopathy Screening Program, vtDR – vision-threatening DR</p>						

## Appendix 3 — Summary and appraisal of individual studies

The included test accuracy studies are summarised in tables 3 to 10 in the main text. Here we report the extracted data on questions 2 – 4 and the results from the methodological quality assessment of the included clinical effectiveness studies (question 2).

**Table 24 Methodological quality assessment of the included studies reporting on the clinical effectiveness and impact of ARIASs in DESPs using the Downs and Black checklist**

Downs and Black Quality Assessment Checklist	Study	
	Keel 2018 (36)	Liu 2020 (33)
<b>Reporting</b>		
1. Is the hypothesis/aim/objective of the study clearly described?	Yes	Yes
2. Are the main outcomes to be measured clearly described in the Introduction or Methods section?	Yes	Yes
3. Are the characteristics of the patients included in the study clearly described?	No (duration of diabetes, previous diagnosis and other risk factors not provided)	Yes
4. Are the interventions of interest clearly described?	Yes	Yes
5. Are the distributions of principal confounders in each group of subjects to be compared clearly described?	N/a ( within-subjects comparison)	No
6. Are the main findings of the study clearly described?	Yes (but some data missing, e.g. cross-tabulation of ARIAS and manual grading, to allow for comparative accuracy estimates)	Yes
7. Does the study provide estimates of the random variability in the data for the main outcomes?	No	Yes
8. Have all important adverse events that may be a consequence of the intervention been reported?	Yes	Yes
9. Have the characteristics of patients lost to follow-up been described?	No	Yes

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

10. Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?	No	Yes
<b>External validity</b>		
11. Were the subjects asked to participate in the study representative of the entire population from which they were recruited?	Unclear	Unclear (number of patients invited to participate not reported)
12. Were those subjects who were prepared to participate representative of the entire population from which they were recruited?	Unclear	Unclear
13. Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive?	Unclear (but not representative of the UK DESP)	Unclear (but not representative of the UK DESP)
<b>Internal validity – bias</b>		
14. Was an attempt made to blind study subjects to the intervention they have received?	No	Unclear
15. Was an attempt made to blind those measuring the main outcomes of the intervention?	No (unclear if the ophthalmologist who did the reference grading was part of the research team)	Unclear
16. If any of the results of the study were based on “data dredging”, was this made clear?	N/a	N/a
17. In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls?	Yes (within-subjects comparison)	Yes
18. Were the statistical tests used to assess the main outcomes appropriate?	N/a (no statistical tests used)	Yes
19. Was compliance with the intervention/s reliable?	Yes	Yes
20. Were the main outcome measures used accurate (valid and reliable)?	Unclear (accuracy was based on a single ophthalmologist’s grading)	Unclear (accuracy was based on a single [out of 5] retina specialist’s grading)
<b>Internal validity – confounding (selection bias)</b>		
21. Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population?	N/a (within-subjects design)	Yes
22. Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time?	N/a (within-subjects design)	No (historical controls)
23. Were study subjects randomised to intervention groups?	No	No
24. Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable?	N/a (within-subjects design)	N/a
25. Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?	N/a (within-subjects design)	No (patients in the ARIAS cohort received 3



		telephone calls and a letter to encourage them to attend; unclear if the same level of engagement was used in the historical control)
26. Were losses of patients to follow-up taken into account?	No (loss to follow up reported as % but patient characteristics not reported)	Yes
27. Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?	N/a (within-subjects design)	Unclear (sample size calculation not reported)

**Table 25 Question 3: Evidence map of UK-based health economic evaluations of AI in DESPs**

Study	Country	Study type	Objectives	Components of the study	Outcomes	Authors' conclusions
<b>Prescott 2014 &amp; Olsen 2013</b>	England (Birmingham, Liverpool and Oxford)  Scotland (Aberdeen, Dundee, Dunfermline, Edinburgh)	CUA  Markov microsimulation model (20 years time-horizon, 2009/2010 cost year).	Cost-effectiveness analysis of methods of identifying diabetic macular oedema from retinal photographs including the role of automated grading	P: Retinal screening programmes in England and Scotland I: Fully automated grading (a version of iGradingM) C: Three manual grading strategies: English (strategy 1), Scottish (strategy 2), alternative to English manual grading (having similar sensitivity and higher specificity) O: Costs, cases identified and QALYs	<b><i>Vs English manual grading</i></b> <u>Cost per case detected</u> Fully automated strategy dominates strategy 1 (English manual grading): automated is cheaper and identifies more cases. <u>Cost per QALY</u> After 20 years modelled, fully automated strategy dominates strategy 1: fully automated is £97 cheaper and provides incremental QALYs of 0.0001 than English manual grading (strategy 1),  <b><i>Vs Scottish manual grading</i></b> <u>Cost per case detected</u> Fully automated dominates Scottish manual grading (strategy 2) based on English screening and referral costs. It was estimated to cost £43,000 less and identified 1 more case than the Scottish manual grading system. When patient mix was adjusted to reflect expected frequency within a screening programme, fully adjusted was estimated to cost an additional £900 per case detected compared to the Scottish manual system. <u>Cost per QALY</u> At 20 years, the fully automated system is estimated to cost £113 more than the Scottish manual system and provide incremental	“When applying a ceiling ratio of 30,000 per quality adjusted life years (QALY) gained, Scotland's scheme was preferred. Assuming automated grading could be implemented without increasing grading costs, automation produced a greater number of QALYS for a lower cost than England's scheme, but was not cost effective, at the study's operating point, compared with Scotland's.”

Study	Country	Study type	Objectives	Components of the study	Outcomes	Authors' conclusions
					<p>QALYS of 0.0005, thus the fully automated system has an ICER of £222,210 compared to the Scottish system.</p>	
					<p>It is assumed that this analysis is adjusted for patient mix.</p>	
<p>Tufail 2016 &amp; 2017 (Liew 2014 and Egan 2016)</p>	<p>England (London)</p>	<p>CEA  Decision tree model (1 year time-horizon, 2013/ 2014 cost year) [from full-text]</p>	<p>Can ARIAS be safely introduced into DR screening pathways to replace human graders</p>	<p>P: Patients attending routine annual diabetic eye screening at one London hospital between June 1, 2012, and November 4, 2013. I: RetmarkerSR and EyeArt v1 as a replacement for initial human grading (strategy 1) and as a filter prior to primary human grading (strategy 2) [iGradingM not included in CEA] C: Manual grading O: Cost per appropriate screening outcome (defined as disease present when the reference human grade indicated the presence of potentially sight-threatening retinopathy or technical failure, and disease absent when the reference human grade indicated absence of retinopathy or background retinopathy only)</p>	<p>For both ARIAS and strategies, automated is less costly, but less effective than manual grading.</p> <p><b>Strategy 1</b> <b>Manual vs EyeArt</b> Incr costs: £101,820 Incr appr outcomes: 14,257 ICER: £7.14 (cost reduction per additional appropriate outcome missed)</p> <p><b>Manual vs RetmarkerSR</b> Incr costs: £167,251 Incr appr outcomes: 8,953 ICER: £18.68 (cost reduction per additional appropriate outcome missed)</p> <p><b>Strategy 2</b> <b>Manual vs EyeArt</b> Incr costs: £120,026 Incr appr outcomes: 14,256 ICER: £8.42 (cost reduction per additional appropriate outcome missed)</p> <p><b>Manual vs RetmarkerSR</b> Incr costs: £137, 521 Incr appr outcomes: 8,924 ICER: £15.37 (cost reduction per additional appropriate outcome missed)</p>	<p>“EyeArt v1 and RetmarkerSR saved costs compared with manual grading both as a replacement for initial human grading and as a filter prior to primary human grading, although the latter approach was less cost-effective.”</p>

Study	Country	Study type	Objectives	Components of the study	Outcomes	Authors' conclusions
Bhaskaranand 2016 (conference abstract)	UK	CMA  Decision tree approach	Comparison of annual population costs	P: UK NHS DESP I: EyeArt v1 C: Similar to the UK NHS DESP O: Annual population costs	Annual population costs  EyeArt: £3,626,974 + £528,000 (=£4,154,974) Fully manual grading: £12,000,709  Reported savings using EyeArt v1: £7,627,054	Automated DR screening using EyeArt v1 can provide significant cost savings in DR screening programs
Scotland 2007	Scotland (Grampians)	CEA  Decision tree model	Cost-effectiveness of replacing level 1 manual grading in the National Screening Programme in Scotland with an automated system	P: Diabetic population of Scotland (Grampian region) I: Automated system (a version of iGradingM) C: Manual grading O: Incr cost per case detected, incr cost per additional appropriate screening outcome	Automated cases detected: 5560 cases (86.9%) Manual cases detected: 5610 cases (87.7%). Incr effects (automated vs manual): -50 cases  Cost savings (NHS per year) for automated vs manual: £201,600  Incr cost per additional referable case detected ( <b>manual vs automated</b> ): £4088 Incr cost per additional appropriate screening outcome ( <b>manual vs automated</b> ): £1990	“Given that automated grading is less costly and of similar effectiveness, it is likely to be considered a cost-effective alternative to manual grading.”
Scotland 2010	Scotland	CEA (and CUA)  Decision tree model (same as that used in	Assess cost-effectiveness of two automated grading algorithms (as that in Scotland 2007 and an	P: Three screening centres in Scotland I: Two algorithms (versions of iGradingM). Algorithm (a) is simpler (using image quality	Algorithm (b) vs manual grading: Incr effects: -123 referable cases Incr effects: -734 appropriate screening outcomes Incr cost: - £212 695	“Algorithm (b) is more cost-effective than the algorithm based on quality assessment and MA/DH detection. With respect to the

Study	Country	Study type	Objectives	Components of the study	Outcomes	Authors' conclusions
		Scotland 2007).	improvement of Scotland 2007) with manual grading	assessment and MA/DH detection than algorithm (b) (combines image quality assessment with detection algorithms for MA, blot haemorrhages and exudates) C: Manual grading O: Incremental costs per: <ul style="list-style-type: none"> <li>• additional case found,</li> <li>• appropriate screening outcome</li> </ul> QALY gained (using a 20-year extrapolation model to assess impact of any missed referable cases)	<b>ICERs (Manual vs algorithm b - manual more expensive and more effective than algorithm)</b> Incr cost per referable case: £1727 Incr cost per additional appropriate screening outcome: £289  The ICER for manual vs algorithm b is between £25,676 and £267,115 per QALY, depending on the probability of algorithm b missing true proliferative cases	value of introducing automated detection systems into screening programmes, automated grading operates within the recommended national standards in Scotland and is likely to be considered a cost-effective alternative to manual disease/no disease grading."

AI – Artificial Intelligence, ARIAS - Automated Retinal Image Analysis System, CEA – Cost Effectiveness Analysis, CMA – Cost Minimisation Analysis, CUA – Cost Utility Analysis, DESP – Diabetic Eye Screening Programme, DH – Dot Haemorrhage, DR – Diabetic Retinopathy, ICER – Incremental Cost Effectiveness Ratio, MA – Microaneurysm, QALY – Quality Adjusted Life Years, PICO – Population Intervention Comparator Outcome

**Table 26 Question 4: Evidence map of studies investigating the social and ethical implications of implementing AI in screening programmes: Primary studies**

Study	Study design and country	Objectives	Components of the study	Outcomes	Authors' conclusions
Alexander 2020	Survey, USA	Survey of US radiologists' workload and use of AI and review of the progress of AI in medical imaging	<b>P:</b> Radiologists <b>I:</b> N/a <b>C:</b> N/a <b>O:</b> Experience	Not available in the abstract	As AI in medical imaging increasingly proves its worth, it is hard to imagine that AI will not ultimately transform radiology

Bourla 2018	Survey, France	To explore psychiatrists' perspectives on different AI-based assessment techniques through the prism of new CDSS	<b>P:</b> Psychiatrists <b>I:</b> AI-based CDSS <b>C:</b> Alternative AI-based CDSS technologies <b>O:</b> Acceptability	Overall acceptability was moderate (n=515). MRI coupled with ML was considered to be the most useful system, and the connected wristband was considered the least. All the systems were described as risky (410/515, 79.6%). Acceptability was strongly influenced by socio-epidemiological variables (professional culture), such as gender, age, and theoretical approach.	Moderate acceptability, mostly due to lack of knowledge about these new technologies rather than a strong rejection. Strong correspondences between acceptability profiles and professional culture profiles. Many medical, forensics, and ethical issues were raised, including therapeutic relationship, data security, data storage, and privacy risk. It is essential for psychiatrists to receive training and become involved in the development of new technologies.
Coppola 2020	Survey, Italy	A nationwide online survey on AI among radiologist members of the Italian Society of Medical and Interventional Radiology (SIRM)	<b>P:</b> Radiologists <b>I:</b> AI in radiology <b>C:</b> N/a <b>O:</b> Perceived advantages of AI and overall opinion about AI	1032 radiologists (9.5% of active SIRM members) joined the survey. Perceived AI advantages included a lower diagnostic error rate (750/1027, 73.0%) and optimization of radiologists' work (697/1027, 67.9%). The risk of a poorer professional reputation of radiologists compared with non-radiologists (617/1024, 60.3%), and increased costs and workload due to AI system maintenance and data analysis (399/1024, 39.0%) were seen as potential issues. Most radiologists stated that specific policies should regulate the use of AI (933/1032, 90.4%) and were not afraid of losing their job due to it (917/1032, 88.9%). Overall, 77.0% of respondents (794/1032) favoured the adoption of AI, whereas 18.0% (186/1032) were uncertain and 5.0% (52/1032) were unfavourable.	Radiologists had mostly positive attitudes toward the implementation of AI in their working practice. They were not concerned that AI will replace them, but rather that it might diminish their professional reputation.

European Society of Radiology 2019	Survey, Europe	To investigate the expectations about AI in 5-10 years among members of the European Society of Radiology (ESR)	<b>P:</b> Radiologists <b>I:</b> AI in radiology <b>C:</b> N/a <b>O:</b> Expectations	<p>675 (2.8%) of ESR members completed the survey. AI impact was mostly expected (<math>\geq 30\%</math> of responders) on breast, oncologic, thoracic, and neuro imaging, mainly involving mammography, computed tomography, and magnetic resonance. Responders foresee AI impact on: job opportunities (375/675, 56%), 218/375 (58%) expecting increase, 157/375 (42%) reduction; reporting workload (504/675, 75%), 256/504 (51%) expecting reduction, 248/504 (49%) increase; radiologist's profile, becoming more clinical (364/675, 54%) and more subspecialised (283/675, 42%). For 374/675 responders (55%) AI-only reports would be not accepted by patients, for 79/675 (12%) accepted, for 222/675 (33%) it is too early to answer. For 275/675 responders (41%) AI will make the radiologist-patient relation more interactive, for 140/675 (21%) more impersonal, for 259/675 (38%) unchanged. If AI allows time saving, radiologists should interact more with clinicians (437/675, 65%) and/or patients (322/675, 48%). For all responders, involvement in AI-projects is welcome, with different roles: supervision (434/675, 64%), task definition (359/675, 53%), image labelling (197/675, 29%). Of 675 responders, 321 (48%) do not currently use AI, 138 (20%) use AI, 205 (30%) are planning to do it. According to 277/675 responders (41%), radiologists will take responsibility for AI outcome,</p>	Generally, responders had a favourable attitude towards AI.
------------------------------------	----------------	---	--	---	---

				while 277/675 (41%) suggest shared responsibility with other professionals.	
Ginestra 2019	Survey (prospective observational study), USA	To assess clinician perceptions of a ML-based early warning system to predict severe sepsis and septic shock (Early Warning System 2.0)	<p><b>P:</b> Nurses and health care providers dealing with non-ICU admissions</p> <p><b>I:</b> Early Warning System</p> <p><b>C:</b> N/a</p> <p><b>O:</b> Change in perception of the patient's risk of sepsis; perceptions of the alert's helpfulness and impact on care</p>	Few (24% nurses, 13% providers) identified new clinical findings after responding to the alert. Perceptions of the presence of sepsis at the time of alert were discrepant between nurses (13%) and providers (40%). The majority of clinicians reported no change in perception of the patient's risk for sepsis (55% nurses, 62% providers). A third of nurses (30%) but few providers (9%) reported the alert changed management. Almost half of the nurses (42%) but less than a fifth of providers (16%) found the alert helpful at 6 hours.	In general, clinical perceptions of Early Warning System 2.0 were poor. Nurses and providers differed in their perceptions of sepsis and alert benefits. These findings highlight the challenges of achieving acceptance of predictive and machine learning-based sepsis alerts.
Gorges 2020 (CA)	Survey, N/A	To assess physicians' and general public perceptions on the use of AI to assist medical decision making; in particular, the notion of uncertainty in outcome predictions, and how this might influence treatment decisions	<p><b>P:</b> Physicians and family members</p> <p><b>I:</b> Use of AI to assist medical decision making</p> <p><b>C:</b> Comparison between the 2 groups and low- and high-risk medical scenarios</p> <p><b>O:</b> Perceptions</p>	26 family members and 21 physicians were included in the analysis. Familiarity with AI varied, yet >90% of participants agreed that AI has the potential to improve medical services. Regarding liability for AI-augmented decisions, both families and physicians agreed that the physician was primarily responsible, yet families also assigned responsibility to AI design companies. In <b>low-risk scenarios</b> , both groups trusted the AI's suggestion and emphasized patient-physician discussions of results: 92% of families and 95% of physicians would follow the AI's recommendation when positive outcomes [40% vs 20% effectiveness] were predicted; 43% of physicians (vs. 67% of families) reverted to physician	These surveys suggested that families were accepting of AI-assisted medical decision making. Physicians were more hesitant in trusting AI predictions in the high-risk scenario; this may be in part due to assumed liability favoring a more conservative clinical approach. Results of this survey may inform the development of AI and decision support systems to make these technologies more acceptable to expert and lay users.



judgment when AI risk assessment showed equal effectiveness. **High-risk scenarios** revealed significant differences between the two groups: only 38% of physicians (vs. 69% of families) would follow an AI's suggested intervention if it was against common practice and only 38% of physicians were likely to discuss options with patients. Both groups would consider AI risk metrics when making treatment decisions: 62% of physicians and 88% of families for short-term differences in outcome [44% vs. 66% mortality]; 52% of physicians and 81% for long-term differences [worse immediate mortality, but 20% vs. 50% improved 5-year outcome]. Physicians strongly reverted to physician judgment when AI generated risk assessment showed equal effectiveness: only 5% would follow the AI recommendation, compared with 52% of families.

Hamilton 2002	Survey, USA	Survey of directors of screening organisations	<p><b>P:</b> Directors of screening organisations</p> <p><b>I:</b> High-throughput screening (HTS) automation</p> <p><b>C:</b> N/a</p> <p><b>O:</b> Perceptions of the current vs. desired state of HTS</p>	Not available in the abstract	Not available in the abstract
Jonmarker 2019	Survey, Sweden	To survey breast cancer screening participants' attitudes towards potential future uses of computerization.	<p><b>P:</b> Women in a breast cancer screening program</p> <p><b>I:</b> Computerization of breast cancer screening</p>	Response rate was 1.3%. Of the submitted surveys, 97.5% were complete; 38% of respondents reported a preference for a computer-only examination. The highest level of confidence was	A high level of trust in computerized decision-making was expressed. Higher age was associated with a lower understanding of technology but did not affect

			<p><b>C:</b> N/a  <b>O:</b> Attitudes towards potential future uses of computerization</p>	<p>given a computer-only reading followed by a physician reading. Participants with &gt; 12 years of education were more likely to prefer a computer-only reading (odds ratio [OR] 1.655, 95% confidence interval [CI] 1.168-2.344), had a greater trust in letting a computer determine screening intervals and the need for a supplemental MRI (OR 1.606, 95% CI 1.171-2.202 and OR 1.577, 95% CI 1.107-2.247, respectively). Age was not found to be a significant predictor.</p>	<p>attitudes to computerized decision-making. A lower level of education was associated with a lower trust in computerization. This may be valuable knowledge for future studies.</p>
Jungmann 2020	Survey, Germany	To investigate the attitudes of radiologists, information technology (IT) specialists, and industry representatives on AI and its future impact on radiological work	<p><b>P:</b> Radiologists, IT specialists and industry  <b>I:</b> AI  <b>C:</b> Comparison between groups  <b>O:</b> Attitudes</p>	<p>The strongest agreement between all respondents occurred with the following: plausibility checks are important to understand the decisions of the AI (93% agreement), validation of AI algorithms is mandatory (91%), and medicine becomes more efficient in the age of AI (86%). In contrast, only 25% of the respondents had confidence in the AI results, and only 17% believed that medicine will become more human through the use of AI. The answers were significantly different between the three professions for four items: relevance for protocol selection in cross-sectional imaging (p=0.034), medical societies should be involved in validation (p=0.028), patients should be informed about the use of AI (p=0.047), and AI should be part of medical education (p=0.026).</p>	<p>Currently, a discrepancy exists between high expectations for the future role of AI and low confidence in the results. This attitude was similar across all three groups. The demand for plausibility checks and the need to prove the usefulness in randomized controlled studies indicate what is needed in future research.</p>
Keel 2018	Survey, Australia	Patient acceptability of a novel AI-based DR screening model within	<p><b>P:</b> Adults with DM undergoing DR screening  <b>I:</b> ARIAS using DL</p>	<p>96 participants were screened for DR and the mean assessment time for automated screening was 6.9 minutes. 96% reported that</p>	<p>AI-based DR screening in endocrinology outpatient settings appears to be</p>

		endocrinology outpatient settings	algorithm and real-time reporting of results <b>C:</b> Usual practice, results received in 2 weeks <b>O:</b> Overall satisfaction and preferred model of care	they were either satisfied or very satisfied with the automated screening model and 78% reported that they preferred the automated model over manual. The sensitivity and specificity of the DLA for correct referral was 92.3% and 93.7%, respectively.	feasible and well accepted by patients.
Koh 2019 (CA)	Survey, UK	To understand the current attitudes and perceptions of radiologists to AI and ML in cancer imaging (online international survey)	<b>P:</b> Radiologists <b>I:</b> AI and ML in cancer imaging <b>C:</b> N/a <b>O:</b> Attitudes and perceptions	664 responses from radiologists across more than 40 countries, a range of practice backgrounds; > 66% of the responders indicated that the benefits of AI and ML are much bigger or slightly bigger than the risks for cancer imaging; > 86% felt that AI tools would be used in at least some areas of work that would add value to cancer imaging within the next 5 years. The participants had good agreement with the perceived positive effects of utilising AI and ML; but there was more disagreement about the possible negative effects. The responders to the survey indicated the importance of AI in specific areas.	Not available in the abstract.
Meyer 2020	Survey, USA	To examine patients' experiences using an AI-assisted online symptom checker	<b>P:</b> Users of the Isabel Symptom Checker [online tool] <b>I:</b> Isabel Symptom Checker <b>C:</b> N/a <b>O:</b> Experiences of symptom checker use, experiences discussing results with physicians, and prior personal history of experiencing a	329 usable responses obtained. Patients most commonly used the symptom checker to better understand the causes of their symptoms (232/304, 76.3%), followed by for deciding whether to seek care (101/304, 33.2%) or where (eg, primary or urgent care: 63/304, 20.7%), obtaining medical advice without going to a doctor (48/304, 15.8%), and understanding their diagnoses better (39/304, 12.8%). Most patients reported receiving useful information for their health	Despite ongoing concerns about symptom checker accuracy, a large patient-user group perceived an AI-assisted symptom checker as useful for diagnosis. Formal validation studies evaluating symptom checker accuracy and effectiveness in real-world practice could provide additional useful information about their benefit.

diagnostic error were collected. problems (274/304, 90.1%), with half reporting positive health effects (154/302, 51.0%). Most patients perceived it to be useful as a diagnostic tool (253/301, 84.1%), as a tool providing insights leading them closer to correct diagnoses (231/303, 76.2%), and reported they would use it again (278/304, 91.4%). Patients who discussed findings with their physicians (103/213, 48.4%) more often felt physicians were interested (42/103, 40.8%) than not interested in learning about the tool's results (24/103, 23.3%) and more often felt physicians were open (62/103, 60.2%) than not open (21/103, 20.4%) to discussing the results. Compared with patients who had not previously experienced diagnostic errors (missed or delayed diagnoses: 123/304, 40.5%), patients who had previously experienced diagnostic errors (181/304, 59.5%) were more likely to use the symptom checker to determine where they should seek care (15/123, 12.2% vs 48/181, 26.5%; P=.002), but they less often felt that physicians were interested in discussing the tool's results (20/34, 59% vs 22/69, 32%; P=.04).

Nadarzynski 2019	Survey and semi-structured interviews, UK	To explore participants' willingness to engage with AI-led health chatbots	<p><b>P:</b> General public  <b>I:</b> AI-based chatbot systems  <b>C:</b> N/a  <b>O:</b> Acceptability and perceived utility and trustworthiness</p>	<p>Three broad themes: 'Understanding of chatbots', 'AI hesitancy' and 'Motivations for health chatbots' were identified, outlining concerns about accuracy, cyber-security, and the inability of AI-led services to empathise. The survey showed moderate acceptability (67%), correlated</p>	<p>Most internet users would be receptive to using health chatbots, although hesitancy regarding this technology is likely to compromise engagement.</p>
------------------	---	--	---	--	--

				negatively with perceived poorer IT skills OR=0.32 [95% CI: 0.13–0.78] and dislike for talking to computers OR=0.77 [95% CI: 0.60–0.99] as well as positively correlated with perceived utility OR¼ 5.10 [CI95%:3.08–8.43], positive attitude OR=2.71 [95% CI: 1.77–4.16] and perceived trustworthiness OR=1.92 [95% CI: 1.13–3.25].	
Ooi 2019	Survey, Singapore	To assess the attitudes and learner needs of radiology residents and faculty radiologists regarding AI and ML in radiology.	<b>P:</b> Radiologists <b>I:</b> AI and ML <b>C:</b> N/a <b>O:</b> Attitudes and learner needs	<p>125 respondents (86 male, 39 female; 70 residents, 55 faculty radiologists) completed the questionnaire. The majority agreed that AI/ML will drastically change radiology practice (88.8%) and makes radiology more exciting (76.0%), and most would still choose to specialise in radiology if given a choice (80.0%). 64.8% viewed themselves as novices in their understanding of AI/ML, 76.0% planned to further advance their AI/ML knowledge and 67.2% were keen to get involved in an AI/ML research project. An overwhelming majority (84.8%) believed that AI/ML knowledge should be taught during residency, and most opined that this was as important as imaging physics and clinical skills/knowledge curricula (80.0% and 72.8%, respectively). More than half thought that their residency programme has not adequately implemented AI/ML teaching (59.2%). In subgroup analyses, male and tech-savvy respondents were more involved in AI/ML activities, leading to better technical understanding.</p>	A growing optimism of radiology undergoing technological transformation and AI/ML implementation has led to a strong demand for AI/ML curriculum in residency education.

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

Ooms 2019 (CA)	Survey, USA	To compare experience and perceptions of adults screened by ARIAS and receiving direct specialist contact via tele-presence robot (TR) and usual practice	<p><b>P:</b> Adults screened for DR</p> <p><b>I:</b> ARIAS interfaced with TR</p> <p><b>C:</b> Usual practice</p> <p><b>O:</b> Preferences</p>	A paired t-test did not suggest that those who interacted with the TR had a preference between the TR and a certified reader (4.57 vs 4.79, p=.19). An unequal variances t-test suggested that interacting with the TR increased the likelihood of wanting robots involved in one's healthcare (3.50 vs 2.48, p=.02).	The AI and TR combination presents a novel approach to improving access to ophthalmic care. The TR, having received positive feedback from participants, could facilitate direct interaction with specialists when AI detects DR.
Palmisciano 2020	Survey, UK	To evaluate attitudes of patients and their relatives regarding use of AI in neurosurgery	<p><b>P:</b> Patients and families</p> <p><b>I:</b> AI in neurosurgery</p> <p><b>C:</b> n/a</p> <p><b>O:</b> Attitudes and perceptions</p>	In the first stage, 20 participants responded. Five themes were identified: interpretation of imaging (4/20; 20%), operative planning (5/20; 25%), real-time alert of potential complications (10/20; 50%), partially autonomous surgery (6/20; 30%), and fully autonomous surgery (3/20; 15%). In the second stage, 107 participants responded. Most thought it appropriate and acceptable to use AI for imaging interpretation (76.7%; 66.3%), operative planning (76.7%; 75.8%), real-time alert of potential complications (82.2%; 72.9%), and partially autonomous surgery (58%; 47.7%). Conversely, most did not think that fully autonomous surgery was appropriate (27.1%) or acceptable (17.7%). Demographics did not have a significant influence on perception	Most patients and their relatives believed that AI has a role in neurosurgery and found it acceptable. Notable exceptions were fully autonomous systems, with most wanting the neurosurgeon ultimately to remain in control.
Paul 2006	Survey, India	To assess patient satisfaction levels and factors influencing it during teleophthalmology consultation in India [NN-based ARIAS also mentioned as part of the new technologies]	<p><b>P:</b> Patients</p> <p><b>I:</b> Tele-ophthalmology screening</p> <p><b>C:</b> N/a</p> <p><b>O:</b> Satisfaction with tele-ophthalmology screening</p>	348 respondents in total; 56.4% were males; the mean age of was 50 +/- 17 years. Age ranged from 2 years to 83 years. 44.4% of the respondents were satisfied with teleophthalmology screening (95% CI: 38.58%-49.42%). No association was found between age, gender, education, and occupation, respectively, with	This study highlights sentiments of the rural subjects when they underwent teleophthalmology consultations. This study provides valuable insights about patient's preferences and satisfaction levels with this newer technology.

				satisfaction levels. We found that patients who asked questions during the screening were 2.18 times more likely to be satisfied with teleophthalmology than those who did not (odds ratio [OR] = 2.19, 95% CI 1.37-3.5).	
Waymel 2019	Survey, France	To assess the perception, knowledge, wishes and expectations of a sample of French radiologists towards the rise of artificial intelligence (AI) in radiology	<p><b>P:</b> Radiologists</p> <p><b>I:</b> AI in radiology</p> <p><b>C:</b> N/a</p> <p><b>O:</b> Perception, knowledge, wishes and expectations</p>	A total of 70 radiology residents and 200 senior radiologists participated in the survey (43.8% (270/617) response rate). 73.3% (198/270) estimated they had received insufficient previous information on AI; 94.4% (255/270; ) would consider attending a generic continuous medical education in this field and 69.3% (187/270) a technically advanced training on AI; 79.3% (214/270) thought that AI will have a positive impact on their future practice. The highest expectations were the lowering of imaging-related medical errors (81%, 219/270), followed by the lowering of the interpretation time of each examination (74.4%, 201/270) and the increase in the time spent with patients (52.2%, 141/270).	While respondents had the feeling of receiving insufficient previous information on AI, they are willing to improve their knowledge and technical skills on this field. They share an optimistic view and think that AI will have a positive impact on their future practice. A lower risk of imaging-related medical errors and an increase in the time spent with patients are among their main expectations.
Xiang 2020	Survey, China	To investigate public perceptions of and demands regarding the implementation of medical AI	<p><b>P:</b> General public (including healthcare vs other lines of work)</p> <p><b>I:</b> Implementation of medical AI</p> <p><b>C:</b> N/a</p> <p><b>O:</b> Perceptions and demands</p>	2,780 participants from 22 provinces were enrolled. There was no significant difference between the healthcare workers (54.3% of all participants) and non-healthcare workers in the high proportion (99 %) of participants expressing acceptance of AI (p = 0.8568), but remarkable distributional differences were observed in demands (p<0.001 for both demands for AI assistance and the desire for AI improvements) and perceptions (p<0.001 for safety, validity, trust,	The public exhibits a high level of receptivity regarding the implementation of medical AI. There is a strong demand for intelligent assistance in many medical areas, including imaging and pathology departments, outpatient services, and surgery.

and expectations). High levels of receptivity (approximately 100 %), demands (approximately 80 %), and expectations (100 %) were expressed among different age groups. The receptivity of medical AI among the non-healthcare workers was associated with gender, educational qualifications, and demands and perceptions of AI. There was a very large gap between current availability of and public demands for intelligence services ( $p < 0.001$ ); >90 % of healthcare workers expressed a willingness to devote time to learning about AI and participating in AI research.

AI – Artificial Intelligence, ARIAS - Automated Retinal Image Analysis System, CDSS – Clinical Decision Support System, DESP – Diabetic Eye Screening Programme, DL – Deep Learning, DM – Diabetes Mellitus, DR – Diabetic Retinopathy, ML – Machine Learning, MRI – Magnetic Resonance Imaging, NN – Neural Networks, PICO – Population Intervention Comparator Outcome, TR – Tele-presence Robot

**Table 27 Question 4: Evidence map of studies investigating the social and ethical implications of implementing AI in screening programmes: Review and opinion papers**

Study	Study type	Objectives	Focus of the study
<b>Anonymous 2018</b>	Editorial	Discussion of the emerging role of AI in eye conditions and CVD	AI in eye conditions and CVD
<b>Anonymous 2019</b>	Editorial	Discussion of ML in medicine	ML in medicine
<b>Rajalakshmi 2020</b>	Editorial	ARIAS in DESP in India	ARIAS in DESP
<b>Shaban-Nejad 2018</b>	Editorial	Review of advances of AI in healthcare	AI in healthcare
<b>Sosale 2019</b>	Editorial	AI in DR screening: discussion of its effectiveness and cost-effectiveness	ARIAS in DESP
<b>Sivaprasad 2020</b>	Report	The ORNATE India Project: to build research capacity and capability in India and the UK to tackle global burden of diabetes-related visual impairment.	ARIAS in DESP
<b>Abramoff 2010</b>	Review	To survey the methods, potential benefits and limitations of ARIAS in order to better manage translation into clinical practice	ARIAS in DESP
<b>Abramoff 2020</b>	Review	To perform a literature review of bioethical principles for AI, and derived evaluation rules for autonomous AI, grounded in bioethical principles	AI in healthcare



<b>Balyen 2019</b>	Review	Impact of AI on the early detection and treatment of ophthalmic diseases	AI in ophthalmology
<b>Berens 2020</b>	Review	"This article discusses to what extent the application of AI algorithms can contribute to quality assurance in the field of ophthalmology	AI in ophthalmology
<b>Broome 2020</b>	Review	To review the current status of ML in various aspects of diabetes care (incl. DR screening) and identify key challenges that must be overcome to leverage ML to its full potential.	ML in diabetes
<b>Carter 2020</b>	Review	This paper proceeds in three parts. In part one, the authors consider what AI is, and examples of its development and evaluation for potential use in breast cancer care. In part two, they outline the ethical, legal and social issues of AI. In the final section, they anticipate future directions for AI in breast cancer care and draw some conclusions.	AI in breast cancer care (focus on social, legal and ethical issues)
<b>Channa 2020</b>	Review	Using the example of DR screening to highlight some important aspects to be considered by developers, policymakers, and end users when bringing autonomous AI algorithms into clinical practice.	ARIAS in DESP
<b>Chee 2018</b>	Review	An update and overview of the literature on current telemedicine applications in retina	DL in DR in the context of telemedicine
<b>Fatehi 2020</b>	Systematic review	To investigate the characteristics and usability features of tele-ophthalmology for the elderly population (including AI-based technology)	Tele-ophthalmology for the elderly population
<b>Francolini 2020</b>	Review	Overview of AI and its use in radiotherapy	AI in Radiotherapy
<b>Graham 2019</b>	Review	A review of recent research on AI in healthcare and mental health	AI in healthcare and mental health
<b>Halamka 2019</b>	Review	Review of the role of AI in family medicine using DR and colon cancer as examples	AI and ML in DR and colon cancer
<b>Jheng 2020</b>	Review	To discuss the issues of computing resource consumption and performance of the mobile device-based AI systems, and highlight recent research regarding the feasibility and future potential of application of the mobile device-based AI systems in telemedicine	AI in eye disease screening using smart phones
<b>Kapoor 2019</b>	Review	Review of the application and use of AI in OCT imaging in ophthalmology	AI in OCT in ophthalmology
<b>Kapoor 2019</b>	Review	Review of the role of AI in the diagnosis and management of glaucoma	AI in glaucoma
<b>Keskinbora 2020</b>	Review	To review the developments and potential practices regarding the use of AI in the field of ophthalmology, and the related topic of medical ethics	AI in ophthalmology
<b>Larson 2020</b>	Theoretical paper	The authors propose an ethical framework for using and sharing clinical data for the development of AI applications	AI in healthcare
<b>Liew 2019</b>	Review	Consider 3 core questions relating to AI in radiology, and the barriers to the widespread adoption of AI in radiology and propose solutions and describe a "Centaur" model as a promising avenue for enabling the interfacing between AI and radiologists.	AI in radiology
<b>O'Connor 2019</b>	Review	Discuss the potential reason for the slow adoption of machine learning tools into systematic reviews in healthcare	ML in systematic reviews
<b>Padhy 2019</b>	Review	Review of ARIAS in DESP	ARIAS in DESP
<b>Patel 2007</b>	Review	Review of application of ANN in healthcare	ANN in healthcare

<b>Rahimy 2018</b>	Review	"To describe the emerging applications of DL in ophthalmology	DL in ophthalmology
<b>Ruamviboonsuk 2020</b>	Review	ARIAS in DESP: recent developments and contribution of Asia in DR and other areas	ARIAS in DESP
<b>Scott 2019</b>	Review	Prospects and pitfalls of ML in clinical decision making	ML in clinical decision making
<b>Stolte 2020</b>	Review	Overview of AI in DR including some social/ethical aspects	ARIAS in DESP
<b>Ting 2019</b>	Review	Describes global eye disease burden, unmet needs and common conditions of public health importance for which AI and DL systems may be applicable	AI and DL in eye disease
<b>Ting 2020</b>	Review	Overview of AI in eye disease	AI in eye disease
<b>Vollmer 2020</b>	Review	A proposed framework to inform design, conduct and reporting of AI in healthcare	AI in healthcare
<b>Wang 2012</b>	Review	A short introduction to machine learning and survey its applications in radiology	AI in radiology
<b>Wong 2019</b>	Review	Overview of the field of DR including screening and AI	ARIAS in DESP
<b>Wong 2020</b>	Review	Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence	ARIAS in DESP
<b>Ting 2019</b>	Review	Review of DL applications in ophthalmology	DL in ophthalmology

AI – Artificial Intelligence, ANN – Artificial Neural Networks, ARIAS - Automated Retinal Image Analysis System, CVD – Cardio Vascular Disease, DESP – Diabetic Eye Screening Programme, DL – Deep Learning, DM – Diabetes Mellitus, DR – Diabetic Retinopathy, ML – Machine Learning , OCT – Optical Coherence Tomography

### Appraisal for quality and risk of bias

The results from the methodological quality assessment of the included test accuracy studies (question 1) is reported in the main text. Here we provide additional information on the specific criteria used in the assessment.

**Table 28 QUADAS-2 checklist with definitions of the signalling questions**

Domain	Signalling questions
Patient selection	<ol style="list-style-type: none"> <li>1. Was a consecutive or random sample of patients enrolled? Yes, if clear from the paper</li> <li>2. Was a case-control design avoided? Yes, if clear from the paper</li> <li>3. Did the study avoid inappropriate exclusions? Yes, if no patients that would normally be included in the EDESP were excluded</li> </ol> <p>Risk of bias: Low if all of the above are answered 'yes'</p> <p>Applicability concerns: Low only if the study is conducted in the relevant UK population; Unclear, if conducted in a non-UK population, unless there is a clear indication that the population is different from the target population (e.g. a mixture of diabetic and non-diabetic patients)</p>
Index test	<ol style="list-style-type: none"> <li>1. Were the index test results interpreted without knowledge of the results of the reference standard? In most cases, this will be 'yes', unless specific reason is given</li> </ol>

	<p>2. If a threshold was used, was it prespecified? Yes, if clear from the paper                  Risk of bias: Low, if all of the above are answered 'yes'                  Applicability concerns: High, if different from the intended EDESP use of the system</p>
Reference standard	<p>1. Is the reference standard likely to correctly classify the target condition? Yes, the reference standard involves a panel of retinal specialists or ophthalmologists or similar experts independently reading the images; Yes, if the final grading from a national screening programme with multilevel grading, such as EDESP, which has clear training and quality assurance protocols, is used*; No, if a single grader determines the ground truth, or the final grade from a national screening programme is used without external adjudication                  2. Were the reference standard results interpreted without knowledge of the results of the index test? Yes, if clear from the paper                  Risk of bias: Low, if all of the above are answered 'yes'                  Applicability concerns: High, if the definition of the target condition is different from the one used in the EDESP (e.g. maculopathy is not included)</p>
Flow and timing	<p>1. Did all patients receive a reference standard?                  2. Did all patients receive the same reference standard?                  3. Were all patients included in the analysis? No, if ungradable images have been excluded                  Risk of bias: Low, if all of the above are answered 'yes'</p>
Additional questions	<p>1. Were the study methods prespecified?                  2. Was the study funded/financially supported by the manufacturer or the authors declared conflict of interest?                  3. Is the algorithm publically available?                  4. Is the output of the model Interpretable and can it be interrogated (visualization of decisions)?                  5. Are differential diagnoses and estimates of confidence provided?</p>
<p>*Initially, this criterion also included external adjudication of disagreements between the final manual grade and the result from ARIAS; however, after discussion with experts we decided to drop the requirement for external adjudication and accept the final grade from an established national screening programme sufficient to ensure low risk of bias, provided the programme has clear training and quality assurance protocols.</p> <p>EDESP – English Diabetic Eye Screening Programme</p>	

**Table 29 QUADAS-C checklist: Comparative accuracy**

Domain	Signalling questions
Patient selection*	<p>C1.1 Was the risk of bias for this domain judged 'low' for all index tests?                      C1.2 Was the intention for patients either to receive all index tests or to be randomly allocated to index tests?                      C1.3 If patients were randomized, was the allocation sequence random?                      C1.4 If patients were randomized, was the allocation sequence concealed until patients were enrolled and assigned to index tests?</p>
Index test	<p>C2.1 Was the risk of bias for this domain judged 'low' for all index tests?                      C2.2 If patients received multiple index tests, were test results interpreted without knowledge of the results of the other index test(s)?                      C2.3 If patients received multiple index tests, is undergoing one index test unlikely to affect the performance of the other index test(s)?</p>

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

	C2.4 Were differences in the conduct or interpretation between the index tests unlikely to advantage one of the tests?
Reference standard	C3.1 Was the risk of bias for this domain judged 'low' for all index tests? C3.2 Did the reference standard avoid incorporating any of the index tests?
Flow and timing	C4.1 Was the risk of bias for this domain judged 'low' for all index tests? C4.2 Was there an appropriate interval between the index tests? C4.3 Was the same reference standard used for all index tests? C4.4 Are the proportions and reasons for missing data similar across index tests? C4.5 Could the patient flow have introduced bias in the comparison?

\*The QUADAS-C tool was applied according to the guidance provided by Bada Yang ([b.d.yang@amsterdamumc.nl](mailto:b.d.yang@amsterdamumc.nl)) and the QUADAS-C group.

## Appendix 4 – Additional information

Links to the manufacturer/developer’s website for each of the ARIASs included in the accuracy analysis (or a brief description if no website was found)

### **iGradingM (traditional ML), no website found**

We failed to find the manufacturer website or any information about the most up-to-date version of the system. Tufail 2016 (1) was the most recent study reporting on the system and evaluated iGradingM v 1.1. We assumed that the data reported by Philip 2017(17) (conference abstract) also relates to iGradingM, but the authors did not specify and simply referred to the system as ‘the autograder’.

iGradingM was developed by the University of Aberdeen and, during the study, was purchased from the Medalytix Group Ltd, Manchester, UK (which funded Fleming 2010a) by Digital Healthcare, Cambridge, UK, at the initiation of the study, and purchased in turn by EMIS UK, Leeds, UK, after its conclusion (1). The system is designed to remove normal images from manual grading queue by 1) checking the image quality, and 2) detecting early signs of retinopathy by looking for a) microaneurysms only or b) microaneurysms, haemorrhages and exudates (as two alternative modes of operation) (5).

### **RetmarkerSR (traditional ML), <https://www.retmarker.com/>**

We contacted the manufacturer who confirmed that: 1) they are not aware of additional titles; 2) the system has been undergoing continuous improvement and the results reported in older studies may not reflect the current version of the system; 3) that the software has not been upgraded to DL-based; 4) audit-based reports on the use and performance of the system within the Portuguese DESP are not publicly available as each local authority organises and audits its own screening programme. More detailed description of the system is reported in Tufail 2016 (1).

### **Retinalyze (traditional ML), <https://www.retinalyze.com/>**

We contacted the manufacturer who confirmed that: 1) the system has not been upgraded to DL; 2) they are not aware of additional titles; 3) the most recent evaluation is the one reported in Bouhaimed 2008 (40) which evaluated the Retinalyze v.1.0.6.1.

### **EyeArt v2 (DL): <https://www.eyenuk.com/en/products/eyear2/>**

**EyeGrade (DL):** Healgoo Interactive Medical Technology Co. Ltd, Guangzhou, China (no valid link found)

### **IDx-DR v2 (DL): <https://dxs.ai/products/idx-dr/idx-dr-overview-2/>**

**Google AI (DL): <https://health.google/for-clinicians/ophthalmology/>**

**RedCAD (DL): <https://www.delft.care/retcad/>**

**SELENA (DL): <https://www.eyris.io/index.cfm>**

**VUNO (DL): <https://www.vuno.co/en/>**

## Systematic reviews investigating the accuracy of AI algorithms for detection of diabetic retinopathy

**Table 30 Systematic reviews investigating the accuracy of AI algorithms for detection and grading of diabetic retinopathy**

Study	Aim	Inclusion and exclusion criteria	Searches	QUADAS-2	Results
<b>Systematic reviews (most criteria for a systematic review met)</b>					
Islam 2020 (136)	To investigate the performance of DL algorithms for automated detection of DR in retinal colour fundus photographs	<b>Inclusion criteria:</b> 1) published in English and peer- reviewed; 2) provided an outcome of DL algorithms and DR detection; 3) report accuracy; 4) database and number of images reported; 5) definition of DR provided; 6) clearly described DL algorithms and process used in the DR detection. <b>Exclusion criteria:</b> Editorials, short reports, traditional methods for detecting DR were excluded	<b>Databases searched:</b> EMBASE, PubMed, Google Scholar, Scopus, Web of Science and reference lists <b>Search period:</b> 01.01.2000 to 31.03.2019	<b>High or unclear RoB</b> PS: 13/23 IT: 0/23 RS: 5/23 F&T: 6/23 <b>Applicability concerns</b> PS: 20/23 IT: 0/23 RS: 9/23	<b>Studies included in the review:</b> 23 studies (6 with external validation) included in the review <b>Studies included in meta-analysis:</b> 20 <b>Accuracy:</b> The pooled AUROC was 0.97 (95%CI: 0.95–0.98), SE was 0.83 (95%CI: 0.83–0.83), and SP 0.92 (95%CI: 0.92–0.92); LR+ and LR- were 14.11 (95%CI: 9.91–20.07), and 0.10 (95%CI: 0.07–0.16), respectively <b>Subgroup analysis:</b> SE and SP for vision-threatening DR was 0.92 (95%CI:0.90–0.94), and 0.91 (95%CI: 0.90–0.92)
Nagendran 2020 (137)	To systematically examine the design, reporting standards, RoB, and claims of studies comparing the performance of diagnostic DL algorithms for medical imaging with that of expert clinicians.	<b>Inclusion criteria:</b> 1) a peer reviewed scientific report of original research; 2) English language; 3) assessed a DL algorithm applied to a clinical problem in medical imaging; 4) compared algorithm performance with a contemporary human group not involved in establishing the ground truth (RS); 5) at least one human in the group was considered an expert; 6) the aim was to use medical imaging for predicting absolute risk of existing disease or classification into diagnostic groups. <b>Exclusion criteria:</b> informal publication types, such as commentaries, letters to the editor, editorials and meeting abstracts.	<b>Databases:</b> Medline, Embase, CENTRAL and WHO-ICTRP; manual screening of the reference lists of relevant publications <b>Search period:</b> 2010 to June 2019	<b>PROBAST</b> used to assess RoB, but results not reported: 5/7 externally validated; 1/7 and 3/7 stated the need of prospective studies in abstract and discussion, respectively; 1/7 recommended clinical use	<b>Studies included in the review:</b> 7 of the included 91 studies met our inclusion criteria (of the 91 studies 10 were trial registrations, none of which was relevant, and 81 non-randomised studies); <b>Studies included in meta-analysis:</b> n/a <b>Accuracy:</b> SE ranged from 86.2% to 97.1% for AI and from 61.1% to 91.2% for clinical experts; sensitivity of AI was higher than that of experts in 4/7 studies (n=7); in the 3 studies for which specificity was reported it ranged from 93.4% to 95.2% for AI and from 95.9% to 98.7% for clinical experts. Two studies reported as a primary measure AUC: 0.936 in the first study and ranging from 0.894 to 0.972 in the second; and one study reported the quadratic weighted kappa: 0.84 for AI and 0.82 for clinical experts. Abramoff 2018 reported SE of 87.2% (95% CI, 81.8–91.2%) and specificity of 90.7% (95% CI, 88.3–92.7%) for AI which were omitted in the review.

Nielsen 2019 (138)	To review the diagnostic performance of DL-based algorithms in screening patients with diabetes for DR	<p><b>Inclusion criteria:</b> DTA studies of DL—either a complete algorithm or DL features—to classify full-scale DR in retinal fundus images of patients with diabetes. Studies had to state a grading scale, have a human grader as RS, and provide a performance score for the DL method. No restrictions were made as to the performance score or grading scale, or the profession of the human grader.</p> <p><b>Exclusion criteria:</b> Studies examining only lesions or subcategories of the disease (e.g. mild DR); duplicates, animal studies, reviews, editorials, CA, and unpublished articles.</p>	<p><b>Databases:</b> Medline (via PubMed) and Embase (via OvidSP) and reference lists</p> <p><b>Search period:</b> up to 05.04.2018</p>	<p><b>High or unclear RoB</b></p> <p>PS: 8/11 IT: 0/11 RS: 1/11 F&amp;T: 2/11</p> <p><b>Applicability concerns</b></p> <p>PS: 9/11 IT: 0/11 RS: 6/11</p>	<p><b>Studies included in the review:</b> 11</p> <p><b>Studies included in meta-analysis:</b> 20</p> <p><b>Accuracy:</b> 8 studies reported SE of 80.28% to 100.0%; SP of 84.0% to 99.0%; 2 report overall accuracy of 78.7% and 81.0%, respectively; and one study reported AUC of 0.955</p>
Norgaard 2017 (139)	To identify studies with methodology and design that are similar to or replicate actual screening scenarios using ARIAS for detection of DR	<p><b>Inclusion criteria:</b> 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 non-mydriatic) images from consecutively recruited patients with any form of DM 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</p>	<p><b>Databases:</b> PubMed, Cochrane Library, and Embase; In addition, (1) Review studies identified in database searching were screened for references; (2) Authors with several publications that involved ARIAS were identified, and a search was performed to reveal additional publications; (3) Google search</p> <p><b>Search period:</b> [up to] 21.10.2016</p>	<p><b>N/a</b></p> <p>Methodological concerns discussed throughout the paper and included: an analysis based on limited number of episodes, accuracy estimates with no CIs and no clear grading scale</p>	<p><b>Studies included in the review:</b> 7</p> <p><b>Studies included in meta-analysis:</b> n/a</p> <p><b>Accuracy:</b> the detection of DR had high SE (87.0 – 95.2%) but lower SP (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 DME</p>
Simoes 2019 (140)	To evaluate the accuracy of DSS in diagnosing DR	<p><b>Inclusion criteria:</b> DTA studies evaluating the accuracy of DSS in diagnosing DR in patients with DM (both type 1 and 2</p>	<p><b>Databases searched:</b> 10 databases including Medline</p>	<p><b>Performed,</b> but not results reported and unclear if</p>	<p><b>Studies included in the review:</b> 18</p> <p><b>Studies included in the meta-analysis:</b> 18</p>

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

		included); the RS had to be fundus examination or funduscopy. <b>Exclusion criteria:</b> Other types of diabetes, and studies not reporting sufficient data to allow the reconstruction of the 2x2 table.	via PubMed, Embase, grey literature and reference lists <b>Search period:</b> 1970 - 2018	considered in the analysis	<b>Accuracy:</b> The pooled SE was 97.7% (95% CI: 97.5%-97.9%) and the pooled SP 90.3% (95% CI: 90.0%-90.6%) <b>Meta-regression:</b> Clinical and technological co-factors had no effect on accuracy (no details reported) <b>Subgroup and sensitivity analysis:</b> heterogeneity remained high (no details reported)
Wang 2020 (141)	To estimate the SE and SP of NN in DR grading	<b>Included:</b> Studies evaluating the accuracy of NN to detect referable DR (incl. DME) comparing the IT with ophthalmologists' diagnosis as RS, based on fundus photography without assistance of other medical records, and providing sufficient information for quantitative data synthesis.	<b>Databases searched:</b> Medline, Embase, IEEE Xplore, and Cochrane Library <b>Search period:</b> up to 23.07.2019	<b>High or unclear RoB</b> PS: 13/24 IT: 7/24 RS: 10/24 F&T: 0/24 <b>Applicability concerns</b> PS: 0/24 IT: 0/24 RS: 9/24	<b>Studies included in the review:</b> 24 <b>Studies included in meta-analysis:</b> 24 <b>Accuracy:</b> Pooled SE of 91.9% (95% CI: 89.6% to 94.3%) and SP of 91.3% (95% CI: 89.0% to 93.5%). <b>Subgroup analyses and meta-regression</b> did not provide any statistically significant findings for the heterogeneous diagnostic accuracy in studies with different image resolutions, sample sizes of training sets, architecture of CNN, or diagnostic criteria.
AI – artificial intelligence, AUC - area under the [ROC] curve, CENTRAL - Cochrane Central Register of Controlled Trials, CI – confidence interval, CNN – convolutional neural networks, DL – deep learning, DME – diabetic macular oedema, DR – diabetic retinopathy, DSS – decision support system, F&T – flow and timing, IT – index test, ML – machine learning, NN – neural networks, PROCAST - prediction model risk of bias assessment tool, RS – reference standard, RoB – risk of bias, SE – sensitivity, SP – specificity, WHO-ICTRP - World Health Organization International Clinical Trials Registry Platform					

**Table 31 Quality assessment of the identified systematic reviews**

Study	AMSTAR II Questions															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Islam 2020	No	n/a	Yes	n/a*	Yes	Yes	Yes	n/a*	Yes	No	Yes	No	No	Yes	n/a*	Yes
Nagendran 2020	No	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No MA	No MA	Yes	Yes	No MA	No
Nielsen 2019	Yes	No	No	Partial Yes	Yes	NR	Yes	Partial Yes	Yes	No	No MA	No MA	Yes	Yes	No MA	Yes
Norgaard 2017	Yes	No	Yes	No	NR	NR	Yes	Yes	No	No	No MA	No MA	NA	No	No MA	No
Simoes 2019	Yes	No	Yes	Partial Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	NR	No
Wang 2020	Yes	No	No	Partial Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	Yes

\*We were unable to access the supplementary file.

AMSTAR II Questions:  
1. Did the research questions and inclusion criteria for the review include the components of PICO?



UK NSC external review – Automated grading in the Diabetic Eye Screening Programme

2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?
3. Did the review authors explain their selection of the study designs for inclusion in the review?
4. Did the review authors use a comprehensive literature search strategy?
5. Did the review authors perform study selection in duplicate?
6. Did the review authors perform data extraction in duplicate?
7. Did the review authors provide a list of excluded studies and justify the exclusions?
8. Did the review authors describe the included studies in adequate detail?
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?
10. Did the review authors report on the sources of funding for the studies included in the review?
11. If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?
13. Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?
15. If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

## Factors reported to affect the performance of ARIAS

### Population

Tufail 2016 (1) reported that the accuracy of EyeArt v1 was not affected by ethnicity and sex, but sensitivity was marginally lower with increasing patient age. The accuracy of RetmarkerSR was affected by patient age and ethnicity. There was considerable variation in the mean age and racial composition of the cohorts across the included studies.

Some studies included only patients with type 2 diabetes (e.g. van der Heijden 2018 (21)), but most of the cohorts included both type 1 and type 2 patients. Time since diagnosis and diabetic control are associated with the progression of retinopathy; if cohorts vary on these parameters, this is likely to affect the performance of the ARIAS.

Tufail 2016 (1) excluded "...patients whose photographs were ungradable at their previous screening episodes, for example because of a known cataract that degraded the quality of retinal photography, were 'technically failed' and underwent slit-lamp biomicroscopy by optometrists in a clinic adjacent to the photographic screening clinic." (p. 7). Variation in the way such patients are handled could affect the performance of ARIAS as well as the inclusion of cohorts from opportunistic rather than organised screening. Most studies which excluded patients with ungradable images according to the reference grading did not report if this included patients known to have conditions that affect image quality.

### Index test

#### Setting and test operator

The setting and the test operator varied across studies and are likely to affect the results from the evaluation of ARIAS. Most of the UK studies were done within a national screening programme. Many studies were conducted in primary care (e.g. van der Heijden 2018) or primary care and ophthalmology clinics (Lim 2019). The qualification and experience of the test operators also varied (e.g. van der Heijden 2018; Verbraak 2019; Abramoff 2018 in the assessment of IDx-DR; Lim 2019, Heydon 2020 in the assessment of EyeArt).

#### Photographic protocols

There was a significant variation in the fundus photography protocols including (but probably not limited to):

- The number of fields and the area of the retina covered: Studies used 1-field (e.g. SDESP-based studies), 2-field (e.g. EDESP-based studies) or 3-field images (e.g. Bhaskaranand 2019 evaluating EyeArt v2 using the EyePACS protocol). Goatman 2011 (5) showed that iGradingM has slightly better performance when 2-field

(EDESP) rather than 1-field (SDESP) images are used; but the system failed in Tufail 2016, due to failure to read disc-centred images. Yip 2020 (113) compared the performance of SELENA using 1-, 2- and 7-field images and showed that the system performed better when the number of fields is increased. Also, in some retrospective studies not all of the images were used (e.g. Krause 2018 (34)).

- Multiple images: Some studies reported that the operator taking the images could decide to take more than the specified number of images, to make sure that images of sufficient quality are available for each eye; in other studies re-imaging was not allowed. Van der Heijden 2018 (21) , who evaluated IDx-DR in primary care, reported that the image quality feedback of the system was underutilized and, as a result, a considerable proportion of the patients were referred due to ungradable images. As the study shows, this could affect the overall performance of the system and the impact of context-related factors need to be considered in the evaluation (e.g. busy primary care vs. dedicated screening programme; the attitude of those performing the imaging).
- Use of pupil dilation: studies varied according to whether mydriasis was used routinely (e.g. EDESP-based studies, Shah 2020a), not used at all (e.g. Liu 2020) or only when deemed necessary by the test operator; the latter group also varied with regards to the proportion of patients in whom pupil dilation was used.
- Camera type: Tufail 2016 (1) reported that the accuracy of RetmarkerSR was marginally affected by camera type while the accuracy of EyeArt v1 was not.
- Image size: A number of studies showed that image size affects ARIAS performance with larger images (up to a specific threshold) leading to better accuracy (Krause 2018 (34), Yip 2020 (113))
- Cataract surgery: Yip 2020 (113) showed that SELENA achieved higher accuracy in pseudophakic eyes compared to phakic eyes (AUC 0.918 vs 0.833,  $p < 0.001$ ).

## **Variation in ARIASs**

Use of data from previous patient episodes: Some algorithms use data from previous screening episodes to improve the classification of the images from the index visit. An example of such ARIAS is RetmarkerSR which is based on traditional ML and has been implemented in the Portuguese DESP. More information on this approach and examples of DL-based algorithms are provided in Stolte 2020 (103).

Algorithms looking for more than one condition: Some ARIASs are designed for detection of more than one condition, such as DR and AMD (e.g. Gonzalez-Gonzalo 2020 and Son 2020).

## Comparator

The accuracy of manual grading may vary across different DESPs and therefore the comparative accuracy of ARIASs versus human graders may not be generalizable from one DESP to another (see Table 10). Also, there might be variation in the performance of human graders across different sites within the same DESP. This means that studies, in which the comparison between ARIAS and manual grading has been done in a single site may not be representative of the whole programme. For instance, Gulshan 2019 (27) reported considerable variation in the performance of manual grading at 2 different sites in India.

## Reference standard and target condition

Grading system: Different systems were used across studies. As already discussed in relation to IDx-DR (21), this could affect the reference grading and, from there, the performance of ARIAS.

Graders qualification and experience: The qualification and experience of graders could affect grading but, as demonstrated by Gulshan 2019 (27), the direction of this effect is not always easy to predict.

A screening programme manual grading with or without external adjudication: A number of studies used the final grades from the local DESPs as the ground truth for the evaluation of ARIASs. In some cases a proportion of the images (e.g. higher grade images, disagreements between the programme and ARIAS) were adjudicated externally. It is difficult to judge the importance of this but, as discussed earlier in relation to Tufail 2016 (1), it is another potential source of heterogeneity across studies.

Method of adjudication: Krause 2018 (34) showed that adjudication by 3 retinal specialists (who first graded all images independently) could lead to different accuracy results depending on whether the final decision was based on majority voting or discussion until consensus is reached.

Variation in the technology: As discussed earlier, Abramoff 2018 (20) showed that the performance of IDx-DR differed when the reference standard was stereo wide-field fundus photography (4W-D stereo protocol) and the combination of 4W-D and OCT. Also, both reference standards are superior to the standard non-stereo fundus photography used in most of the studies.

## Flow and timing

Handling of ungradable images: Studies handled images judged to be 'ungradable' by the reference graders and/or the system differently. In some studies, all such images were excluded from analysis; in other studies only the images considered to be 'ungradable' by

the reference graders were excluded, while those defined as ‘ungradable’ by the ARIAS but ‘gradable’ by the reference graders were reported separately, or treated as ‘referrals’; and some studies reported separately the agreement between the system and the reference graders in terms of ‘ungradable images’. In a number of studies images were excluded because the results from the reference grading were missing.

## Appendix 5 – UK NSC reporting checklist for evidence summaries

All items on the UK NSC Reporting Checklist for Evidence Summaries have been addressed in this report. A summary of the checklist, along with the page or pages where each item can be found in this report, is presented in Table 32.

**Table 32 UK NSC reporting checklist for evidence summaries**

	Section	Item	Page no.
<b>1.</b>	<b>TITLE AND SUMMARIES</b>		
<b>1.1</b>	Title sheet	Identify the review as a UK NSC evidence summary.	Title page
<b>1.2</b>	Plain English summary	Plain English description of the executive summary.	8
<b>1.3</b>	Executive summary	Structured overview of the whole report. To include: the purpose/aim of the review; background; previous recommendations; findings and gaps in the evidence; recommendations on the screening that can or cannot be made on the basis of the review.	10
<b>2.</b>	<b>INTRODUCTION AND APPROACH</b>		
<b>2.1</b>	Background and objectives	Background – Current policy context and rationale for the current review – for example, reference to details of previous reviews, basis for current recommendation, recommendations made, gaps identified, drivers for new reviews	22
		Objectives – What are the questions the current evidence summary intends to answer? – statement of the key questions for the current evidence summary, criteria they address, and number of studies included per question, description of the overall results of the literature search.	33
		Method – briefly outline the rapid review methods used.	34
<b>2.2</b>	Eligibility for inclusion in the review	State all criteria for inclusion and exclusion of studies to the review clearly (PICO, dates, language, study type, publication type, publication status etc.) To be decided <i>a priori</i> .	34
<b>2.3</b>	Appraisal for quality/risk of bias tool	Details of tool/checklist used to assess quality, e.g. QUADAS 2, CASP, SIGN, AMSTAR.	37
<b>3.</b>	<b>SEARCH STRATEGY AND STUDY SELECTION (FOR EACH KEY QUESTION)</b>		
<b>3.1</b>	Databases/sources searched	Give details of all databases searched (including platform/interface and coverage dates) and date of final search.	37

<b>3.2</b>	Search strategy and results	<p>Present the full search strategy for at least one database (usually a version of Medline), including limits and search filters if used.</p> <p>Provide details of the total number of (results from each database searched), number of duplicates removed, and the final number of unique records to consider for inclusion.</p>	82
<b>3.3</b>	Study selection	State the process for selecting studies – inclusion and exclusion criteria, number of studies screened by title/abstract and full text, number of reviewers, any cross checking carried out.	34
<b>4.</b>	<b>STUDY LEVEL REPORTING OF RESULTS (FOR EACH KEY QUESTION)</b>		
<b>4.1</b>	Study level reporting, results and risk of bias assessment	<p>For each study, produce a table that includes the full citation and a summary of the data relevant to the question (for example, study size, PICO, follow-up period, outcomes reported, statistical analyses etc.).</p> <p>Provide a simple summary of key measures, effect estimates and confidence intervals for each study where available.</p> <p>For each study, present the results of any assessment of quality/risk of bias.</p>	<p>Q1: 40 – 68, 117</p> <p>Q2: 70, 126</p> <p>Q3: 129</p> <p>Q4: 132 - 145</p>
<b>4.2</b>	Additional analyses	Describe additional analyses (for example, sensitivity, specificity, PPV, etc.) carried out by the reviewer.	N/a
<b>5.</b>	<b>QUESTION LEVEL SYNTHESIS</b>		
<b>5.1</b>	Description of the evidence	For each question, give numbers of studies screened, assessed for eligibility, and included in the review, with summary reasons for exclusion.	<p>Q1: 41</p> <p>Q2: 69</p> <p>Q3: 72</p> <p>Q4: 75</p>
<b>5.2</b>	Combining and presenting the findings	Provide a balanced discussion of the body of evidence which avoids over reliance on one study or set of studies. Consideration of four components should inform the reviewer’s judgement on whether the criterion is ‘met’, ‘not met’ or ‘uncertain’: quantity; quality; applicability and consistency.	<p>Q1: 66</p> <p>Q2: 70</p> <p>Q3: 72</p> <p>Q4: 75</p>
<b>5.3</b>	Summary of findings	<p>Provide a description of the evidence reviewed and included for each question, with reference to their eligibility for inclusion.</p> <p>Summarise the main findings including the quality/risk of bias issues for each question.</p> <p>Have the criteria addressed been ‘met’, ‘not met’ or ‘uncertain’?</p>	<p>Q1: 68</p> <p>Q2: 71</p> <p>Q3: 72</p> <p>Q4: 75</p>
<b>6.</b>	<b>REVIEW SUMMARY</b>		
<b>6.1</b>	Conclusions and implications for policy	<p>Do findings indicate whether screening should be recommended?</p> <p>Is further work warranted?</p>	77

Are there gaps in the evidence highlighted by the review?			
<b>6.2</b>	Limitations	Discuss limitations of the available evidence and of the review methodology if relevant.	80



## References

[Please if possible use EndNote for references]

1. 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.
2. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. *Acta Diabetol*. 2017;54(6):515-25.
3. Wilkinson CP, Ferris FL, 3rd, Klein RE, Lee PP, Agardh CD, Davis M, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110(9):1677-82.
4. Aldington SJ, Kohner EM, Meuer S, Klein R, Sjølie AK. Methodology for retinal photography and assessment of diabetic retinopathy: the EURODIAB IDDM complications study. *Diabetologia*. 1995;38(4):437-44.
5. 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(12):e27524.
6. WHO. Diabetes 2020 [Available from: <https://www.who.int/news-room/factsheets/detail/diabetes>].
7. UK D. Facts & Figures 2020 [Available from: <https://www.diabetes.org.uk/>].
8. ETDRS. Treatment Techniques and Clinical Guidelines for Photocoagulation of Diabetic Macular Edema: Early Treatment Diabetic Retinopathy Study Report Number 2. *Ophthalmology*. 1987;94(7):761-74.
9. 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(2):e004015.
10. Yau JWY, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global Prevalence and Major Risk Factors of Diabetic Retinopathy. *Diabetes Care*. 2012;35(3):556.
11. Cheloni R, Gandolfi SA, Signorelli C, Odone A. Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. *BMJ Open*. 2019;9(3):e022188.
12. ICO. ICO Guidelines for Diabetic Eye Care 2017 [Available from: [http://www.icoph.org/enhancing\\_eyecare/diabetic\\_eyecare.html](http://www.icoph.org/enhancing_eyecare/diabetic_eyecare.html)].
13. PHE. NHS Diabetic Eye Screening Programme Overview of patient pathway, grading pathway, surveillance pathways and referral pathways. In: England PH, editor.: PHE; 2017.
14. UK NSC. Extending diabetic eye screening intervals for people at low risk of developing sight threatening retinopathy: summary. 2016.
15. RCO. The Royal College of Ophthalmologists. Diabetic Retinopathy Guidelines. 2012.
16. NICE. Evidence standards framework for digital health technologies. In: Excellence NIfHaC, editor.: NICE; 2021.
17. Philip S, Lee N, Black M, Sharp P, Olson J. Impact of introducing automated grading into the Scottish national diabetic retinopathy screening programme. *Diabetic Medicine*. 2017;34 (Supplement 1):172.

18. 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*. 2014.
19. Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. *Curr Diab Rep*. 2019;19(9):72.
20. Abramoff 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*. 2018;1:39.
21. 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 (Oxf)*. 2018;96(1):63-8.
22. Verbraak FD, Abramoff MD, Bausch GCF, Klaver C, Nijpels G, Schlingemann RO, et al. Diagnostic Accuracy of a Device for the Automated Detection of Diabetic Retinopathy in a Primary Care Setting. *Diabetes Care*. 2019;42(4):651-6.
23. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *Jama*. 2017;318(22):2211-23.
24. Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. *Ophthalmology*. 2020;127(1):85-94.
25. Soto-Pedre E, Navea A, Millan S, Hernaez-Ortega MC, Morales J, Desco MC, et al. Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload. *Acta Ophthalmol*. 2015;93(1):e52-6.
26. Raumviboonsuk P, Krause J, Chotcomwongse P, Sayres R, Raman R, Widner K, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj digit*. 2019;2:25.
27. 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 Ophthalmology*. 2019;13:13.
28. Zachariah S, Wykes W, Yorston D. The Scottish Diabetic Retinopathy Screening programme. *Community Eye Health*. 2015;28(92):s22-s3.
29. Lim J BM, Ramachandra C, Bhat S, Solanki K, Sadda S. Artificial Intelligence Screening for Diabetic Retinopathy: Analysis from a Pivotal Multi-Center Prospective Clinical Trial. *ARVO Imaging in the Eye Conference 2019*; Vancouver, BC, Canada: ARVO; 2019.
30. FDA. EyeArt FDA approval letter, 510(k) Number: K200667 2020.
31. 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. *The British journal of ophthalmology*. 2020.
32. Olvera-Barrios A, Heeren TF, Balaskas K, Chambers R, Bolter L, Egan C, et al. Diagnostic accuracy of diabetic retinopathy grading by an artificial intelligence-enabled algorithm compared with a human standard for wide-field true-colour confocal scanning and standard digital retinal images. *British Journal of Ophthalmology*. 2020;06:06.
33. Liu J, Gibson E, Ramchal S, Shankar V, Piggott K, Sychev Y, et al. Diabetic Retinopathy Screening with Automated Retinal Image Analysis in a Primary Care Setting Improves Adherence to Ophthalmic Care. *Ophthalmol Retina*. 2020;17:17.

34. 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.
35. Gonzalez-Gonzalo C, Sanchez-Gutierrez V, Hernandez-Martinez 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 (Oxf)*. 2020;98(4):368-77.
36. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, Maclsaac RJ, et al. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci*. 2018;8(1):4330.
37. Philip S, Fleming AD, Goatman KA, Fonseca S, McNamee P, Scotland GS, et al. The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. *British Journal of Ophthalmology*. 2007;91(11):1512-7.
38. 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. *British Journal of Ophthalmology*. 2010;94(12):1606-10.
39. Hansen AB, Hartvig NV, Jensen MS, Borch-Johnsen K, Lund-Andersen H, Larsen M. Diabetic retinopathy screening using digital non-mydratic fundus photography and automated image analysis. *Acta Ophthalmol Scand*. 2004;82(6):666-72.
40. Bouhaimed M, Gibbins R, Owens D. Automated detection of diabetic retinopathy: results of a screening study. *Diabetes Technol Ther*. 2008;10(2):142-8.
41. Larsen N, Godt J, Grunkin M, Lund-Andersen H, Larsen M. Automated detection of diabetic retinopathy in a fundus photographic screening population. *Invest Ophthalmol Vis Sci*. 2003;44(2):767-71.
42. Larsen M, Gondolf T, Godt J, Jensen MS, Hartvig NV, Lund-Andersen H, et al. Assessment of automated screening for treatment-requiring diabetic retinopathy. *Curr Eye Res*. 2007;32(4):331-6.
43. Oliveira CM, Cristovao LM, Ribeiro ML, Abreu JR. Improved automated screening of diabetic retinopathy. *Ophthalmologica*. 2011;226(4):191-7.
44. Olson J, Sharp P, Goatman K, Prescott G, Scotland G, Fleming A, et al. Improving the economic value of photographic screening for optical coherence tomography-detectable macular oedema: a prospective, multicentre, UK study. *Health Technol Assess*. 2013;17(51):1-142.
45. Prescott G, Sharp P, Goatman K, Scotland G, Fleming A, Philip S, et al. Improving the cost-effectiveness of photographic screening for diabetic macular oedema: a prospective, multi-centre, UK study. *British Journal of Ophthalmology*. 2014;98(8):1042-9.
46. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology*. 2017;124(3):343-51.
47. Bhaskaranand M, Ramachandra C, Bhat S, Solanki K. Cost savings enabled by automated diabetic retinopathy screening in a UK-like screening program. *Investigative Ophthalmology and Visual Science*. 2016;57 (12):5584.
48. Scotland GS, McNamee P, Philip S, Fleming AD, Goatman KA, Prescott GJ, et al. Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *British Journal of Ophthalmology*. 2007;91(11):1518-23.
49. Scotland GS, McNamee P, Fleming AD, Goatman KA, Philip S, Prescott GJ, et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *British Journal of Ophthalmology*. 2010;94(6):712-9.

50. Hamilton SD. HTS automation study: Results from a 2001 survey of the current vs. desired state of HTS automation. *JALA - Journal of the Association for Laboratory Automation*. 2002;7(2):78-83.
51. Jonmarker O, Strand F, Brandberg Y, Lindholm P. The future of breast cancer screening: what do participants in a breast cancer screening program think about automation using artificial intelligence? *Acta Radiol Open*. 2019;8(12):2058460119880315.
52. Ooms A, Caterfino A, Prasad N, Khouri P, Wilson L, Szirth B. Robotics and artificial intelligence in the management of vision threatening disease. *Investigative Ophthalmology and Visual Science Conference*. 2019;60(9).
53. Paul PG, Raman R, Rani PK, Deshmukh H, Sharma T. Patient satisfaction levels during teleophthalmology consultation in rural South India. *Telemed J E Health*. 2006;12(5):571-8.
54. Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, Pavan K, et al. Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Crit Care Med*. 2019;47(11):1477-84.
55. Fatehi F, Jahedi F, Tay-Kearney ML, Kanagasingam Y. Teleophthalmology for the elderly population: A review of the literature. *International Journal of Medical Informatics*. 2020;136:104089.
56. Carter S, Win K, Wang L, Rogers W, Richards B, Houssami N. Ethical, legal and social implications of artificial intelligence systems for screening and diagnosis. *BMJ Evidence-Based Medicine*. 2019;24 (Supplement 2):A37-A8.
57. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. *Radiology*. 2020;295(3):675-82.
58. Abramoff MD, Niemeijer M, Russell SR. Automated detection of diabetic retinopathy: barriers to translation into clinical practice. *Expert Rev Med Devices*. 2010;7(2):287-96.
59. Abramoff MD, Tobey D, Char DS. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *Am J Ophthalmol*. 2020;214:134-42.
60. Alexander A, Jiang A, Ferreira C, Zurkiya D. An Intelligent Future for Medical Imaging: A Market Outlook on Artificial Intelligence for Medical Imaging. *J*. 2020;17(1 Pt B):165-70.
61. Anonymous. All eyes are on AI. *Nat*. 2018;2(3):139.
62. Anonymous. Ascent of machine learning in medicine. *Nat Mater*. 2019;18(5):407.
63. Balyen L, Peto T. Promising Artificial Intelligence-Machine Learning-Deep Learning Algorithms in Ophthalmology. *Asia Pac J Ophthalmol (Phila)*. 2019;8(3):264-72.
64. Berens P, Waldstein SM, Ayhan MS, Kummerle L, Agostini H, Stahl A, et al. [Potential of methods of artificial intelligence for quality assurance]. *Ophthalmologe*. 2020;117(4):320-5.
65. 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.
66. Bourla A, Ferreri F, Ogorzelec L, Peretti CS, Guinchard C, Mouchabac S. Psychiatrists' Attitudes Toward Disruptive New Technologies: Mixed-Methods Study. *JMIR Ment Health*. 2018;5(4):e10240.
67. Broome DT, Hilton CB, Mehta N. Policy Implications of Artificial Intelligence and Machine Learning in Diabetes Management. *Curr Diab Rep*. 2020;20(2):5.
68. Channa R, Wolf R, Abramoff MD. Autonomous Artificial Intelligence in Diabetic Retinopathy: From Algorithm to Clinical Application. *J Diabetes Sci Technol*. 2020:1932296820909900.

69. Chee RI, Darwish D, Fernandez-Vega A, Patel S, Jonas K, Ostmo S, et al. Retinal Telemedicine. *Current Ophthalmology Reports*. 2018;6(1):36-45.
70. Coppola F, Faggioni L, Regge D, Giovagnoni A, Golfieri R, Bibbolino C, et al. Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol Med (Torino)*. 2020;29:29.
71. Egan C, Rudisill C, Owen C, Rudnicka A, Sadda SR, Taylor P, et al. Automated diabetic retinopathy image assessment softwares: Diagnostic accuracy and cost-effectiveness compared to human graders. *Ophthalmologica*. 2016;236 (Supplement 1):40-1.
72. European Society of R. Impact of artificial intelligence on radiology: a EuroAIM survey among members of the European Society of Radiology. *Insights imaging*. 2019;10(1):105.
73. Figueiredo IN, Kumar S, Oliveira CM, Ramos JD, Engquist B. Automated lesion detectors in retinal fundus images. *Comput Biol Med*. 2015;66:47-65.
74. Fleming AD, Goatman KA, Philip S, Williams GJ, Prescott GJ, Scotland GS, et al. The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *British Journal of Ophthalmology*. 2010;94(6):706-11.
75. Francolini G, Desideri I, Stocchi G, Salvestrini V, Ciccone LP, Garlatti P, et al. Artificial Intelligence in radiotherapy: state of the art and future directions. *Med Oncol*. 2020;37(6):50.
76. Gorges M, Sung J, Portales-Casamar E. Perceptions of expert and lay users on trust in the use of artificial intelligence for medical decision-making and risk prediction. *Anesthesia and Analgesia*. 2020;130 (5 Supplement 1):49-50.
77. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Curr Psychiatry Rep*. 2019;21(11):116.
78. Halamka J, Cerrato P. An FP's guide to AI-enabled clinical decision support. *Journal of Family Practice*. 2019;68(9):486;8;90;92.
79. Jheng YC, Chou YB, Kao CL, Yarmishyn AA, Hsu CC, Lin TC, et al. A Novelty Route for Smartphone-based Artificial Intelligence Approach to Ophthalmic Screening. *J Chin Med Assoc*. 2020;09:09.
80. Jungmann F, Jorg T, Hahn F, Pinto Dos Santos D, Jungmann SM, Duber C, et al. Attitudes Toward Artificial Intelligence Among Radiologists, IT Specialists, and Industry. *Acad Radiol*. 2020;12:12.
81. Kapoor R, Whigham BT, Al-Aswad LA. Artificial Intelligence and Optical Coherence Tomography Imaging. *Asia Pac J Ophthalmol (Phila)*. 2019;8(2):187-94.
82. Kapoor R, Whigham BT, Al-Aswad LA. The Role of Artificial Intelligence in the Diagnosis and Management of Glaucoma. *Current Ophthalmology Reports*. 2019;7(2):136-42.
83. Keskinbora K, Guven F. Artificial intelligence and ophthalmology. *Turkish Journal of Ophthalmology*. 2020;50(1):37-43.
84. Koh DM. Attitudes and perception of artificial intelligence and machine learning in oncological imaging. *Cancer Imaging Conference: 19th Meeting and Annual of the International Cancer Imaging Society Italy*. 2019;19(Supplement 1).
85. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. *Diabetes Care*. 2018;41(12):2509-16.
86. Liew G, Egan CA, Rudnicka A, Owen C, Kapetan V, Sim DA, et al. Evaluation of automated software grading of diabetic retinopathy and comparison with manual image grading-an accuracy and cost effectiveness study. *Investigative Ophthalmology and Visual Science*. 2014;55 (13):2293.

87. Liew CJ, Krishnaswamy P, Cheng LT, Tan CH, Poh AC, Lim TC. Artificial Intelligence and Radiology in Singapore: Championing a New Age of Augmented Imaging for Unsurpassed Patient Care. *Ann Acad Med Singapore*. 2019;48(1):16-24.
88. Meyer AND, Giardina TD, Spitzmueller C, Shahid U, Scott TMT, Singh H. Patient Perspectives on the Usefulness of an Artificial Intelligence-Assisted Symptom Checker: Cross-Sectional Survey Study. *J Med Internet Res*. 2020;22(1):e14679.
89. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *DIGITAL HEALTH*. 2019;5:2055207619871808.
90. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst*. 2019;8(1):143.
91. Ooi SKG, Makmur A, Soon AYQ, Fook-Chong S, Liew C, Sia SY, et al. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. *Singapore Med J*. 2019;04:04.
92. Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: A natural step to the future. *Indian J Ophthalmol*. 2019;67(7):1004-9.
93. Palmisciano P, Jamjoom AAB, Taylor D, Stoyanov D, Marcus HJ. Attitudes of Patients and Their Relatives Toward Artificial Intelligence in Neurosurgery. *World Neurosurg*. 2020;138:e627-e33.
94. Patel JL, Goyal RK. Applications of artificial neural networks in medical science. *Current Clinical Pharmacology*. 2007;2(3):217-26.
95. Rahimy E. Deep learning applications in ophthalmology. *Current Opinion in Ophthalmology*. 2018;29(3):254-60.
96. Rajalakshmi R. The impact of artificial intelligence in screening for diabetic retinopathy in India. *Eye*. 2020;34(3):420-1.
97. Ruamviboonsuk P, Cheung CY, Zhang X, Raman R, Park SJ, Ting DSW. Artificial Intelligence in Ophthalmology: Evolutions in Asia. *Asia Pac J Ophthalmol (Phila)*. 2020;9(2):78-84.
98. Scott IA, Cook D, Coiera EW, Richards B. Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*. 2019;211(5):203-5.
99. Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. *npj digit*. 2018;1:53.
100. Shah A, Clarida W, Amelon R, Hernaez-Ortega MC, Navea A, Morales-Olivas J, et al. Validation of Automated Screening for Referable Diabetic Retinopathy With an Autonomous Diagnostic Artificial Intelligence System in a Spanish Population. *J Diabetes Sci Technol*. 2020:1932296820906212.
101. Sivaprasad S, Raman R, Conroy D, Mohan t, Wittenberg R, Rajalakshmi R, et al. The ORNATE India Project: United Kingdom-India Research Collaboration to tackle visual impairment due to diabetic retinopathy. *Eye*. 2020;34(7):1279-86.
102. Sosale AR. Screening for diabetic retinopathy-is the use of artificial intelligence and cost-effective fundus imaging the answer? *International Journal of Diabetes in Developing Countries*. 2019;39(1).
103. Stolte S, Fang R. A survey on medical image analysis in diabetic retinopathy. *Med Image Anal*. 2020;64:101742.
104. Ting DSW, Carin L, Abramoff MD. Observations and Lessons Learned From the Artificial Intelligence Studies for Diabetic Retinopathy Screening. *JAMA Ophthalmology*. 2019;13:13.

105. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*. 2019;103(2):167.
106. Ting DSJ, Foo VH, Yang LWY, Sia JT, Ang M, Lin H, et al. Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology. *British Journal of Ophthalmology*. 2020;12:12.
107. Vollmer S, Mateen BA, Bohner G, Kiraly FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *The BMJ*. 2020;368 (no pagination)(l6927).
108. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal*. 2012;16(5):933-51.
109. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagnostic and Interventional Imaging*. 2019;100(6):327-36.
110. Wong TY, Sabanayagam C. The War on Diabetic Retinopathy: Where Are We Now? *Asia Pac J Ophthalmol (Phila)*. 2019;8(6):448-56.
111. Wong TY, Sabanayagam C. Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence. *Ophthalmologica*. 2020;243(1):9-20.
112. Xiang Y, Zhao L, Liu Z, Wu X, Chen J, Long E, et al. Implementation of artificial intelligence in medicine: Status analysis and development suggestions. *Artificial Intelligence in Medicine*. 2020;102:101780.
113. Yip MYT, Lim G, Lim ZW, Nguyen QD, Chong CCY, Yu M, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *npj digit*. 2020;3:40.
114. Abramoff 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.
115. Araujo T, Aresta G, Mendonca L, Penas S, Maia C, Carneiro A, et al. DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Med Image Anal*. 2020;63:101715.
116. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health*. 2019;1(1):e35-e44.
117. Bhaskaranand M, Ramachandra C, Bhat S, Cuadros J, Nittala MG, Sadda S, et al. Automated Diabetic Retinopathy Screening and Monitoring Using Retinal Fundus Image Analysis. *J Diabetes Sci Technol*. 2016;10(2):254-61.
118. Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology*. 2017;124(7):962-9.
119. 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.
120. Harangi B, Toth J, Baran A, Hajdu A. Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features. *Conf Proc IEEE Eng Med Biol Soc*. 2019;2019:2699-702.
121. He J, Cao T, Xu F, Wang S, Tao H, Wu T, et al. Artificial intelligence-based screening for diabetic retinopathy at community hospital. *Eye*. 2020;34(3):572-6.

122. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of Artificial Intelligence-Based Grading of Diabetic Retinopathy in Primary Care. *JAMA Network Open*. 2018;1(5):e182665.
123. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci*. 2017;7(1):17816.
124. Liu YP, Li Z, Xu C, Li J, Liang R. Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network. *Artificial Intelligence in Medicine*. 2019;99:101694.
125. Nazir T, Irtaza A, Shabbir Z, Javed A, Akram U, Mahmood MT. Diabetic retinopathy detection through novel tetragonal local octa patterns and extreme learning machines. *Artificial Intelligence in Medicine*. 2019;99:101695.
126. Pires R, Avila S, Wainer J, Valle E, Abramoff MD, Rocha A. A data-driven approach to referable diabetic retinopathy detection. *Artificial Intelligence in Medicine*. 2019;96:93-106.
127. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. *Clin Experiment Ophthalmol*. 2018;46(4):412-6.
128. Romero-Aroca P, Verges-Puig R, de la Torre J, Valls A, Relano-Barambio N, Puig D, et al. Validation of a Deep Learning Algorithm for Diabetic Retinopathy. *Telemed J E Health*. 2019;04:04.
129. Roychowdhury S, Koozekanani DD, Parhi KK. DREAM: diabetic retinopathy analysis using machine learning. *IEEE j*. 2014;18(5):1717-28.
130. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*. 2019;126(4):552-64.
131. Shah P, Mishra DK, Shanmugam MP, Doshi B, Jayaraj H, Ramanjulu R. Validation of Deep Convolutional Neural Network-based algorithm for detection of diabetic retinopathy - Artificial intelligence versus clinician for screening. *Indian J Ophthalmol*. 2020;68(2):398-405.
132. Al Turk L, Wang S, Krause P, Wawrzynski J, Saleh GM, Al Sawadi H, et al. Evidence Based Prediction and Progression Monitoring on Retinal Images from Three Nations. *Transl*. 2020;9(2):44-.
133. Usher D, Dumskyj M, Himaga M, Williamson TH, Nussey S, Boyce J. Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening. *Diabetic Medicine*. 2004;21(1):84-90.
134. Wang XN, Dai L, Li ST, Kong HY, Sheng B, Wu Q. Automatic Grading System for Diabetic Retinopathy Diagnosis Using Deep Learning Artificial Intelligence Software. *Curr Eye Res*. 2020:1-6.
135. Zago GT, Andreato RV, Dorizzi B, Teatini Salles EO. Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Comput Biol Med*. 2020;116:103537.
136. Islam MM, Yang HC, Poly TN, Jian WS, Jack Li YC. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Comput Methods Programs Biomed*. 2020;191:105320.
137. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *The BMJ*. 2020;368 (no pagination)(m689).
138. Nielsen KB, Lautrup ML, Andersen JKH, Savarimuthu TR, Grauslund J. Deep Learning-Based Algorithms in Screening of Diabetic Retinopathy: A Systematic Review of Diagnostic Performance. *Ophthalmol Retina*. 2019;3(4):294-304.
139. Nørgaard MF, Grauslund J. Automated Screening for Diabetic Retinopathy - A Systematic Review. *Ophthalmic Res*. 2018;60(1):9-17.



140. Simoes PW, Dos Passos MG, Amaral LL, Garcia D, Vicente RB, de Abreu LLT, et al. Meta-Analysis of the Sensitivity of Decision Support Systems in Diagnosing Diabetic Retinopathy. *Studies in Health Technology & Informatics*. 2019;264:878-82.
141. Wang S, Zhang Y, Lei S, Zhu H, Li J, Wang Q, et al. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: a systematic review and meta-analysis of diagnostic test accuracy. *Eur*. 2020;183(1):41-9.