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: 4.0

Author: WeBS-ESG

Date: 25 March 2026

This report was commissioned by the National Institute for Health and Care Research (NIHR) Evidence Synthesis Programme. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care

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

Contents

Automated grading in the Diabetic Eye Screening Programme 1
Contents 2
About the UK National Screening Committee (UK NSC) 4
Plain English summary 5
Executive summary 7
Purpose of the review 7
Background 7
Focus of the review 8
Recommendation under review 9
Findings and gaps in the evidence of this review 10
Recommendations on screening 13
Limitations 14
Evidence uncertainties 15
Introduction and approach 16
Background 16
Current policy context and previous reviews 17
Objectives 23
Methods 25
Eligibility for inclusion in the review 25
Appraisal for quality/risk of bias tool 27
Databases/sources searched 28
Results 29
Question level synthesis 30
Criteria 4 and 5 — Accuracy of ARIASs 30
Summary of Findings Relevant to Criteria 4 and 5: met (for some, but not all, ARIASs) 63
Criterion 11 — Effectiveness of DESPs using ARIASs at level 1 grading 64
Summary of Findings Relevant to Criterion 11: not met. 73
Criterion 14 – Cost-effectiveness of ARIASs 74
Summary of Findings Relevant to Criterion 14: not met. 75
Review summary 76
Conclusions and implications for policy 76

Appendix 1 — Search strategy 80
Electronic databases 80
Search Terms 80
Appendix 2 — Included and excluded studies 91
Publications included after review of full-text articles 92
Publications excluded after review of full text articles 96
Appendix 3 — Summary and appraisal of individual studies 112
Data Extraction 112
Appraisal for quality and risk of bias 150
Appendix 4 – Supplementary ‘Methods’ and ‘Results Tables and Figures’ 151
Supplementary Methods 151
Supplementary Tables and Figures (Question 1) 153
Appendix 5 – UK NSC reporting checklist for evidence summaries 158
Appendix 6 – Patient and Public Involvement and Engagement (PPIE) (reporting based on GRIPP2† guidelines) 161
References 168

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, SW1H 0EU

www.gov.uk/uknsc

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

For queries relating to this document, please contact: uknsc@dhsc.gov.uk.

Plain English summary

Diabetes is a health condition which makes a person's blood sugar level high. This can harm different body parts, including the retina located at the back of the eyes. The damage to the retina is called diabetic retinopathy. It can cause loss of eyesight if not found and treated early.

In the UK, people aged 12 and over who have diabetes are invited for an eye screening test every 1 to 2 years. This test takes pictures of the back of the eyes. Trained staff, known as primary (or first) graders, look for changes or signs of damage in the pictures. Pictures that show eye damage or are hard to read go to a second grader for a second opinion. If the 2 graders do not agree, the picture goes to a third grader to make a final decision. Patients without eye damage and most with mild changes can return in 1 to 2 years for screening. Patients with worse damage may need more tests and treatment. The screening process takes a lot of staff time. As the number of people with diabetes gets bigger, so will the workload of staff and the costs.

An Automated Retinal Image Analysis System (ARIAS) is a type of computer programme. It can check eye pictures to find signs of eye disease using artificial intelligence. In some countries, ARIAS is already used to help with diabetic eye screening.

In 2021, the UK National Screening Committee (UK NSC) looked at evidence about whether ARIAS could replace the primary human grader in UK Diabetic Eye Screening Programmes. They found there was not enough good evidence to recommend the use of ARIAS over current care.

There has now been another review. This updated review looked at the new evidence from June 2020 on:

- Whether ARIAS are as good as humans at finding diabetic eye disease
- How ARIAS use affects patients, staff, and the screening programme
- Whether replacing human primary graders with ARIAS is good value for money

We found 2 studies from the UK and 5 from similar countries (including Europe and the USA) that compared ARIAS to human graders' performance to detect diabetic eye disease. They found:

- Most ARIAS are as good as human graders at finding people with more severe diabetic eye disease
- Most ARIAS are not as good as human graders at finding people who do not have eye disease at all
- The performance of ARIAS is different depending on which company made it, and where it was used
- Some ARIAS may perform differently based on patient ethnicity or age

Four studies looked at how ARIAS use affects patients, staff, and the screening programme. These were carried out in Spain and the USA and looked at the proportion of people getting an eye screening test, the proportion referred to secondary grading, the proportion of people referred to follow-up eye exams and the adherence to these follow-up exams. Because of how the studies were carried out, we cannot be sure if the results are what would happen if ARIAS would be introduced in the UK. No studies looked at any other outcomes of using ARIAS.

We found no UK-based studies that looked at whether ARIAS were good value for money.

The findings of the current review suggest that ARIAS that are as good as the first human grader could be tested in the NHS. This should look at how good ARIAS are in the real world and collect information about the effect of ARIAS on patients, staff, and costs.

Executive summary

Purpose of the review

The purpose of this review is to inform the UK National Screening Committee's (UK NSC) consideration of a major modification proposal to the UK Diabetic Eye Screening Programme (DESP). The proposal involved replacing the primary human grader in the current screening pathway with an automated retinal analysis system (ARIAS), with the aim of triaging patients prior to human grading. The review addressed 3 key questions:

- What is the diagnostic accuracy of Automated Retinal Image Analysis Systems (ARIASs) at detecting diabetic eye disease in patients with diabetes mellitus?
- What is the wider clinical impact of diabetic eye screening programmes using ARIASs for primary grading compared with screening programmes using fully manual grading, in terms of patient outcomes, service delivery, and grading workload?
- What is the cost-effectiveness of using ARIASs in diabetic eye screening programmes compared with programmes employing manual grading?

Background

Diabetic eye disease encompasses a range of conditions affecting people with diabetes mellitus (DM), the most common of which is diabetic retinopathy (DR). DR progresses through 3 stages: background DR, in which microaneurysms form and sometimes leak; pre-proliferative DR, involving more widespread vascular changes and bleeding; and proliferative DR, where ischaemia leads to the growth of fragile new vessels prone to bleeding and scarring, sometimes causing retinal detachment. Diabetic maculopathy occurs when DR affects the macula, causing swelling (macular oedema), resulting in blurred or distorted vision.

In the UK, diabetic eye screening aims to detect diabetic eye disease during the asymptomatic early stages, enabling timely treatment to reduce the risk of vision loss. Screening is offered every 1 to 2 years to individuals with DM aged 12 years and over, except women with gestational diabetes. There are some variations in the screening protocols and grading schemes used in the DESPs across the 4 UK nations. The DESPs in England, Wales and Northern Ireland involve taking 2-field 45° colour fundus photographs using pupil dilation (mydriasis), whereas the programme in Scotland takes 1-field images with mydriasis performed only if small pupils caused image quality to be inadequate.

The UK DESPs use a multi-level grading pathway: in England for instance, all images are assessed by primary human graders, with abnormal or ungradable images and a random 10% of

normal images reviewed by secondary graders; disagreements and complex cases are referred to arbitration graders. In Scotland, a machine-learning based ARIAS (iGradingM) triages image sets classified as ungradable or containing any DR for level 1 human grading. Images confirmed by the level 1 human grader as abnormal or ungradable are referred to level 2 graders who send images showing more-than-mild retinopathy or maculopathy to level 3 graders. The final grade determines screening intervals or referral to either Digital Surveillance clinics or the Hospital Eye Service.

Final screening grades used in the UK DESPs are:

- 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)
- ungradable images (U).

The large volume of images generated for grading across the UK, combined with rising diabetes prevalence, makes the DESPs labour-intensive and places increasing demands on the NHS. One possible solution to manage the volume of images is to use ARIAS. ARIASs are artificial intelligence-based algorithms designed to read retinal images and classify them as ‘disease’, ‘no disease’ or ‘ungradable’. ARIASs are already in use for primary grading in Scotland and Portugal and have been considered for clinical use in several other countries.

In 2024, the UK NSC received a proposal to modify the English DESP by replacing the primary human grader with an ARIAS, triaging patients into low-risk and high-risk cases. Images flagged as positive (high risk) or ungradable, along with a random 10% of negative (low risk) cases, would proceed to secondary grading by a human grader as well as arbitration grading, if necessary, as in the existing grading pathway. The proposal suggested that ARIAS use could reduce the number of human grading sessions required, lowering programme costs.

Focus of the review

The aim of the current review is to identify and synthesise evidence on a major modification proposal to the English DESP, which involves the use of an ARIAS for primary grading to triage patients prior to manual grading. While the proposal was initially developed for England, the UK NSC considers the evidence in the context of all 4 UK nations.

UK NSC external review – Automated grading in the Diabetic Eye Screening Programme, [Date of review completion]

The review aimed to address the following questions:

Question 1 (criteria 4 & 5): What is the diagnostic accuracy of ARIASs for detecting diabetic eye disease in patients with diabetes mellitus?

(The term 'diagnostic accuracy' in this review did not imply that ARIASs were used to diagnose diabetic retinopathy or any other condition. The only role of ARIASs investigated here was 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, as part of a multi-level screening programme such as the English DESP.)

Question 2 (criterion 11): What is the wider clinical impact of diabetic eye screening programmes using ARIAS for primary grading compared with diabetic eye screening programmes using fully manual grading?

Question 3 (criterion 14): What is the cost-effectiveness of ARIASs in diabetic eye screening programmes compared with diabetic eye screening programmes using manual grading?

All 3 questions were addressed through an evidence summary using the UK NSC evidence review approach. For question 1, the review prioritised UK-based studies and studies conducted in comparable populations and screening settings (Europe, USA, Canada, Australia and New Zealand), evaluating accuracy of commercially available and CE-marked or FDA-approved ARIASs compared to human grading. For question 2, the review prioritised primary studies conducted in the UK or UK-similar countries as defined above. For question 3, only UK-based economic evaluations were included.

A comprehensive literature search was conducted to address all 3 questions simultaneously, including searches of electronic databases, clinical trial registries, as well as examination of reference lists of included studies and relevant systematic reviews. The search focused on evidence published from 2020 onwards.

Recommendation under review

The UK NSC currently does not recommend the incorporation of automated grading into the UK DESPs. A previous evidence summary commissioned in 2021 to evaluate the use of ARIAS in the DESP concluded that, although ARIASs showed promise, the evidence base was insufficient to support widespread implementation of ARIAS at that time. Key limitations identified included: limited real-world evidence from large-scale clinical studies providing robust diagnostic accuracy data; highly context-specific results that could not be generalised beyond their evaluation settings; absence of randomised controlled trials or prospective cohort studies directly comparing clinical outcomes between manual and ARIAS-based grading pathways; limited and insufficient economic evaluations lacking long-term data and not accounting for the evolving na-

ture of AI technologies; and inadequately explored social and ethical implications, including acceptance by health professionals and patients. The 2021 review report stated there was evidence that the ARIAS EyeArt v2.1 was safe and cost-effective, but emphasised the need for further independent, high-quality research including diagnostic accuracy in relevant clinical contexts, clinical outcome comparisons, comprehensive economic analyses, and considerations of social acceptability before ARIAS could be confidently implemented in the UK DESPs. A subsequent UK NSC position statement on evidence requirements for ARIAS for diabetic eye screening[14] emphasised the importance of focussing on not missing referable retinopathy, and considering 3 types of evidence, retrospective accuracy evidence including ethnic, age, and sex biases, prospective evidence on workflow and workload (and associated modelling of cost savings) and prospective evidence on the wider clinical impact.

Findings and gaps in the evidence of this review

We included 72 articles for question 1 (diagnostic accuracy) and 11 articles for question 2 (wider clinical impact); no relevant study was identified for question 3 (cost effectiveness). Of these, 7 and 5 publications were prioritised for data extraction and analysis for question 1 and question 2, respectively, according to prioritisation rules pre-specified in the protocol.

Question 1

Seven studies from the UK (England, Scotland) or UK-similar countries (Czech Republic, Denmark, Germany, 2 studies from the US) evaluated the accuracy of single or multiple commercially available ARIASs as stand-alone grader compared to human grading (25 direct comparisons in total). **Table 1** shows how each of the ARIASs compared to human graders. Only one study (from England) provided comparative test accuracy data for detection of referable DR with additional information on performance by individual retinopathy grades and subgroups (age, ethnicity). Within this, 3 ARIASs (EyeArt v3.0.0, EyeCheckup AI, NEC) displayed sensitivity that was comparable to human graders for referable DR (including across age and ethnicity subgroups). However, specificity for each of these ARIASs was below that of human graders, suggesting more suitability to triaging which cases the primary grader reviews, or, if they are used as the primary grader, to measuring the impact on behaviour of a secondary grader caused by increased referrals from ARIAS.

Table 1. Question 1 - Summary of findings (25 direct comparisons, 7 prioritised studies)

ARIAS name / Manufacturer / Threshold (if applicable)	Study	Sensitivity similar or higher than human grading	Specificity similar or higher than human grading	Algorithmic fairness – Ethnicity	Algorithmic fairness – Age
Aireen (Aireen a.s.)	Sin 2025[60]	Any DR: Yes	No DR: Yes	NR	NR
RetinaLyze (RetinaLyze System A/S) (optimal threshold)	Nissen 2023[59]	Any DR: Yes	No DR: Yes	NR	NR
Masked ARIAS G	Lee 2021[58]	Any DR: Yes	No DR: Yes	NR	NR
EyeArt v3.0.0 (Eyenuk Inc.) (Threshold 1)	Rudnicka 2025[61]	RDR: Yes	NR	Sensitivity RDR: Yes	Sensitivity RDR: Yes
NEC (NEC Software Solutions) (Threshold 1)	Rudnicka 2025[61]	RDR: Yes	NR	Sensitivity RDR: Yes	Sensitivity RDR: Yes
EyeCheckup AI (EyeCheckup) (Threshold 1)	Rudnicka 2025[61]	RDR: Yes	NR	Sensitivity RDR: Yes	Sensitivity RDR: Yes
RetinaLyze (RetinaLyze System A/S) (lower threshold)	Nissen 2023[59]	Any DR: Yes	No DR: No	NR	NR
Masked ARIAS C, D, E, F (NR)	Lee 2021[58]	Any DR: Yes	No DR: No	NR	NR
IDx-DR (now: LumineticsCore) (Digital Diagnostics)	Dow 2023a[57]	RDR: Yes	Non-RDR: No	NR	NR
IDx-DR (now: LumineticsCore) (Digital Diagnostics)	Poschkamp 2025[62]	RDR: Yes	Non-RDR: No	NR	NR
RedCAD (Thirona) (manufacturer threshold)	Poschkamp 2025[62]	RDR: Yes	Non-RDR: No	NR	NR
RedCAD (Thirona) (Youden adjusted threshold)	Poschkamp 2025[62]	RDR: Yes	Non-RDR: No	NR	NR
RedCAD (Thirona) (customised referral threshold)	Poschkamp 2025[62]	RDR: Yes	Non-RDR: No	NR	NR
iGradingM (Medalytix) (stand-alone)	Fleming 2024[15]	Any DR: No	NA	NA	NA
iGradingM (Medalytix) (final DESP grade)	Fleming 2024[15]	Any DR: No	NA	NA	NA
DRISTi 2.0 (Artelus Ltd) (Threshold 1)	Rudnicka 2025[61]	RDR: No	NA	NA	NA

ARIAS name / Manufacturer / Threshold (if applicable)	**Study**	**Sensitivity similar or higher than human grading**	**Specificity similar or higher than human grading**	**Algorithmic fairness – Ethnicity**	**Algorithmic fairness – Age**
MONA (MONA.health) (Threshold 1)	Rudnicka 2025[61]	RDR: No	NA	NA	NA
OphtAI 2.3 (Evolucare Technologies SAS) (Threshold 1)	Rudnicka 2025[61]	RDR: No	NA	NA	NA
Remidio (Remidio Innovative Solutions Pvt Ltd) (Threshold 1)	Rudnicka 2025[61]	RDR: No	NA	NA	NA
Retmarker (Retmarker SA) (Threshold 1)	Rudnicka 2025[61]	RDR: No	NA	NA	NA
Masked ARIAS A and B (NR)	Lee 2021[58]	Any DR: No	NA	NA	NA

ARIAS, Automated retinal analysis system; DESP, Diabetic Eye Screening Programme; DR, Diabetic retinopathy; NA, Not applicable as sensitivity worse than human grading; non-RDR, Non-referable diabetic retinopathy (here: R0M0 or R1M0); NR, Not reported; RDR, Referable diabetic retinopathy.

Ethnic groups: White, Black, South Asian, Other / missing.
Age groups: <30 years, 30-45 years, 45-60 years, 60-75 years, 75+ years.

Risks of bias and applicability concerns were present in all studies. Only one prospective study was identified that integrated an ARIAS as primary grader triaging images for up to 3 levels of human grading. No study used ARIAS as primary grader, followed by human secondary and arbitration grading, as it would be used in the proposed UK screening practice, and then compared the accuracy of the final grade between pathways with human vs ARIAS-based primary grading.

Question 2

We did not identify any relevant RCTs from the UK or UK-similar countries comparing DESPs using an ARIAS as primary grader in a multi-level grading pathway to DESPs with fully manual grading in terms of clinical outcomes and overall impact. The 4 prioritised before-after studies (published in 5 articles) evaluated the effect of ARIAS use on screening rates, the proportion of people referred to secondary grading as well as to a follow-up eye examination, and follow-up appointment uptake. No study was identified that assessed potential patient health benefits (e.g. fewer people with vision loss, better health-related quality of life). Methodological flaws were present in all studies, for instance non-random sampling in all 4 studies, no concurrent control group in 2 studies, and no appropriate statistical analysis that controlled for confounding in 3 studies. In addition, the prioritised studies were not conducted within national or regional organised DESPs, they did not use a 3-level grading pathway comparable to the UK, and study populations were different to the UK in 2 US studies (for example, ethnicity); these limitations reduce the validity and reliability of the findings and limit generalisability to the UK.

Question 3

We did not identify any eligible evidence for this question.

Recommendations on screening

There is sufficient evidence to undertake an in-service evaluation in clinical practice. Two requirements identified in the UK NSC position statement on evidence requirements for ARIASs in diabetic eye screening[14] have been met, namely that ARIASs have shown sufficient sensitivity to accurately detect sight-threatening retinopathy, and UK-based retrospective research from one screening centre in North East London that has identified that some (but not all) ARIASs have high sensitivity for referable diabetic retinopathy without evidence of ethnic or age biases. However, there is no prospective study or evidence on the clinical and cost effectiveness of replacing the primary human grader with an ARIAS in a multi-level grading pathway. Therefore, ARIASs with sufficient sensitivity for referable DR and especially higher retinopathy grades compared to human grading, no ethnicity/age biases and relevant approval for use could be trialled in an in-service evaluation as primary grader or filter before the primary human grader to address the identified evidence gaps and to confirm satisfactory performance in a real 3-tier setting as well as generalisability of the findings to different screening centres and population subgroups.

Future research: in-service evaluation

ARIASs that have been identified in the reported studies to reach sufficient sensitivity for referable DR and especially higher retinopathy grades compared to human grading, that are without ethnicity/age biases and that have the required approval for use could be trialled in an in-service evaluation as primary grader or filter before primary human grading. In both use cases, the ARIAS triages high-risk images or ungradable images to subsequent reading by human graders. With ARIAS deployed in a primary grader role, no images would be assessed by primary human graders, and secondary human graders would only screen images that have been flagged by the ARIAS as 'Disease' or 'Ungradable', along with a small proportion of episodes classed as 'No disease' for quality assurance. With ARIAS deployed in a filter role before primary graders, only the subset of images flagged by ARIAS as 'Disease' or 'Ungradable', along with a small proportion of those images classed as 'No disease', would be seen by primary human graders.

The in-service evaluation should assess ARIAS performance when integrated into screening practice, effect on the next-level human grader behaviour and performance, the performance of the multi-level grading pathway as a whole, and whether ARIAS implementation reduces costs without impacting outcomes. As currently up to 90% of images sets classed as 'No disease' by the primary grader are not assessed by a second grader, the proportion of ARIAS screen-negative cases that go to next-level human grading for quality assurance could be (at least temporarily) increased as a safety measure to ensure no cases of sight-threatening DR are missed by the ARIAS. Disagreements between the ARIAS and next-level human graders that go to adjudication could be used to analyse error rates and type of errors of ARIAS compared to human graders. This in-service evaluation should test and report ARIAS performance by screening centre and in subgroups stratified by age, sex, and ethnicity to confirm that a satisfactory performance is reached across different screening centres and population subgroups (algorithmic fairness and equity). We recommend also evaluating feasibility of IT integration, acceptability to clinicians and patients, and the practical impact of incorporating ARIASs into screening practice, such as their impact on staff training, human grading workload, reporting times, number of referrals to follow-up eye examinations as well as referrals to treatment. Most ARIASs undergo regular update, which may involve changes in the AI-derived algorithm. Ongoing monitoring of ARIAS performance and service provision should therefore be implemented. Updated cost effectiveness analyses are rapidly needed to compare current practice with triage (primary grader or filter role) by different ARIASs.

Limitations

This review was conducted as a rapid evidence assessment, which involved methodological adjustments compared with a full systematic review. Only peer-reviewed publications in English were included, which may have led to the omission of some relevant publications in other lan-

guages or grey literature. These choices are standard in rapid review methodology and are unlikely to have excluded pivotal studies, especially as the searches covered the full period since the previous UK NSC external review. However, some uncertainty remains due to the limited availability of recent UK-specific data, particularly in relation to the wider clinical impact and cost-effectiveness of modern ARIASs. Screening, data extraction, and quality appraisal were primarily conducted by a single reviewer, with only 20% cross-checked, which increases the risk of error. Identified publications for questions 1 and 2 were prioritised for data extraction and analysis. A hierarchy of evidence was applied, meaning conclusions were drawn primarily from the most applicable studies. However, deprioritised studies may still contain useful information.

Evidence uncertainties

There remain important uncertainties in the evidence base for using ARIAS as primary grader in diabetic eye screening programmes. For question 1 (diagnostic accuracy), although several studies using real-world data compared commercially available ARIASs with human grading, many were conducted outside the UK. In these studies, the photographic protocols, human grader experience and speciality, as well as quality assurance procedures differed from UK DESPs, limiting generalisability. In the largest and most applicable study, which included 202,886 consecutive screening encounters from the North East London DESP, the reference standard was biased in favour of the primary human grader. Despite this, the ARIAS performance was still considered adequate for detecting moderate to severe diabetic eye disease.

For question 2 (wider clinical impact), there is very limited evidence comparing ARIAS-based pathways with fully manual pathways. The available studies are few, small, and methodologically weak, and no robust randomised controlled trials or large-scale prospective studies were found that directly assess patient-related outcomes, safety, or the impact on screening programmes and health professionals. Furthermore, evidence on key intermediate outcomes—such as how secondary graders respond to additional ARIAS referrals—is also limited, making it difficult to predict the downstream consequences for patients and screening services.

For question 3 (cost-effectiveness), there is an absence of contemporary UK-based economic evaluations published since 2020; therefore, this review cannot estimate cost-effectiveness or long-term resource impacts for the UK. Non-UK evaluations exist but were outside scope for inclusion in the present review.

Introduction and approach

Background

Diabetes

Diabetes mellitus (DM) is a chronic, metabolic condition characterised by elevated levels of blood glucose (hyperglycaemia). In people with DM, the body cannot produce enough of a hormone called insulin (which regulates the blood glucose levels), or the produced insulin is not effective. As a result, glucose cannot enter the cells and accumulates in the blood. There are two main types of DM: In 'Type 1 DM', the body's immune system attacks and destroys the cells in the pancreas that produce insulin. In 'Type 2 DM', the body cannot produce enough insulin, or it does not use the insulin properly. Other, less common, types include gestational diabetes (which develops during pregnancy), type 3c (due to damage to the pancreas), maturity-onset diabetes of the young (MODY), neonatal diabetes, and latent autoimmune diabetes in adults.[1] Over time, hyperglycaemia can lead to serious damage of the heart, blood vessels, eyes, kidneys and nerves.

According to the World Health Organization, the number of people living with DM globally rose from 200 million in 1990 to 830 million in 2020.[2] In 2022, the DM prevalence was 14% among adults aged 18 years and older, an increase from 7% in 1990. In 2021, DM and kidney disease due to DM caused over 2 million deaths worldwide. Additionally, around 11% of cardiovascular deaths were caused by hyperglycaemia.[2] In the UK, the number of adults (20 – 79 years) with diabetes rose from 1.5 million in 2000 to 4.5 million in 2024, and numbers are expected to increase to 4.9 million by 2050.[3]

Diabetic eye disease

Diabetic eye disease encompasses a group of eye problems that can affect people with DM. Of which, diabetic retinopathy (DR) is the most common. Hyperglycaemia caused by DM can damage the blood vessels in the back of the eyes (the retina). As a result, they can become blocked, leak, or grow incorrectly. If left undiagnosed and untreated, DR can cause progressive damage to the retina and lead to vision loss. DR develops in stages:[4,5]

- Stage 1 - Background DR involves tiny bulges (microaneurysms) in retinal blood vessels that may leak small amounts of blood but usually does not affect eyesight.
- Stage 2 - Pre-proliferative DR presents more severe and widespread changes in retinal blood vessels, including significant bleeding, with high risk of vision compromise.
- Stage 3 - Proliferative DR occurs when many retinal blood vessels become damaged or blocked, leading to insufficient blood supply (ischaemia). The body compensates by growing new, abnormal blood vessels that are weak and prone to bleeding, potentially causing retinal detachment and very high risk of irreversible vision loss.

There is also a type of DR called diabetic maculopathy, which affects the macula (the central part of the retina), leading to blurred or distorted central vision when blood vessels near this area leak or become blocked, causing fluid buildup and macular swelling (macular oedema).

In 2017 in the UK, the prevalence of DR, sight-threatening retinopathy, and diabetic maculopathy among people with DM aged 12 years and above were 33.8%, 12.3% and 7.9%, respectively.[6] DR or diabetic maculopathy was the second leading cause of blindness among working age adults in England and Wales in 2009/2010, causing 14.4% cases of blindness during that time period.[7]

Current policy context and previous reviews

Diabetic eye screening in the UK

Screening for DR aims to detect people with stage 2 (pre-proliferative) DR, stage 3 (proliferative) DR or with diabetic maculopathy during its common asymptomatic stage, so that treatment can be administered when it is most effective and the risk of vision loss can be reduced.[8] Diabetic Eye Screening programmes (DESPs) were introduced in England, Scotland and Wales in 2003[9] and in Northern Ireland in 2008. Screening is offered to individuals with DM aged 12 years and over (excluding women with gestational diabetes). The differences between DESPs across the UK 4 nations are summarised in **Table 2**.

Table 2. Differences between DESPs across the 4 UK nations
(modified from Zhelev et al. 2021[10])

	England	**Wales**	**Scotland**	**Northern Ireland**
Software	No national system, 2 suppliers	Single commissioned National system	Single commissioned National system	Single commissioned National system
Automation	None	None	Automated primary grading	None
Images	2-field	2-field	1-field	2-field
Mydriasis	Yes	Yes	Some cases	Yes
Extended screening intervals for people at lowest risk	Yes	Yes	Yes	Yes

In the English DESP for instance, the screening test involves taking two 45° digital retinal photographs per eye (one macula-centred and one disc-centred) taken with pupil dilation (mydriasis) that are manually graded, whereas the programme in Scotland takes 1-field images with mydriasis performed only if small pupils caused image quality to be inadequate. Possible grading results are reported in **Table 3**.[11]

Table 3. Retinal image grading in the English Diabetic Eye Screening Programme

Level of DR or maculopathy	Grade	Possible interpretation	
No observable retinopathy	R0	**No DR**	**No DR**
No maculopathy	M0		
Background retinopathy	R1	**Non-referable DR**	**Any DR**
Maculopathy – low risk or high risk	M1	**Referable DR**	
Pre-proliferative retinopathy - low risk (R2L) or high risk (R2H)	R2		
Proliferative retinopathy - active (R3A) or treated and stable (R3S)	R3		
Ungradable	U		

Table modified from Incubator for AI and Digital Healthcare (2025).[12]
DR, Diabetic retinopathy.

A simplified version of the multi-level manual grading pathway of retinal photographs[13] used in England is shown in **Figure 1**. All retinal images are assessed by primary graders. Images graded as R3, U or with other non-DR diseases identified go straight to the referral outcome grader. Secondary graders review all images graded as R1, R2 or M1 by primary graders. For quality control, secondary graders also review a random 10% of images that were graded as R0M0 by primary graders. Anyone receiving a R3, U or agreed R2 or M1 between primary and secondary graders is referred to a Digital Surveillance clinic for a rescreen after 3-6 months or the hospital eye service. Disagreements between primary and secondary graders (other than R3 and U) are assessed by arbitration graders. Based on the final grades - the specific level of DR (R0, R1, R2, or R3) and the presence/absence of diabetic maculopathy (M0 or M1) – the frequency of screening and the need for referral, respectively, are determined.[13] Patients with referable DR (R2, R3 and/or M1) or ungradable images (U) are reviewed by the Referral outcome grader to confirm referral to a Hospital Eye Service for specialist assessment and treatment if required.

Figure 1 depicts the flow of retinal images through the current manual multi-level grading pathway of the English diabetic eye screening programme.

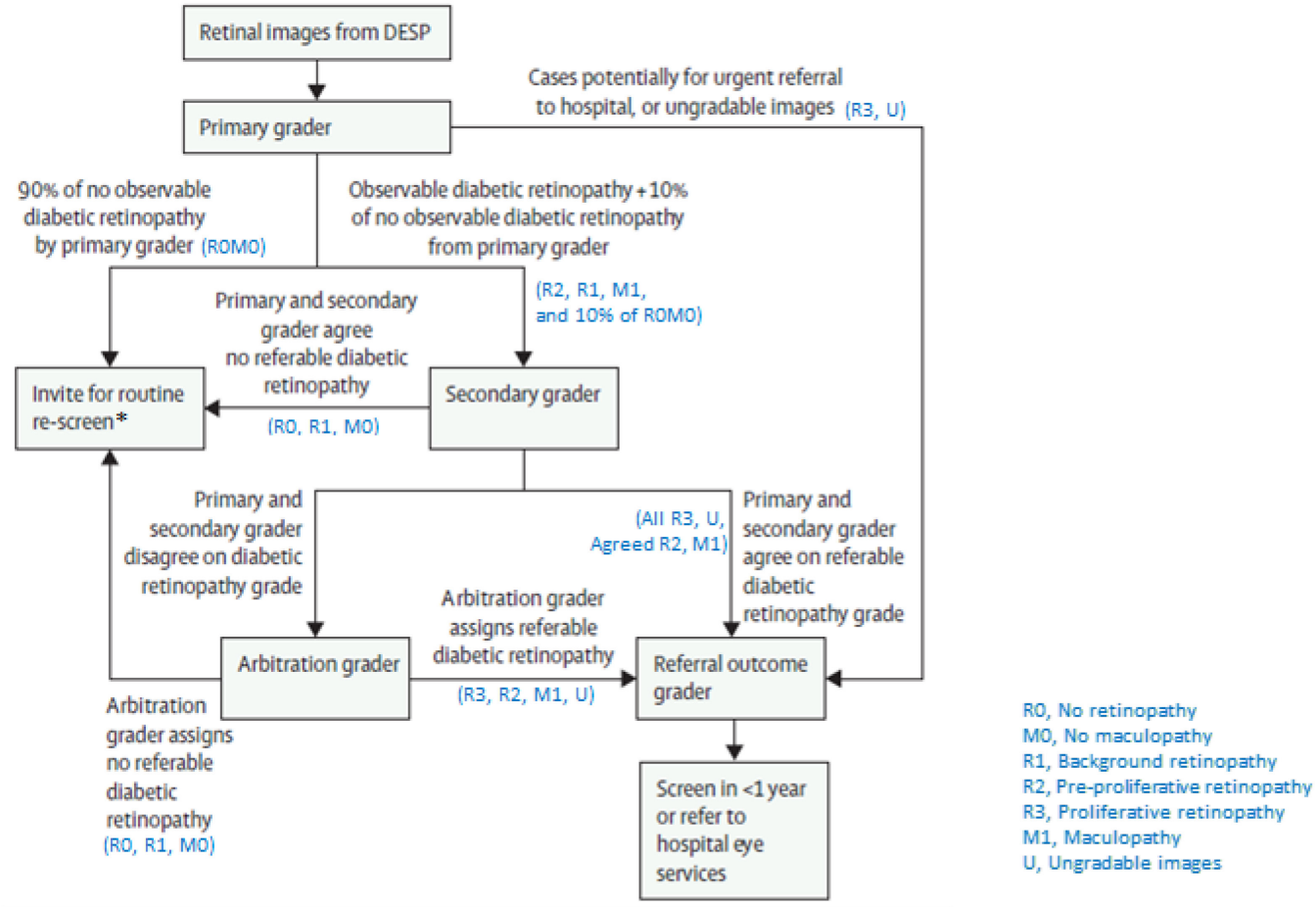


Figure 1. Current grading practice in the English DESP.
(modified from MacDonald et al. 2025[14])

* Annual rescreen after a single negative screening result (R0M0 or R1M0);
biennial rescreen after 2 consecutive negative screening results (R0M0 and R1M0).

Patients graded R0M0 (no DR) or R1M0 (non-referable DR) in the worse eye are invited to return for rescreening 12 months after a single negative result. After 2 consecutive negative screening results (R0M0 or R1M0), patients are invited for a rescreen after 24 months. Patients graded R2L, R3S and/or low risk M1 are referred to a Digital Surveillance clinic and are kept under surveillance and screened more frequently (every 3 to 6 months). Patients graded R3A or those who have higher risk maculopathy (M1) are referred to the Hospital Eye Service for management and treatment of DR. Patients with ungradable images (U) are referred to undergo a second test called slit lamp biomicroscopy.

In Scotland, a machine-learning based ARIAS (iGradingM) triages image sets classified as ungradable or containing any DR for level 1 human grading.[15] Images confirmed by the level 1 human grader as any DR or ungradable are referred to level 2 graders who send images showing more-than-mild retinopathy or maculopathy to level 3 graders. The final grade determines screening intervals or referral to either Digital Surveillance clinics or the Hospital Eye Service.

Just over 1 million people were screened in 2020/2021 in England; 1.49 million people were offered diabetic eye screening (DES), among whom the uptake was 67.9%.[16] As each screen creates at least 4 retinal images that require assessing by up to 3 trained human graders, the current English DESP is very labour intensive. This represents a major challenge for the NHS, especially as the numbers of retinal photographs that need grading are expected to escalate in the future since both prevalence and incidence of DM are expected to increase markedly.[17]

Automated Retinal Image Analysis Systems

Automated retinal image analysis systems (ARIASs) are artificial intelligence (AI)-based algorithms designed to read digital retinal images and classify them into ‘disease’, ‘no disease’ and ‘ungradable images’. AI broadly refers to “machines that perform tasks normally performed by human intelligence, especially when the machines learn from data how to do those tasks”.[18] Machine learning (ML) is a subset of AI, and deep learning (DL) is a subfield of ML. **Figure 2** depicts concentric circles showing how artificial intelligence, machine learning and deep learning topics sit within each other.

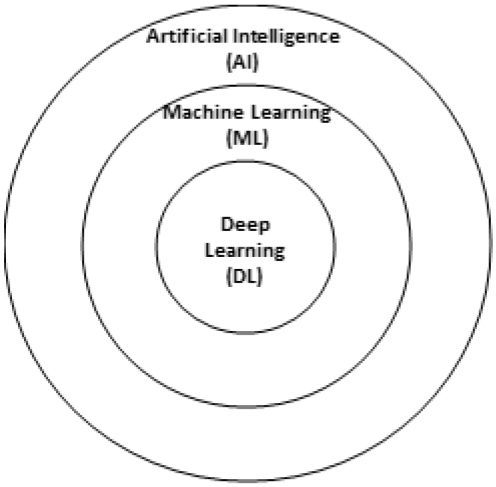


Figure 2. Hierarchical organisation of artificial intelligence, machine learning and deep learning

Non-ML-based algorithms (also called symbolic AI) extract pre-specified ‘hand-crafted’ features (e.g. microaneurysms) from the retinal photographs and use them to classify the image sets. Most of the recently developed ARIASs use ML-based algorithms that do not depend on pre-

specified features; instead, these algorithms use labelled input data and learn to distinguish 'disease' from 'no disease' images on their own. DL is a subset of ML that uses artificial neural networks with multiple layers to analyse data. DL algorithms are trained using large amounts of data and can learn and improve over time, becoming more accurate as they process more data.

The development of ML-based algorithms typically comprises 3 stages: training, fine-tuning (internal validation) and testing (external validation).[19] During internal validation, the AI algorithm is evaluated in the same dataset that was used for development and training, e.g. using cross validation or split sample validation. This can substantially overestimate the performance of the algorithm.[19] External validation studies, on the other hand, assess the AI performance with separate and independent data not used for model development. Here, studies using images collected from different sites (geographical validation) should be preferred as they allow an assessment of the generalisability of the ARIAS performance across potentially different technical parameters (such as different machines) and operating personnel.[14,19]

ARIASs have already been introduced in the DESPs in Scotland[20] and Portugal[21] and are now considered for clinical use in other countries.[22] In Scotland, a ML-based ARIAS (iGradingM) is used as primary grader, with images showing any DR or classed as ungradable being assessed by up to 3 levels of human grading.[15] In 2024, the UK National Screening Committee (UK NSC) received a proposal to modify the DESP by using an ARIAS for primary grading to triage patients into low-risk and high-risk cases (**Figure 3**). The proposal suggested that using ARIAS for DES would reduce human grader requirements, thereby reducing screening costs.

Under the proposed screening pathway with ARIAS as primary grader, any patient images denoted as ARIAS test-positive (or considered as technical failures) and a random 10% of those classed as ARIAS test-negative would proceed to secondary grading (**Figure 3**). The remaining pathway would then work exactly the same as for the manual grading arm.

Figure 3 depicts the flow of retinal images through the proposed multi-level grading pathway, replacing the primary grader with an ARIAS.

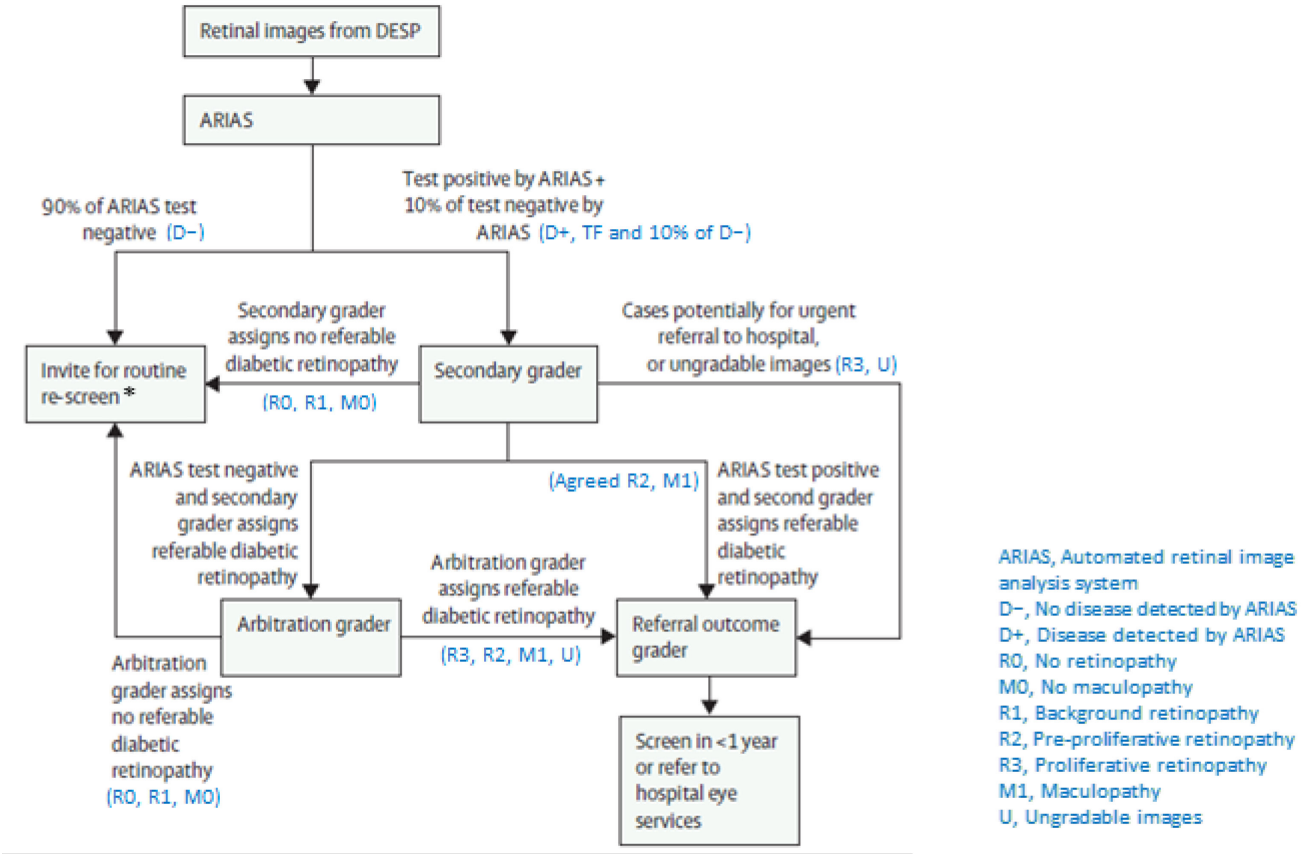


Figure 3. Proposed grading pathway replacing the primary grader with an automated retinal image analysis system (ARIAS) (modified from MacDonald et al. 2025[14])

* Annual rescreen after a single negative screening result (R0M0 or R1M0); biennial rescreening after 2 consecutive negative screening results (R0M0 and R1M0).

UK NSC recommendation

DESPs were implemented in England, Scotland and Wales in 2003,[9] and in Northern Ireland in 2008. The UK NSC currently does not recommend the incorporation of automated grading into the UK DESPs.

In 2025, the UK NSC published a position statement outlining its approach to evaluating the evidence required to support the introduction of AI technologies, specifically ARIAS, into the DESP.[14] The UK NSC recognised that ML-based ARIAS have demonstrated strong diagnostic performance and could potentially reduce workload and improve cost-effectiveness within the screening pathway. However, it concluded that current evidence remains limited with respect to their real-world clinical impact, implementation feasibility, and wider social and ethical implications. The UK NSC emphasised the need for robust, independent evidence – including large-

scale retrospective accuracy studies and well-designed prospective interventional trials - to establish the safety, effectiveness and fairness of ARIAS across diverse population subgroups. It also highlighted that any future implementation should demonstrate not only high diagnostic accuracy but also measurable benefits for patients, the screening service and the wider NHS, ensuring that adoption of AI is safe, equitable and delivers public value.

Subsequently, a Target Product Profile (TPP) for AI technologies in diabetic eye screening was developed in 2025 through a multi-stakeholder Delphi consensus process.[23] The TPP defines the essential and desirable specifications that ARIAS systems must meet for potential adoption within the English DESP. These include demonstrating a sensitivity for sight-threatening DR equal to or greater than the current state of the art, maintaining or improving clinical outcomes for people with DM compared with the existing screening pathway, and ensuring adequate diagnostic performance across key subgroups (age, gender, ethnicity, and diabetes type). Additional essential criteria include demonstrating cost-effectiveness, supporting compliance with NHS DESP service specifications and quality standards, and maintaining alignment with UK medical device regulations, including post-marketing surveillance, data governance, and cybersecurity requirements. Collectively, these specifications operationalise the UK NSC's evidence requirements and provide a transparent framework to guide the safe, effective, and equitable evaluation and potential implementation of ARIAS within the national screening programme.

Objectives

The aim of the current review was to identify and synthesise evidence on the major modification proposal to the English Diabetic Eye Screening Programme (EDESP), which involves the use of an Automated Retinal Image Analysis System (ARIAS) for primary grading to triage patients into low-risk and high-risk cases. Although the proposal related specifically to the EDESP, the UK NSC would consider the evidence reviewed here in the context of the whole UK.

The key questions the review attempts to answer are presented in **Table 4** below against the relevant UK NSC screening criteria.

Table 4. Key questions for the evidence summary and relationship to the UK NSC screening criteria

	Criterion	Key questions	Articles Included (Prioritised)
THE TEST			
4	There should be a simple, safe, precise and validated screening test.	Question 1: What is the diagnostic accuracy* of Automated Retinal Image Analysis Systems at detecting diabetic eye disease in patients with diabetes mellitus?	72 (7)
5	The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.		
THE SCREENING PROGRAMME			
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" (eg. 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 wider clinical impact of diabetic eye screening programmes with the use of Automated Retinal Image Analysis Systems for primary grading compared with diabetic eye screening programmes with fully manual grading?	11 (5)
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 (ie. 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. What is the cost-effectiveness of Automated Retinal Image Analysis Systems in diabetic eye screening programmes compared with diabetic eye screening programmes with manual grading?	0

* The term 'diagnostic accuracy' does not imply that the system is used to diagnose diabetic retinopathy or any other condition; the only use of automated retinal image systems investigated in the current review is as a first line screening test designed to identify patients with 'no disease' or 'non-referable disease', as part of a multi-level screening programme, such as the English Diabetic Eye Screening Programme.

Methods

The current review was conducted in keeping with the UK NSC evidence review process.[24] A comprehensive literature search was conducted to address all 3 questions simultaneously, including searches of electronic databases, clinical trial registries, as well as examination of reference lists of included studies and relevant systematic reviews. The search focused on evidence published from 2020 onwards. The original database searches were conducted on 17th July 2025; 4 update searches covered the period up to 2nd December 2025.

More details on the search strategy and other review methods (i.e., prioritisation of included studies, data extraction, analysis & synthesis) can be found in **Appendix 1** and **Appendix 4**, respectively.

Eligibility for inclusion in the review

The following review process was followed:

1. After removing duplicates, the records identified in the searches were imported into the Rayyan platform. Titles and abstracts of records identified by the searches were screened against the inclusion/exclusion criteria by one reviewer. A second independent reviewer assessed a random 20% sample of the titles/abstracts.
2. Full-text articles considered potentially relevant by either reviewer were retrieved.
3. Each full-text article was assessed against the eligibility criteria by one reviewer, who determined whether the article was relevant to one or more of the review questions. A second reviewer assessed a random 20% sample independently. Any disagreements were resolved by consensus, or through discussion with a third reviewer. Records rejected at full text stage were documented (including reasons for exclusion).

Inclusion criteria for each question are presented in **Table 5**.

Table 5. Inclusion criteria for the key questions

Key question	Population	Target condition	Intervention	Comparator (and Reference standard)	Outcome	Study type
1. What is the diagnostic accuracy* of ARIAS at detecting diabetic eye disease in patients with DM?	People with type 1 and type 2 DM ≥12 years, including rarer forms of DM such as MODY, who underwent standard fundus photography for DR screening	'Any DR' (R1 or higher and/or M1) or Referable DR' (R2, R3) and/or maculopathy (M1)	ARIAS, alone or as part of the workflow	Manual grading, no comparator, head-to-head comparison of ARIAS. Reference standard: Any	Overall accuracy measures for 'Any DR' (R1 or higher and/or M1; prioritised) or 'Referable DR' (R2 or higher and/or M1; deprioritised) and, if reported, for each grade of retinopathy	RCTs, prospective or retrospective single-gate studies, two-gate studies, systematic reviews** and meta-analyses (studies were prioritised by design, see **Appendix 4** for details)
2. What is the wider clinical impact of DESPs with the use of ARIAS for primary grading compared with DESPs with fully manual grading?	People with type 1 and type 2 DM ≥12 years, including rarer forms of DM such as MODY	NA	DESP for detecting DR and/or maculopathy using ARIAS on fundus photographs for primary grading followed by manual grading for secondary and arbitration grading	DESP for detecting DR and/or maculopathy using human manual grading on fundus photographs at all levels of grading	Any clinical utility outcomes, such as: Vision loss, health-related quality of life. Any patient management and practical implications outcomes, such as: Workforce (e.g. workload), inequalities.	RCTs, comparative prospective and retrospective cohort studies, and systematic reviews** and meta-analyses of these (studies were prioritised by design, see **Appendix 4** for details)
3. What is the cost-effectiveness of an ARIAS in DESP compared with DESP with manual grading?	People with type 1 and type 2 DM ≥12 years, including rarer forms of DM such as MODY; UK setting	NA	as in 2.)	as in 2.)	Any cost-effectiveness or modelled clinical outcomes	UK-based economic evaluations (of any type) and reviews of these

ARIAS, Automated Retinal Image Analysis System; DESP, Diabetic eye screening programme; DM, Diabetic mellitus; DR, Diabetic retinopathy; MODY, Maturity-onset diabetes of the young; NA, Not applicable; RCT, Randomised controlled trial.

* The term 'diagnostic accuracy' does not imply that the system is used to diagnose diabetic retinopathy or any other condition; the only use of automated retinal image systems investigated in the current review is as a first line screening test designed to identify patients with 'no disease' or 'non-referable disease', as part of a multi-level screening programme, such as the English Diabetic Eye Screening Programme.

** Systematic reviews will be defined as per Centre for Reviews and Dissemination (CDR) Database of Abstract of Reviews of Effects (DARE) criteria.[25]

Papers that fulfilled the following criteria were excluded:

Qualitative studies, studies without relevant outcomes, studies where more than 10% of the population did not meet the inclusion criteria or where outcomes are not reported separately, any algorithms in development, internal validations of ARIAS, articles not available in the English language, articles published prior to 2020, single case studies (one patient or one family), letters, reviews, editorials, communications, commentaries, conference abstracts and other grey literature, studies that have been retracted, studies where it could not be established if the inclusion criteria were met.

Publications were prioritised for data extraction and analysis for question 1 and question 2, respectively, according to prioritisation rules pre-specified in the protocol (i.e., Question 1: UK-based studies and studies conducted in UK-similar countries [”UK-similar countries”; e.g. North-Western European, USA, Canada, Australia], evaluating accuracy of commercially available and CE-marked or FDA-approved ARIASs compared to human grading. Question 2: primary studies conducted in the UK or UK-similar countries as defined above).

More details on the review methods (i.e., prioritisation of included studies, data extraction, analysis & synthesis) can be found in **Appendix 4.**

Appraisal for quality/risk of bias tool

The quality appraisal tools used for each study design are reported in **Table 6**. For question 1, quality appraisal of test accuracy studies was conducted using a modified QUADAS-2 tool[26] with signalling questions tailored for critical appraisal of machine learning studies.[27] Comparative test accuracy studies were additionally assessed using QUADAS-C.[28] For question 2, quality appraisal of before-after studies was conducted using Joanna Briggs Institute (JBI) Checklist for Quasi-Experimental Studies.[29] Quality appraisal for all prioritised studies was conducted by one reviewer, with a random 20% independently assessed by a second reviewer. Disagreements were resolved by discussion and, if necessary, by involving a third reviewer. No quality assessment was performed for deprioritised studies.

Table 6. Risk of bias/quality appraisal tools

Study design	Tool
Systematic reviews with / without meta-analysis	A MeaSurement Tool to Assess systematic Reviews (AMSTAR 2)[30]
Randomised controlled trials	Cochrane Risk of Bias 2 tool (RoB 2)[31]
Test accuracy studies	QUADAS-2[26] and QUADAS-C[28]
Quasi-experimental studies	Joanna Briggs Institute Checklist for Quasi-Experimental Studies[29]
Cohort studies	Joanna Briggs Institute Checklist for Cohort Studies[32]
Economic evaluations	Joanna Briggs Institute Checklist for Economic Evaluations[33]

Databases/sources searched

Systematic literature searches were undertaken using terms for diabetic retinopathy, screening and ARIAS which identified evidence for all review questions. The search strategy from the previous UK NSC review[10] was used as a starting point, with the addition of MeSH headings for Early Diagnosis and Mass Screening. The search was adapted to include 25 product and manufacturer names of commercially available ARIASs with regulatory approval as identified by a 2025 scoping review.[34] The search strategies were developed in MEDLINE (Ovid).

The search was adapted for Embase (OVID) and the Cochrane Library (Wiley). Embase and the Cochrane Library included the clinical trials harvested from ClinicalTrials.gov. Searches were also undertaken in the International Clinical Trials Registry Platform (ICTRP) database from the WHO. The search strategies used in each of the databases are provided in **Appendix 1**.

The search strategy comprised the following elements:

1. Searching of electronic bibliographic databases,
2. Searching of the ICTRP database,
3. Scrutiny of the references of included studies and relevant systematic reviews.

Database searches were run in Medline (OVID), Embase (OVID) and The Cochrane Library (Reviews only) and the ICTRP databases on the 17th July 2025. All searches were limited to English only. A total of 6,373 references were retrieved. After the removal of pre-June 2020 references and duplicate references, 4,493 papers were screened.

Update searches were run at the end of August, September, October and November, retrieving 269, 190, 115 and 122 additional references after deduplication, respectively.

Patient and Public Involvement and Engagement (PPIE)

A patient and public involvement and engagement (PPIE) group of 8 contributors was created to discuss review findings at strategic points in the process (see **Appendix 6**, GRIPP2 reporting for PPIE, for full details) and to provide feedback on the Plain English Summary. Group recruitment was targeted through known networks and aimed to include participants of different backgrounds and experiences with diabetes and diabetic eye screening. Four online meetings were conducted (September to November 2025), where the group discussed, (1) the background, rationale and search strategy, (2) results relating to questions 2 and 3, (3) results relating to question 1, and (4), summary of findings and consolidation of feedback. The PPIE group will also be updated on any feedback to this report.

Results

The original search addressing questions 1 to 3 identified 6,373 records of which 118 were published pre-June 2020 and 1,762 were excluded as duplicates, leaving 4,493 records that were screened at title and abstract level. Update searches were run on 29th August 2025, 3rd October 2025, 4th November 2025 and 2nd December 2025 retrieving 269, 190, 115 and 122 unique records, respectively (**Appendix 1, Table 17**). Across the 5 database searches combined, 243 records were selected for full text assessment, plus another 2 identified through other sources. After full text assessment, 72 articles were judged to be relevant to question 1, 11 articles for question 2 and no eligible articles were identified for question 3. Details of the selection process are provided in **Figure 12.**

For question 1, the review prioritised UK-based studies and studies conducted in comparable populations and screening settings (”UK-similar countries”; e.g. North-Western European, USA, Canada, Australia), evaluating accuracy of commercially available and CE-marked or FDA-approved ARIASs compared to human grading (7 studies). For question 2, the review prioritised primary studies conducted in the UK or UK-similar countries as defined above (4 studies published in 5 articles).

All publications included in the review and the questions they are relevant to are listed in **Appendix 2**, **Table 18.** Reasons for exclusion at full text level are provided in **Appendix 2**, **Table 19** (original search), **Table 20** (August update), **Table 21** (September update), **Table 22** (October update) and **Table 23** (November update). Reasons for deprioritisation of included studies are reported in **Appendix 3**, **Table 24** and **Table 25**.

We also identified 22 systematic reviews with aims similar to ours.[35-56] Since their inclusion criteria differed from ours (e.g., including studies published prior to 2020 or only including prospective studies or studies in real-world settings), we decided to use these systematic reviews for reference list screening only and to include their primary studies if eligible rather than to rely on the results from previous systematic reviews.

Question level synthesis

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

The previous review concluded that *"There is no high quality evidence for the accuracy of the other ARIASs* [i.e., other than EyeArt v2.1, iGradingM and RetmarkerSR] *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."*

This review therefore updates the published evidence on the diagnostic accuracy of ARIASs at detecting diabetic eye disease in patients with DM.

Eligibility for inclusion in the review

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

- Population: People with type 1 and type 2 DM ≥12 years, including rarer forms of DM such as MODY, who underwent standard fundus photography for DR screening.
- Target condition: 'Any DR' (R1 or higher and/or M1) or 'Referable DR' (R2, R3 and/or M1).
- Index tests: ARIAS, alone or as part of the workflow.
- Comparator: Manual grading, no comparator, head-to-head comparison of ARIASs.
- Reference standard: Any, as defined by the authors.

- Outcome measures: Overall accuracy measures for 'Any DR' (R1 or higher and/or M1; prioritised) or 'Referable DR' (R2 or higher and/or M1; deprioritised) and, if reported, for each grade of retinopathy.
- Study design: RCTs, prospective or retrospective single-gate studies, two-gate studies, systematic reviews and meta-analyses (studies were prioritised by design, see below and **Appendix 4**).

Papers that fulfilled the following criteria were excluded:

- Qualitative studies,
- studies without relevant outcomes,
- studies where more than 10% of the population do not meet our inclusion criteria and outcomes are not reported separately,
- any algorithms in development,
- internal validations of ARIAS,
- articles not available in the English language,
- articles published prior to 2020,
- single case studies (one patient or one family), letters, reviews, editorials, communications, commentaries, conference abstracts and other grey literature,
- studies that have been retracted,
- studies where it cannot be established if the inclusion criteria are met.

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

- ARIAS in the UK or a UK-similar country (North-Western European, USA, Canada, Australia),
- Performance of ARIAS compared to human grading,
- Commercially available ARIASs,
- With (or pending) CE-mark and/or FDA-approval.

The rest of the section is structured as follows:

1) we describe the volume and type of evidence relevant to this question,
2) we report the result of the quality appraisal (bias and applicability),

3) we summarise the findings from the prioritised studies reporting accuracy to detect 'Any DR',
4) we summarise findings from prioritised studies reporting accuracy to detect 'Referable DR',
5) we report subgroup analyses by retinopathy grade, ethnicity and age, and
6) we discuss the evidence reviewed in this section and state our conclusions regarding criteria 4 and 5.

The only role of ARIASs investigated in this review is to identify patients at low risk defined as 'no disease' or 'non-referable disease' thus distinguishing them from those at high risk defined as 'any disease' or 'referable disease', prior to manual grading and as part of a multi-level DESP.

Description of the evidence

Database searches (all searches combined) yielded 5,189 unique results, of which 70 were judged to be relevant to this question. Two relevant articles were identified through other sources, so 72 articles were ultimately included in this review.

Appendix 2 contains a full PRISMA flow diagram (**Figure 12**), along with a table listing the included publications and indicating which research questions each publication addresses (**Table 18**).

Of the 72 included studies, 7 studies from the UK or UK-similar countries that included commercially available ARIASs and a human comparator were prioritised and included in the narrative synthesis.[15,57-62] Characteristics and findings of deprioritised studies and reason for deprioritisation are reported in **Appendix 3**, **Table 24**.

Study characteristics of prioritised studies are reported in **Table 7** and in **Appendix 3**, **Table 24**. Four prioritised studies from Scotland, Denmark, the Czech Republic and the USA each compared the accuracy of a single commercially available ARIAS (Aireen[60], IDx-DR [now: LumineticsCore],[57] iGradingM[15] and RetinaLyze,[59] respectively) to human grading. The Scottish study[15] also reported the accuracy of the final DES grade (iGrading as primary grader followed by up to 3 levels of human grading). In this study, findings on the accuracy of a second, newly developed ARIAS (deep learning autograder, DLAG) were deprioritised as this ARIAS is not commercially available (see **Appendix 3**, **Table 24** for accuracy results of DLAG).

The remaining 3 studies compared multiple ARIASs to human grading: One study from Germany[62] compared the accuracy of IDx-DR (now: LumineticsCore) and RetCAD to ophthalmologist grading. One US study[58] compared the performance of 7 different ARIASs from 5 manufacturers (OphtAI, AirDoc, Eyenuk, Retina-AI Health, and Retmarker) to the original teleretinal grades (might include images graded by both an ARIAS and a human grader in a semiautomated fashion). The remaining study was performed within the EDESP (North East London

DESP) and compared the accuracy of 8 ARIASs (DRISTi 2.0, EyeArt v3.0.0, EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio, Retmarker) to the original primary human grade.[61]

In total, 25 unique direct ARIAS-human comparisons were performed across the 7 studies, with some studies evaluating an ARIAS at different thresholds, and 2 studies evaluating the same ARIASs (IDx-DR; now: LumineticsCore).[62,63] The identities of the 7 tested algorithms in the study by Lee et al.[58] were masked, but 3 manufacturers (OphtAI/Evolucare Technologies SAS, Eyenuk and Retmarker) overlapped with manufacturers that participated in the evaluation by Rudnicka et al.[61]

All studies were external validation studies: 6 studies used external, geographically separate data.[15,58,59,61-63] For one study[60] it was unclear if data were external temporal or external geographical. Four studies evaluated the ARIAS prospectively within DESPs,[15,60,62,63] with human grading performed at the same time,[60] retrospectively,[15,63] or the timing being unclear.[62] Poschkamp et al. evaluated a second ARIAS (RetCAD) retrospectively.[62] The other 3 studies evaluated ARIAS retrospectively on curated datasets from DESPs, usually compared to the original human grading.[58,59,61]

The study populations ranged from 744 images[15] to ~1.2 million images.[61] Retinopathy prevalence varied widely and was 14.8% in consecutive DM patients from Atlanta[58] and 90.7% in an enriched, quality assurance dataset from Scotland.[15] Only 2 studies[57,62] used the same photographic protocol (one macula-centred image and one optic disc-centred image for each eye without mydriasis); all other studies varied in the number of views and use of mydriasis.

Original human grading results from a DESP were used as a comparator in 3 studies,[58,59,61] whereas Fleming et al. used the external quality assurance performance from level 1 and level 2 graders of the Scottish DESP,[15] and Dow et al. used human consensus overread by 2 retina specialists with involvement of a 3rd or 4th ophthalmologist in case of disagreement as comparator.[57] In the study by Poschkamp et al. it is unclear if the ophthalmologist graded the images prospectively as part of clinical practice or retrospectively.[62] In the remaining study, accuracy of ARIAS was compared to retina specialists and general ophthalmologists (without retina subspeciality training) in a prospective, cross-sectional diagnostic study.[60]

The reference standard was majority consensus of at least 3 retinal specialists, ophthalmologists or similar experts in 3 studies.[15,58,59] The remaining 3 studies used the following reference standards: results of a follow-up in-person examination,[57] ophthalmologists' mydriatic fundoscopy with image analysis,[62] final human grade in the worst eye from up to 3 trained human graders within the NHS DES[61] and agreement between the 3 index tests, with only disagreements assessed by a 3-member high-level expert committee.[60]

Table 7. Study characteristics of prioritised studies – ARIAS vs human grading (7 studies)

Study	Country	Setting	Study design	1. Participants (Study) 2. Participants (Analysis) 3. DR prevalence	Photo-graphic protocol	Human comparator	Reference standard
Aireen							
Sin 2025[60]	Czech Republic	4 primary diabetes-care centres	External validation; prospective	1. 1,273 consecutive patients. 2. 1,154 patients. 3. 31.9% DR.	1-field, No mydriasis	1. General ophthalmologists – no subspeciality training in retina, 2. Retina specialists	Agreement between the 3 index tests (n = 734); Majority consensus of 3-member high-level expert committee (n = 420)
IDx-DR (now: LumineticsCore)							
Dow 2023a[57]	USA	DESP at 7 primary care sites in the San Francisco Bay Area (STATUS programme)	External geographical; ARIAS prospective; Human grading retrospective	1. 1,222 consecutive patients (12-months AI-phase) 2. ARIAS: 80 patients, Human: 122 patients. 3. ARIAS: 27.5% mtmDR (22/80); Human: 18.9% mtmDR (23/122)	2-field, No mydriasis	Human consensus overread: 2 fellowship-trained retina specialists at the Stanford Ophthalmic Reading Center, 3rd or 4th ophthalmologist for adjudication	In-person examination at Byers Eye Institute: dilated fundus examination, SD-OCT imaging, other retinal imaging. If diagnosis different to index test, a 2nd reader adjudicated the evaluation
iGradingM (stand-alone or as part of DES multi-level grading pathway)[a]							
Fleming 2024[15]	Scotland	External QA of Scottish DESP	External geographical; ARIAS prospective; Human grading retrospective	1. 744 images enriched for difficult and referable cases. 2. 743 images. 3. 90.7% DR	1-field, Reflex dilation only	Level 1 and 2 graders of Scottish DESP	Panel of 7 to 9 ophthalmologists, majority consensus
RetinaLyze							
Nissen 2023[59]	Denmark	National DESP (Steno Diabetes Center, North Jutland)	External geographical; ARIAS retrospective; Original human grading	1. 1,001 images, unclear how selected. 2. ARIAS: 967 images, Human: 964 images. 3. 58.2% DR.	5-field, With mydriasis	Routine grading 2019-2020; 10 ophthalmologist consultants and senior registrars	Majority consensus of 3 independent ophthalmologists

Study	Country	Setting	Study design	1. Participants (Study) 2. Participants (Analysis) 3. DR prevalence	Photo-graphic protocol	Human comparator	Reference standard
IDx-DR (now: LumineticsCore) and RetCAD							
Poschkamp 2025[62]	Germany	1 specialised diabetes centre (Klinikum Karlsburg)	External geographical; IDx-DR: prospective; RetCAD: retrospective; Human grading: unclear	1. 1,802 consecutive patients. 2. IDx-DR: 1,145 patients; RetCAD: 1,618 patients, Human: 1,217 patients. 3. 39.0% DR; 18.2% mtmDR.	2-field, No mydriasis	1 of 6 ophthalmologists (independent from fundoscopy analysis)	Ophthalmologist mydriatic fundoscopy with image analysis (1 of 6 experienced ophthalmologists)
7 different ARIASs from 5 manufacturers (OphtAI, AirDoc, Eyenuk, Retina-AI Health, Retmarker)[b]							
Lee 2021[58]	USA	2 U.S. Veterans Affairs hospitals (Seattle and Atlanta)	External geographical; ARIAS retrospective; Original human grading	1. 311,604 images from 23,724 consecutive patients. 2. 7,379 images from 735 encounters (enriched for DR and ungradable). 3. Seattle: 29.95% DR; Atlanta: 14.79% DR	≥4 images per eye, With mydriasis (Atlanta) or no routine mydriasis (Seattle)	Original teleretinal grades (2006 – 2018), no information on human graders, might include images graded by both an ARIAS and a human grader	Double-masked arbitration by 2 clinical experts, 3rd retina specialist for arbitration
DRISTi 2.0, EyeArt v3.0.0, EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio, Retmarker							
Rudnicka 2025[61]	England	North East London DESP (Homerton Healthcare NHS Foundation Trust)	External geographical; ARIAS retrospective; Original human grading	1. 202,886 consecutive encounters (any screening pathway), 126,365 patients. 2. 201,438 encounters. 3. 34.6% DR; 14.9% referable DR.	2-field, With mydriasis	NHS DESP primary graders	Final human grade in the worst eye per encounter (up to 3 trained human graders)

ARIAS, Automated Retinal Image Analysis System; DES, Diabetic Eye Screening; DESP, Diabetic Eye Screening Programme; DR, Diabetic retinopathy; mtmDR, More-than-mild diabetic retinopathy; NHS, National Health Service; QA, Quality assurance; SD-OCT, Spectral-domain optical coherence tomography; STATUS, Stanford Teleophthalmology Autonomous Testing and Universal Screening.

[a] Fleming et al.[15] also report on the accuracy of a newly developed deep learning autograder (DLAG) but this ARIAS was deprioritised as it was not commercially available (see Appendix 3, Table 24 for findings).

[b] One of the 7 ARIASs has FDA approval, 4 are in clinical use outside the U.S. and have been submitted to the FDA for approval, and several have a CE marking.

Methodological quality of the evidence

The methodological quality of the 7 prioritised comparative studies was assessed using tailored QUADAS-2[26] and QUADAS-C[28] tools. Assessment results are summarised in **Table 8**.

Risk of bias

Risk of bias was considered high in at least one of the 4 QUADAS-C domains in all 7 studies.

The study by Lee et al.[58] was rated as high risk of bias in the *flow & timing domain* as not all patients received a reference standard and only 7,379 of 311,604 images (enriched for DR and ungradable images) were included in the analysis.

Three studies were at high risk of bias in 2 QUADAS-C domains.[15,59,61] The study by Fleming et al.[15] did not enrol a consecutive or random sample of DM patients and used separate sampling schemes ("gates") for diseased (cases) and non-diseased individuals (controls) (*population domain*), and human graders' decisions were not made in a clinical practice context (*index test domain*). Nissen et al.[59] did not use a pre-specified ARIAS threshold (*index test domain*) and not all patients received a reference standard (*flow & timing domain).* In the study by Rudnicka et al.[61] the reference standard incorporated at least one of the index tests (*reference standard domain),* not all patients received a reference standard, and not all patients received the same reference standard (*flow & timing domain*).

Two studies were rated as being at high risk of bias in 3 QUADAS-C domains.[60,62] Poschkamp et al.[62] did not use a pre-specified threshold for one ARIAS (*index test domain*), the reference standard did not involve a panel of retinal specialists or ophthalmologists or similar experts independently reading the fundus images (*reference standard domain*) and fewer than 90% of patients were included in the analysis (*flow & timing domain*). In the study by Sin et al.,[60] human graders' decisions were not made in a clinical practice context (*index test domain*), the reference standard incorporated at least one of the index tests and did not involve a panel of retinal specialists or ophthalmologists or similar experts independently reading the fundus images (*reference standard domain*) and not all patients received the same reference standard (*flow & timing domain*). This study was also rated at high risk of bias in the added *role of sponsor domain* as one author received personal fees from an ARIAS vendor.

The remaining study by Dow et al.[57] was at high risk of bias in all 4 QUADAS-C domains as this study did not use a fully paired or randomised comparative design (*population domain*), human graders' decisions were not made in a clinical practice context (*index test domain*), the reference standard incorporated at least one of the index tests (*reference standard domain*), not all patients received a reference standard and fewer than 90% of patients were included in the analysis (*flow & timing domain*). This study was also rated at high risk of bias in the added *role of sponsor domain* as one author declared that they held stock from an ARIAS vendor.

Applicability concerns

Applicability concerns were considered high in at least one domain in all 7 studies (**Table 8**).

In 4 studies, applicability concerns were high in one of the 3 QUADAS-2 domains.[15,58,60,61] For 3 studies, there were significant concerns regarding the applicability of the research identified to the UK screening *population* as the photographic protocol (e.g. number of views, use of mydriasis) was different to that of the UK NHS DESPs[15,58,60] and the image dataset was enriched for referable and difficult examples.[15] In the study by Lee et al,[58] it is also unclear if the original teleretinal grades included images graded by both an ARIAS and a human grader in a semiautomated fashion. In the study by Rudnicka et al.,[61] applicability concerns regarding the *reference standard* were rated as high as the target condition was 'Referable DR' (R2 or higher and/or M1) and not 'Any DR' (R1 or higher and/or M1). The applicability of the *population* was rated as unclear as the included screening encounters were based on the whole programme (all screening pathways), and the generalisability of the North East London diabetic population to the rest of the UK was unclear.

In the study by Nissen et al.,[59] applicability concerns were high in 2 of the 3 QUADAS-2 domains. There were significant concerns regarding the applicability of the research identified to the UK screening *population* as the photographic protocol (e.g. number of views, use of mydriasis) was different to that of the UK NHS DESPs. The *index test domain* was rated as having high applicability concerns as there was no pre-specified ARIAS threshold.

Two studies had applicability concerns in all 3 QUADAS-2 domains.[62,63] In the studies by Poschkamp et al. and Dow et al., there were significant concerns regarding the applicability of the research identified to the UK screening *population* as the photographic protocol (e.g. number of views, use of mydriasis) was different to that of the UK NHS DESPs. Concerns regarding the applicability of the *index test(s)* to the situation in the UK were classified as high as there was no pre-specified ARIAS threshold in the study by Poschkamp et al.[62] and the human comparator was not applicable to UK screening practice (human consensus overread by 2 fellowship-trained retina specialists, with a 3rd or 4th ophthalmologist involvement to resolve disagreements) in the study by Dow et al.,[57] respectively. Concerns regarding the applicability of the *reference standard* were rated as high as the target condition was 'Referable DR' (R2 or higher and/or M1) and not 'Any DR' (R1 or higher and/or M1).

Taken together, the least biased and most applicable evidence comes from the studies by Lee et al.[58] (high risk of bias and applicability concerns in one domain each), Rudnicka et al.[61] and Fleming et al.[15] (high risk of bias in 2 domains and high applicability concerns in one domain). However, in the study by Rudnicka et al.,[61] the reference standard is biased towards the primary human grader as only a subset of images classed as no observable DR (R0M0) by the primary human grader were graded by a second grader, so ARIAS accuracy might have been underestimated and primary human grader's performance overestimated.

Table 8. Quality assessment results based on QUADAS-2 and QUADAS-C tools – ARIAS vs human grading (7 studies)

Study	Test	Risk of bias (QUADAS-2)					Applicability concerns (QUADAS-2)			Risk of bias (QUADAS-C)			
		P	I	R	FT	S	P	I	R	P	I	R	FT
Dow 2023[57]	IDx-DR	Low	Low	Unclear	High	High	High	Low	High	High	High	High	High
	Human	Low	High	Unclear	High		High	High	High				
Fleming 2024[15]	iGradingM	High	Low	Low	Low	Unclear	High	Low	Low	High	High	Low	Unclear
	Human	High	High	Low	Unclear		High	Low	Low				
Lee 2021[58]	ARIAS A	Low	Low	Low	High	Low	High	Low	Low	Low	Low	Low	High
	ARIAS B	Low	Low	Low	High		High	Low	Low				
	ARIAS C	Low	Low	Low	High		High	Low	Low				
	ARIAS D	Low	Low	Low	High		High	Low	Low				
	ARIAS E	Low	Low	Low	High		High	Low	Low				
	ARIAS F	Low	Low	Low	High		High	Low	Low				
	ARIAS G	Low	Low	Low	High		High	Low	Low				
	Human	Low	Low	Low	High		High	Unclear	Low				
Nissen 2023[59]	RetinaLyze	Unclear	High	Low	High	Unclear	High	High	Low	Unclear	High	Low	High
	Human	Unclear	Low	Low	High		High	Unclear	Low				
Poschkamp 2025[62]	IDx-DR	Low	Low	High	High	Low	High	Low	High	Low	High	High	High
	RetCAD[a]	Low	Low	High	High		High	Low	High				
	RetCAD[b]	Low	High	High	High		High	High	High				
	Human	Low	Unclear	High	High		High	Unclear	High				
Rudnicka 2025[61]	DRISTi 2.0	Low	Low	Low	High	Low	Unclear	Low	High	Low	Low	High	High
	EyeArt v3.0.0	Low	Low	Low	High		Unclear	Low	High				
	EyeCheckup AI	Low	Low	Low	High		Unclear	Low	High				
	MONA	Low	Low	Low	High		Unclear	Low	High				
	NEC	Low	Low	Low	High		Unclear	Low	High				
	OphtAI 2.3	Low	Low	Low	High		Unclear	Low	High				
	Remidio	Low	Low	Low	High		Unclear	Low	High				
	Retmarker	Low	Low	Low	High		Unclear	Low	High				
	Human	Low	Low	High	High		Unclear	Low	High				
Sin 2025[60]	Aireen	Low	Low	High	High	High	High	Low	Low	Low	High	High	High
	GO	Low	High	High	High		High	Unclear	Low				
	RS	Low	High	High	High		High	Unclear	Low				

ARIAS, Automated Retinal Image Analysis System; FT, Flow & timing; GO, General ophthalmologists; I, Index test; P, Population; R, Reference standard, RS, Retina specialists; S, Role of sponsor. [a] Manufacturer threshold. [b] Optimised threshold, not pre-specified.

Test accuracy to detect 'Any retinopathy'

According to a modified Delphi consensus study, the TPP for an ARIAS for use in the EDESP *"… must demonstrate a sensitivity for sight threatening diabetic retinopathy equal to or greater than that of the current state of the art, and a specificity for any diabetic retinopathy and un-gradable images resulting in improved service outcomes (see 'Cost effectiveness')."* (classed as 'essential') with *"… a sensitivity for any diabetic retinopathy greater than or equal to the current state of the art"* classed as 'desirable'.[23]

We assessed the overlap of the 95% confidence intervals (CIs) of ARIAS vs human graders in direct comparisons to investigate if ARIAS demonstrated similar or greater sensitivity and specificity than the human grader.

Comparative test accuracy to detect 'Any DR' was reported in 4 studies (**Table 9** and **Figure 4**). ARIAS sensitivity ranged from significantly less ($p \leq 0.001$) than 82.2% (ARIAS A and ARIAS B, exact numbers not reported)[58] to 96.8% (95%CI 93.3 – 98.2) (RetinaLyze, lower threshold),[59] whereas the sensitivity of the human grader ranged from 82.2% (95%CI 80.8 – 83.6)[58] to 95.8% (95%CI 95.6 – 96.0).[15] Across the 4 studies, the specificity of ARIAS to detect R0M0 ranged from 51.7% (95%CI 46.8 – 56.6) (RetinaLyze at lower threshold)[59] to 90.7% (88.7 – 92.7) (Aireen),[60] and the specificity of human graders ranged from 75.3 (95%CI 74.4 – 76.1)[15] to 85.3% (95%CI 81.8 – 88.6).[59]

<u>ARIAS compared to human graders</u>

ARIAS sensitivity for 'Any DR' was similar or higher than human grader's sensitivity for 7 ARIASs (9 comparisons, 3 studies) (**Figure 4** and **Table 9**):

- RetinaLyze (lower threshold),[59] RetinaLyze (optimal threshold),[59]
- Aireen (compared to retina specialists),[60] Aireen (compared to general ophthalmologists)[60] and
- masked ARIASs C, D, E, F and G from the study by Lee et al.[58]

Of these, 3 ARIASs (Aireen,[60] RetinaLyze [optimal threshold][59] and masked ARIAS G from the study by Lee et al.[58]) also demonstrated specificities equal or greater than human graders.

ARIAS sensitivity was lower than the human grader's sensitivity for 3 ARIASs (4 comparisons, 2 studies):

- iGradingM as stand-alone grader,[15]
- iGradingM as part of a multi-level grading pathway[15] and
- masked ARIASs A and B from the study by Lee et al.[58]

Figure 4 plots sensitivity (y-axis) against 1-specificity (x-axis) for the detection of 'Any DR' for the various ARIASs (triangle) and human grading (circle). Estimates from the same study are connected with a line. Estimates closer to the top and left corner mean higher true positive cases and lower false positive cases, respectively.

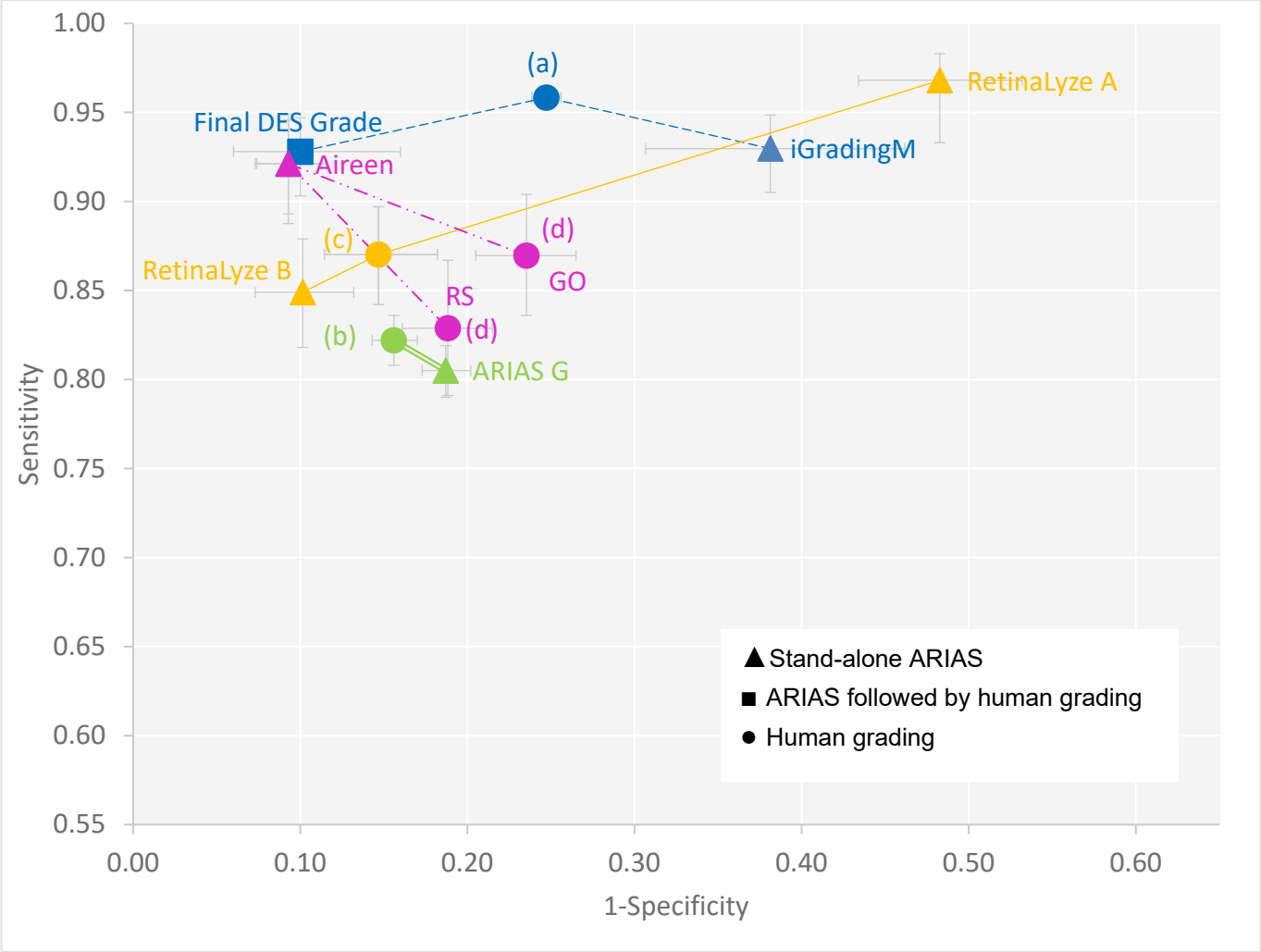


Figure 4. Comparative accuracy to detect 'Any DR' in prioritised studies – ARIAS vs human (4 studies)

Estimates connected with a line are from the same study. (a) Fleming 2024[15] (blue, – – –); (b) Lee 2021,[58] ARIAS G (light green, ═══), for ARIASs A-F, point estimates were not reported for both sensitivity and specificity; (c) Nissen 2023[59] (orange, ———); (d) Sin 2025[60] (purple, – · · –).

Error bars are 95% confidence intervals.

ARIAS, Automated retinal image analysis system; DES, Diabetic Eye Screening; DR, Diabetic retinopathy; GO, General ophthalmologists; RS, Retina specialists.

RetinaLyze A: Lower threshold (≥1 red lesion, combined image grading).

RetinaLyze B: Optimal threshold, not pre-specified (≥3 red lesions, combined image grading).

Table 9. Comparative diagnostic accuracy to detect 'Any DR' in prioritised studies – ARIAS vs human grading (4 studies)

Study, country, # analysis	Target condition	Index test	2x2 table				Sensitivity, % (95% CI)	Specificity, % (95% CI)	Exclusions from analysis / Ungradable images
			TP	TN	FP	FN			
Fleming 2024,[15] Scotland n = 743 / 46,443	Any DR: Any disease or ungradable	iGradingM	542	99	61	41	92.97 (90.50 – 94.85)	61.88 (53.84 – 69.33)	1 image missing for iGradingM; Unclear for human grading.
		Final DESP grade[a]	541	144	16	42	92.80 (90.31 – 94.70)	90.00^ (84.01 – 93.99)	
		All human grading	35,443	7,120	2,338	1,542	95.83 (95.62 – 96.03)	75.28 (74.40 – 76.14)	
Nissen 2023,[59] Denmark, ARIAS: n = 967 Human: n = 964	Any DR: Any disease	RetinaLyze[b]	545	209	195	18	96.8^ (93.3 – 98.2)	51.7 (46.8 – 56.6)	11 no majority consensus; 4 ungradable by retinal specialists; 13 ungradable by ARIAS; 6 ungradable by ARIAS and retinal specialists. + 3 routine grades N/A.
		RetinaLyze[c]	478	363	41	85	84.9^ (81.8 – 87.9)	89.9^ (86.8 – 92.7)	
		Routine grading	489	343	59	73	87.0 (84.2 - 89.7)	85.3 (81.8 – 88.6)	
Lee 2021,[58] USA, n = 7,379	Any DR: ICDR grades 1–4) or unreadable	ARIAS A	NR	NR	NR	NR	↓ than human (p<0.05)	90.0^ (88.9 – 91.1)	304,225 (97.6%) not selected for arbitration subset (no reference standard). Subset combined 1) consecutive sampling of images, and 2) 50 images from each retinopathy level and ungradable class according to original teleretinal grades.
		ARIAS B	NR	NR	NR	NR	↓ than human (p<0.05)	↓ than human (p<0.05)	
		ARIAS C	NR	NR	NR	NR	Similar to human (p>0.05)^	↓ than human (p<0.05)	
		ARIAS D	NR	NR	NR	NR	Similar to human (p>0.05)^	↓ than human (p<0.05)	
		ARIAS E	NR	NR	NR	NR	92.7^ (91.8 – 93.7)	↓ than human (p<0.05)	
		ARIAS F	NR	NR	NR	NR	92.7^ (91.8 – 93.7)	↓ than human (p<0.05)	
		ARIAS G	NR	NR	NR	NR	80.5^ (79.0 – 81.9)	81.3^ (79.8 – 82.7)	
		Original human grader	NR	NR	NR	NR	82.2 (80.8 - 83.6)	84.4 (83.0 – 85.7)	

Study, country, # analysis	Target condition	Index test	2x2 table				Sensitivity, % (95% CI)	Specificity, % (95% CI)	Exclusions from analysis / Ungradable images
			TP	TN	FP	FN			
Sin 2025,[60] Czech Republic, n = 1,154	Any DR: Any DR or ungradable	Aireen	339	713	73	29	92.1^ (89.3 – 94.9)	90.7^ (88.7 – 92.7)	119 (9.3%) non-gradable images of both eyes.
		GO	320†	601†	185†	48†	87.0 (83.6 – 90.4)	76.5 (73.5 – 79.5)	
		Retina specialists	305†	638†	148†	63†	82.9 (79.1 – 86.7)	81.2 (78.5 – 83.9)	

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; DESP, Diabetic Eye Screening Programme, DR, Diabetic retinopathy; FN, False negative; FP, False positive; GO, General ophthalmologist; ICDR, International Clinical Diabetic Retinopathy; N/A, Not available; NR, Not reported; TN, True negative; TP, True positive; U, Ungradable.

[a] iGradingM as primary grader followed by up to 3 levels of human grading.
[b] Low threshold: ≥1 red lesion, combined image grading.
[c] Optimal threshold, not pre-specified: ≥3 red lesions, combined image grading.
[d] Threshold 1: ARIAS test positive defined as disease present or ungradable/technical failure.
^ ARIAS sensitivity/specificity same or higher than human grader sensitivity/specificity.

Test accuracy to detect 'Referable retinopathy'

Three comparative studies did not report on the accuracy to detect 'Any DR' (prioritised outcome) but on the accuracy to detect 'Referable DR' (grades M1, R2 and R3 combined, or more-than-mild DR) only (**Figure 5** and **Table 10**).[57,61,62] A forest plot is provided in **Appendix 4**, **Figure 13**.

As for 'Any DR' above, we assessed the overlap of the 95% CIs of the ARIAS vs human graders in direct comparisons to investigate if ARIAS demonstrated a similar or greater sensitivity and specificity than the human grader. In addition, the NHS quality assurance standards specify that a national threshold less than 85% for sensitivity and less than 80% specificity for referable DR (R2, R3, M1) would attract a warning flag.[64] We therefore assessed which ARIAS achieved a minimum sensitivity and specificity of 85% and 80%, respectively.

Sensitivity for 'Referable DR' ranged from 83.7% (95%CI 83.2 – 84.1) to 98.7% (95%CI 98.5 – 98.8) across the 8 ARIASs in the study by Rudnicka et al.[61] and was 95.5% (95%CI 86.7 – 100)[57] and 90.2% (95%CI 84.7 – 93.9),[62] respectively, for IDx-DR. RetCAD achieved a sensitivity of 70.4% (95%CI 64.5 – 75.7) using the pre-specified manufacturer threshold and higher sensitivities of 86.9% (95%CI 82.1 – 90.6) and 93.6% (95%CI 89.8 – 96.1), respectively, with customised referral threshold modifications.[62] Human grader's sensitivity was 69.6% (95%CI 50.7 – 88.3),[57] 78.2% (95%CI 71.6 – 83.7),[62] and 94.6% (95%CI 94.3 – 94.9).[61]

Specificity of primary human graders for no or mild DR was 97.0% (95%CI 93.5 – 100)[57] and 99.4% (95%CI 98.7 – 99.8)[62] and was not reported for the study by Rudnicka et al.[61] ARIAS specificity varied widely and ranged from 30.5% (95%CI 30.3 – 30.7)[61] to 93.86% (95%CI 92.41 – 95.05).[62]

<u>ARIAS compared to human graders</u>

ARIAS sensitivity for 'Referable DR' was similar or higher than human grader's sensitivity for 5 ARIASs (3 studies) (**Figure 5** and **Table 10**):

- IDx-DR,[57,62]
- RetCAD with manufacturer threshold,[62] RetCAD with Youden adjusted threshold,[62] RetCAD with customised referral threshold,[62] and
- EyeArt v3.0.0, EyeCheckup AI and NEC.[61]

Of these, none demonstrated specificities equal or greater than human graders. To note, specificity of the human grader in the study by Rudnicka et al.[61] was not reported.

ARIAS sensitivity for 'Referable DR' was lower than the human grader's sensitivity for 5 ARIASs (1 study) (**Figure 5** and **Table 10**):

- DRISTi 2.0, MONA, OphtAI 2.3, Remidio and Retmarker.[61]

ARIAS performance compared to minimum quality assurance standards

A minimum sensitivity of 85% and minimum specificity of 80% was achieved by 3 ARIASs (2 studies) (**Figure 5**, top and left of the dashed lines, and **Table 10**):

- IDx-DR in one study[62] but not in a second study,[57]
- Remidio,[61]
- RetCAD at the Youden adjusted threshold.[62]

Figure 5 plots sensitivity (y-axis) against 1-specificity (x-axis) for the detection of 'Referable DR' for the various ARIASs (triangle) and human grading (circle). Estimates from the same study are connected with a line. Estimates closer to the top and left corner mean higher true positive cases and lower false positive cases, respectively.

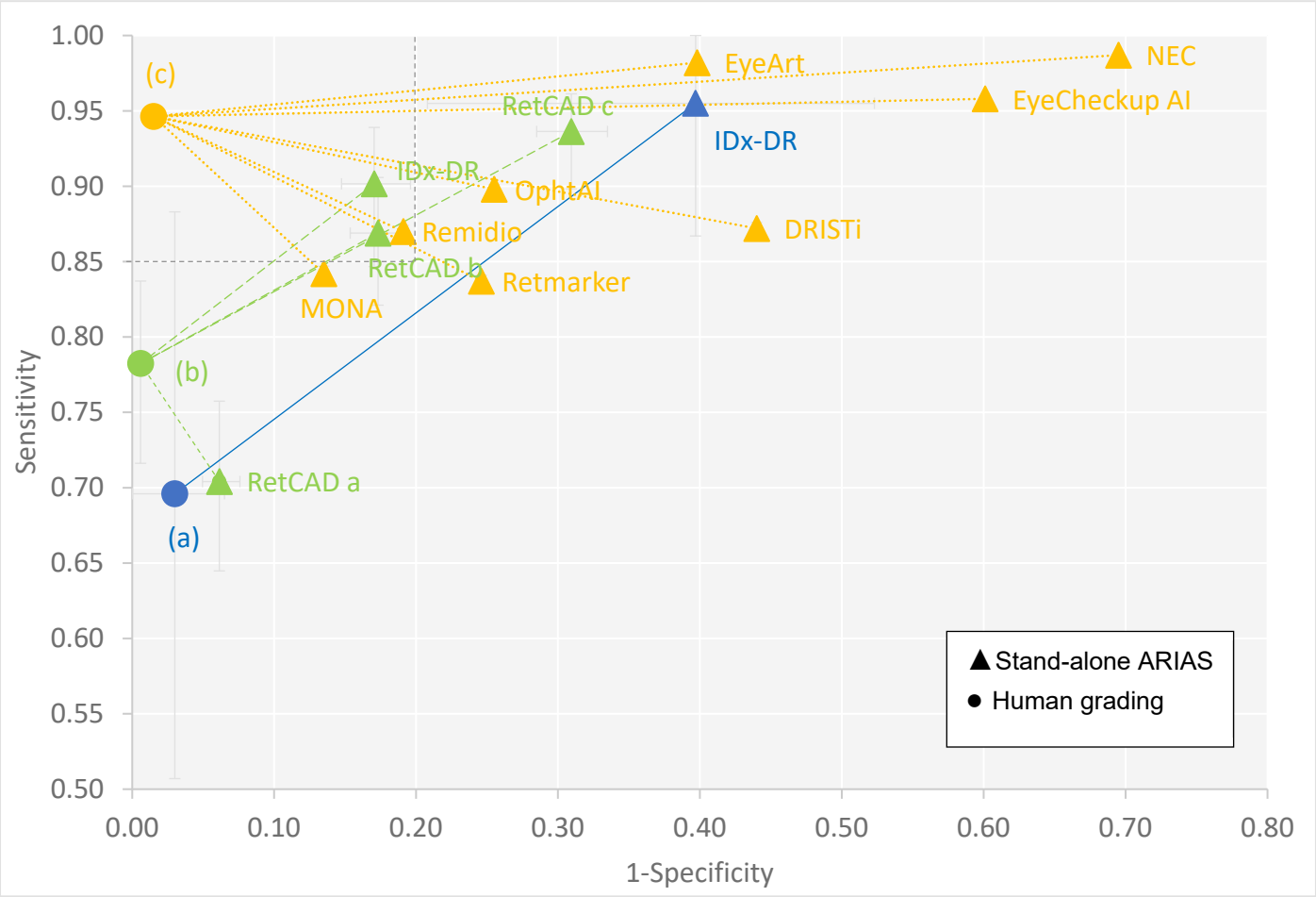


Figure 5. Comparative accuracy to detect 'Referable DR' in prioritised studies – ARIAS vs human grading (3 studies)

Estimates connected with a line are from the same study: (a) Dow 2023a,[57] (blue; solid line); (b) Poschkamp 2025[62] (green; dashed line); (c) Rudnicka 2025[61] (orange, dotted line).
Error bars are 95% confidence intervals.
Grey dashed line is at 85% sensitivity and 80% specificity, respectively (quality assurance standards for human graders in the English DESP[64]).
ARIAS, Automated retinal image analysis system; DR, Diabetic retinopathy.
RetCAD a, Manufacturer threshold (≥2); RetCAD b, Youden adjusted threshold (>1.02);
RetCAD c, Greifswald modification (>0.1).

Table 10. Comparative diagnostic accuracy to detect 'Referable DR' in prioritised studies – ARIAS vs human grading (3 studies)

Study, country, # analysis	Target condition	Index test	2x2 table				Sensitivity, % (95% CI)	Specificity, % (95% CI)	Exclusions from analysis / Ungradable images
			TP	TN	FP	FN			
Dow 2023a[57], USA, IDx-DR: n = 80; Human: n = 122	Referable DR: mtmDR	IDx-DR	21	35	23	1	95.5^ (86.7 – 100)	60.3 (47.7 – 72.9)	1,042 from AI phase who did not have in-person examination (mostly ARIAS-negative). ARIAS: 100 without AI evaluation; Human consensus: 58 without human consensus.
		Human consensus overread	16	96	3	7	69.6 (50.7 – 88.3)	97.0 (93.5 – 100)	
Poschkamp 2025[62], Germany, IDx-DR: n = 1,145; RetCAD: n = 1,618; Human: n = 1,217	Referable DR: mtmDR	IDx-DR	165†	798†	164†	18†	90.16^ (84.67 – 93.90)†	82.95 (80.39 – 85.25)†	11 declined permission; 75 no fundoscopy examination by ophthalmologist; 98 image acquisition unsuccessful; IDx-DR: 35 incomplete images; IDx-DR: 438 not analysable; Ophthalmologist: 401 not analysable.
		RetCAD[a]	188†	1,268†	83†	79†	70.41^ (64.48 – 75.74)†	93.86 (92.41 – 95.05)†	
		RedCAD[b]	232†	1,117†	234†	35†	86.89^ (82.11 – 90.58)†	82.68 (80.53 – 84.64)†	
		RetCAD[c]	250†	933†	418†	17†	93.63^ (89.82 – 96.13)†	69.06 (66.51 – 71.50)†	
		Ophthalmologist	151†	1,018†	6†	42†	78.24† (71.62 – 83.71)†	99.41 (98.66 – 99.76)†	

Study, country, # analysis	Target condition	Index test	2x2 table				Sensitivity, % (95% CI)	Specificity, % (95% CI)	Exclusions from analysis / Ungradable images
			TP	TN	FP	FN			
Rudnicka 2025[61], England, n = 195,816	Referable DR (M1, R2 and R3 combined, excluding U)	DRISTi 2.0, Threshold 1[d]	NR	NR	NR	NR	87.2 (86.8 – 87.6)	56.0 (55.8 – 56.2)†	Total 7,070: 39 camera test encounters; 12 did not have human grading outcomes; 55 were repeat encounters on the same day; 1,342 omitted the code for 'one-eye encounters'; 5,622 with U grade.
		EyeArt v3.0.0, Threshold 1[d]	NR	NR	NR	NR	98.2^ (98.1 – 98.4)	60.2 (59.9 – 60.4)†	
		EyeCheckup AI, Threshold 1[d]	NR	NR	NR	NR	95.8^ (95.5 – 96.0)	39.9 (39.6 – 40.1)†	
		MONA, Threshold 1[d]	NR	NR	NR	NR	84.2 (83.7 – 84.7)	86.5 (86.3 – 86.6)†	
		NEC, Threshold 1[d]	NR	NR	NR	NR	98.7^ (98.5 – 98.8)	30.5 (30.3 – 30.7)†	
		OphtAI 2.3, Threshold 1[d]	NR	NR	NR	NR	89.8 (89.4 – 90.1)	74.5 (74.3 – 74.7)†	
		Remidio, Threshold 1[d]	NR	NR	NR	NR	87.0 (86.6 – 87.4)	80.9 (80.7 – 81.0)†	
		Retmarker, Threshold 1[d]	NR	NR	NR	NR	83.7 (83.2 – 84.1)	75.4 (75.1 – 75.6)†	
		Primary human grader[e]	NR	NR	NR	NR	94.6 (94.3 – 94.9)	NR	

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; DR, Diabetic retinopathy; FN, False negative; FP, False positive; M0, No observable diabetic maculopathy; M1, Any diabetic maculopathy; mtmDR, More-than-mild diabetic retinopathy; NR, Not reported; R0, No observable retinopathy; R1, Mild non-proliferative retinopathy; R2, Moderate-to-severe non-proliferative (pre-proliferative) retinopathy; R3, Proliferative retinopathy; TN, True negative; TP, True positive; U, Ungradable.

[a] Manufacturer threshold (≥2). [b] Youden adjusted threshold (>1.02). [c] Customised referral threshold modification (>0.1). [d] Threshold 1: ARIAS test positive defined as disease present or ungradable/technical failure. [e] Classified as referable diabetic retinopathy by primary human grader.

† Numbers were calculated by reviewers. For Poschkamp et al. (2025),[62] 95% CIs were calculated using Wilson score interval with continuity correction. For Rudnicka et al. (2025),[61] 95% CI of specificity = 100% minus the reported 95% CIs for the false positive rate (%).

^ ARIAS sensitivity/specificity at least human grader sensitivity/specificity.

Sensitivity to detect specific grades of retinopathy

Three studies reported sensitivity by specific grade of retinopathy.[58,61,62] All studies compared multiple commercially available ARIASs to human graders' performance (**Table 11**).

As above, we assessed the overlap of the 95% CIs of the ARIAS vs human graders in direct comparisons to investigate if ARIAS demonstrated a similar or greater sensitivity than the human grader. We also assessed which ARIAS achieved a minimum sensitivity of 85%.

ARIAS sensitivity compared to human graders

For moderate non-proliferative DR or worse, ARIAS sensitivity was similar to that of human graders for 3 ARIASs (1 study) (**Table 11**):

- masked ARIASs E, F and G in the study by Lee et al.[58]

Lower sensitivities were observed for 4 ARIASs (1 study):

- masked ARIASs A, B, C and D in the study by Lee et al.[58]

For moderate-to-severe non-proliferative DR (R2M0/M1), ARIAS sensitivity was similar or higher than human grader's sensitivity for all 8 ARIASs (1 study) (**Table 11**):

- DRISTi 2.0, EyeArt v3.0.0, EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio and Retmarker.[61]

For severe non-proliferative DR or worse, ARIAS sensitivities were similar of higher than human graders' for 7 ARIASs (2 studies) (**Table 11**):

- masked ARIASs C, D, E, F and G in the study by Lee et al.[58]
- IDx-DR[62] and
- RetCAD at manufacturer threshold, Youden adjusted threshold and customised referral threshold, respectively.[62]

Lower sensitivities were observed for 2 ARIASs (1 study):

- masked ARIASs A and B in the study by Lee et al.[58]

For proliferative DR (R3M0/M1), ARIAS sensitivities were similar or higher than human graders' for 14 ARIASs (2 studies) (**Table 11**):

- masked ARIASs B, C, D, E, F and G in the study by Lee et al.[58] and
- DRISTi 2.0, EyeArt v3.0.0, EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio and Retmarker.[61]

A lower sensitivity was observed for one ARIAS (1 study):

- masked ARIAS A in the study by Lee et al.[58]

ARIAS performance compared to minimum quality assurance standards

A minimum sensitivity of at least 85% for the evaluated retinopathy grades (ranging from moderate non-proliferative DR to proliferative DR) was achieved by 16 ARIASs (3 studies) (**Table 11**, point estimates were not reported for all ARIASs in the study by Lee et al.[58] but were estimated from a figure in the article):

- masked ARIASs B, C, D, E, F and G from the study by Lee et al.,[58]
- IDx-DR,[62]
- RetCAD at Youden adjusted threshold and customised referral threshold[62] and
- DRISTi 2.0, EyeArt v3.0.0, EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio and Retmarker.[61]

Table 11. Sensitivity to detect specific grades of retinopathy (3 studies)

Study, country,	DR grade	Index test	DR cases (n)	Sensitivity, % (95% CI)
Lee 2021,[58] USA	Moderate NPDR or worse	ARIAS A	NR	Lower than human (p<0.05)
		ARIAS B	NR	Lower than human (p<0.05)
		ARIAS C	NR	Lower than human (p<0.05)
		ARIAS D	NR	Lower than human (p<0.05)
		ARIAS E	NR	Similar to human (p>0.05)^
		ARIAS F	NR	Similar to human (p>0.05)^
		ARIAS G	NR	Similar to human (p>0.05)^
		Human grader	NR	100.00 (NR)
	Severe NPDR or worse	ARIAS A	NR	Lower than human (p<0.05)
		ARIAS B	NR	Lower than human (p<0.05)
		ARIAS C	NR	Similar to human (p>0.05)^
		ARIAS D	NR	Similar to human (p>0.05)^
		ARIAS E	NR	Similar to human (p>0.05)^
		ARIAS F	NR	Similar to human (p>0.05)^
		ARIAS G	NR	Similar to human (p>0.05)^
		Human grader	NR	100.00 (NR)
	PDR	ARIAS A	NR	74.42 (NR) (p<0.05)
		ARIAS B	NR	Similar to human (p>0.05)^
		ARIAS C	NR	Similar to human (p>0.05)^
		ARIAS D	NR	Similar to human (p>0.05)^
		ARIAS E	NR	Similar to human (p>0.05)^
		ARIAS F	NR	Similar to human (p>0.05)^
		ARIAS G	NR	Similar to human (p>0.05)^
		Human grader	NR	100.00 (NR)
Poschkamp 2025,[62] Germany	Severe NPDR or worse	IDx-DR	78	93.59 (85.03 – 97.62)†^
		RetCAD, threshold 1[a]	105	62.86 (52.83 – 71.93)†^
		RetCAD, threshold 2[b]	105	90.48 (82.78 – 95.09)†^
		RetCAD, threshold 3[c]	105	97.14 (91.27 – 99.26)†^
		Ophthalmologist	83	77.11 (66.34 – 85.32)†
Rudnicka 2025,[61] England	Moderate-to-severe NPDR (R2M0 + R2M1)	DRISTi 2.0[d]	3,881	98.0 (97.5 - 98.4)^
		EyeArt v3.0.0[d]	3,881	99.8 (99.6 - 99.9)^
		EyeCheckup AI[d]	3,881	99.6 (99.4 - 99.8)^
		MONA[d]	3,881	98.4 (98.0 - 98.8)^
		NEC[d]	3,881	99.8 (99.7 - 99.9)^
		OphtAI 2.3[d]	3,881	98.5 (98.1 - 98.9)^
		Remidio[d]	3,881	98.9 (98.5 - 99.2)^
		Retmarker[d]	3,881	96.7 (96.1 - 97.2)^
		Human grader[e]	3,881	95.4 (94.7 - 96.1)
	PDR (R3M0 + R3M1)	DRISTi 2.0[d]	1,978	98.7 (98.1 - 99.1)^
		EyeArt v3.0.0[d]	1,978	99.4 (98.9 - 99.7)^
		EyeCheckup AI[d]	1,978	97.5 (96.7 - 98.1)^
		MONA[d]	1,978	96.8 (95.9 - 97.5)^

Study, country,	DR grade	Index test	DR cases (n)	Sensitivity, % (95% CI)
		NEC[d]	1,978	99.5 (99.1 - 99.8)^
		OphtAI 2.3[d]	1,978	95.8 (94.8 - 96.6)^
		Remidio[d]	1,978	97.1 (96.2 - 97.8)^
		Retmarker[d]	1,978	97.7 (96.9 - 98.3)^
		Human grader[e]	1,978	95.8 (94.8 - 96.6)

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; DR, Diabetic retinopathy; NPDR, Non-proliferative diabetic retinopathy; M0, No observable diabetic maculopathy; M1, Any diabetic maculopathy; NR, Not reported; PDR, Proliferative retinopathy; R2, Moderate-to-severe non-proliferative (pre-proliferative) retinopathy; R3, Proliferative retinopathy.

[a] Manufacturer threshold (≥2).

[b] Youden adjusted threshold (>1.02).

[c] Customised referral threshold modification (>0.1).

[d] Threshold 1 = ARIAS test positive defined as disease present or ungradable/technical failure.

[e] Classified as referable diabetic retinopathy by primary human grader.

† Numbers were calculated by reviewers. 95% CIs were calculated using Wilson score interval with continuity correction.

^ Sensitivity similar or higher than human grader.

Subgroup analyses by ethnic group

One study reported accuracy of ARIAS and human graders by ethnic group (that is, White, Black, South Asian, Other / missing).[61] Algorithmic fairness was assessed by graphically examining and summarising the absolute percentage differences in ARIAS and primary human grader performance across ethnic groups.

Figure 6 on the next page displays the sensitivity (top) and specificity (bottom) (point estimates with 95% CIs) of the 8 ARIASs and the primary human grader for the detection of 'Referable DR' overall and by 4 ethnic groups (White, Black, South Asian, Other / missing).

Across the 4 ethnic subgroups, primary human graders' sensitivity for 'Referable DR' (R2, R3 and/or M1) ranged from 93.5% (95%CI 92.9 - 94.2) to 95.3% (95%CI 94.5 - 96.1) and was homogenous across ethnic subgroups.

For the 8 ARIASs, sensitivity for 'Referable DR' ranged from 80.3% (Retmarker, Black ethnicity) to 99.1% (NEC, White ethnicity), and ARIAS performance varied by ethnic subgroup for some ARIASs by over 7 percentage points (e.g. DRISTi 2.0 and Retmarker). The observed heterogeneity in sensitivity for 'Referable DR' in some ARIASs was driven by variability in sensitivity for R1M1, not R2 or R3 (see **Figure 7** below).

Sensitivities for 'Referable DR' of DRISTi 2.0, MONA, OphtAI 2.3, Remidio and Retmarker were lower than human graders' sensitivity for all ethnic groups, whereas EyeArt v3.0.0, EyeCheckup AI and NEC reached similar or higher sensitivities than the primary human grader for all subgroups.

Specificity for R0M0 or R1M0 was not reported by ethnic group for the primary human grader. For the 8 ARIASs, specificities ranged from 29.5% (NEC, White ethnicity) to 90.0% (MONA, White ethnicity) and varied by ethnic group for most ARIASs. Absolute percentage differences between ethnic subgroups were between 2 and 6 percentage points for 5 ARIASs (EyeCheckup AI, MONA, NEC, OphtAI 2.3, Remidio) but were 12 to 17 percentage points for DRISTi 2.0, EyeArt v3.0.0 and Retmarker.

All data are also reported in **Table 27**.

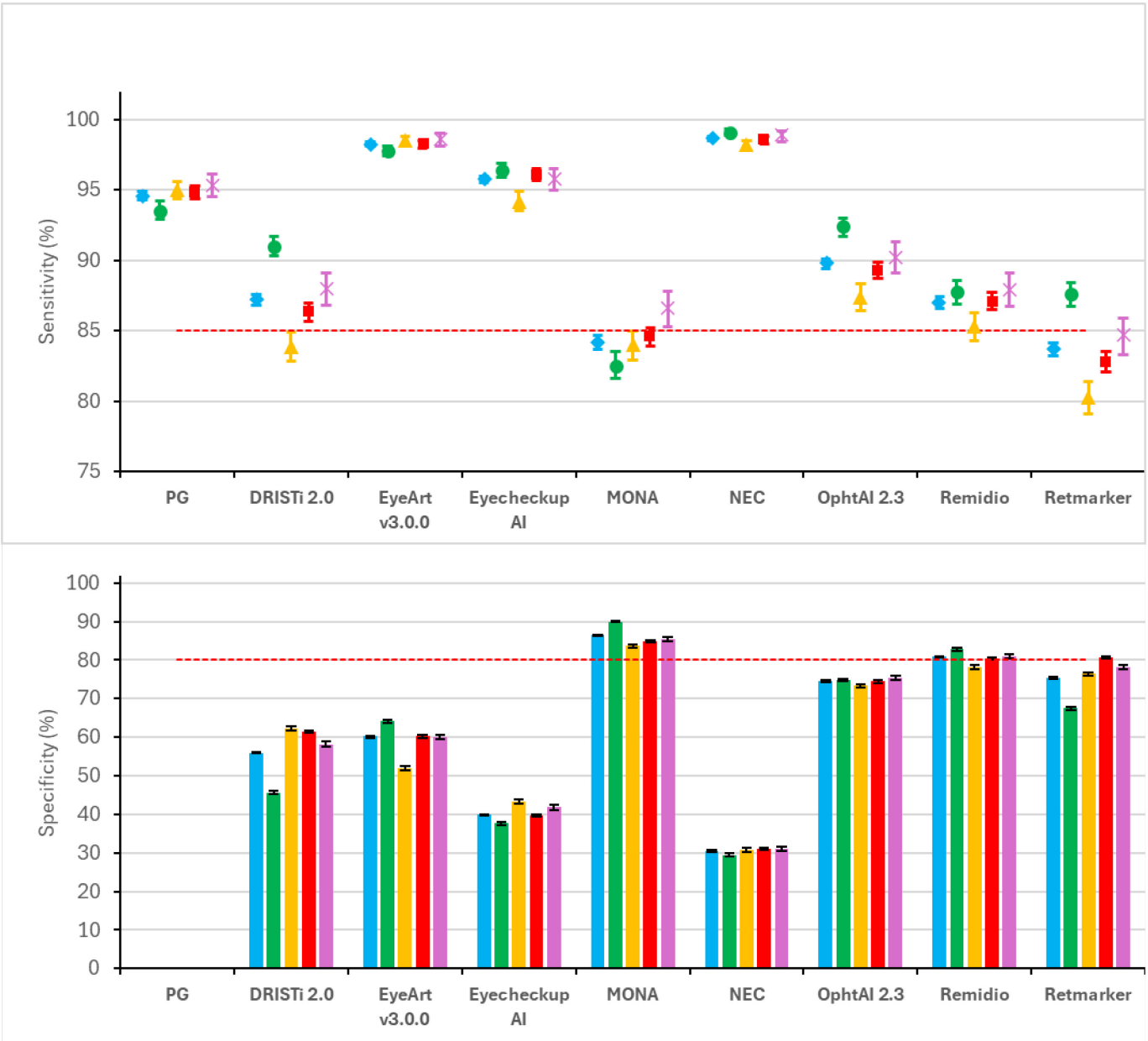


Figure 6. Accuracy for 'Referable DR' (R2, R3, M1), by ethnic group.
(modified from Rudnicka et al.[61])

PG, Primary human grader (threshold: referable diabetic retinopathy).
Blue (diamond) = All ethnicities combined; Green (circle) = White ethnicity; Yellow (triangle) = Black ethnicity; Red (square) = South Asian ethnicity, Purple (cross) = Other / missing ethnicity.

Error bars are 95% confidence intervals.
Red dashed line is at 85% sensitivity and 80% specificity, respectively (quality assurance standards for human graders in the English DESP.[64]
Specificity values were calculated by reviewers from the reported false positive rate: Specificity (%) = 100% – False positive rate (%).

Figure 7 on the next page displays the sensitivity for the detection of moderate-to-severe non-proliferative DR (R2M0 + R2M1) (top) and for the detection of proliferative DR (R3M0 + R3M1) (bottom) (point estimates with 95% CIs) of the 8 ARIASs and the primary human grader and overall and by 4 ethnic groups (White, Black, South Asian, Other / missing).

For moderate-to-severe non-proliferative DR (R2M0 + R2M1) and for proliferative DR (R3M0 + R3M1), primary human graders sensitivity ranged from 94.8% (95%CI 93.3 - 96.0) to 96.2% (95%CI 94.2 - 97.7) and from 94.1% (95%CI 92.0 - 95.7) to 97.6% (95%CI 95.5 - 98.9), respectively, and sensitivities were homogenous across ethnicity subgroups.

For the 8 ARIASs, sensitivities ranged from 94.4% (95%CI 92.4 - 96.0) to 100.0% (95%CI 99.3 - 100.0) for moderate-to-severe non-proliferative DR (R2) and from 93.7% (95%CI 90.0 - 96.4) to 100.0% (95% CI 99.0 - 100.0) for proliferative DR (R3); as for the human grader, sensitivities for these retinopathy grades were homogenous across ethnicity subgroups. Within-ARIAS variation in sensitivity across ethnic groups was typically less than 2 percentage points.

Sensitivities for moderate-to-severe non-proliferative DR (R2M0 + R2M1) and for proliferative DR (R3M0 + R3M1) of all 8 ARIASs were similar or higher than that of primary human graders for all ethnic subgroups.

Exact data are provided in **Table 27**.

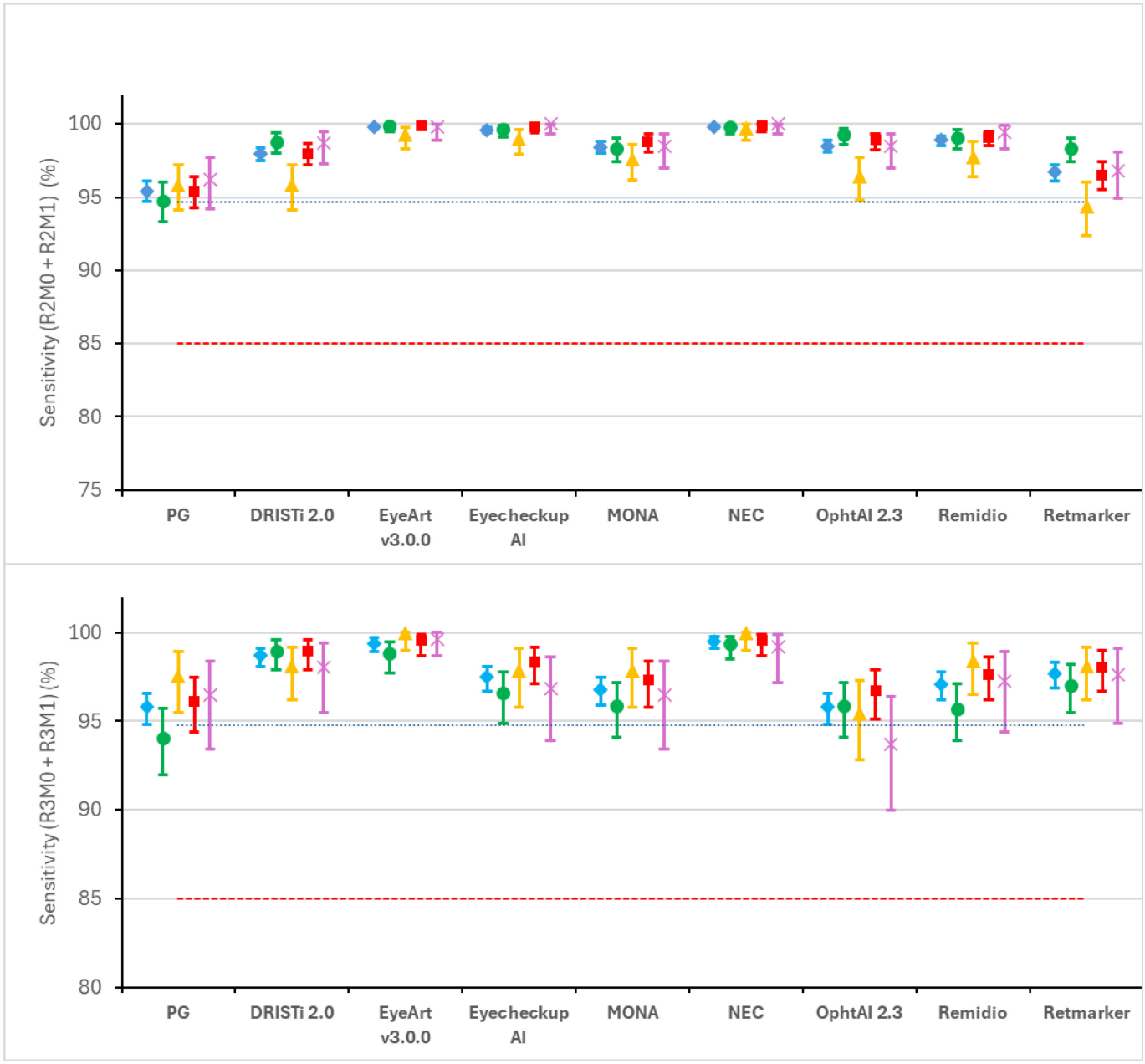


Figure 7. Sensitivity to detect individual grades R2 and R3, by ethnic group.
(modified from Rudnicka et al.[61])

PG, Primary human grader (threshold: referable diabetic retinopathy).
Blue (diamond) = All ethnicities combined; Green (circle) = White ethnicity; Yellow (triangle) = Black ethnicity; Red (square) = South Asian ethnicity, Purple (cross) = Other / missing ethnicity.

Error bars are 95% confidence intervals.
Red dashed line is at 85% (current quality assurance standard for human graders in the English DESP[64]).
Blue dashed line is the lower 95% confidence limit for primary human graders' sensitivity for all subgroups combined.

Subgroup analyses by age group

One study reported accuracy of ARIAS and human graders by age group (that is, <30 years, 30 to <45 years, 45 to <60 years, 60 to <75 years, and 75+ years).[61] Algorithmic fairness was assessed by graphically examining and summarising the absolute percentage differences, in ARIAS and primary human grader performance across age groups.

Figure 8 on the next page displays the sensitivity (top) and specificity (bottom) (point estimates with 95% CIs) of the 8 ARIASs and the primary human grader for the detection of 'Referable DR' overall and by 5 age groups (<30 years, 30 to <45 years, 45 to <60 years, 60 to <75 years, and 75+ years).

Across the 5 age groups, primary human graders' sensitivity for 'Referable DR' ranged from 92.1% (95%CI 91.1 - 93.0) to 95.7% (95%CI 95.2 - 96.1). For the 8 ARIASs, sensitivity ranged from 76.2% (MONA, age group 75+ years) to 99.8% (NEC, age group <30 years).

Declining sensitivities with rising patient age were observed for the primary human grader as well as most ARIASs. For 5 out of the 8 ARIASs, sensitivity for 'Referable DR' varied by 8 to 14 percentage points (that is, DRISTi 2.0, MONA, OphtAI 2.3, Remidio, Retmarker) between age groups. Sensitivities of ARIASs were lower than human graders' sensitivity for all age groups for DRISTi 2.0 and Retmarker, and lower for all but the youngest age group (<30 years) for ARIASs MONA, OphtAI 2.3 and Remidio. EyeArt v3.0.0, Eyecheckup AI and NEC reached similar or higher sensitivities than the primary human grader for all age subgroups.

Specificity for R0M0 or R1M0 was not reported by age group for the primary human grader. For the 8 ARIASs, specificities varied widely from 16.5% (95%CI 15.3 – 17.1) to 88.6% (95%CI 88.1 – 89.0) and was usually lowest for the oldest subgroup (75+ years). ARIAS specificity varied by age groups for all vendors. The absolute percentage differences were typically less than 10 percentage points, but larger variation was observed for the youngest and oldest age groups.

Exact data can be found in **Table 28.**

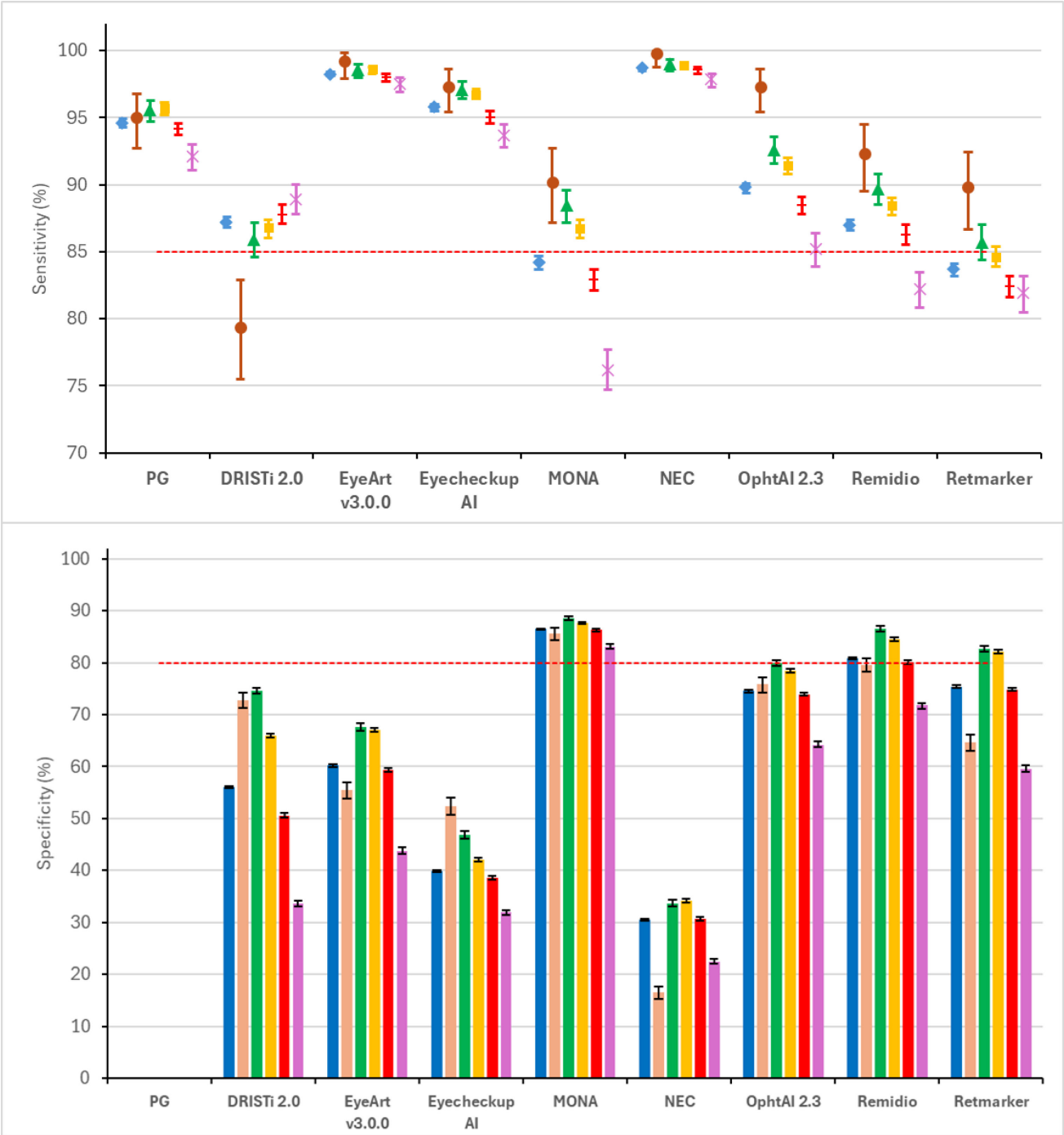


Figure 8. Accuracy for 'Referable DR' (R2, R3, M1) by age group
(modified from Rudnicka et al.[61])

PG, Primary human grader (threshold: referable diabetic retinopathy).
Blue (diamond) = All ages combined; Brown (circle) = <30 years; Green (triangle) = 30 to <45 years; Yellow (square) = 45 to <60 years; Red (line) = 60 to <75 years, Purple (cross) = 75+ years.
Error bars are 95% confidence intervals.
Red dashed line is at 85% sensitivity and 80% specificity, respectively (current quality assurance standards for human graders in the English DESP[64]). Specificity values were calculated by reviewers from the reported false positive rate: Specificity (%) = 100% – False positive rate (%).

Figure 9 on the next page displays the sensitivity for the detection of moderate-to-severe non-proliferative DR (R2M0 + R2M1) (top) and for the detection of proliferative DR (R3M0 + R3M1) (bottom) (point estimates with 95% CIs) of the 8 ARIASs and the primary human grader overall and by 5 age groups (<30 years, 30 to <45 years, 45 to <60 years, 60 to <75 years, and 75+ years).

For moderate-to-severe non-proliferative DR (R2M0 + R2M1), the sensitivity of primary human graders ranged from 92.7% (95%CI 89.3 – 95.3) to 96.4% (95%CI 94.5 - 97.8) and for proliferative DR (R3M0 + R3M1) from 94.5% (95%CI 92.6 – 96.0) to 100.0% (95%CI 93.3 – 100.0). Human graders' sensitivity declined with rising patient age for grades R2 and R3, respectively.

For the 8 ARIASs, the sensitivity for moderate-to-severe non-proliferative DR (R2) ranged from 96.2% (95%CI 95.0 – 97.1) to 100.0% (99.3 – 100.0). For proliferative DR (R3), the lowest sensitivity was 93.0% (95%CI 89.3 – 95.2; OphtAI 2.3, 75+ years); all 8 ARIASs had a sensitivity of 100% for patients under 30 years. A trend of declining sensitivities for grades R2 and R3 with rising patient age was observed for some ARIASs. Within-ARIAS variation in sensitivity across age groups was typically less than 2 percentage points with higher variation (up to 7%) associated with the older age groups.

Sensitivities for moderate-to-severe non-proliferative DR (R2) and for proliferative DR (R3) of all 8 ARIASs were similar or higher than that of primary human graders for all age subgroups.

Exact data are provided in **Table 28**.

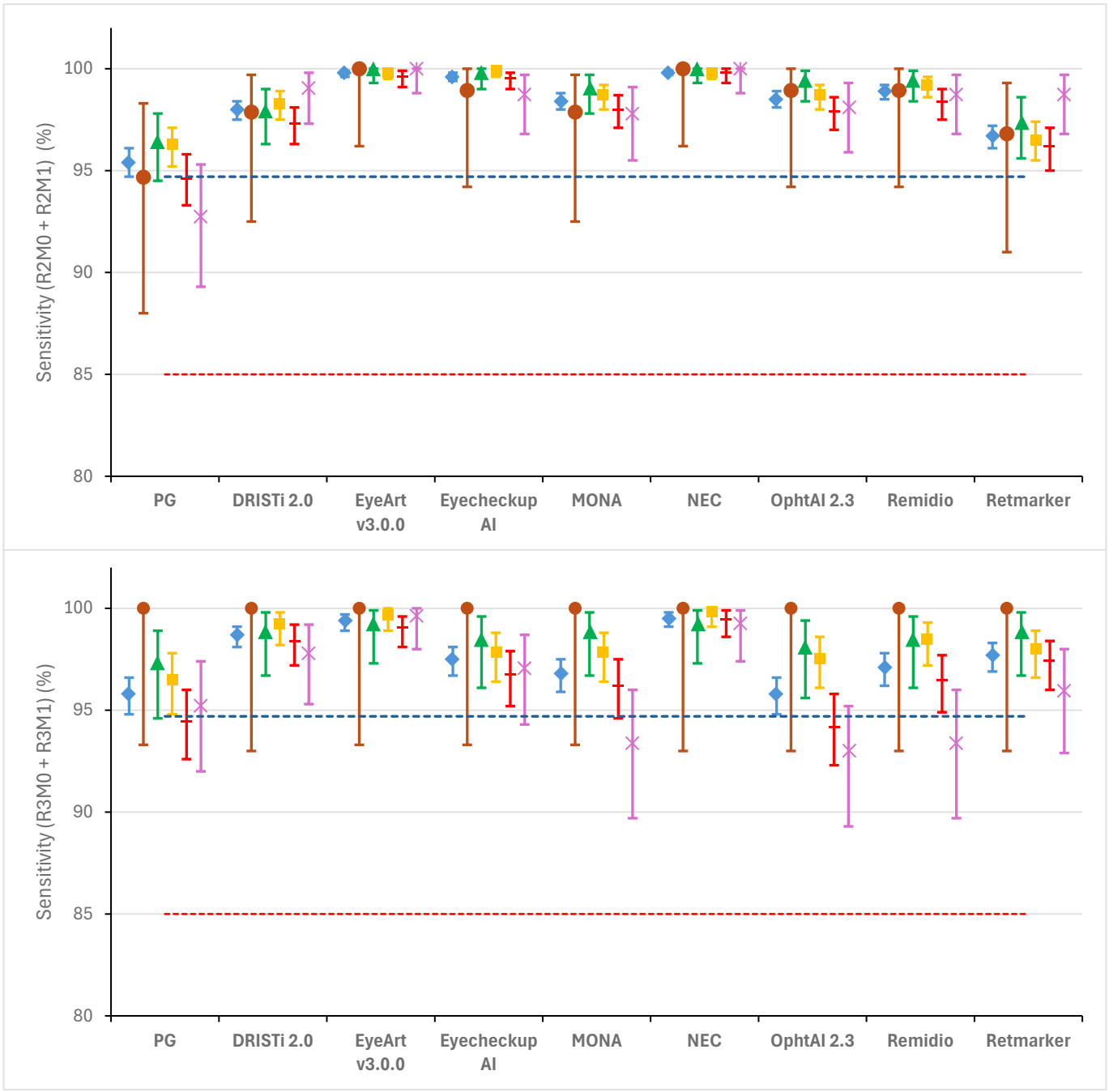


Figure 9. Sensitivity to detect individual grades R2 and R3, by age group.
(modified from Rudnicka et al.[61])

PG, Primary human grader (threshold: referable diabetic retinopathy).
Blue (diamond) = All ages combined; Brown (circle) = <30 years; Green (triangle) = 30 to <45 years; Yellow (square) = 45 to <60 years; Red (line) = 60 to <75 years, Purple (cross) = 75+ years.
Error bars are 95% confidence intervals.
Red dashed line is at 85% (current quality assurance standard for human graders in the English DESP[64]).
Blue dashed line is the lower 95% confidence limit for primary human graders' sensitivity for all subgroups combined.

Discussion of findings

A study-level summary of data extracted from each included publication is presented in 'Summary and appraisal of individual studies **Appendix 3**' (**Table 24**). In **Appendix 3** publications are stratified by question.

Seven studies from the UK or UK-similar countries considered test performance of single or multiple commercially available ARIASs compared to human grading. However, these studies compared their accuracy in isolation without considering the multi-level grading pathway, i.e. they did not investigate the final DES grade between pathways with human vs ARIAS-based primary grading. The evaluated ARIASs showed significant performance differences between the studies/settings and ARIAS vendors, on the one hand how good they were to identify cases with and without DR compared to a reference standard, and on the other hand, how they performed relative to the human graders' performance. Studies found higher, similar or lower ARIAS sensitivities for 'Any DR' or 'Referable DR' compared to human grading. ARIAS specificities varied widely and were lower (and sometimes a lot lower) than human graders' specificities in 12/17 direct comparisons. Only 2 ARIASs (3 comparisons) had higher specificities than human graders. ARIAS sensitivities for higher retinopathy grades (e.g. R2 and R3) were usually equal to or greater than that of human graders. Some ARIASs showed variability in sensitivity and specificity by ethnic group and age group.

Risk of bias

All included studies had limitations in at least one QUADAS-2 / QUADAS-C domain (see **Table 8**), e.g. no consecutive or random sampling was used,[15] the human grading was not performed in clinical practice,[15,57,60] no pre-specified ARIAS threshold was used,[59,62] the reference standard incorporated at least one of the index tests,[57,60,61] not all patients received a reference standard,[57-59,61,62] or they did not receive the same reference standard.[60,61]

Applicability

Applicability concerns were present in all 7 studies, limiting generalisability of the findings (see **Table 8**). High concerns on the study population were present in 6 out of 7 prioritised studies as the photographic protocol (e.g. number of images per eye, use of mydriasis) was different to UK DESP practice[15,57-60,62] which could affect the ARIAS test failure rate as well as ARIAS performance. In the remaining study,[61] the generalisability of the North East London diabetic population to the rest of the UK was unclear, and encounters from all screening pathways (routine digital screening and digital surveillance) were included. Concerns regarding the applicability of the human comparator (e.g. speciality, experience, quality assurance) were high[57] or unclear[58-60,62] for the 5 studies not performed in the UK. The target condition was 'Referable DR' (R2 or higher and/or M1) and not 'Any DR' (R1 or higher and/or M1) in 3 studies,[57,61,62] it is therefore unclear how ARIAS would perform in the proposed role as primary grader sorting patients into 'No disease' and 'Disease'. However, patients with R1M0 disease in the worse eye are classed as

non-referable DR and invited to return for rescreening eventually, so being missed by the primary grader does not change the patient management substantially.

Consistency

Findings were inconsistent. The performance of the evaluated ARIASs varied significantly between the studies/settings and ARIAS vendors, on the one hand how good they were to identify cases with and without DR compared to a reference standard, and on the other hand, how they performed relative to human graders' performance.

Figure 10 consists of stacked bars displaying the number of ARIASs with higher (green), similar (yellow) or lower (red) performance when directly compared to human grading from 25 direct comparisons (7 studies). ARIAS sensitivities for 'Any DR' (prioritised outcome) or 'Referable DR' (top 2 bars combined) were higher in 10/25 comparisons, similar in 6/25 comparisons or lower in 9/25 comparisons. ARIAS sensitivities for higher retinopathy grades (e.g. R2 and R3) were more consistent and for all but 3 comparisons similar or higher than that of human graders (middle 3 bars). ARIAS specificities varied widely and were usually lower (12/17 comparisons) than human graders' specificities (bottom 2 bars).

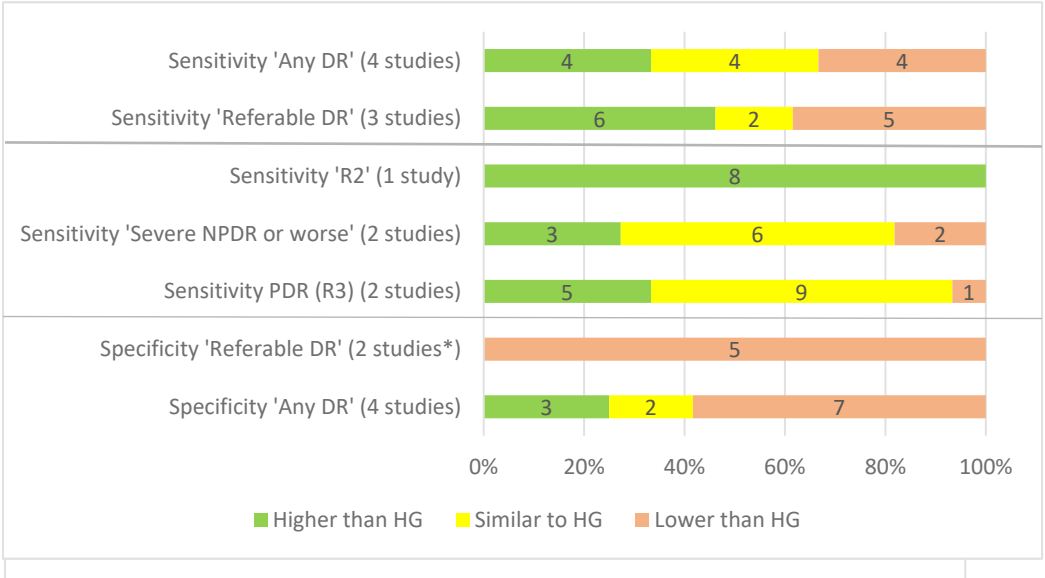


Figure 10. Number of ARIASs with higher, similar or lower performance compared to human grading (direct comparisons from 7 studies)

DR, Diabetic retinopathy; HG, Human grader; PDR, Proliferative diabetic retinopathy.
Any DR, Mild non-proliferative retinopathy or higher and/or maculopathy; R2, Moderate-to-severe non-proliferative retinopathy; R3, Proliferative retinopathy; Referable DR, More-than-mild non-proliferative retinopathy and/or maculopathy.
Numbers in the coloured bars refer to numbers of ARIAS; some studies investigated more than one ARIAS.
* The study by Rudnicka et al.[61] did not report specificity of the primary human grader.

Sensitivity for ‘Referable DR’ of some ARIASs varied by ethnic group; however, within-ARIAS variation for the detection of higher retinopathy grades (e.g. R2 and R3) was largely consistent. Sensitivity of most ARIASs declined with increasing patient age for ‘Referable DR’ but also for higher DR grades like R2 and R3, and false-positive rates within vendors varied across subgroups of age and ethnicity. The observed differences in ARIAS performance between studies might be at least partly explained by differences in patient age and ethnicity. Heterogeneity in the photographic protocol (e.g. mydriasis use, number of views), image quality, coexisting disease as well as human graders’ experience, speciality and grading conditions might have contributed to the observed inconsistencies. One study[62] showed that adjusting manufacturer’s ARIAS threshold with real-world, setting-specific data significantly improved ARIAS sensitivity.

Twelve ARIASs have shown potential to have equal or greater sensitivity for the detection of ‘Any DR’ or ‘Referable DR’ than the human grader. These were RetinaLyze,[59] Aireen,[60] IDx-DR (now: LumineticsCore),[62,63] RetCAD,[62] EyeArt v3.0.0,[61] EyeCheckup AI,[61] NEC,[61] and masked ARIASs C, D, E, F and G from the study by Lee et al.[58] Of these, only 3 ARIASs (Aireen, RetinaLyze [optimal threshold] and masked ARIAS G from the study by Lee et al.[58]) had both comparable or higher sensitivity as well as specificity to the human grader. Equal sensitivity across subgroups by age and ethnicity have only been demonstrated so far by 3 ARIASs: EyeArt v3.0.0,[61] EyeCheckup AI[61] and NEC.[61]

Summary of Findings Relevant to Criteria 4 and 5: met (for some, but not all, ARIASs)

No prospective study was identified that integrated an ARIAS as primary grader, followed by human secondary and arbitration grading, into the DESP grading pathway (as it would be used in the proposed UK screening practice) and then compared the accuracy of the final DES grade between pathways with human vs ARIAS-based primary grading. Seven studies were prioritised that evaluated the accuracy of single or multiple commercially available ARIASs as a standalone grader compared to human grading in UK or UK-similar countries. Two studies were conducted in the UK. Risks of bias and applicability concerns were present throughout the evidence base.

One study (examining multiple ARIASs) was conducted in a UK setting (North East London DESP, England) and provided comparative test accuracy data for 'Referable DR' while also providing data on their performance in age and ethnic subgroups. Within this study, 3 ARIASs (EyeArt v3.0.0, EyeCheckup AI, NEC) were comparable to human graders in terms of sensitivity to detect referable DR (including across age and ethnic subgroups). However, their specificity was lower than human graders.

ARIASs which reach sufficient sensitivity for referable DR and especially higher retinopathy grades in comparison to a human grader and do not display ethnicity or age biases could be trialled in an in-service evaluation to assess their performance as a replacement to the primary human grader or as filter prior to the primary human grader when integrated into screening practice. In both use cases, the ARIAS would triage high-risk images or ungradable images to subsequent reading by human graders. These in-service evaluations should measure the impact of ARIAS on the next-level human grader behaviour, and any additional responsibilities that arise from deploying ARIAS in a primary grader role or filter role. They should consider ongoing training pathways and work plans of human staff, and measure diagnostic yield by grade, failure rate and cost effectiveness in a broad geographical population. Monitoring of ARIAS performance should also be implemented.

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 – What is the wider clinical impact of diabetic eye screening programmes with the use of Automated Retinal Image Analysis Systems for primary grading compared with diabetic eye screening programmes with fully manual grading?

The previous review concluded that "*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."*

This review updated the published evidence on the wider clinical impact of using ARIAS as primary grader in DESPs compared to DESPs 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 and type 2 DM ≥12 years, including rarer forms of DM such as MODY, who underwent standard fundus photography for DR screening.
- Index tests: DESP for detecting DR and/or maculopathy using ARIAS on fundus photographs for primary grading followed by manual grading for secondary and arbitration grading.
- Comparator: DESP for detecting DR and/or maculopathy using human manual grading on fundus photographs at all levels of grading.

- Outcome measures: Any clinical utility outcomes, such as: Vision loss, health-related quality of life. Any patient management and practical implications outcomes, such as: Workforce (e.g. workload), inequalities.
- Study design: RCTs, comparative prospective and retrospective cohort studies, and systematic reviews and meta-analyses of these (studies were prioritised by design, see below).

Papers that fulfilled the following criteria were excluded:

- Qualitative studies,
- studies without relevant outcomes,
- studies where more than 10% of the population do not meet our inclusion criteria and outcomes are not reported separately,
- any algorithms in development,
- internal validations of ARIAS,
- articles not available in the English language,
- articles published prior to 2020,
- single case studies (one patient or one family), letters, reviews, editorials, communications, commentaries, conference abstracts and other grey literature,
- studies that have been retracted,
- studies where it cannot be established if the inclusion criteria are met.

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

- DESPs with vs without ARIAS in the UK or a UK-similar country (that is, North-Western European, USA, Canada, Australia),
- in a primary study.

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 the prioritised primary studies performed in a UK-similar setting, and
3) we discuss the evidence reviewed in this section and state our conclusions regarding criterion 11.

Description of the evidence

Database searches (all searches combined) yielded 5,189 unique results, of which 10 were judged to be relevant to this question. One relevant article was identified through other sources, so 11 articles were ultimately included in this review. **Appendix 2** contains a full PRISMA flow diagram (**Figure 12**) along with a table of the included publications and details of which questions these publications were identified as being relevant to (**Table 18**).

Of the 11 included articles, 4 studies (published in 5 articles) were prioritised and included in the narrative synthesis. Characteristics and findings of deprioritised studies and reason for not being prioritised are reported in **Appendix 3, Table 25**.

Some accuracy studies included for question 1 reported the hypothetical impact in terms of work-load reduction; however, the reported workload reduction was inferred from the reported accuracy estimates, for instance the test negativity rate. As these studies did not report observed outcomes, they were not included for question 2.

The 4 prioritised studies comprised 2 uncontrolled before-after studies from the USA[63,65] and Spain,[66] respectively, and 2 controlled before-after studies from the USA.[67,68] The 4 studies evaluated the impact of ARIAS deployment on screening rates,[65,67,68] proportion of people referred to secondary grading as well as to a follow-up eye examination,[66] and follow-up appointment adherence.[63] No study was identified that assessed a potential patient benefit (e.g. fewer people with vision loss, better health-related quality of life). Study characteristics and findings are summarised in **Table 12.**

The study by Pinto et al. was conducted within the ongoing annual DESP at the University Hospital of Navarre (HUN) in Spain.[66] From March 2015 to June 2020, the DESP used 2-level grading: one of 4 primary care GPs with specific training remotely assessed retinal images (level 1). When they detected signs of referable DR or the image was non-gradable, images were graded by an ophthalmologist (level 2) who decided whether an on-site eye examination was necessary. In July 2020, a custom, in-house ARIAS (NaIA-RD) was introduced. In the new workflow, level 1 graders reviewed the ARIAS output before assessing the images, while the rest of the screening steps remained the same. This uncontrolled before-after study included 19,828 patients screened before ARIAS implementation (March 2015 to June 2020) and 22,962 patients screened afterwards (July 2020 to December 2023). NaIA-RD influenced the decisions of the 4 level 1 grading GPs differently. The negative agreement (i.e. ARIAS proposes no referral and GP does not refer) was usually high (range 0.946 to 0.999), whereas the positive agreement (i.e. ARIAS proposes referral and GP refers) ranged from 0.234 to 0.866. After NaIA-RD was introduced as decision support at level 1, the mean proportion of screened patients who were appointed for an on-site eye examination at level 2 rose by roughly 1.5-times from 3.08% in 2018–2019 to 4.65% in 2022–2023. Over the same comparison window, the mean level 1 referral rate decreased by 13.3%. This pattern is consistent with more selective (efficient) triage at level 1 once GPs could view the ARIAS output.

The uncontrolled before-after study published in Chen et al.[65] and Dow et al.[63] was conducted within the Stanford Teleophthalmology Autonomous Testing and Universal Screening (STATUS) programme, a DESP at 7 primary care and endocrinology clinics in the San Francisco Bay Area (USA). Before ARIAS deployment (March 2021), the proportion of DM patients with annual eye examination was 67.2%. After switching to ARIAS-based DR screening, the screening rates were 74.8% at the end of Year 1 (04/2022) and rose gradually during the second year to a peak adherence of 78.1% in April 2023. Dow et al.[63] found that the proportion of patients attending a follow-up appointment at the university eye clinic within 90 days was 3-times greater in the autonomous ARIAS workflow compared to the human workflow or ARIAS-human hybrid workflow where patients received their screening result up to one week later.

The second US study was conducted at more than 30 primary care sites at Johns Hopkins Medicine.[68] Propensity score weighting analysis was used to compare the change in adherence to annual diabetic eye screening from 2019 to 2021 between sites that switched to ARIAS and sites that had not switched (controlled before-after design). ARIAS-switched sites experienced a 7.6 percentage point greater increase in DES than non-ARIAS sites from 2019 to 2021 ($p < 0.001$). Subgroup analysis by ethnicity showed that the adherence rate for Black/African Americans increased by 12.2 percentage points within ARIAS-switched sites but decreased by 0.6% points within non-ARIAS sites ($p < 0.001$). In sites that had switched to ARIAS use, the adherence rate gap between Asian Americans and Black/African Americans was reduced from 15.6% in 2019 (61.1% vs 45.5%) to 3.5% in 2021 (60.9% vs 57.4%).

In the third US study, DR screening rates between December 2019 and June 2023 were tracked at 2 rural primary care centres in Maine.[67] Western Maine Primary Care (WMPC) introduced ARIAS-based DR screening during the study period, whereas Franklin Health Primary Care (FHPC) did not switch to ARIAS. Before ARIAS implementation, patients seen at WMPC were 14% less likely to be screened for DR than at FHPC (Odds ratio [OR] = 0.86, 95% CI 0.74 - 0.99, $p = 0.041$). In June 2023, around 2.5 years after WMPC switched to ARIAS-based DES, patients seen at WMPC were 2.8 times more likely to be screened for DR than those seen at FHPC (OR = 2.82, 95%CI 2.42 - 3.27, $p < 0.001$). Around one year after implementing ARIAS-based DES (March 2022), WMPC screening rates reached target level (that is 67%) for the first time and continued to do so until the last measurement in June 2023.

Table 12. Studies evaluating the impact of implementing ARIAS in a DESP

Study and country	Study design	Setting	Population	Sample size	Intervention	Comparator	Outcomes	Findings	Risk of bias / Applicability
Chen 2025,[65] USA	Before-after study	STATUS DESP at 7 primary care sites in the San Francisco Bay Area	Patients with DM without a prior DR diagnosis or a DR exam in the past 12 months	NR	From 04/2021: ARIAS (IDx-DR, now Luminetics-Core®) as single grader, with human secondary grading of ARIAS-ungradable cases (Hybrid)	09/2019 to 03/2021 (18 months): Human-based teleophthalmology, single retina specialists	% DM patients with annual DES	Human (03/2021): 67.2% ARIAS (04/2022): 74.8% ARIAS (04/2023): 78.1%.	Not randomised; no concurrent control group; potential differences in participants/confounding; follow-up data in community not available; regression to the mean; no statistical analysis. No human secondary grading for test-positive cases; population not applicable to UK (e.g. White Non-Hispanic/Latino 25.0%, Hispanic/Latino 21.2%, Asian 33.1%); DES at primary care practices.
Dow 2023b,[63] USA				Patients with positive or ungradable screening result: ARIAS: 279 Hybrid: 103 Human: 117			% with university follow-up appointment within 90 days	ARIAS (04/2021 to 09/2022): 35.5% (99/279)[a] (+ 33.7% [94/279] community FU) Hybrid (04/2021 to 09/2022): 11.7% (12/103) Human (09/2019 to 03/2021): 12.0% (14/117)[b]	
Heuer 2025,[67] USA	Controlled before-after study	2 rural primary care practices in Maine: WMPC and FHPC	Adult (>18 years) with DM at WMPC and FHPC from December 2019 to June 2023	NR	WMPC: Switch to ARIAS-based DES (EyeArt, Eyenuk) in January 2021	FHPC: No ARIAS-based DES during study period. All diabetics referred to an eye care centre annually.	% DM patients with annual DES	WMPC vs FHPC: Before ARIAS: OR = 0.86 (95%CI 0.74 - 0.99) (p=0.041). After ARIAS switch at WMPC: OR = 2.82 (95%CI 2.42 - 3.27) (p<0.001)	Not randomised; potential differences in participants/confounding, no appropriate statistical analysis. No human secondary grading for test-positive cases; population applicability unclear, DES at 2 rural primary care practices.

Study and country	Study design	Setting	Population	Sample size	Intervention	Comparator	Outcomes	Findings	Risk of bias / Applicability
Huang 2024,[68] USA	Controlled before-after study	Johns Hopkins Medicine (>30 primary care sites)	Patients with DM who were managed at primary care sites of Johns Hopkins Medicine in the calendar years 2019 and 2021	2019 (prior ARIAS switch): ARIAS-switched sites: 5,505 patients; Non-ARIAS sites: 12,169 patients. 2021 (post ARIAS switch): ARIAS-switched sites: 5,580 patients; Non-ARIAS sites: 12,010 patients.	Switch to autonomous ARIAS (IDx-DR, now Luminetics-Core®) as single grader in 2020	No use of autonomous ARIAS from 2019 to 2021, no other information on graders and grading pathway	% DM patients with annual DES	Change 2019 to 2021 (propensity score weighting analysis): ARIAS-switched: +7.4% Non-ARIAS: -0.3% Difference ARIAS-switched vs Non-ARIAS sites: +7.6% (p<0.001). Also reports change in adherence rates by gender, age and race.	Not randomised; regression to the mean. No human secondary grading for test-positive cases; population not applicable to UK (e.g. 45-48% White, 37-41% Black, 5-6% Asian); DES at primary care practices.

Study and country	Study design	Setting	Population	Sample size	Intervention	Comparator	Outcomes	Findings	Risk of bias / Applicability
Pinto 2025,[66] Spain	Before-after study	University Hospital of Navarre (HUN), ongoing annual DESP	Patients assigned to HUN and diagnosed with Type 2 diabetes who underwent DES from 2015 to 2023	Before: 19,828 patients screened. After: 22,962 patients screened.	July 2020 to 2023: Custom, in-house AI tool (NaIA-RD) as level 0 grader. Level 1 grader (GPs) review screening proposal of NaIA-RD before assessing the images. Remaining pathway as in 'Comparator'.	2015 to July 2020: 2-level grading: Level 1: Team of 4 primary care GPs with specific training remotely assess images. Referable DR or ungradable images referred to level 2 grading by ophthalmologist.	1. Annual % of patients referred to onsite eye examination. 2. Mean referral rate by level 1 graders	1. 2018/2019: mean 3.08%. 2022/2023: mean 4.65% (1.5 times increase). 2. 2022/2023: 13.3% fewer patients referred to level 2 grading (mean referral rate) compared to 2018–2019.	Lack of concurrent control group; potential differences in participants/confounding; regression to the mean; no statistical analysis. ARIAS not as primary grader.

AI, Artificial intelligence; ARIAS, Automated retinal image analysis system; CI, Confidence interval; DES, Diabetic eye screening; DESP, Diabetic eye screening programme; DM, Diabetes mellitus; DR, Diabetic retinopathy; FHPC, Franklin Health Primary Care; FU, Follow-up; GP, General practitioner; HUN, University Hospital of Navarre; NR, Not reported; OR, Odds ratio; STATUS, Stanford Teleophthalmology Autonomous Testing and Universal Screening; WMPC, Western Maine Primary Care.
[a,b] Groups with different letters significantly different.

Discussion of findings

A study-level summary of data extracted from each included publication is presented in 'Summary and appraisal of individual studies **Appendix 3**' (**Table 25**). In **Appendix 3** publications are stratified by question.

Overall, 4 studies (published in 5 articles) compared the wider clinical impact of DR screening using ARIAS for primary grading with DR screening using manual grading at all levels. The 4 before-after studies evaluated the effect of ARIAS use on screening rates (3 studies), the proportion of people referred to secondary grading and to a follow-up eye examination, respectively, (1 study) and follow-up appointment adherence (1 study). No study was identified that assessed a potential patient benefit (e.g. fewer people with vision loss, better health-related quality of life). Three US studies reported increased annual screening rates in primary care sites after implementing ARIAS-based DR screening;[65,67,68] one of them also found an increased uptake of the follow-up eye examination after autonomous ARIAS-based DR screening.[63] One Spanish study found that the decisions of the 4 level 1 grading GPs were influenced differently by the provision of the ARIAS result.[66] After ARIAS deployment, level 1 graders referred fewer patients to level 2 grading; however, the proportion of patients that were eventually referred to a follow-up eye examination increased, suggesting a more selective (efficient) triage at level 1 once GPs could view the ARIAS output.

Risk of bias

The stacked bars in **Figure 11** report the number of studies, outcomes and results, respectively, by answer (yes / no / unclear) for each signalling question, with 'No' answers (red colour of the bar) indicating methodological flaws.

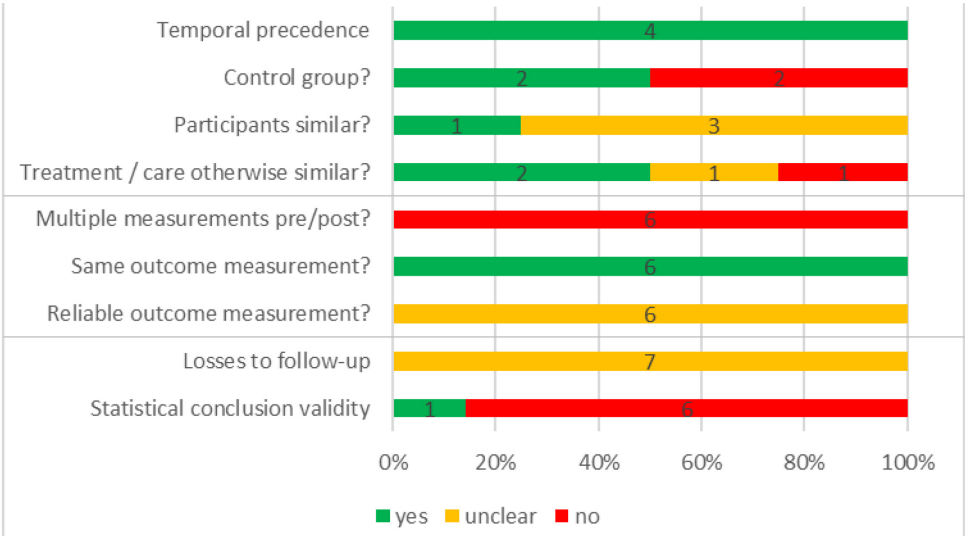


Figure 11. Critical appraisal results based on JBI tool for quasi-experimental studies (4 studies / 6 outcomes / 7 timepoints)

All studies had key methodological issues identified in the quality assessment (see **Appendix 3**, **Table 26** for details).

In all 4 studies, patients were not randomly allocated to a DESP with or without ARIAS use, and the human grading period preceded the period with ARIAS use (before-after study). Two studies (reported in 3 articles) had no concurrent control group,[63,65,66] it is therefore unknown how screening rates, follow-up appointment adherence and proportion referred to level 2 grading as well as to a follow-up eye examination would have fluctuated over time without deployment of ARIAS, in particular as time periods coincided with COVID-19 disruption. In 2 studies, it is unclear if [67] or likely that [63,65] other changes (e.g. a quality improvement intervention to increase ARIAS gradability and patient encounters[63,65]) occurring at the same time as ARIAS deployment contributed to the observed effects. Three studies (reported in 4 papers)[63,65-67] did not control adequately for confounding. Therefore, the observed changes after implementing ARIAS might be due to differences in population characteristics, e.g. a different DR prevalence would affect the proportion of people referred to secondary grading or a follow-up appointment. No study performed multiple measurements before and after ARIAS deployment, meaning that the observed effect could have been caused by naturally occurring changes that are unrelated to the ARIAS use. It is also possible that extreme values (either high or low) measured before ARIAS use would have reverted to less extreme values even in the absence of the intervention (regression to the mean). No statistical analysis[63,65,66] or inadequate statistical analysis (lack of adequate adjustment for potential confounding factors)[67] was performed for 5 outcomes (6 timepoints / 3 studies).

Applicability

Generalisability to the UK is limited in 2 US studies[63,65,68] as patients' ethnicity is different to that in the UK. Furthermore, there are high applicability concerns regarding the intervention and comparator in all 4 studies as they were not performed in regional or national organised DESPs. The observed increase in screening rates in US primary care settings[65,67,68] might be due to increased screening capacity with ARIAS-based DR screening and is likely not transferable to the UK situation. The 3 US studies also did not use a 3-level grading pathway comparable to the UK NHS DESPs but used ARIAS as a single grader. The observed faster notification time after an ARIAS-positive screening test result (which was speculated to be the reason for the observed higher uptake of the follow-up appointment) is therefore not generalisable to the UK DESPs where ARIAS-positive results would be referred for human secondary grading. The Spanish HUN DESP[66] used 2-level grading before ARIAS deployment (level 1: trained GPs, level 2: ophthalmologist). After switching to ARIAS use, the GPs would review the ARIAS result before assessing the images, with the remaining pathway staying the same. It is questionable if the observed decrease in referrals to level 2 grading and increase in the proportion of DM patients referred to a follow-up eye examination after introducing ARIAS as decision support for levels 1 graders could be transferred to stand-alone ARIAS primary grading as proposed for the UK NHS DESPs.

Consistency

Three US studies consistently reported increased annual screening rates in primary care sites after implementing ARIAS-based DR screening,[65,67,68] and one US study also found an increased uptake of the follow-up appointment;[63] however, findings are not applicable to the UK DESPs. One Spanish study observed a reduction in level 1 grader referrals and an increase in the proportion of patients that are referred to a follow-up eye examination by the level 2 grader after ARIAS was introduced as decision support for levels 1 graders.[66]

Summary of Findings Relevant to Criterion 11: not met.

We did not identify any relevant RCTs from the UK or UK-similar countries comparing DR screening with primary ARIAS grading followed by manual grading for secondary and arbitration grading with manual grading at all levels in terms of clinical outcomes and overall impact. The 4 prioritised before-after studies evaluated the effect of ARIAS use on DR screening rates, the proportion of people referred to level 2 grading or to a follow-up eye examination as well as follow-up appointment uptake in UK-similar countries. No study was identified that assessed a potential patient benefit (e.g. fewer people with vision loss, better health-related quality of life). Methodological flaws were present in all studies, reducing the validity and reliability of the findings. In addition, the prioritised studies were not conducted within national or regional organised DESPs, they did not use a 3-level grading pathway comparable to the UK, and study populations were different to the UK in 2 US studies (for example, ethnicity). These applicability concerns limit generalisability of the findings to the UK. The only (at least partly) applicable study conducted in a Spanish DESP found that level 1 grading GPs escalated a reduced proportion of images to level 2 graders, yet a greater share of those referred were ultimately judged by the ophthalmologist to need a follow-up eye examination.

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 – What is the cost-effectiveness of Automated Retinal Image Analysis Systems in diabetic eye screening programmes compared with diabetic eye screening programmes with manual grading?

This question was previously examined in the 2021 UK NSC review which identified 5 studies that evaluated the cost-effectiveness of 3 ARIASs (EyeArt v1, RetmarkerSR and iGradingM, all based on traditional ML algorithms). These studies generally found that ARIASs were less effective but less costly compared to manual grading, with performance driven down by relatively high false positive error rates while being comparable to human graders in detecting cases with disease. However, these studies were published between 2007 and 2017 and evaluated older ARIAS technologies. The review concluded that these studies needed updating to incorporate new data and capture cost-effectiveness over time. Given the rapid evolution of AI technologies since 2020, with newer DL-based systems potentially offering different cost-effectiveness profiles than the traditional ML algorithms evaluated previously, re-examining this question is essential to inform current policy decisions about ARIAS implementation in the NHS.

Eligibility for inclusion in the review

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

Population: People with type 1 and type 2 DM ≥12 years, including rarer forms of DM such as MODY.

Intervention: DESP for detecting DR and/or maculopathy using ARIAS on fundus photographs for primary grading followed by manual grading for secondary and arbitration grading.

Comparator: DESP for detecting DR and/or maculopathy using human manual grading on fundus photographs at all levels of grading.

Outcomes: Any cost-effectiveness or modelled clinical outcomes.

Study design: Economic evaluations (of any type) and reviews of these.

Country: UK-based evaluations only.

Publication date and language: Studies published in English after 2020.

Papers that fulfilled the following criteria were excluded: qualitative studies, studies without relevant outcomes, studies where more than 10% of the population do not meet inclusion criteria and outcomes are not reported separately, any algorithms in development, internal validations of ARIAS, articles not available in the English language, articles published prior to 2020, single case studies, letters, reviews, editorials, communications, commentaries, conference abstracts and other grey literature, studies that have been retracted, studies where it cannot be established if the inclusion criteria are met, non-UK-based evaluations.

Description of the evidence

Database searches (original, August, September, October and November updates combined) yielded 5,189 unique results. No study met the inclusion criteria for this question. Database searches yielded very few economic evaluations related to ARIAS in diabetic eye screening programmes published since 2020, and all of those identified were non-UK based (Brazil: n = 1; China: n = 3; USA: n = 1). Additional exclusions comprised studies in ineligible populations (paediatric; n = 1).

Discussion of findings

The absence of recent UK-based cost-effectiveness evidence represents a significant evidence gap since the previous UK NSC review in 2021.[10] The lack of recent economic evidence is concerning given the rapid advancement in AI technologies and the availability of newer, potentially more accurate systems since 2020. Modern DL-based ARIASs may have different cost-effectiveness profiles compared to the traditional ML-based systems evaluated in earlier studies. Additionally, the cost structure of screening programmes, staff wages, and technology costs have changed since the earlier evaluations. Without current economic evidence, it is not possible to determine whether implementation of modern ARIAS would represent value for money within the NHS context. The 2021 review noted that previous economic evaluations were limited and insufficient, lacking long-term data and not adequately accounting for the evolving nature of AI technologies. The continued absence of updated economic evidence means these limitations persist. This represents a critical evidence gap that would need to be addressed through high-quality, UK-specific health economic research before ARIAS implementation could be recommended on cost-effectiveness grounds.

Summary of Findings Relevant to Criterion 14: not met.

We did not identify any new evidence published since 2020 on the cost-effectiveness of ARIAS in DESPs compared with DESPs with manual grading in a UK setting.

Review summary

Conclusions and implications for policy

Seven studies from the UK or UK-similar countries evaluated the accuracy of single or multiple commercially available ARIASs as stand-alone grader compared to human grading. Except for one study, the evidence on the test accuracy of ARIAS as primary grader lacked applicability to the UK context. In the largest and most applicable study performed within the North East London DESP, the reference standard was biased in favour of the primary human grader as only a subset of images classed as 'No disease' received higher level grading. No prospective study compared the accuracy of ARIAS-based vs human primary grading in a 3-level screening pathway. We also did not identify any evidence on the clinical and cost effectiveness of replacing the primary human grader with an ARIAS in a multi-level grading pathway. However, the UK DESPs are very sensitive and few sight-threatening DR cases are missed, so a trial with rare outcomes like vision loss would possibly be underpowered to prove a difference. In addition, DR progresses slowly and health outcomes occur late, so use of comparative evidence on ARIAS vs human grading accuracy might be a reasonable compromise.

Study findings suggest that at least some of the commercially available ARIASs match or exceed primary human graders' sensitivity for 'Any DR' or 'Referable DR', and most ARIASs achieved similar or higher sensitivities for the detection of moderate-to-severe non-proliferative (R2) or proliferative (R3) DR. ARIAS' specificities varied widely and were mostly lower than human graders' specificity, leading to higher false positive results. However, when ARIAS is used as first pass to triage patients into low-risk (that is 'No DR') and high-risk ('Any DR') diabetic eye disease cases, false positives can be overturned by the next level of human grading. A lower specificity might still be cost effective and increase human grading capacity, but this depends on how secondary graders manage the additional referrals.

Based on findings from our review, potential implications for service provision include:

- ARIASs that have been identified in the reported studies to match or exceed primary human grader sensitivity for referable DR and especially higher retinopathy grades, that showed no ethnicity or age biases and that have the required approval for use could be trialled in an in-service evaluation to evaluate their performance as triage test prior to human grading when integrated into screening practice, and to assess test failure rate.
- ARIAS could be deployed in a primary grader role or as filter role before the primary grader. In both use cases, the ARIAS triages high-risk images or ungradable images to subsequent reading by human graders. In the former strategy, no images would be assessed by primary human graders, and secondary human graders would only screen images that have been flagged by the ARIAS as 'Disease' or 'Ungradable', along with a small proportion of episodes classed as 'No disease' for quality assurance. In the latter strategy, only the subset of images flagged by ARIAS as 'Disease' or 'Ungradable', along

with a small proportion of those images classed as ‘No disease’, would be seen by primary human graders.

- The in-service evaluation should also look at effects on next-level grader’s behaviour and performance as well as the performance of the multi-level grading pathway as a whole, and test and report ARIAS performance by screening centre and in subgroups stratified by age, sex, and ethnicity to confirm that a satisfactory performance is reached across different screening centres and population subgroups (algorithmic fairness and equity) in the eventual deployment setting.
- As currently up to 90% of image sets classed as ‘No disease’ by the primary grader are not assessed by a second grader, the proportion of ARIAS screen-negative cases that go to next-level human grading for quality assurance could be increased (temporarily or permanently) as a safety measure to ensure no cases of sight-threatening DR are missed.
- Disagreements between ARIAS and next-level human grading that go to adjudication could be used to analyse error rates and type of errors of ARIAS compared to human grading.
- Feasibility of IT integration, acceptability to clinicians and patients and the practical impact of incorporating ARIASs into screening practice, such as their impact on staff training, human grading workload, reporting times, number of referrals to follow-up eye examinations as well as referrals to treatment need to be evaluated.
- Updated cost effectiveness analyses are urgently needed to assess value for money and to compare current practice with triage (primary grader or filter role) by different ARIASs.
- Most ARIASs undergo regular update, which may involve changes in the AI-derived algorithm. Ongoing audit of the potential impact of these updates on test accuracy and service provision may be desirable. The DESP should consider new monitoring provision over time that exceeds the manufacturer’s requirements for post-market surveillance and includes assessment of changes introduced as part of updates.

The discussions of the PPIE group largely concurred with the main conclusions and recommendations. The high sensitivity (and comparative sensitivity to human graders) of some ARIASs trialled in a UK setting show promise for use in screening (see **Appendix 6**). However, the group were concerned about the lack of UK data in general, and the disparities between different ARIASs in the large London-based study. They felt that more information is needed, including on how ARIAS operate across different geographical regions with differing socio-demographic characteristics. They suggested that this could be achieved through pilot work, integrating ARIAS within (and comparing it to) standard practice. The group stressed the importance of adequate communication with screening patients and the public about the use of ARIAS, including their potential benefits and limitations.

Limitations

Limitations of the available evidence

Of the 72 included articles for question 1 (accuracy), only one included a consecutive sample from the UK and compared commercially available ARIAS to human grading. No UK-based studies reporting direct evidence on clinical effectiveness (question 2) or cost-effectiveness (question 3) were found.

The most applicable evidence to address question 1 comes from studies where ARIAS is integrated into the screening pathway, as it would be used in screening practice. These studies need to report the change of the whole pathway when ARIAS is replacing the primary human grader. Only one study from Scotland reported the accuracy of a multi-level grading pathway with ARIAS as primary reader followed by up to 3 levels of human grading; however, the accuracy was not compared to a multi-level grading pathway without ARIAS. Therefore, the downstream effects of a potentially lower ARIAS specificity on human grading performance are unclear. All other prioritised studies compared the accuracy of ARIAS and humans as stand-alone grader only: 4 studies evaluated ARIASs prospectively within DESPs, with the human grading performed at the same time or retrospectively. The other 3 studies evaluated ARIAS retrospectively on curated datasets from DESPs, usually compared to the original human grading.

Most studies had applicability concerns regarding the target population as they used a different photographic protocol (e.g. use of mydriasis, number of views) to UK DESP practices and/or were enriched for difficult or ungradable images.

The reference standard (final human grade) in the most applicable and largest test accuracy study comparing 8 ARIAS with (pending) CE-mark and human primary graders within a UK setting was biased in favour of the primary human graders because of the selective higher-level grading in the EDESP.

All the issues related to risk of bias and applicability presented in ‘Discussion of findings’ sections and highlighted above increase the uncertainty and limit generalisability of the study findings.

Limitations of the review

This review was conducted as an enhanced rapid evidence assessment, involving methodological adaptations compared with a full systematic review. Such approaches are widely used internationally by policymakers to provide timely evidence for decision-making, though they carry some inherent limitations.[69,70] All stages of the review process - including title and abstract screening, full-text assessment, data extraction, and quality appraisal - were undertaken by a single reviewer, with approximately 20% of each task independently checked by a second reviewer. Consequently, around 80% of review tasks were not verified in duplicate, increasing the

risk of error and the possibility that some eligible studies were missed. These limitations are typical of rapid evidence synthesis approaches and have been documented in previous methodological comparisons.[71]

Only peer-reviewed studies published in English were included, meaning that potentially relevant publications in non-English language or grey literature may have been missed.

The search strategy combined MeSH headings and keyword searches limited to the 'Title' and 'Abstract' fields, consistent with the approach used in the previous UK NSC review. However, one recently published paper[60] was not identified because it had not yet been indexed and the term *"artificial intelligence"* appeared only as an author keyword rather than in the title or abstract. This highlights the general challenge of identifying newly published or pre-indexed studies in database searches.

During the reference list screening of included studies and relevant systematic reviews, the number of potentially eligible full texts assessed as well as reasons for exclusion at full-text level were not recorded**.** This represents a limitation because it prevents a full accounting of potentially eligible studies identified through supplementary sources and limits transparency regarding exclusions at the full-text screening stage. Consequently, reproducibility and completeness of the study selection process cannot be fully verified.

Studies were prioritised for extraction and analysis based on study design and applicability to the UK setting. Although this hierarchy focused on the most relevant evidence, some deprioritised studies may still contain useful insights.

While the tailored QUADAS-2 tool[26] as well as the QUADAS-C[28] tool were used to assess study quality of comparative test accuracy studies, the QUADAS-2 adaptation for AI-based tests needs further refinement taking into consideration the revised tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI).[72] A revised version of the QUADAS tool (i.e., QUADAS-3[73]) as well as a new QUADAS-AI[72] are currently under development.

Overall, these limitations are standard for rapid review methodologies.[69,70] They are unlikely to have excluded pivotal studies, especially as the searches covered the full period since the previous UK NSC external review.

Appendix 1 — Search strategy

Electronic databases

The search strategy included searches of the databases shown in **Table 13**: MEDLINE ALL (OVID), Embase (OVID) and The Cochrane Library (Wiley).

Table 13. Summary of electronic database searches and dates

Database	Platform	Searched on date	Date range of search
MEDLINE ALL	**Ovid SP**	**17 July 2025**	**1946 to present**
Embase	**Ovid SP**	**17 July 2025**	**1974 to 2016 July 01**
The Cochrane Library, including: — Cochrane database of systematic reviews (CDSR) — Cochrane Central Register of Controlled Trials (CENTRAL) — Database of Abstracts of Reviews of Effects (DARE)	**Wiley Online**	**17 July 2025**	**CDSR**

Search Terms

Search terms included combinations of free text and subject headings (Medical Subject Headings [MeSH] for MEDLINE, and Emtree terms for Embase), grouped into the following categories:

disease area: eye diseases, retinopathy, eye pathology, maculopathy, diabetic eye, diabetic macular, retinal fundus, fundus photograph*, fundus camera*, retinal photograph*, retinal camera*

study design: Ophthalmological diagnostic techniques, computer-assisted diagnosis, "sensitivity and specificity", diagnostic test*, diagnostic accuracy, diagnostic performance, screening, imaging, sensitivity or specificity, reference standard, optical coherence tomography, early diagnosis, mass screening

other term group: Artificial Intelligence, deep learning, neural network*, automated retinal image analysis system, automated grading, automated level, (automated tool* or automated technique* or automated identification or automated detection, ARIAS, GradingM, EyeArt, IDx-DR, DR-RACS, RetinaLyze, RetmarkerSR DR, Singapore Eye Lesion Analyzer, RetinaVue, TRIAD network, LumineticsCore, Retmarker, DAIRET, eyRIS, Automated Disease Assessment, ARDA, Medios, OphtAI, RetCAD, DeepDee, MONA DR, AEYE-DS, EyeCheckup, Reti-EyeReti, DrNoon, ITOS Mass Screening, Nexy AI, EyeWisdom, iPredict System, TeleMedC DR grader, Eyetelligence system, DRISTi, VUNO Med, AutoGrader, UMI DR

Search terms for MEDLINE ALL (OVID) and Embase (OVID) are shown in **Table 14** and **Table 15**, respectively, and search terms for the Cochrane Library databases are shown in **Table 16**.

Table 14. Search strategy for MEDLINE ALL (OVID).

Term Group	#	Search terms	Results
Disease area	1	exp eye diseases/	675,229
Disease area	2	retinopathy.ti,ab.	60,893
Disease area	3	eye pathology.ti,ab.	371
Disease area	4	maculopathy.ti,ab.	5,852
Disease area	5	diabetic eye.ti,ab.	1,082
Disease area	6	diabetic macular.ti,ab.	6,435
Disease area	7	retinal fundus.ti,ab.	681
Disease area	8	fundus photograph*.ti,ab.	8,422
Disease area	9	fundus camera*.ti,ab.	1,248
Disease area	10	retinal photograph*.ti,ab.	1,425
Disease area	11	retinal camera*.ti,ab.	331
Disease area	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	699,475
Study Design	13	exp Diagnostic Techniques, Ophthalmological/	204,203
Study Design	14	exp Diagnosis, Computer-Assisted/	91,432
Study Design	15	"Sensitivity and Specificity"/	380,834
Study Design	16	diagnostic test*.ti,ab.	64,670
Study Design	17	diagnostic accuracy.ti,ab.	74,475
Study Design	18	diagnostic performance.ti,ab.	34,884
Study Design	19	screening.ti,ab.	768,195
Study Design	20	imaging.ti,ab.	1,178,551
Study Design	21	(Sensitivity or specificity).ti,ab.	1,403,535
Study Design	22	reference standard.ab.	22,002
Study Design	23	optical coherence tomography.ti,ab.	57,612
Study Design	24	early diagnosis/	32,178

Term Group	#	Search terms	Results
Study Design	25	Mass Screening/	122,073
Study Design	26	or/13-25	3,640,230
Other term group	27	exp Artificial Intelligence/	243,216
Other term group	28	artificial intelligence.ti,ab.	66,235
Other term group	29	deep learning.ti,ab.	77,052
Other term group	30	neural network*.ti,ab.	128,064
Other term group	31	automated retinal image analysis system.ti,ab.	7
Other term group	32	automated grading.ti,ab.	160
Other term group	33	automated level.ti,ab.	13
Other term group	34	(automated adj (tool* or technique* or identification or detection)).ti,ab.	7,468
Other term group	35	ARIAS.ti,ab.	377
Other term group	36	iGradingM.ti,ab.	3
Other term group	37	EyeArt.ti,ab.	25
Other term group	38	IDx-DR.ti,ab.	28
Other term group	39	DR-RACS.ti,ab.	0
Other term group	40	RetinaLyze.ti,ab.	3
Other term group	41	RetmarkerSR DR.ti,ab.	0
Other term group	42	Singapore Eye Lesion Analyzer.ti,ab.	0
Other term group	43	RetinaVue.ti,ab.	6
Other term group	44	TRIAD network.ti,ab.	2
Other term group	45	LumineticsCore.ti,ab.	3
Other term group	46	Retmarker.ti,ab.	7
Other term group	47	DAIRET.ti,ab.	2
Other term group	48	eyRIS.ti,ab.	0
Other term group	49	(Automated Disease Assessment or ARDA).ti,ab.	93
Other term group	50	Medios.ti,ab.	840
Other term group	51	OphtAI.ti,ab.	1
Other term group	52	RetCAD.ti,ab.	8
Other term group	53	DeepDee.ti,ab.	0

Term Group	#	Search terms	Results
Other term group	54	MONA DR.ti,ab.	1
Other term group	55	AEYE-DS.ti,ab.	1
Other term group	56	EyeCheckup.ti,ab.	1
Other term group	57	Reti-EyeReti.ti,ab.	0
Other term group	58	DrNoon.ti,ab.	0
Other term group	59	ITOS Mass Screening.ti,ab.	0
Other term group	60	Nexy AI.ti,ab.	0
Other term group	61	EyeWisdom.ti,ab.	3
Other term group	62	iPredict System.ti,ab.	0
Other term group	63	TeleMedC DR grader.ti,ab.	0
Other term group	64	Eyetelligence system.ti,ab.	0
Other term group	65	DRISTi.ti,ab.	9
Other term group	66	VUNO Med.ti,ab.	5
Other term group	67	AutoGrader.ti,ab.	2
Other term group	68	UMI DR.ti,ab.	0
Other term group	69	or/27-68	370,017
	70	12 and 26 and 69	4,166
	71	limit 70 to ed=20200625-20250717	2,638
	72	limit 70 to dt=20200625-20250717	2,944
	73	limit 70 to ez=20200625-20250717	2,933
	74	limit 70 to ep=20200625-20250717	2,362
	75	71 or 72 or 73 or 74	3,174
	76	limit 75 to english language	3,122

Table 15. Search strategy for Embase (OVID).

Term Group	#	Search Terms	Results
Disease area	1	exp eye disease/di	189172
Disease area	2	retinopathy.ti,ab.	90023
Disease area	3	eye pathology.ti,ab.	526
Disease area	4	maculopathy.ti,ab.	7507
Disease area	5	diabetic eye.ti,ab.	1787
Disease area	6	diabetic macular.ti,ab.	10178
Disease area	7	retinal fundus.ti,ab.	1007
Disease area	8	eye photography/	8227
Disease area	9	fundus photograph*.ti,ab.	11944
Disease area	10	fundus camera*.ti,ab.	2056
Disease area	11	retinal camera*.ti,ab.	597
Disease area	12	retinal photograph*.ti,ab.	2189
Disease area	13	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12	288676
Study design	14	exp computer assisted diagnosis/	1701633
Study design	15	exp "sensitivity and specificity"/	557703
Study design	16	diagnostic test*.ti,ab.	95701
Study design	17	diagnostic accuracy.ti,ab.	107844
Study design	18	diagnostic performance.ti,ab.	47886
Study design	19	screening.ti,ab.	1127625
Study design	20	imaging.ti,ab.	1667795
Study design	21	(Sensitivity or specificity).ti,ab.	1860346
Study design	22	reference standard.ab.	31090
Study design	23	optical coherence tomography.ti,ab.	79048
Study design	24	early diagnosis/	153520
Study design	25	mass screening/	71650
Study design	26	or/14-25	5896411
Other term group	27	exp Artificial Intelligence/	149256
Other term group	28	artificial intelligence.ti,ab.	78566
Other term group	29	deep learning.ti,ab.	88120
Other term group	30	neural network*.ti,ab.	148059
Other term group	31	automated retinal image analysis system.ti,ab.	9
Other term group	32	automated grading.ti,ab.	230
Other term group	33	automated level.ti,ab.	13
Other term group	34	(automated adj (tool* or technique* or identification or detection)).ti,ab.	9869
Other term group	35	ARIAS.ti,ab.	756
Other term group	36	iGradingM.ti,ab.	4
Other term group	37	EyeArt.ti,ab.	60

Other term group	38	IDx-DR.ti,ab.	47
Other term group	39	DR-RACS.ti,ab.	1
Other term group	40	RetinaLyze.ti,ab.	8
Other term group	41	RetmarkerSR DR.ti,ab.	0
Other term group	42	Singapore Eye Lesion Analyzer.ti,ab.	0
Other term group	43	RetinaVue.ti,ab.	16
Other term group	44	TRIAD network.ti,ab.	3
Other term group	45	LumineticsCore.ti,ab.	4
Other term group	46	IDx-DR.ti,ab.	47
Other term group	47	Retmarker.ti,ab.	21
Other term group	48	DAIRET.ti,ab.	2
Other term group	49	eyRIS.ti,ab.	2
Other term group	50	(Automated Disease Assessment or ARDA).ti,ab.	123
Other term group	51	Medios.ti,ab.	130
Other term group	52	OphtAI.ti,ab.	2
Other term group	53	RetCAD.ti,ab.	8
Other term group	54	DeepDee.ti,ab.	0
Other term group	55	MONA DR.ti,ab.	2
Other term group	56	AEYE-DS.ti,ab.	7
Other term group	57	EyeCheckup.ti,ab.	5
Other term group	58	Reti-EyeReti.ti,ab.	0
Other term group	59	DrNoon.ti,ab.	0
Other term group	60	ITOS Mass Screening.ti,ab.	0
Other term group	61	EyeWisdom.ti,ab.	3
Other term group	62	Nexy AI.ti,ab.	0
Other term group	63	iPredict System.ti,ab.	3
Other term group	64	TeleMedC DR grader.ti,ab.	1
Other term group	65	Eyetelligence system.ti,ab.	0
Other term group	66	DRISTi.ti,ab.	8
Other term group	67	VUNO Med.ti,ab.	10
Other term group	68	AutoGrader.ti,ab.	5
Other term group	69	UMI DR.ti,ab.	0
Other term group	70	or/27-69	340989
	71	13 and 26 and 70	4043
	72	limit 71 to dd=20200625-20250717	3080
	73	limit 71 to dc=20200625-20250717	3097
	74	72 or 73	3105
	75	limit 74 to english language	3005

Table 16. Search strategy for The Cochrane Library (Wiley). Limited to English only and 2020-

Term Group	#	Search Terms	Results
Disease area	#1	MeSH descriptor: [Eye Diseases] explode all trees	27235
Disease area	#2	retinopathy:ti,ab	6276
Disease area	#3	eye pathology:ti,ab	472
Disease area	#4	maculopathy:ti,ab	371
Disease area	#5	diabetic eye:ti,ab	2053
Disease area	#6	diabetic macular:ti,ab	3135
Disease area	#7	retinal fundus:ti,ab	1171
Disease area	#8	fundus NEXT photograph*:ti,ab	976
Disease area	#9	fundus NEXT camera*:ti,ab	82
Disease area	#10	retinal NEXT photograph*:ti,ab	122
Disease area	#11	retinal NEXT camera*:ti,ab	18
Disease area	#12	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11	33146
Study design	#13	MeSH descriptor: [Diagnostic Techniques, Ophthalmological] explode all trees	11013
Study design	#14	MeSH descriptor: [Diagnosis, Computer-Assisted] explode all trees	2495
Study design	#15	MeSH descriptor: [Sensitivity and Specificity] explode all trees	21860
Study design	#16	(diagnostic NEXT test*):ti,ab	3331
Study design	#17	(diagnostic NEXT accuracy):ti,ab	3779
Study design	#18	(diagnostic NEXT performance):ti,ab	1353
Study design	#19	screening:ti,ab	80675
Study design	#20	imaging:ti,ab	53760

Term Group	#	Search Terms	Results
Study design	#21	(Sensitivity or specificity):ti,ab	66039
Study design	#22	(reference NEXT standard):ab	1273
Study design	#23	(optical NEXT coherence NEXT tomography):ti,ab	4332
Study design	#24	MeSH descriptor: [Early Diagnosis] this term only	789
Study design	#25	MeSH descriptor: [Mass Screening] this term only	5177
Study design	#26	#13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25	218025
Other term group	#27	MeSH descriptor: [Artificial Intelligence] explode all trees	3606
Other term group	#28	"artificial intelligence":ti,ab	2513
Other term group	#29	"deep learning":ti,ab	1240
Other term group	#30	(neural NEXT network*):ti,ab	1958
Other term group	#31	"automated retinal image analysis system":ti,ab	1
Other term group	#32	"automated grading":ti,ab	4
Other term group	#33	"automated level":ti,ab	1
Other term group	#34	(automated NEXT (tool* or technique* or identification or detection)):ti,ab	190
Other term group	#35	ARIAS:ti,ab	23
Other term group	#36	iGradingM:ti,ab	0
Other term group	#37	EyeArt:ti,ab	2
Other term group	#38	IDx-DR:ti,ab	0
Other term group	#39	DR-RACS:ti,ab	0

Term Group	#	Search Terms	Results
Other term group	#40	RetinaLyze:ti,ab	0
Other term group	#41	RetmarkerSR:ti,ab	0
Other term group	#42	"Singapore Eye Lesion Analyzer":ti,ab	0
Other term group	#43	RetinaVue:ti,ab	4
Other term group	#44	"TRIAD network":ti,ab	0
Other term group	#45	LumineticsCore:ti,ab	0
Other term group	#46	Retmarker:ti,ab	3
Other term group	#47	DAIRET:ti,ab	0
Other term group	#48	eyRIS:ti,ab	0
Other term group	#49	((Automated NEXT Disease NEXT Assessment) or (ARDA)):ti,ab	11
Other term group	#50	Medios:ti,ab	21
Other term group	#51	OphtAI:ti,ab	0
Other term group	#52	RetCAD:ti,ab	0
Other term group	#53	DeepDee:ti,ab	0
Other term group	#54	MONA NEXT DR:ti,ab	0
Other term group	#55	AEYE-DS:ti,ab	0
Other term group	#56	EyeCheckup:ti,ab	0
Other term group	#57	Reti-EyeReti:ti,ab	0
Other term group	#58	DrNoon:ti,ab	0
Other term group	#59	ITOS NEXT Mass NEXT Screening:ti,ab	0
Other term group	#60	EyeWisdom:ti,ab	0
Other term group	#61	Nexy NEXT AI:ti,ab	0
Other term group	#62	iPredict NEXT System:ti,ab	0
Other term group	#63	TeleMedC NEXT DR NEXT grader:ti,ab	0

Term Group	#	Search Terms	Results
Other term group	#64	Eyetelligence:ti,ab	0
Other term group	#65	DRISTi:ti,ab	5
Other term group	#66	VUNO NEXT Med:ti,ab	1
Other term group	#67	AutoGrader:ti,ab	0
Other term group	#68	UMI NEXT DR:ti,ab	0
Other term group	#69	#27 or #28 or #29 or #30 or #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38 or #39 or #40 or #41 or #42 or #43 or #44 or #45 or #46 or #47 or #48 or #49 or #50 or #51 or #52 or #53 or #54 or #55 or #56 or #57 or #58 or #59 or #60 or #61 or #62 or #63 or #64 or #65 or #66 or #67 or #68	7409
	#70	#12 and #26 and #69	168

Results were imported into Endnote and duplicates removed.

Table 17 reports the results from each database searched (total retrieved, duplicated removed and final number of unique records used for title and abstract screening) for the original database search and the 4 update searches.

Table 17. Results from each database searched (total retrieved, duplicates removed, final number of unique records)

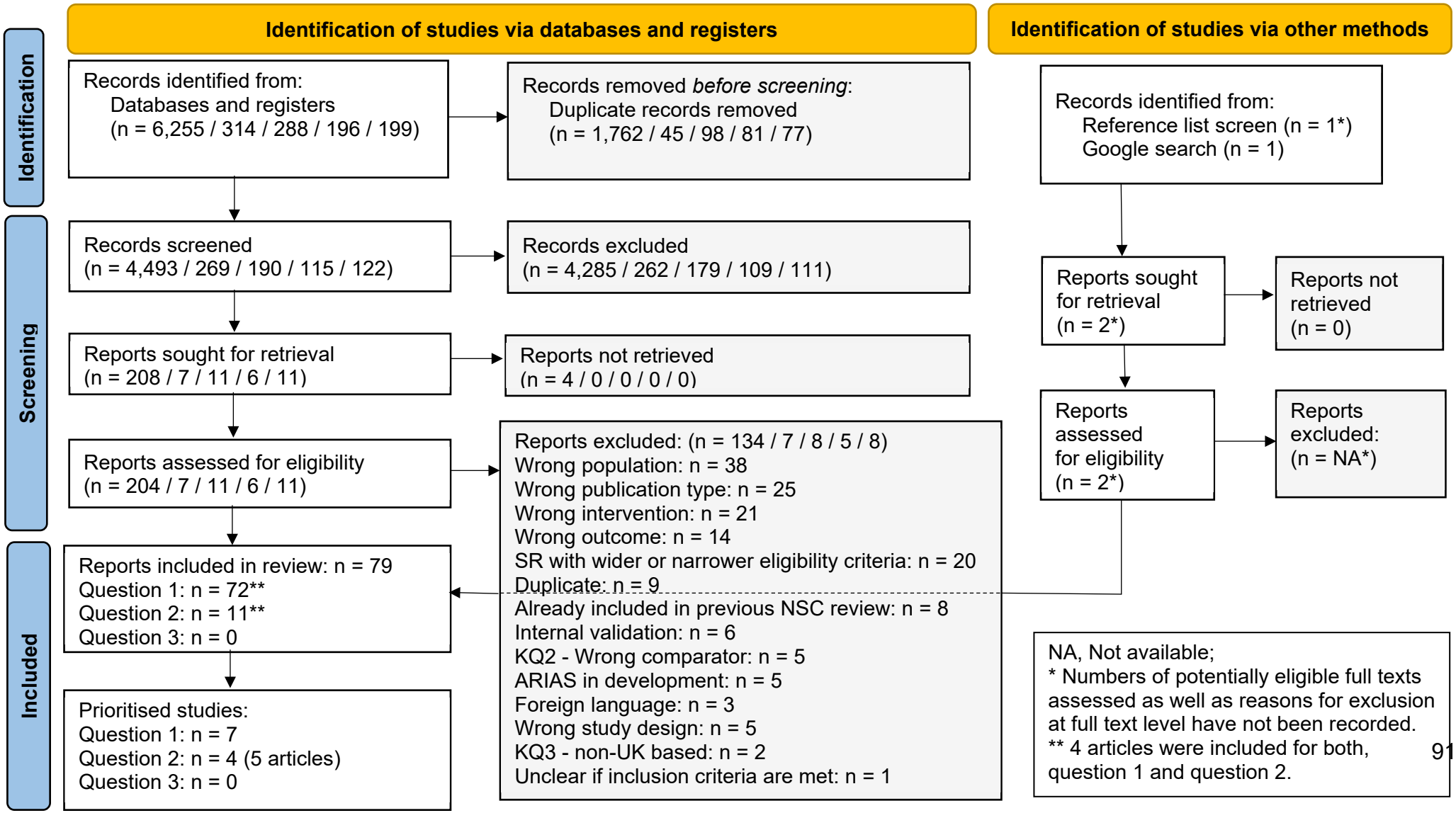
Search and search date	Medline ALL	Embase, inclusive CT.gov	Cochrane Reviews	Cochrane Trials	ICTRP	TOTAL
Original search 17/07/25	3,122	3,005	2	123	121	**6,373**
Original search 17/07/25: From June 2020	3,061	2,948	2	123	121	**6,255**
Original search 17/07/25: Deduplicated	3,055	1,300	0	32	106	**4,493**
August Update 29/08/25	94	218 (excl CT.gov)	0	2	NA	**314**
August Update 29/08/25: Deduplicated	88	179	0	2	NA	**269**
August (2) Update 03/10/25	57	36	0	NA	NA	**93**
August (2) Update 03/10/25: Deduplicated	57	28	0	NA	NA	**85 (80)***
September Update 03/10/25	114	81	0	NA	NA	**195**
September Update 03/10/25: Deduplicated	90	30	0	NA	NA	**120 (110)***
October Update 04/11/25	108	85	3	NA	NA	**196**
October Update 04/11/25: Deduplicated	88	26	1	NA	NA	**115**
November Update 02/12/25	108	91	0	NA	NA	**199**
November Update 02/12/25: Deduplicated	86	36	0	NA	NA	**122**

CT.gov, ClinicalTrials.gov; excl, Exclusive; ICTRP, International Clinical Trials Registry Platform; NA, Not available.

*August (2) update and September update further deduplicated against first August update search and further duplicates removed. Results for August (2) and September updates combined in the PRISMA flow diagram (80 + 110 = 190).

Appendix 2 — Included and excluded studies

Figure 12. PRISMA diagram. Summary of publications included and excluded at each stage of the review (original searches / August update search / September update search / October update search / November update search)



Publications included after review of full-text articles

The 79 publications included after review of full texts are summarised in **Table 18** below.

Table 18. Summary of publications included after review of full-text articles, and the question(s) each publication was identified as being relevant to the questions

Study	Question 1 (Accuracy)	Question 2 (Clinical impact)	Question 3 (Cost effectiveness)	Comments (Prioritised / deprioritised)
Abreu-Gonzalez 2025[74]	Yes	No	No	Deprioritised
Al-Turk 2022[75]	Yes	Yes	No	Deprioritised
Antaki 2024[76]	Yes	No	No	Deprioritised
Arenas-Cavalli 2022[77]	Yes	No	No	Deprioritised
Baget-Bernaldiz 2024[78]	Yes	No	No	Deprioritised
Baget-Bernaldiz 2021[79]	Yes	No	No	Deprioritised
Brant 2025[80]	Yes	No	No	Deprioritised
Burlina 2024[81]	Yes	No	No	Deprioritised
Chen 2025[65]	No	Yes	No	Prioritised
Chia 2023[82]	Yes	No	No	Deprioritised
Chotcomwongse 2025[83]	No	Yes	No	Deprioritised
Curran 2023[84]	Yes	No	No	Deprioritised
Dai 2021[85]	Yes	No	No	Deprioritised
Dogan 2024[86]	Yes	No	No	Deprioritised
Dong 2022[87]	Yes	No	No	Deprioritised
Dow 2023a[57]	Yes	No	No	Prioritised
Dow 2023b[63]	No	Yes	No	Prioritised
Duggal, 2025[88]	Yes	No	No	Deprioritised
Fleming 2023[15]	Yes	No	No	Prioritised
Grzybowski 2021[89]	Yes	No	No	Deprioritised
Grzybowski 2024[90]	Yes	No	No	Deprioritised
Grzybowski 2025[91]	Yes	No	No	Deprioritised
Grzybowski 2023[92]	Yes	No	No	Deprioritised
Guedes 2025[93]	Yes	No	No	Deprioritised
Hao 2022a[94]	Yes	No	No	Deprioritised
Hao 2022b[95]	Yes	No	No	Deprioritised
Heuer 2025[67]	No	Yes	No	Prioritised

Study	Question 1 (Accuracy)	Question 2 (Clinical impact)	Question 3 (Cost effectiveness)	Comments (Prioritised / deprioritised)
Hu 2024[96]	Yes	No	No	Deprioritised
Huang 2024[68]	No	Yes	No	Prioritised
Ipp 2021[97]	Yes	No	No	Deprioritised
Karabeg 2024[98]	Yes	No	No	Deprioritised
Karabeg 2025[99]	Yes	No	No	Deprioritised
Lee 2021[58]	Yes	No	No	Prioritised
Li 2021a[100]	Yes	No	No	Deprioritised
Li 2021b[101]	Yes	Yes	No	Deprioritised
Lim 2023[102]	Yes	No	No	Deprioritised
Limwattanayingyong 2020[103]	Yes	No	No	Deprioritised
Mathenge 2022[104]	No	Yes	No	Deprioritised
Mbaye 2025[105]	Yes	No	No	Deprioritised
Mehra 2022[106]	Yes	No	No	Deprioritised
Meredith 2023[107]	Yes	No	No	Deprioritised
Ming 2021[108]	Yes	No	No	Deprioritised
Mokhashi 2022[109]	Yes	No	No	Deprioritised
Musetti 2025[110]	Yes	No	No	Deprioritised
Nissen 2023[59]	Yes	No	No	Prioritised
Noriega 2021[111]	Yes	No	No	Deprioritised
Parravano 2025[112]	Yes	No	No	Deprioritised
Peeters 2023[113]	Yes	No	No	Deprioritised
Pei 2022[114]	Yes	No	No	Deprioritised
Peris-Martinez 2021[115]	Yes	No	No	Deprioritised
Piatti 2024[116]	Yes	No	No	Deprioritised
Piatti 2025[117]	Yes	No	No	Deprioritised
Pinto 2025[66]	Yes	Yes	No	Prioritised
Poschkamp 2005[62]	Yes	No	No	Prioritised
Rahmati 2025[118]	Yes	No	No	Deprioritised
Rao 2022[119]	Yes	No	No	Deprioritised
Riotto 2024[120]	Yes	No	No	Deprioritised
Ruamviboonsuk 2022[121]	Yes	Yes	No	Deprioritised
Ruan 2022[122]	Yes	No	No	Deprioritised
Rudnicka 2025[61]	Yes	No	No	Prioritised
Sarao 2020[123]	Yes	No	No	Deprioritised

Study	Question 1 (Accuracy)	Question 2 (Clinical impact)	Question 3 (Cost effectiveness)	Comments (Prioritised / deprioritised)
Scheetz 2021[124]	Yes	No	No	Deprioritised
Sedova 2022[125]	Yes	No	No	Deprioritised
Sin 2025[60]	Yes	No	No	Prioritised
Skevas 2024[126]	Yes	No	No	Deprioritised
Tan-Torres 2025[127]	Yes	No	No	Deprioritised
Teoh 2023[128]	Yes	No	No	Deprioritised
Tsai 2022[129]	Yes	No	No	Deprioritised
Vaghefi 2022[130]	Yes	No	No	Deprioritised
Van 2024[131]	Yes	No	No	Deprioritised
Wang 2024[132]	Yes	No	No	Deprioritised
Whitestone 2024[133]	Yes	No	No	Deprioritised
Wintergerst 2022[134]	Yes	No	No	Deprioritised
Wongchaisuwat 2021[135]	Yes	No	No	Deprioritised
Xu 2025[136]	Yes	No	No	Deprioritised
Yang 2022[137]	Yes	Yes	No	Deprioritised
Yao 2024[138]	Yes	No	No	Deprioritised
Zhang 2022[139]	Yes	No	No	Deprioritised
Zhang 2020[140]	Yes	No	No	Deprioritised

Studies were prioritised for extraction and data synthesis for questions 1 and 2. It was planned a priori that the following approach would be taken to prioritise studies for extraction:

Question 1

The highest priority study would be a comparative UK-based study that evaluates the performance of commercially available, CE-marked and/or FDA approved ARIAS(s) compared to human primary graders to detect ‘Any retinopathy’ (R1 or higher and/or M1) in dilated eyes (with mydriasis) using an external, geographically separate study population.

In the absence, or minimal volume, of such studies, we prioritised studies using a combination of the following criteria in discussion with the UK NSC:

- Studies from comparable countries (e.g. North-Western European, USA, Canada, Australia);
- Studies assessing ARIAS with (or pending) CE-mark and/or FDA-approval;
- Studies assessing the performance of ARIAS in undilated eyes;

- Studies reporting ‘Referable retinopathy’ (R2 or higher and/or M1) only;
- Studies using an external, temporal validation.

Question 2

The ideal study would be a systematic review of UK studies. In the absence of a suitable systematic review, primary UK-based RCTs or comparative prospective cohort studies were prioritised next.

In the absence, or minimal volume, of such studies, we prioritised studies as follows:

- Studies from comparable countries (North-Western European, America, Australia);
- Retrospective cohort studies or other study designs.

Publications excluded after review of full text articles

Of the 243 publications included after the review of titles and abstracts, 166 were ultimately judged not to be relevant to this review. These publications, along with reasons for exclusion, are listed in **Table 19** for the original search (n = 138), **Table 20** for the August search update (n = 7), **Table 21** for the September search update (n = 8), **Table 22** for the October search update (n = 5) and **Table 23** for the November search update (n = 8).

Table 19. Publications excluded after review of full text articles from original search (n = 138, including 4 articles with full text not available)

Reference		Reason for exclusion
1.	Acharyya M, Moharana B, Jain S, Tandon M. A double-blinded study for quantifiable assessment of the diagnostic accuracy of AI tool "ADVEN-i" in identifying diseased fundus images including diabetic retinopathy on a retrospective data. Indian Journal of Ophthalmology. 2024;72:S46-S52.	Wrong population
2.	Alqahtani AS, Alshareef WM, Aljadani HT, Hawsawi WO, Shaheen MH. The efficacy of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis. International Journal of Retina and Vitreous. 2025;11(1):48.	Wrong study design
3.	Anton N, Doroftei B, Curteanu S, Catalin L, Ilie O-D, Tarcoveanu F, et al. Comprehensive Review on the Use of Artificial Intelligence in Ophthalmology and Future Research Directions. Diagnostics 2022;13(1).	Duplicate
4.	Anton N, Doroftei B, Curteanu S, Catalin L, Ilie OD, Tarcoveanu F, et al. Comprehensive Review on the Use of Artificial Intelligence in Ophthalmology and Future Research Directions. Diagnostics. 2023;13(1):100.	Systematic review including studies published before 2020
5.	Arunga S, Morley KE, Kwaga T, Morley MG, Nakayama LF, Mwavu R, et al. Assessment of Clinical Metadata on the Accuracy of Retinal Fundus Image Labels in Diabetic Retinopathy in Uganda: Case-Crossover Study Using the Multimodal Database of Retinal Images in Africa. JMIR Formative Research. 2024;8:e59914.	Wrong intervention
6.	Bajwa A, Nosheen N, Talpur KI, Akram S. A Prospective Study on Diabetic Retinopathy Detection Based on Modify Convolutional Neural Network Using Fundus Images at Sindh Institute of Ophthalmology & Visual Sciences. Diagnostics 2023;13(3).	Wrong intervention, in development
7.	Beuse A, Grohmann C, Schadwinkel HM, Skevas C, Spitzer MS. Artificial Intelligence for the Detection of Diabetic Retinopathy. Kunstliche-Intelligenz-Systeme zur Erkennung der Diabetischen Retinopathie. 2025.	Foreign language

Reference		Reason for exclusion
8.	Bhambhwani V, Whitestone N, Patnaik JL, Ojeda A, Scali J, Cherwek DH. Feasibility and Patient Experience of a Pilot Artificial Intelligence-Based Diabetic Retinopathy Screening Program in Northern Ontario. Ophthalmic Epidemiology. 2024:1-7.	Wrong comparator
9.	Bhuiyan A, Govindaiah A, Alauddin S, Otero-Marquez O, Smith RT. Combined automated screening for age-related macular degeneration and diabetic retinopathy in primary care settings. Annals of Eye Science. 2021;6.	Wrong population
10.	Bhuiyan A, Govindaiah A, Deobhakta A, Gupta M, Rosen R, Saleem S, et al. Development and Validation of an Automated Diabetic Retinopathy Screening Tool for Primary Care Setting. Diabetes Care. 2020;43(10):e147-e8.	Unclear if inclusion criteria met
11.	Bhuiyan A, Govindaiah A, Deobhakta A, Hossain M, Rosen R, Smith T. Automated diabetic retinopathy screening for primary care settings using deep learning. Intelligence-based Medicine. 2021;5.	Wrong intervention
12.	Bhulakshmi D, Rajput DS. A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. PeerJ Computer Science. 2024;10:e1947.	Wrong publication
13.	Bi Z, Li J, Liu Q, Fang Z. Deep learning-based optical coherence tomography and retinal images for detection of diabetic retinopathy: a systematic and meta analysis. Frontiers in Endocrinology. 2025;16:1485311.	Systematic review including studies published before 2020
14.	Boyle J, Vignarajan J, Saha S. Automated Diabetic Retinopathy Diagnosis for Improved Clinical Decision Support. Studies in Health Technology and Informatics. 2024;310:1490-1.	Wrong publication type
15.	Cao S, Zhang R, Jiang A, Kuerban M, Wumaier A, Wu J, et al. Application effect of an artificial intelligence-based fundus screening system: evaluation in a clinical setting and population screening. Biomedical Engineering Online. 2023;22(1):38.	Wrong population
16.	Chawla R, Karkhanis P, Shah M, Das A, Sharma R, Almaula D, et al. Artificial intelligence for advancing eye care in resource-poor settings: Assessing the predictive accuracy of an AI-model for diabetic retinopathy screening in India. Global Epidemiology. 2025;9:100209.	Wrong population
17.	Chen D, Lee S, Elgin C, Zhou R, Geevarghese A, Al-Aswad LA. Assessment of AI Algorithms for Diabetic Retinopathy Classification Using Model Cards. Investigative Ophthalmology and Visual Science. 2022;63(7):3005 EP - F0275.	Wrong publication type
18.	Chen EM, Chen D, Chilakamarri P, Lopez R, Parikh R. Economic Challenges of Artificial Intelligence Adoption for Diabetic Retinopathy. Ophthalmology. 2021;128(3):475-7.	Non-UK based
19.	Ciapponi A, Ballivian J, Gentile C, Mejia JR, Ruiz-Baena J, Bardach A. Diagnostic utility of artificial intelligence software through non-mydriatic digital retinography in the screening of diabetic retinopathy: an overview of reviews. Eye (Basingstoke). 2025;39(10):2083 EP - 9.	Systematic review including studies published before 2020
20.	Cicinelli MV, Gravina S, Rutigliani C, Checchin L, La Franca L, Lattanzio R, et al. Assessing Diabetic Retinopathy Staging With AI: A Comparative Analysis Between Pseudocolor and LED Imaging. Translational Vision Science & Technology. 2024;13(3):11.	Wrong population

Reference	Reason for exclusion
21. Cleland CR, Bascaran C, Makupa W, Shilio B, Tufail A, Egan CA, et al. Head-to-Head Comparative Evaluation of Four Commercially Available Artificial Intelligence Systems for Detecting Referable Diabetic Retinopathy in an African Population. SSRN. 2025.	Wrong outcome
22. Cleland CR, Makupa WU, Shilio BR, Rwiza J, Macleod D, Bascaran C, et al. Implementing an artificial intelligence system into a diabetic eye screening programme in Tanzania. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2025;119(6):567-9.	Full text not available
23. Cruz MNC, Dampil OAC, Untalan PGGC, Agustin ND, Chao PMG. Diagnostic Accuracy of Point-of-Care MediosTM Artificial Intelligence Aided Fundus Photography in Detecting Diabetic Retinopathy among Filipino Patients with Type 2 Diabetes Mellitus. Phillippine Journal of Internal Medicine. 2025;63(1):7 EP - 15.	Wrong population
24. Cuadros J. The Real-World Impact of Artificial Intelligence on Diabetic Retinopathy Screening in Primary Care. Journal of Diabetes Science and Technology. 2021;15(3):664-5.	Wrong publication type
25. Dai L, Sheng B, Chen T, Wu Q, Liu R, Cai C, et al. A deep learning system for predicting time to progression of diabetic retinopathy. Nature Medicine. 2024;30(2):584-94.	Wrong intervention
26. Daley JR, Wang X, Ngo N, Khoo CL, Heydon P, Liew G, et al. Development and Validation of a Deep Learning System for the Provision of a District-Wide Diabetes Retinal Screening Service. Clinical & Experimental Ophthalmology. 2025.	Wrong population
27. Djoumessi K, Huang Z, Kuhlewein L, Rickmann A, Simon N, Koch LM, et al. An Inherently Interpretable AI model improves Screening Speed and Accuracy for Early Diabetic Retinopathy. medRxiv. 2024.	Wrong intervention
28. Djoumessi K, Huang Z, Kuhlewein L, Rickmann A, Simon N, Koch LM, et al. An inherently interpretable AI model improves screening speed and accuracy for early diabetic retinopathy. PLOS digital health. 2025;4(5):e0000831.	Duplicate
29. Donaghue KC, Liew G. Measuring Outcomes of Diabetic Retinopathy Screening: What Is Important? Diabetes Care. 2024;47(6):930 EP - 2.	Wrong publication type
30. Dos Reis MA, Kunas CA, da Silva Araujo T, Schneiders J, de Azevedo PB, Nakayama LF, et al. Advancing healthcare with artificial intelligence: diagnostic accuracy of machine learning algorithm in diagnosis of diabetic retinopathy in the Brazilian population. Diabetology & Metabolic Syndrome. 2024;16(1):209.	Internal validation
31. Esmaeilkhanian H, Gutierrez KG, Myung D, Fisher AC. Detection Rate of Diabetic Retinopathy Before and After Implementation of Autonomous AI-based Fundus Photograph Analysis in a Resource-Limited Area in Belize. Clinical Ophthalmology (Auckland, NZ). 2025;19:993-1006.	Wrong population
32. Farahat Z, Zrira N, Souissi N, Bennani Y, Bencherif S, Benamar S, et al. Diabetic retinopathy screening through artificial intelligence algorithms: A systematic review. Survey of Ophthalmology. 2024;69(5):707-21.	Wrong outcome
33. Farooq MS, Arooj A, Alroobaea R, Baqasah AM, Jabarulla MY, Singh D, et al. Untangling Computer-Aided Diagnostic System for Screening Diabetic Retinopathy Based on Deep Learning Techniques. Sensors (Basel, Switzerland). 2022;22(5).	Systematic review including studies published before 2020

Reference	Reason for exclusion
34. Font O, Torrents-Barrena J, Royo D, Garcia SB, Zarranz-Ventura J, Bures A, et al. Validation of an autonomous artificial intelligence-based diagnostic system for holistic maculopathy screening in a routine occupational health checkup context. Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie. 2022;260(10):3255-65.	Wrong population
35. Fuller SD, Hu J, Liu JC, Gibson E, Gregory M, Kuo J, et al. Five-Year Cost-Effectiveness Modeling of Primary Care-Based, Nonmydriatic Automated Retinal Image Analysis Screening Among Low-Income Patients With Diabetes. Journal of Diabetes Science and Technology. 2022;16(2):415-27.	Non-UK based
36. Gognieva D, Durzhinskaya M, Vorobyeva I, Chomakhidze P, Suvorov A, Kuznetsova N, et al. Diabetic Retinopathy Diagnosis based on Convolutional Neural Network in the Russian Population: A Multicenter Prospective Study. Current diabetes reviews. 2024;20(8):77-83.	Wrong population
37. Goh JHL, Ang E, Srinivasan S, Lei X, Loh J, Quek TC, et al. Comparative Analysis of Vision Transformers and Conventional Convolutional Neural Networks in Detecting Referable Diabetic Retinopathy. Ophthalmology science. 2024;4(6):100552.	Wrong intervention
38. Goldstein J, Weitzman D, Lemerond M, Jones A. Determinants for scalable adoption of autonomous AI in the detection of diabetic eye disease in diverse practice types: key best practices learned through collection of real-world data. Frontiers in digital health. 2023;5:1004130.	Wrong comparator, wrong study design
39. 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 ophthalmologica. 2020;98(4):368-77.	Already included in previous UK NSC review
40. Grzybowski A, Brona P, Krzywicki T, Gaca-Wysocka M, Berlinska A, Swiech A. Variability of Grading DR Screening Images among Non-Trained Retina Specialists. Journal of clinical medicine. 2022;11(11).	Wrong outcome
41. Grzybowski A, Jin K, Zhou J, Pan X, Wang M, Ye J, et al. Retina Fundus Photograph-Based Artificial Intelligence Algorithms in Medicine: A Systematic Review. Ophthalmology and therapy. 2024;13(8):2125-49.	Wrong outcome
42. Gu C, Wang Y, Jiang Y, Xu F, Wang S, Liu R, et al. Application of artificial intelligence system for screening multiple fundus diseases in Chinese primary healthcare settings: a real-world, multicentre and cross-sectional study of 4795 cases. The British journal of ophthalmology. 2024;108(3):424-31.	Wrong population
43. Han R, Cheng G, Zhang B, Yang J, Yuan M, Yang D, et al. Validating automated eye disease screening AI algorithm in community and in-hospital scenarios. Frontiers in public health. 2022;10:944967.	Wrong population
44. Han R, Yu W, Chen H, Chen Y. Using artificial intelligence reading label system in diabetic retinopathy grading training of junior ophthalmology residents and medical students. BMC medical education. 2022;22(1):258.	Wrong intervention

Reference	Reason for exclusion
45. Hao Z, Cui S, Zhu Y, Shao H, Huang X, Jiang X, et al. Application of non-mydriatic fundus examination and artificial intelligence to promote the screening of diabetic retinopathy in the endocrine clinic: an observational study of T2DM patients in Tianjin, China. Therapeutic advances in chronic disease. 2020;11:2040622320942415.	Wrong comparator, wrong study design
46. 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 (London, England). 2020;34(3):572-6.	Already included in previous UK NSC review
47. He S, Bulloch G, Zhang L, Xie Y, Wu W, He Y, et al. Cross-camera Performance of Deep Learning Algorithms to Diagnose Common Ophthalmic Diseases: A Comparative Study Highlighting Feasibility to Portable Fundus Camera Use. Current eye research. 2023;48(9):857-63.	Wrong population
48. 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. 2021;105(5):723-8.	Already included in previous UK NSC review
49. Hsieh Y-T, Chuang L-M, Jiang Y-D, Chang T-J, Yang C-M, Yang C-H, et al. Application of deep learning image assessment software VeriSee TM for diabetic retinopathy screening. Journal of the Formosan Medical Association = Taiwan yi zhi. 2021;120(1):165-71.	Internal validation
50. Huber SL, Parzer V, Ludvik B, Pollreisz A, Mahnert N, Brix JM. Evaluation of IDx-DR software for diabetic retinopathy screening in outpatient clinics: Efficacy, safety, and feasibility in a real-world setting. Journal of diabetes and its complications. 2025;39(10):109120.	Wrong outcome
51. Joseph S, Selvaraj J, Mani I, Kumaragurupari T, Shang X, Mudgil P, et al. Diagnostic Accuracy of Artificial Intelligence-Based Automated Diabetic Retinopathy Screening in Real-World Settings: A Systematic Review and Meta-Analysis. American journal of ophthalmology. 2024;263:214-30.	Wrong study design
52. Kalavar M, Al-Khersan H, Sridhar J, Gorniak RJ, Lakhani PC, Flanders AE, et al. Applications of Artificial Intelligence for the Detection, Management, and Treatment of Diabetic Retinopathy. International ophthalmology clinics. 2020;60(4):127-45.	Wrong publication type
53. Karabeg M, Petrovski G, Hertzberg SN, Erke MG, Fosmark DS, Russell G, et al. A pilot cost-analysis study comparing AI-based EyeArt R and ophthalmologist assessment of diabetic retinopathy in minority women in Oslo, Norway. International journal of retina and vitreous. 2024;10(1):40.	Duplicate
54. Katz O, Presil D, Cohen L, Nachmani R, Kirshner N, Hoch Y, et al. Evaluation of a New Neural Network Classifier for Diabetic Retinopathy. Journal of diabetes science and technology. 2022;16(6):1401-9.	In development
55. Khan Z, Gaidhane AM, Singh M, Ganesan S, Kaur M, Sharma GC, et al. Diagnostic Accuracy of IDX-DR for Detecting Diabetic Retinopathy: A Systematic Review and Meta-Analysis. American journal of ophthalmology. 2025;273:192-204.	Wrong intervention
56. Kiran Mayee M, Humera Khanam M. Artificial Intelligence and Deep Learning Methods for Ophthalmic Image Processing: A Review. NeuroQuantology. 2022;20(10):8021 EP - 31.	Wrong publication type

Reference	Reason for exclusion
57. Komatsu K, Sano K, Fukai K, Nakagawa R, Nakagawa T, Tatemichi M, et al. Associated factors of diabetic retinopathy by artificial intelligence evaluation of fundus images in Japan. Scientific reports. 2023;13(1):19742.	Wrong outcome
58. Komatsu K, Sano K, Fukai K, Nakagawa R, Nakagawa T, Tatemichi M, et al. Standardized evaluation of diabetic retinopathy using artificial intelligence and its association with metabolic dysfunction-associated steatotic liver disease in Japan: A cross-sectional study. PloS one. 2024;19(12):e0315752.	Wrong outcome
59. Krogh M, Hentze M, Jensen MSA, Jensen MB, Nielsen MG, Vorum H, et al. Valuable insights into general practice staff's experiences and perspectives on AI-assisted diabetic retinopathy screening-An interview study. Frontiers in medicine. 2025;12:1565532.	Wrong outcome
60. Kummerle D, Beals D, Simon L, Rogers F, Pogroszewski S. Revolutionizing Diabetic Retinopathy Screening: Integrating AI-Based Retinal Imaging in Primary Care. Journal of CME. 2025;14(1):2437294.	Wrong intervention
61. Kurian DE, Kalra S, Kapoor N. Screening for diabetic retinopathy in primary care: Future prospects in low-middle income countries. JPMA The Journal of the Pakistan Medical Association. 2021;71(12):2826-7.	Wrong publication type
62. Labib KM, Ghumman H, Jain S, Jarstad JS. A Review of the Utility and Limitations of Artificial Intelligence in Retinal Disorders and Pediatric Ophthalmology. Cureus. 2024;16(10):e71063.	Wrong publication type
63. Lakshminarayanan V, Kheradfallah H, Sarkar A, Jothi Balaji J. Automated Detection and Diagnosis of Diabetic Retinopathy: A Comprehensive Survey. Journal of imaging. 2021;7(9).	Systematic review including studies published before 2020
64. Larsen TJ, Pettersen MB, Nygaard Jensen H, Lynge Pedersen M, Lund-Andersen H, Jorgensen ME, et al. The use of artificial intelligence to assess diabetic eye disease among the Greenlandic population. International journal of circumpolar health. 2024;83(1):2314802.	In development
65. Lee AY, Lee CS, Hunt MS, Yanagihara RT, Blazes M, Boyko EJ. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. Diabetes Care 2021;44:XXXX-XXXX. Diabetes care. 2021;44(5):e108-e9.	Wrong publication type
66. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, realworld validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care. 2021;44(5):1168 EP - 75.	Duplicate
67. Li F, Pan J, Yang D, Wu J, Ou Y, Li H, et al. A Multicenter Clinical Study of the Automated Fundus Screening Algorithm. Translational vision science & technology. 2022;11(7):22.	Wrong population
68. Lin S, Ma Y, Jiang Y, Li W, Peng Y, Yu T, et al. Service Quality and Residents' Preferences for Facilitated Self-Service Fundus Disease Screening: Cross-Sectional Study. Journal of Medical Internet Research. 2024;26(1):e45545.	Wrong comparator

Reference	Reason for exclusion
69. Li Q, Drinkwater JJ, Woods K, Douglas E, Ramirez A, Turner AW. Implementation of A New, Mobile Diabetic Retinopathy Screening Model Incorporating Artificial Intelligence in Remote Western Australia. The Australian journal of rural health. 2025;33(2):e70031.	Wrong population
70. 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. Ophthalmology Retina. 2021;5(1):71-7.	Already included in previous UK NSC review
71. Liu R, Wang X, Wu Q, Dai L, Fang X, Yan T, et al. DeepDRiD: Diabetic Retinopathy-Grading and Image Quality Estimation Challenge. Patterns (New York, NY). 2022;3(6):100512.	Internal validation
72. Liu R, Li Q, Xu F, Wang S, He J, Cao Y, et al. Application of artificial intelligence-based dual-modality analysis combining fundus photography and optical coherence tomography in diabetic retinopathy screening in a community hospital. Biomedical engineering online. 2022;21(1):47.	In development
73. Liu TYA, Wolf RM. Autonomous Artificial Intelligence for Diabetic Eye Disease Testing Improves Access and Equity in the Pediatric and Adult Populations: The Johns Hopkins Medicine Experience. Diabetes spectrum : a publication of the American Diabetes Association. 2025;38(1):19-22.	Wrong publication type
74. Li Z, Guo X, Zhang J, Liu X, Chang R, He M. Using deep leaning models to detect ophthalmic diseases: A comparative study. Frontiers in medicine. 2023;10:1115032.	Wrong population
75. Mellor J, Jiang W, Fleming A, McGurnaghan SJ, Blackbourn LAK, Styles C, et al. Prediction of retinopathy progression using deep learning on retinal images within the Scottish screening programme. The British journal of ophthalmology. 2024;108(6):833-9.	Internal validation, wrong intervention
76. Menia NK, Diwan S, Mehndiratta A, Venkatesh P. Machine learning and its current and future applications in the management of vitreoretinal disorders. Expert Review of Ophthalmology. 2024;19(3):227 EP - 42.	Wrong publication type
77. Midena E, Zennaro L, Lapo C, Torresin T, Midena G, Pilotto E, et al. Handheld Fundus Camera for Diabetic Retinopathy Screening: A Comparison Study with Table-Top Fundus Camera in Real-Life Setting. Journal of clinical medicine. 2022;11(9).	Wrong intervention
78. Moannaei M, Jadidian F, Doustmohammadi T, Kiapasha AM, Bayani R, Rahmani M, et al. Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: a meta-analysis. Biomedical engineering online. 2025;24(1):34.	Internal validation
79. Nanegrungsunk O, Ruamviboonsuk P, Grzybowski A. Prospective studies on artificial intelligence (AI)-based diabetic retinopathy screening. Annals of translational medicine. 2022;10(24):1297.	Systematic review including studies published before 2020

Reference	Reason for exclusion
80. Nemcansky J, Studnicka J, Vyslouzilova D, Ernest J, Nemec P. Screening diabeticke retinopatie a diabetickeho makularniho edemu, Diabetic Retinopathy and Diabetic Macular Edema - Screening. Ceska a slovenska oftalmologie : casopis Ceske oftalmologicke spolecnosti a Slovenske oftalmologicke spolecnosti. 2023;79(5):250 EP - 5.	Wrong publication type
81. Nolan B, Daybranch ER, Barton K, Korsen N. Patient and Provider Experience with Artificial Intelligence Screening Technology for Diabetic Retinopathy in a Rural Primary Care Setting. Journal of Maine Medical Center. 2023;5(2).	Wrong outcome
82. Nunez do Rio JM, Nderitu P, Bergeles C, Sivaprasad S, Tan GSW, Raman R. Evaluating a Deep Learning Diabetic Retinopathy Grading System Developed on Mydriatic Retinal Images When Applied to Non-Mydriatic Community Screening. Journal of clinical medicine. 2022;11(3).	Wrong population
83. Ogunyemi OI, Gandhi M, Lee M, Teklehaimanot S, Daskivich LP, Hindman D, et al. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. JAMIA open. 2021;4(3):ooab066.	Wrong intervention
84. 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. The British journal of ophthalmology. 2021;105(2):265-70.	Already included in previous UK NSC review
85. Paul SK, Kim CU, Shieh D, Zhou XY, Pan I, Mehra AA, et al. Impact of an Artificial Intelligence Algorithm on Diabetic Retinopathy Grading by Ophthalmology Residents. medRxiv. 2023.	Wrong intervention
86. Pawar B, Lobo SN, Joseph M, Jegannathan S, Jayraj H. Validation of Artificial Intelligence Algorithm in the Detection and Staging of Diabetic Retinopathy through Fundus Photography: An Automated Tool for Detection and Grading of Diabetic Retinopathy. Middle East African journal of ophthalmology. 2021;28(2):81-6.	Wrong population
87. Pietris J, Lam A, Bacchi S, Gupta AK, Kovoor JG, Chan WO. Health Economic Implications of Artificial Intelligence Implementation for Ophthalmology in Australia: A Systematic Review. Asia-Pacific journal of ophthalmology (Philadelphia, Pa). 2022;11(6):554-62.	Systematic review including studies published before 2020
88. Porwal P, Pachade S, Kokare M, Deshmukh G, Son J, Bae W, et al. IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge. Medical image analysis. 2020;59:101561.	Internal validation
89. Poschkamp B, Stahl A. Application of deep learning algorithms for diabetic retinopathy screening. Annals of translational medicine. 2022;10(24):1298.	Wrong publication type
90. Ramoutar RR. An Economic Analysis for the Use of Artificial Intelligence in Screening for Diabetic Retinopathy in Trinidad and Tobago. Cureus. 2024;16(3):e55745.	Systematic review including studies published before 2020
91. Rao DP, Savoy FM, Sivaraman A, Dutt S, Shahsuvaryan M, Jrbashyan N, et al. Evaluation of an AI algorithm trained on an ethnically diverse dataset to screen a previously unseen population for diabetic retinopathy. Indian journal of ophthalmology. 2024;72(8):1162-7.	Wrong population

Reference	Reason for exclusion
92. Rego S, Dutra-Medeiros M, Soares F, Monteiro-Soares M. Screening for Diabetic Retinopathy Using an Automated Diagnostic System Based on Deep Learning: Diagnostic Accuracy Assessment. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2021;244(3):250-7.	In development, internal validation
93. 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. Telemedicine journal and e-health : the official journal of the American Telemedicine Association. 2020;26(8):1001-9.	Already included in previous UK NSC review
94. Ruamviboonsuk P, Chantra S, Seresirikachorn K, Ruamviboonsuk V, Sangroongruangsri S. Economic Evaluations of Artificial Intelligence in Ophthalmology. Asia-Pacific journal of ophthalmology (Philadelphia, Pa). 2021;10(3):307-16.	Wrong publication type
95. Salongcay RP, Jacoba CMP, Salva CMG, Rageh A, Aquino LAC, Saunar AV, et al. One-field, two-field and five-field handheld retinal imaging compared with standard seven-field Early Treatment Diabetic Retinopathy Study photography for diabetic retinopathy screening. The British journal of ophthalmology. 2024;108(5):735-41.	Wrong intervention
96. Savoy M. IDx-DR for Diabetic Retinopathy Screening. American family physician. 2020;101(5):307-8.	Wrong publication type
97. Selvachandran G, Quek SG, Paramesran R, Ding W, Son LH. Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods. Artificial intelligence review. 2023;56(2):915-64.	Wrong publication type
98. 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. Journal of diabetes science and technology. 2021;15(3):655-63.	Already included in previous UK NSC review
99. 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 journal of ophthalmology. 2020;68(2):398-405	Already included in previous UK NSC review
100. Shin JY, Son J, Kong ST, Park J, Park B, Park KH, et al. Clinical Utility of Deep Learning Assistance for Detecting Various Abnormal Findings in Color Retinal Fundus Images: A Reader Study. Translational vision science & technology. 2024;13(10):34.	Wrong population
101. Similie DE, Andersen JKH, Dinesen S, Savarimuthu TR, Grauslund J. Grading of diabetic retinopathy using a pre-segmenting deep learning classification model: Validation of an automated algorithm. Acta ophthalmologica. 2025;103(2):215-21.	In development
102. Skevas C, Weindler H, Levering M, Engelberts J, van Grinsven M, Katz T. Simultaneous screening and classification of diabetic retinopathy and age-related macular degeneration based on fundus photos-a prospective analysis of the RetCAD system. International journal of ophthalmology. 2022;15(12):1985-93.	Wrong population

Reference	Reason for exclusion
103. Srinivasan R, Surya J, Ruamviboonsuk P, Chotcomwongse P, Raman R. Influence of Different Types of Retinal Cameras on the Performance of Deep Learning Algorithms in Diabetic Retinopathy Screening. Life (Basel, Switzerland). 2022;12(10).	Wrong population
104. Srinivasan S, Ji H, Chen DZ, Wong W, Soh ZD, Goh JHL, et al. Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? BMJ open ophthalmology. 2025;10(1).	Wrong population
105. Stranak Z, Pencak M, Veith M. VYUZITI UMĚLE INTELIGENCE V ZACHYTU DIABETICKE RETINOPATIE. PŘEHLED, ARTEFICIAL INTELLIGENCE IN DIABETIC RETINOPATHY SCREENING. A REVIEW. Ceska a slovenska oftalmologie : casopis Ceske oftalmologicke spolecnosti a Slovenske oftalmologicke spolecnosti. 2021;77(5):224 EP - 31.	Wrong publication type
106. Surya J, Garima, Pandy N, Hyungtaek Rim T, Lee G, Priya MNS, et al. Efficacy of deep learning-based artificial intelligence models in screening and referring patients with diabetic retinopathy and glaucoma. Indian journal of ophthalmology. 2023;71(8):3039-45.	Wrong population
107. Tahir HN, Ullah N, Tahir M, Domnic IS, Prabhakar R, Meerasa SS, et al. Artificial intelligence versus manual screening for the detection of diabetic retinopathy: a comparative systematic review and meta-analysis. Frontiers in medicine. 2025;12:1519768.	Systematic review including studies published before 2020
108. Taylor JR, Drinkwater J, Sousa DC, Shah V, Turner AW. Real-world evaluation of RetCAD deep-learning system for the detection of referable diabetic retinopathy and age-related macular degeneration. Clinical & experimental optometry. 2025;108(5):601-6.	Wrong population
109. Tham Y-C, Anees A, Zhang L, Goh JHL, Rim TH, Nusinovici S, et al. Referral for disease-related visual impairment using retinal photograph-based deep learning: a proof-of-concept, model development study. The Lancet Digital health. 2021;3(1):e29-e40.	Wrong intervention
110. Tomic M, Vrabec R, Hendelja D, Kolaric V, Bulum T, Rahelic D. Diagnostic Accuracy of Hand-Held Fundus Camera and Artificial Intelligence in Diabetic Retinopathy Screening. Biomedicines. 2023;12(1).	Wrong population
111. Uy H, Fielding C, Hohlfeld A, Ochodo E, Opare A, Mukonda E, et al. Diagnostic test accuracy of artificial intelligence in screening for referable diabetic retinopathy in real-world settings: A systematic review and meta-analysis. PLOS global public health. 2023;3(9):e0002160.	Systematic review including prospective studies only
112. Vaghefi E, Yang S, Xie L, Hill S, Schmiedel O, Murphy R, et al. THEIA TM development, and testing of artificial intelligence-based primary triage of diabetic retinopathy screening images in New Zealand. Diabetic medicine : a journal of the British Diabetic Association. 2021;38(4):e14386.	In development, internal validation
113. Valkova J, Adam M, Hlavacek J. Artificial intelligence in diabetic retinopathy screening: from idea to a medical device in clinical practice. Umela inteligence ve screeningu diabeticke retinopatie: od napadu po zdravotnicky prostredek v klinicke praxi. 2024;162(7):290-3.	Foreign language

Reference	Reason for exclusion
114. Valkova J, Adam M, Hlavacek J. Umela inteligence ve screeningu diabeticke retinopatie: od napadu po zdravotnicky prostredek v klinicke praxi, Artificial intelligence in diabetic retinopathy screening: from idea to a medical device in clinical practice. Casopis lekaru ceskych. 2024;162(7):290 EP - 3.	Foreign language
115. Vought R, Vought V, Shah M, Szirth B, Bhagat N. EyeArt artificial intelligence analysis of diabetic retinopathy in retinal screening events. International ophthalmology. 2023;43(12):4851-9.	Wrong population
116. Vujosevic S, Aldington SJ, Silva P, Hernandez C, Scanlon P, Peto T, et al. Screening for diabetic retinopathy: new perspectives and challenges. The lancet Diabetes & endocrinology. 2020;8(4):337-47.	Wrong publication
117. Wang H, Meng X, Tang Q, Hao Y, Luo Y, Li J. Development and Application of a Standardized Testset for an Artificial Intelligence Medical Device Intended for the Computer-Aided Diagnosis of Diabetic Retinopathy. Journal of healthcare engineering. 2023;2023:7139560.	Wrong population, wrong intervention
118. Wang X-N, Dai L, Li S-T, Kong H-Y, Sheng B, Wu Q. Automatic Grading System for Diabetic Retinopathy Diagnosis Using Deep Learning Artificial Intelligence Software. Current eye research. 2020;45(12):1550-5.	Wrong intervention
119. Wang Y, Yu M, Hu B, Jin X, Li Y, Zhang X, et al. Deep learning-based detection and stage grading for optimising diagnosis of diabetic retinopathy. Diabetes/metabolism research and reviews. 2021;37(4):e3445.	Wrong intervention
120. Wang Y, Shi D, Tan Z, Niu Y, Jiang Y, Xiong R, et al. Screening Referable Diabetic Retinopathy Using a Semi-automated Deep Learning Algorithm Assisted Approach. Frontiers in medicine. 2021;8:740987.	Wrong population
121. Wang Z, Li Z, Li K, Mu S, Zhou X, Di Y. Performance of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis of prospective studies. Frontiers in endocrinology. 2023;14:1197783.	Systematic review including prospective studies only
122. Weinreb RN, Lee AY, Baxter SL, Lee RWJ, Leng T, McConnell MV, et al. Application of Artificial Intelligence to Deliver Healthcare From the Eye. JAMA ophthalmology. 2025;143(6):529-35.	Wrong publication type
123. Wei Q, Chi L, Li M, Qiu Q, Liu Q. Practical Applications of Artificial Intelligence Diagnostic Systems in Fundus Retinal Disease Screening. International journal of general medicine. 2025;18:1173-80.	Wrong population
124. Weng CY, Maguire MG, Flaxel CJ, Jain N, Kim SJ, Patel S, et al. Effectiveness of Conventional Digital Fundus Photography-Based Teleretinal Screening for Diabetic Retinopathy and Diabetic Macular Edema: A Report by the American Academy of Ophthalmology. Ophthalmology. 2024;131(8):927-42.	Wrong intervention
125. Wewetzer L, Held LA, Steinhauser J. Diagnostic performance of deep-learning-based screening methods for diabetic retinopathy in primary care-A meta-analysis. PloS one. 2021;16(8):e0255034.	Systematic review including studies published before 2020
126. Wolf RM, Channa R, Liu TYA, Zehra A, Bromberger L, Patel D, et al. Autonomous artificial intelligence increases screening and follow-up for diabetic retinopathy in youth: the ACCESS randomized control trial. Nature communications. 2024;15(1):421.	Wrong comparator

Reference	Reason for exclusion
127. Wolf RM, Liu TYA, Thomas C, Prichett L, Zimmer-Galler I, Smith K, et al. The SEE Study: Safety, Efficacy, and Equity of Implementing Autonomous Artificial Intelligence for Diagnosing Diabetic Retinopathy in Youth. Diabetes care. 2021;44(3):781-7.	Wrong population
128. Wu H, Jin K, Yip CC, Koh V, Ye J. A systematic review of economic evaluation of artificial intelligence-based screening for eye diseases: From possibility to reality. Survey of ophthalmology. 2024;69(4):499-507.	Systematic review on Q3 with wider inclusion criteria
129. Xie Y, Gunasekeran DV, Balaskas K, Keane PA, Sim DA, Bachmann LM, et al. Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. Translational vision science & technology. 2020;9(2):22.	Wrong publication type
130. Yang Q, Bee YM, Lim CC, Sabanayagam C, Yim-Lui Cheung C, Wong TY, et al. Use of artificial intelligence with retinal imaging in screening for diabetes-associated complications: systematic review. EClinicalMedicine. 2025;81:103089.	Wrong outcome
131. Yogesan S, Gupta A, Tan J, Xiao D, Martyn S, Kanagasingam Y. Harnessing AI to Overcome Health Disparities: AI-Driven Eye Screening for Underserved Indigenous Communities. SSRN. 2024.	Full text not available
132. Yuan A, Lee AY. Artificial intelligence deployment in diabetic retinopathy: the last step of the translation continuum. The Lancet Digital Health. 2022;4(4):e208 EP - e9.	Wrong publication type
133. Zhang G, Lin J-W, Wang J, Ji J, Cen L-P, Chen W, et al. Automated multidimensional deep learning platform for referable diabetic retinopathy detection: a multicentre, retrospective study. BMJ open. 2022;12(7):e060155.	Wrong intervention
134. Zhelev Z, Peters J, Rogers M, Allen M, Kijauskaite G, Seedat F, et al. Test accuracy of artificial intelligence-based grading of fundus images in diabetic retinopathy screening: A systematic review. Journal of medical screening. 2023;30(3):97-112.	Systematic review including studies published before 2020
135. Zhu A, Tailor P, Verma R, Zhang I, Schott B, Ye C, et al. Implementation of deep learning artificial intelligence in vision-threatening disease screenings for an underserved community during COVID-19. Journal of telemedicine and telecare. 2024;30(10):1590-7.	Wrong population
136. Zulu N, Piotie PN, Webb EM, Maphenduka WG, Cook S, Rheeder P. Screening for diabetic retinopathy at a health centre in South Africa: A cross-sectional study. Journal of public health in Africa. 2025;16(1):681.	Wrong outcome
137. Analysis of Diabetic Retinopathy, Glaucoma and Macular Degeneration Diagnosis Via Digital Fundus Images With Artificial Intelligence A1 - Anonymous. clinicaltrialsgov. 2024.	Full text not available
138. Assessing the Impact of Using Autonomous Artificial Intelligence (AI) for Pre-screening of Diabetic Retinopathy (DR) and Diabetic Macular Edema on Physician Productivity in Bangladesh A1 - Anonymous. clinicaltrialsgov. 2021.	Full text not available

Table 20. Publications excluded after review of full text articles from August update search (n = 7)

Reference	Reason for exclusion
1. Chen KM, Zhao CS, Knapp A, Dow E, Phadke A, Tan M, et al. Targeted interventions lead to quality improvement in year 2 of an artificial intelligence-based diabetic retinopathy detection program in Northern California. Retina. 2025;45(8):1469 EP - 80.	Duplicate
2. Ciapponi A, Ballivian J, Gentile C, Mejia JR, Ruiz-Baena J, Bardach A. Diagnostic utility of artificial intelligence software through non-mydriatic digital retinography in the screening of diabetic retinopathy: an overview of reviews. Eye (London, England). 2025;39(10):2083-9.	Duplicate
3. Huber SL, Parzer V, Ludvik B, Pollreisz A, Mahnert N, Brix JM. Evaluation of IDx-DR software for diabetic retinopathy screening in outpatient clinics: Efficacy, safety, and feasibility in a real-world setting. Journal of Diabetes and its Complications. 2025;39(10):109120.	Duplicate
4. Parrey MUR, Bhatti MOA, Abdul-Latif MM, Rehman S, Ismail MM, Hamid OA. Meta-analysis of AI algorithm performance in detecting retinal diseases. Anaesthesia, Pain and Intensive Care. 2025;29(4):400 EP - 6.	Systematic review including studies published before 2020
5. Rahmati M, Smith L, Piyasena MP, Bowen M, Boyer L, Fond G, et al. Artificial Intelligence improves follow-up appointment uptake for diabetic retinal assessment : a systematic review and meta-analysis. Eye (London, England). 2025;39(12):2398-406.	Duplicate
6. Sacchini F, Mancin S, Cangelosi G, Palomares SM, Caggianelli G, Gravante F, et al. The role of artificial intelligence in diabetic retinopathy screening in type 1 diabetes: A systematic review. Journal of diabetes and its complications. 2025;39(10):109139.	Systematic review with narrower inclusion criteria
7. Saleh I, El-Den NN, Elsharkawy M, Mahmoud A, Sewelam A, Wang W, et al. AI-based methods for diagnosing and grading diabetic retinopathy : A comprehensive review. Artificial intelligence in medicine. 2025;168:103221.	Wrong intervention

Table 21. Publications excluded after review of full text articles from September update search (n = 8)

Reference	Reason for exclusion
1. Cao F, Guo X, Li M, Li S, Peng X. Development and validation of a deep learning model for early detection and screening of diabetic retinopathy. BMC medical informatics and decision making. 2025;25(1):315.	Internal validation
2. Dejene FM, Debelee TG, Schwenker F, Ayano YM, Feyisa DW. Diabetic retinopathy screening using machine learning: a systematic review. BMC biomedical engineering. 2025;7(1):12.	Wrong intervention
3. Joseph S, Wang Y, Drinkwater JJ, Jan CL, Sundar B, Zhu Z, et al. Effectiveness of artificial intelligence-based diabetic retinopathy screening in primary care and endocrinology settings in Australia: a pragmatic trial. The British journal of ophthalmology. 2025.	Wrong population
4. Liang X, Bao Y, Du Y, Kong N. AI-Assisted Screening for Diabetic Retinopathy and Fundus Abnormalities in a Large-Scale Physical Examination Population. Clinical ophthalmology (Auckland, NZ). 2025;19:2889-900.	In development
5. Macdonald TB, Dinnes J, Maniatopoulos G, Solebo AL, Hogg HDJ, Alderman JE, et al. Development of a Target Product Profile for an Artificial Intelligence for Use in English Diabetic Eye Screening, a Modified Delphi Consensus Study. SSRN. 2025.	Wrong outcome
6. Sacchini F, Mancin S, Cangelosi G, Palomares SM, Caggianelli G, Gravante F, et al. The role of artificial intelligence in diabetic retinopathy screening in type 1 diabetes: A systematic review. Journal of diabetes and its complications. 2025;39(10):109139.	Systematic review with narrower inclusion criteria
7. Segovia C, Salinas-Toro D, Moraga C, Sepulveda M. Amending AI Software Accuracy for Diabetic Retinopathy Detection Using Conditional Probability and the Appropriate Reference Standard. Revista medica de Chile. 2025;153(10):686-94.	Wrong outcome
8. Suganya Devi K, Vasireddi HK, Reddy GR, Satti SK. Unfolding the diagnostic pipeline of diabetic retinopathy with artificial intelligence: A systematic review. Survey of ophthalmology. 2025.	Wrong publication type

Table 22. Publications excluded after review of full text articles from October update search (n = 5)

Reference	Reason for exclusion
1. Huhtinen P, Kubin AM, Piila L, Hautala N. Performance and accuracy of 10 AI-based algorithms in diabetic retinopathy screening in a real-world clinical setting. Acta Ophthalmol. 2025 Nov;103(7):e532-e533.	Letter
2. Kabunga R, Asasira J, Njuki S, Daniel A, Morley K, Morley M, Kaggwa F, Cikomola JC, Simon A. Describing the Performance and the Infrastructure Requirements of the Existing Artificial Intelligence (AI)-Based Diabetic Retinopathy (DR) Screening Algorithms for Diabetic Patients: an Umbrella Review. J Med Syst. 2025 Oct 30;49(1):149.	Review of systematic reviews
3. Teng CW, Patel SD, Barkmeier AJ, Liu TYA, Myung D, Henderer J, Liu J, Hansen E, Al-Aswad LA. Autonomous Artificial Intelligence in Diabetic Retinopathy Testing-Lessons Learned on Successful Health System Adoption. Ophthalmol Sci. 2025 Sep 3;6(1):100935.	Systematic review with narrower inclusion criteria
4. Wang TW, Luo WT, Tu YK, Chou YB, Wu YT. Diagnostic Accuracy of EyeArt for Fundus-Based Detection of Diabetic Retinopathy: A Systematic Review and Meta-analysis. Am J Ophthalmol. 2025 Oct 4;281:465-479.	Systematic review with narrower inclusion criteria
5. Yudistira Y, Hendrawan KA, Andayani A, Suryathi NMA, Ernawati T, Gunawan ACV, Daradila NPKM. Diabetic Retinopathy Screening Approaches in Developing Countries: A Systematic Review and Meta-Analysis. Turk J Ophthalmol. 2025 Oct 27;55(5):260-275.	Systematic review with different inclusion criteria

Table 23. Publications excluded after review of full text articles from November update search (n = 8)

Reference	Reason for exclusion
1. Chauhan A, Vale L, Kankaria A, et al. Reach and implementation of human and AI-assisted diabetic retinopathy screening models in primary healthcare settings in India. Scientific reports 2025;15(1):41355. doi: https://dx.doi.org/10.1038/s41598-025-25402-9	Wrong intervention
2. Huhtinen P, Kubin A-M, Dvorak K, et al. Real-World Evaluation of Artificial Intelligence-Based Diabetic Retinopathy Screening Using the Optomed Aurora Handheld Fundus Camera. *Diabetes technology & therapeutics* 2025;27(12):1023-25. doi: https://dx.doi.org/10.1177/15209156251369886	Wrong population
3. kumar A, Chawla M. Enhancing diabetic retinopathy detection using an optimized MobileNet architecture. *Biomedical Signal Processing and Control* 2026;113:109207. doi: https://dx.doi.org/10.1016/j.bspc.2025.109207	In development
4. Leigh J, Drinkwater J, Turner A, et al. Health Economic Considerations for the Implementation of Artificial Intelligence-Enabled Diabetic Retinopathy Screening: A Review. *Clinical & experimental ophthalmology* 2025 doi: https://dx.doi.org/10.1111/ceo.70016	Wrong population
5. Osman AAF. Explainable AI for Diabetic Retinopathy Detection: A Systematic Review of Machine Learning Approaches. *Vascular and Endovascular Review* 2025;8(10 s):165-78.	Wrong intervention
6. Qureshi ARK, Birla S, Arondekar C, et al. AI and ML Powered Early Detection of Diabetic Retinopathy: A Deep Learning Approach for Improved Clinical Decision-Making. *Vascular and Endovascular Review* 2025;8(6):183-99.	In development
7. Raha A. AUTOMATED DIABETIC RETINOPATHY DETECTION in SMARTPHONE BASED FUNDUS PHOTOGRAPHY USING ARTIFICIAL INTELLIGENCE CONDUCTED at HEALTH CAMP in ASSAM. *International Journal of Diabetes in Developing Countries* 2025;45(Supplement 1):S53. doi: https://dx.doi.org/10.1007/s13410-025-01585-9	Wrong publication type
8. Ran AR, Ding JL, Tang Z, et al. Real-World Prospective Validation and Economic Evaluation of Deep Learning-Based Diabetic Retinopathy Detection From Fundus Photographs: A Systematic Review and Meta-analysis. *Diabetes care* 2025 doi: https://dx.doi.org/10.2337/dc25-1493	Systematic review with different inclusion criteria

Appendix 3 — Summary and appraisal of individual studies

Data Extraction

Studies relevant to criteria 4 and 5 are summarised in **Table 24**, which includes full citations, key outcome measures, effect estimates, confidence intervals, and prioritisation rationale. Studies meeting criterion 11 are presented separately in **Table 25**.

Table 24. Studies meeting inclusion criterion 4 and 5 (Question 1), with prioritisation status and rationale

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
Abreu-Gonzalez 2025[74], Spain	Consecutive adults (≥18 years) with type 1 or 2 DM (N=945; mean age 64.6 years; 55% male). One 45° macula-centred fundus image per patient from Topcon or ZEISS cameras in primary care linked to 5 Spanish hospitals. No mydriasis.	Prospective single-gate	LuxIA AI	More-than-mild DR (mtmDR) per ICDR (primary analysis); Any DR could be calculated.	No human comparator	Mean grading of 3 independent retina specialists (ICDR classification).	Topcon (n=829): Accuracy 95.2%, SE 97.1% (136/140), SP 94.8% (653/689), AUC 0.9596, ZEISS (n=32): Accuracy 90.6%, SE 83.3% (5/6), SP 92.3% (24/26).	Deprioritised: No human comparator
Al-Turk 2022[75], Kenya; Saudi Arabia; China (multi-site evaluation)	Adults with DM from Kenya (N=14,340), Saudi Arabia (N=5,013), and China (N=7,500); unclear how selected. 2 fundus images per eye using 50° field-of-view cameras with varying imaging settings across 3 nations; mydriasis unclear.	Retrospective single-gate	DAPHNE	Any DR; Referable DR; PDR; DMO; progression changes reported	No human comparator	Human expert grading according to ICDRS and/or UK NSC guidelines.	Kenia – Any DR (ICDRS): SE 91.2% (95% CI 90.2-92.2), SP 94.6% (95% CI 94.2-94.8). Saudi Arabia / China – Any DR (UK NSC grading scheme): SE 92.6% (95% CI 91.9–93.3), SP 93.0% (95% CI 92.7–93.4)	Deprioritised: Country, unclear if ARIAS commercially available, no human comparator
Antaki 2024[76], Canada	115 consecutive adults with DM recruited prospectively at a tertiary centre (CHUM, Montreal). Single 45° fovea-centred image per eye on	Prospective single-gate	CARA AI system	Any DR (eye-level) and Referable disease	No human comparator	Retinal specialist grading using the ICDR scale.	Patient level (referable disease): SE 87.5% (95% CI 71.9-95.0), SP 66.2% (95% CI 54.3-76.3). Eye level - Retinopathy: SE 88.2% (95% CI 76.6-94.5),	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	Centervue DRS camera; no mydriasis.			(patient-level); DME separately.			SP 71.4% (95% CI 63.7-78.1). Eye level - DME: SE 100% (95% CI 64.6-100), SP 81.9% (95% CI 75.6-86.8)	
Arenas-Cavalli 2022[77], Chile	National DR tele-ophthalmology programme; 1,123 exams from 5 primary-care centres (Peñalolén, Recoleta, Ñuñoa, Concepción, Providencia); first 2 of every 5 exams performed in each commune selected. Two 45° images/eye (temporal & nasal) per EURODIAB; cameras: Nidek AFC-330, Canon CR-2, Topcon TRC-50ex. Mainly nonmydriatic, dilation if required. Data: Nov 11, 2014–Dec 18, 2016.	Retrospective single-gate	DART	Any DR	No human comparator	Grading of retinal images by one ophthalmologist (from a panel of 8) using the ICDR severity scale.	At optimal operating point (ROC curve): SE 94.6% (95% CI 90.9-96.9), SP 74.3% (95% CI 73.3-75.0), PPV 49.9% (95% CI 48.0-51.1), NPV 98.1% (95% CI 96.8-98.9), AUC 0.915	Deprioritised: No human comparator
Baget-Bernaldiz 2024 [78], Spain and France	Only Messidor-2 (n=1,200 images) data external and eligible for our review; mydriasis unclear.	Retrospective (public dataset)	AIRS	Referable DR, Any DR can be calculated	No human comparator .	Classification provided by MESSIDOR-2	Messidor-2 – Referable DR: SE 96.6%† (373/386), SP 99.4%† (809/814).	Deprioritised: No human comparator
Baget-Bernaldiz 2021 [79], Spain and France	Only Messidor (n=1,200 images) data external and eligible for our review; mydriasis unclear.	Retrospective (public dataset)	Not reported	Any DR + referable DR	No human comparator	Classification provided by MESSIDOR	MESSIDOR - Any DR: Accuracy 94.79%, SE 97.32% (545/560), SP 94.57% (610/645). MESSIDOR - Referable DR: Accuracy 98.78%, SE 94.64% (371/392), SP 99.14% (801/808).	Deprioritised: No human comparator
Brant 2025 [80],	4,537 unique fundus photographs from 4,537	Retrospective single-gate	ARDA (Automa-	Referable DR + sight-	No human comparator	Adjudicated consensus	For severe NPDR or PDR: SE 97.0% (95% CI, 92.6-99.2%),	Deprioritised: Country, no

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
India	unique patients screened between January 1, 2019, and July 31, 2023, across 45 sites; approximate 1% random sample from 475,098 images; nonmydriatic images with dilation of required.		ted Retinal Disease Assess-ment).	threatening DR.		grades from 3 U.S. ophthalmologists	SP 96.4% (95% CI, 95.7-97.0%), PPV 50.7%, NPV 99.9%. For sight-threatening DR (STDR): SE 95.9% (95% CI, 93.0-97.4%), SP 94.9% (95% CI, 94.1-95.7%), PPV 67.9%, NPV 99.5%.	human comparator
Burlina 2024 [81], Italy	958 consecutive adults at ULSS8 Berica (Diabetes & Endocrinology Unit + Eye Unit, Veneto); screening June 2022–June 2023; two 45° fields/eye (disc- and macula-centred) on Nikon Retina Station; no mydriasis.	Retrospective single-gate	DAIRET, the Italian version of Retmarker	Any DR + Referable	No human comparator	Grading by a single certified ophthalmologist, who evaluated all fundus images in a blinded manner using ICDRS	Any DR: SE 84.6%† (292/345), SP 59% (308/522). Referable DR (moderate or worse): SE 100% (14/14).	Deprioritised: No human comparator
Chia 2023 [82], Australia	Retrospective external validation at an Aboriginal Community Controlled Health Service in Perth; 1,682 consecutive single-field 45° macula-centred images from 864 DM patients (2013–2020); Topcon Maestro camera; no mydriasis.	Retrospective single-gate	Developed by Google Health	More-than-mild DR (mtmDR): ≥ moderate NPDR or referable DMO. Vision-threatening DR (vtDR): ≥ severe NPDR or referable DMO. All-cause referable DR: mtmDR or	Retina specialist	A 3-person adjudicated grade by a panel of U.S. board-certified retina specialists.	mtmDR - ARIAS: SE 98.0% (95%CI 96.5-99.4), SP 95.1% (95%CI 93.6-96.4). mtmDR – Human grading: SE 87.1% (95%CI 83.6-90.6), SP 97.0% (95%CI 95.9-98.0). vtDR – ARIAS: SE 96.2% (95%CI 93.4-98.6), SP 95.8% 8 (95%CI 94.6-96.9) vtDR – Human grading: SE 84.4% (95%CI 79.7-89.2) SP 97.8% (95%CI 96.9-98.6 All-cause referable DR - ARIAS: SE 93.7% (95%CI 91.8-95.5) SP 91.7% (95%CI 90.0-93.3) All-cause referable DR – Human grading: SE 74.4% (95%CI 95.2 to 97.4)	Deprioritised: Country (indigenous Australians)

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
				ungradable for mtmDR			SP 96.3% (05%CI 95.2 to 97.4	
Curran 2023 [84], Bangladesh	1,332 consecutive children & young adults (3–26 years) with DM screened at Dhaka BIRDEM-2 hospital (Orbis/BIRDEM program); 1 image centred on the macula and 1 on the optic nerve captured by trained technicians; no mydriasis.	Retrospective single-gate	Orbis Internatio-nal's Cybersight AI	Any DR + Referable DR (moderate NPDR or worse).	No human comparator	A fully qualified optometrist (KC), certified in diabetic retinopathy (DR) grading through the Gloucester-shire Retinal Education Group's Diabetic Retinopathy Grading Course	Any DR: SE 75.5% (95% CI 69.7- 81.3), SP 91.8% (95% CI 90.2-93.5), PPV 64.8% (95% CI 58.8- 70.7), NPV 94.9% (95% CI 93.6-96.3), AUC-ROC 91.2%. Referable DR: SE 84.2% (95% CI 67.8-100), SP 98.9% (95% CI 98.3-99.5), PPV 53.3% (95% CI 35.5-71.2), NPV 99.8% (95% CI 99.5-100), AUC-ROC 97.6%.	Deprioritised: Country, no human comparator
Dai 2021 [85], China and USA	2 external datasets eligible: 1) China National Diabetic Complications Study (CNDCS): 92,672 images from 23,186 patients taken in 2018 (real world); mydriasis unclear. 2) EyePACS: public dataset from the United States; 88,702 fundus images; mydriasis unclear.	Retrospective single-gate (multi-dataset)	Cybersight AI (Efficient-Net-B3 CNN)	Referable DR, Non-DR, Mild NPDR, Moderate NPDR, Severe NPDR, PDR	No human comparator	Consensus grading by senior retinal specialists (expert human graders) for DR and DME classification.	CNDCS – Referable DR: SE 94.0% (95% CI 93.6-94.4), SP 88.3% (95% CI 88.1-88.5), AUC 0.970 (95% CI 0.969–0.971). EyePACS – Referable DR: SE 92.8% (95% CI 92.4-93.1), SP 81.3% (95% CI 81.0-81.6), AUC 0.946 (95% CI 0.945–0.947).	Deprioritised: Country (CNDCS - China), no human comparator
Dogan 2024 [86], Turkey	Endocrinology clinic, Akdeniz University Hospital (Antalya). 900 consecutive adults with DM, no prior DR. Two non-mydriatic fields/eye (macula- and disc-centred) on Canon CR2 AF, Topcon TRC-NW400, Optomed Aurora; if	Prospective single-gate	EyeCheck up AI (deve-loped at Akdeniz University by Ural Tele-commu-	Referable DR (mtmDR), vtDR, CSDMO	No human comparator, comparison of 3 different cameras (Optomed aurora, Canon CR2	A consensus diagnosis by 3 retina specialists who graded dilated, 4 wide-field fundus images according to the American	mtmDR – Optomed aurora: SE 90.48% (95% CI 93.95-94.98), SP 97.21% (95% CI 95.74-98.29). mtmDR - Canon CR2 AF: SE 95.65% (95% CI 90.15-98.57), SP 95.92%	Deprioritised: Country, no human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	≤5 non-mydriatic attempts failed, dilate and recapture.		nication Inc.)		AF, Topcon NW400)	Academy of Ophthalmology's Diabetic Retinopathy Preferred Practice Patterns	(95% CI 93.00-97.37). mtmDR - Topcon NW400: SE 95.19% (95% CI 89.14-98.42), SP 96.46% (95% CI 94.40-97.93).	
Dong 2022 [87], China	443 consecutive adults with DM (type 1 & 2), 3 Chinese community healthcare centres; single-field, macula-centred CFP (50°) on non-mydriatic camera (RetiCam 3100); no mydriasis.	Prospective single-gate	CARE (Compre-hensive Artificial Intelli-gence Retinal Expert) system, developed by Shanghai Eagle-Vision Medical Technolo-gy Co., Ltd. (Airdoc).	Any DR, mtmDR, vtDR, DME	No human comparator	2 independent ophthalmologists (or a third senior specialist if they disagreed), using ICDR classification criteria.	Any DR: SE 75.19% (95% CI 69.4780.17), SP 93.99% (95% CI 91.65–95.71), PPV 85.11% (95% CI 79.76–89.28), NPV 89.23% (95% CI 86.44–91.52). More-than-mild DR: SE 78.97% (95% CI 73.06–83.90), SP 92.52% (95% CI 90.07–94.41), PPV 80.00% (95% CI 74.12–84.85), NPV 92.07% (95% CI 89.58–94.02).	Deprioritised: Country, no human comparator
Dow 2023a[57], USA	DESP at 7 primary care sites in the San Francisco Bay Area (STATUS programme) 1,222 consecutive DM patients (12-months AI-phase); In analysis: Subset with eye examination: ARIAS: 80 patients, Human: 122 patients.	Prospective single-gate	IDx-DR (Digital Diagno-stics Inc.)	Referable (mtmDR)	Human consensus overread: 2 fellowship-trained retina specialists at the Stanford Ophthalmic Reading	In-person examination at Byers Eye Institute: dilated fundus examination, SD-OCT imaging, other retinal imaging. If diagnosis	Referable DR – ARIAS: SE 95.5 (95%CI 86.7 – 100), SP 60.3 (95%CI 47.7 – 72.9) Referable DR – Human grader: SE 69.6 (95%CI 50.7 – 88.3), SP 97.0 (95%CI 93.5 – 100)	Prioritised

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
					Center, 3rd or 4th ophthalmo-logist for adjudication	different to index test, a 2nd reader adjudicated the evaluation		
Duggal 2025 [88], India	1) Validation: 250 DM patients from primary health centre and endocrinology clinic in Punjab (500 eyes); non-mydriatic images. 2) Implementation: 343 DM patients from a community health centre in Punjab (686 eyes).	Prospective single-gate (2 phases: validation and implemen-tation)	1) 3 commer-cial AI algorithms (AI-1 to AI-3) evaluated; 2) AI-3 (Leben Care Health Services) selected for deploy-ment after showing best perfor-mance	Any DR + referable DR + DME	No human comparator	1/2) 2 masked human graders; adjudication by 3rd grader (senior vitreoretinal expert) in case of disagreement	1) Validation (AI-3) – Any DR: SE 68.42% (95% CI 59.71-76.05), SP 96.01% (95% CI 93.24-97.72), PPV 86.67% (95% CI 78.31-92.26). 2) Implementation (AI-3): Any DR: SE 99.6%, SP 64.7%, PPV 87.4%, NPV 98.3%. Referable DR: SE 78.9%, SP 98.1%, PPV 89.6%, NPV 95.7%. DME: SE 26.5%, SP 99.7%, PPV 81.8%, NPV 96.0%.	Deprioritised: Country, no human comparator
Fleming 2024 [15], Scotland	External QA of Scottish DESP, 744 images enriched for difficult and referable cases.	Retrospective 2-gate	DLAG (Deep Learning Auto-grader)	Any DR: Any disease or ungradable	Level 1 and 2 graders of Scottish DESP	Panel of 7 to 9 ophthalmologists, majority consensus	Any DR - DLAG: SE 96.23% (35,589/36,985), SP set to match human (75.28%).	Deprioritised: ARIAS not CE-marked/FDA-approved and

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							SE set to match human (95.83%). SP 78.12% (7,389/9,458).	not commercially available
Fleming 2024 [15], Scotland	External QA of Scottish DESP, 744 images enriched for difficult and referable cases.	Retrospective 2-gate	iGradingM	Any DR: Any disease or ungradable	Level 1 and 2 graders of Scottish DESP	Panel of 7 to 9 ophthalmologists, majority consensus	Any DR – iGradingM: SE 92.97 (542/583), SP 61.88 (99/160). Any DR – Human grading: SE 95.83 (35,443/36,985), SP 75.28 (7,120/9,458)	Prioritised
Grzybowski 2021 [89], Poland	Community-based secondary care diabetic clinic, Poznań (Poland). Two 45° images per eye (macula & disc) on Topcon TRC-NW400; no mydriasis; all referable DR positive screening subjects (n = 60) and a random selection of DR negative patient images (n = 110); dataset curated to 4 good-quality images per patient.	Retrospective 2-gate	IDx-DR (Digital Diagno-stics), Retina-Lyze (Retina-Lyze System A/S) (2 different thresholds used for Retina-Lyze)	Referable DR	No human comparator	Single ophthalmologist with basic DR screening experience, using the ICDRS to classify patients as referable DR positive or negative.	Referable DR - IDx-DR: SE 93.33%, SP 95.45%, PPV 91.80%, NPV 96.33%. Referable DR - RetinaLyze Strategy 1: SE 89.66%, SP 71.82%, PPV 62.65%, NPV 92.94%. Referable DR - RetinaLyze Strategy 2: SE 74.14%, SP 93.64%, PPV 86.00%, NPV 87.29%.	Deprioritised: No human comparator
Grzybowski 2024 [90], Poland	750 consecutive DM patients; non-mydriatic community screening; Topcon NW-400; four 45° images per patient (disc + macula).	Retrospective single-gate	MONA DR v1.0.0 (MONA Health), IDx-DR (Digital Diagno-stics)	Both referable and any DR	No human comparator	Majority consensus grading by 3 masked expert graders using the ICDR classification system (320	Any DR – IDx-DR: SE 98.8% (95% CI 95.9-99.9), SP 44.8% (95% CI 40.5-49.2), PPV 37.4% (95% CI 32.8-41.9), NPV 99.3% (95% CI 97.0-100.0). Any DR – MONA DR:	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
						patients by graders in Poland and 430 by graders in India)	SE 80.4% (95% CI 73.6-86.0), SP 95.6% (95% CI 93.5-97.2), PPV 86.9% (95% CI 79.5-90.8), NPV 93.5% (95% CI 91.2-95.5). Referable DR - IDx-DR: SE 99.1% (95% CI 94.9-100.0), SP 71.5% (95% CI 67.6-75.1), PPV 38.1% (95% CI 32.7-44.5), NPV 99.8% (95% CI 98.7-100.0). Referable DR - MONA DR: SE 93.4% (95% CI 86.9-97.3), SP 89.3% (95% CI 86.5-91.7), PPV 60.3% (95% CI 53.2-68.7), NPV 98.6% (95% CI 97.3-99.5).	
Grzybowski 2025 [91], Poland	758 consecutive DM patients from diabetic clinics in Poznań (Mar 2020–Apr 2021). Non-mydriatic Topcon NW-400; ≥4 images/patient (disc + macula); no mydriasis.	Retrospective single-gate	IDx-DR (Digital Diagno-stics), RetCAD v2.1.1 (Thirona Retina BV)	Both referable and any DR	No human comparator	Majority consensus from 3 masked graders using ICDR classification (320 patients by graders in Poland and 438 by graders in India).	Any DR – IDx-DR: SE 99.11% (95% CI 96.81-99.89), SP 44.19% (95% CI 39.93-48.52), PPV 42.69% (95% CI 40.83-44.57), NPV 99.16% (95% CI 96.73-99.79). Any DR – RetCAD: SE 95.54% (95% CI 91.94-97.84), SP 88.95% (95% CI 85.98-91.48), PPV 78.39% (95% CI 74.00-82.81), NPV 97.94% (95% CI 96.28-98.87).	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							Referable DR - IDx-DR: SE 99.30% (95% CI 96.14-99.98), SP 68.99% (95% CI 65.18-72.63), PPV 42.47% (95% CI 39.60-45.39), NPV 99.77% (95% CI 98.37-99.97). Referable DR - RetCAD: SE 89.44% (95% CI 83.18-93.97), SP 94.81% (95% CI 92.75-96.42), PPV 79.87% (95% CI (73.82-84.82), NPV 97.50% (95% CI 96.02-98.43).	
Grzybowski 2023 [92], Poland	811 consecutive adults with DM at diabetic clinics (Mar 2020–Apr 2021). Non-mydriatic Topcon NW400; four 45° images/patient (disc + macula); no mydriasis.	Retrospective single-gate	IDx-DR (Digital Diagno-stics), Medios AI (Medios Technol-ogies, Remidio Innovative Solutions)	Both referable and any DR	No human comparator	Majority consensus from 3 masked graders using ICDR classification (362 patients by graders in Poland and 491 by graders in India).	Any DR – IDx-DR: SE 99% (95% CI 97-100) SP 44% (95% CI 40-48) PPV 46% (95% CI 42-50) NPV 99% (95% CI 97-100) Any DR – Medio AI: SE 89% (95% CI 85-93) SP 90% (95% CI 87-92) PPV 81% (95% CI 76-85) NPV 94% (95% CI 92-96) Referable DR - IDx-DR: SE 99% (95% CI 91-98), SP 68% (95% CI 64-72), PPV 46% (95% CI 41-51), NPV 100% (95% CI 98-100).	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							Referable DR - Medios AI: SE 95% (95% CI 91-98), SP 80% (95% CI 77-83), PPV 57% (95% CI 51-63), NPV 98% (95% CI 97-99).	
Guedes 2025 [93], Canary Islands, Spain	499 adults with DM (≥18 years) randomly recruited from the Retisalud DR screening program (Canary Islands); non-mydriatic images.	Prospective single-gate	EyeArt® 2.1.0 (Eyenuk Inc.)	Any DR	No human comparator	Ophthalmologist determined degree of DR using ICDR scale based on ophthalmological examination and fundus evaluation.	Any DR: SE: 100% (95% CI 98.1–100) SP: 93.5% (95% CI 90.2–96.0)	Deprioritised: No human comparator
Hao 2022 [94], China	Prospective county-level screening of 3,933 registered adults with DM (Jul–Nov 2021). Two 45° photos/eye (macula + disc) on Zeiss VISUCAM 500; undilated (natural pupils) in a darkroom.	Prospective single-gate	Eye-Wisdom (Zhiyuan Huitu fundus image AI analysis software, Visionary Intelli-gence Ltd.)	Any DR	No human comparator	2 ophthalmologists (>10 years experience) graded independently; disagreements resolved by a senior chief physician using ICDR scale.	Any DR - EyeWisdom: SE 81.2% (95% CI 80.3-82.1), SP 94.3% (95% CI 93.7-94.8), PPV 80.4% (95% CI 78.4-82.4), NPV 94.5% (95% CI 93.9-95.1).	Deprioritised: Country, no human comparator
Hao 2022 [95], China	6,146 adults with T2DM screened at the Metabolic Disease Management Center, Tianjin 4th Central Hospital; 22 Oct 2018–16 Jun 2021; single centre; no mydriasis. 95 patients (118 eyes) referred to fundus fluorescein	Retrospective single-gate	Voxel-Cloud (China)	Both “Any DR” distribution and “Referable DR” (moderate NPDR+).	Ophthalmo-logist’s preliminary report	Fundus fluorescein angiography (FFA) for referred patients.	Severe NPDR as cut-off point: VoxelCloud (eye-level): SE 85.4% (41/48), SP 87.1% (61/70), AUC 0.782 (95%CI 0.644–0.877). Ophthalmologist (eye-level): SE 79.2% (38/48),	Deprioritised: Country

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	angiography (FFA) included for comparative analysis.			DME also graded			SP 87.1% (61/70), AUC 0.754 (95%CI 0.614–0.856).	
Hu 2024 [96], China	120 inpatients with DM (240 eyes) (Dec 2019–Apr 2021); 234 eyes analysed (6 excluded for small pupils); no mydriasis.	Prospective single-gate	"AutoEye" AI reading platform (part of "Shang-gong Eye" series)	no DR; mild/moderate/severe NPDR; PDR. "Any DR" = mild NPDR+; stage-wise distribu-tions reported.	Manual reading by ophthalmologist	FFA (Fundus Fluorescein Angiography)	Any DR – ARIAS (eye-level): SE 97.50% (156/160), SP 97.30% (72/74), PPV 98.73% (156/158), NPV 94.74% (72/76). Any DR – Ophthalmologist (eye-level): SE 96.25% (154/160), SP 95.95% (71/74), PPV 98.09% (154/157), NPV 92.21% (71/77).	Deprioritised: Country
Ipp 2021[97], USA (related to paper by Lim 2023[102])	893 consecutive DM patients from 15 U.S. centres; (1,786 eyes); dilate-if-needed protocol	Prospective single-gate	EyeArt, version 2.1.0 (Eyenuk Inc)	Referable DR (mtmDR and vtDR)	No human comparator	ETDRS grading of 4-wide-field stereoscopic dilated fundus photographs by the Fundus Photograph Reading Center (FPRC): 2 independent masked human graders with adjudication by a third more senior grader.	mtmDR – EyeArt (undilated; 1,526 eyes): SE 95.5% (95% CI 92.4-98.5), SP 85.0% (95% CI 82.6-87.4), PPV 59.5% (95% CI 53.9-63.9), NPV 98.8% (95% CI 98.2-99.4). mtmDR – EyeArt (dilate-if-needed; 1,701 eyes): SE 95.5% (95% CI 92.6-98.4), SP 85.3% (95% CI 83.0-87.5), PPV 59.1% (95% CI 53.8-64.4), NPV 98.8% (95% CI 98.2-99.5).	Deprioritised: No human comparator
Karabeg 2024 [98], Norway	Minority Women's Day Nov 1, 2017; 33 women with DM screened; 66 eyes imaged; 64 eyes eligible; no mydriasis.	Prospective single-gate	EyeArt, version 2.1.0 (Eyenuk Inc)	Any DR + DME	No human comparator	Manual grading by 2 experienced retina specialists (ophthalmolo-gists) using the	Any DR - EyeArt: SE 100% (95% CI 73.5-100), SP 100% (95% CI 93.1-100). DME – Eye Art: SE 100% (95% CI 15.8-100),	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
						ICDR severity scale	SP 100% (95% CI 94.2-100).	
Karabeg 2025 [99], Norway	128 consecutive patients with DM from Oslo (2 centres), 247 eyes; no mydriasis; Oct 2022–Nov 2023.	Prospective single-gate	EyeArt version v2.1.0 (Eyenuk Inc., Los Angeles, CA, USA)	Any DR + Referable DR	No human comparator	Manual consensus grading performed by a multi-disciplinary team consisting of an ophthalmologist, an optometrist, and ophthalmic nurses (certified graders)	Any DR – EyeArt: SE 94.0% (95% CI 91.0-96.9), SP 72.6% (95% CI 67.0-78.1), AUC 83.5% (95% CI 78.3-88.7). Referable DR – EyeArt: SE 89.7% (95% CI 85.9 - 93.4), SP 83.0% (95% CI 78.5 - 87.7), AUC 86.3% (95% CI 79.3-93.4).	Deprioritised: No human comparator
Lee 2021[58], USA	2 U.S. Veterans Affairs hospitals (Seattle and Atlanta), 311,604 images from 23,724 consecutive DM patients; both mydriatic and non-mydriatic images. Only arbitration subset (7,379 images from 735 encounters; enriched for DR and ungradable images) provides comparative data.	Retrospective single-gate	7 AI systems (OphtAI, AirDoc, Eyenuk, Retina-AI Health, Retmarker; one FDA-approved); identities masked (ARIAS A to ARIAS G)	Any DR: ICDR grades 1–4) or unreadable	Original teleretinal grades (2006 – 2018), no information on human graders, might include images graded by both an ARIAS and a human grader	Double-masked arbitration by 2 clinical experts, 3rd retina specialist for arbitration	Any DR – Human grading: SE 82.2% (95%CI 80.8 - 83.6), SP 84.4% (95%CI 83.0 – 85.7). Any DR – ARIASs: ARIAS A: SE lower than human (p<0.05), SP 90.0% (95%CI 88.9 – 91.1). ARIAS B: SE lower than human (p<0.05), SP lower than human (p<0.05). ARIASs C, D: SE similar to human (p>0.05), SP lower than human (p<0.05). ARIASs E, F: SE 92.7% (95%CI 91.8 – 93.7), SP lower than human (p<0.05). ARIAS G: SE 80.5% (95%CI 79.0 – 81.9), SP 81.3% (95%CI 79.8 – 82.7).	Prioritised
Li 2021a[100], China	Prospective clinic-based real-world screening; 1,147	Prospective single-gate	Voxel-Cloud	Referable DR	No human comparator	One certified retinal specialist	Referable DR – VoxelCloud: AUC 0.942	Deprioritised: Country, no

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	consecutive DM patients, 2,286 eyes; Oct 2018–Aug 2019; no mydriasis.		automated DR grading software (China)			with more than 12 years of experience who graded images based on the ICDRS scale	(95% CI 0.920-0.964), SE 85.1% (95% CI 83.4-86.8), SP 95.6% (95% CI 94.6-96.6)	human comparator
Li 2021b[101], Taiwan	Before–after implementation in endocrinology clinic: 716 patients (June 2019, pre) vs 700 (Oct 2019, post); consecutive sampling; no mydriasis.	Uncontrolled before-after study, single-gate	VeriSee™ developed by Acer Inc. (New Taipei City, Taiwan).	Referable DR = moderate NPDR or worse (ICDR ≥2); results presented at image and patient levels	5 endocrino-logists in routine clinical practice.	3 ophthalmologists (disagreements adjudicated by senior specialist; consensus of ≥2 required).	Referable DR – VeriSee: SE 91% (95% CI 83 – 96), SP 90% (95% CI 85 – 92) at image level. SE 91% (95% CI 81 – 97), SP 84% (95% CI 80 – 87) at patient level. Referable DR – Endocrinologists: SE 91% (95% CI 81 – 97), SP 50% (95% CI 45 - 55).	Deprioritised: Country
Lim 2023[102], USA (related to Ipp et al. 2021[97])	521/893 adults with DM (999 eyes) at 10 U.S. centres (subgroup who also underwent dilated ophthalmoscopy); nonmydriatic images with dilation if needed.	Prospective single-gate	EyeArt (Eyenuk, Inc.)	ETDRS-based. mtmDR (referable) = ETDRS ≥35 or CSME; vtDR = ETDRS ≥53 or CSME. Primary endpoint is mtmDR; vtDR also reported.	No human comparator (only compares to dilated ophthalmosc opy)	Wisconsin Fundus Photograph Reading Center (WFPRC) grading of 4-widefield stereoscopic dilated fundus photographs using the ETDRS severity scale. 2 independent masked human graders with adjudication by a	mtmDR – EyeArt: SE 96.4% (95% CI 93.1 - 99.8), SP 88.4% (95% CI 85.8 - 91.1). mtmDR – Dilated ophthalmoscopy (not eligible for our review): SE 27.7% (95% CI 20.1 - 35.2), SP 99.6% (95% CI 99.1 - 100).	Deprioritised: No human comparator (only compares to dilated ophthalmos-copy)

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
						third more senior grader.		
Limwatta-nayingyong 2020[103], Thailand	5,738 randomly selected DM patients screened twice ≈2 years apart (2014/2016 or 2015/2017) within a nationwide screening programme; no mydriasis.	Retrospective single-gate	Google DL system (Google Health)	STDR = severe NPDR, PDR, or DME (ICDR-based). This is a referable threshold.	Trained human graders selected from regional DR graders within the national DR screening program.	Adjudication by 3 international retina specialists (from USA, India, and Thailand) who graded colour fundus photographs according to the International Clinical Classification of DR.	STDR - Google DL system: SE 95.03% (95%CI 93.42-96.63), SP 97.97% (95%CI 97.58-98.36) in first screening. SE 90.05% (95%CI 86.01-94.09) SP 97.87% (95%CI 97.42-98.32) in 2nd screening. STDR - Human graders: SE 73.72% (95%CI 70.47-76.97) SP 98.59% (95%CI 98.26-98.92) in first screening. SE 57.09% (95%CI 51.39-62.8) SP 98.52% (95%CI 98.14-98.89) in 2nd screening.	Deprioritised: Country
Mbaye 2025[105], Senegal - West Africa	156 consecutive DM patients (305 eyes) at Abass NDAO Hospital, Dakar, Senegal; March 1, 2021, to September 30, 2022. Dilated images.	Prospective single-gate	Gaiha Prio Retino +™	Any DR, Referable DR + Maculopathy	No human comparator	Clinical classification by 2 specialist ophthalmologists using American Academy of Ophthalmology criteria.	Any DR – ARIAS: SE 92.06%, SP 100%, PPV 100%, NPV 94.71%, AUC 0.968. Referable DR - ARIAS: SE 92.31%, SP 99%, PPV 97.96%, NPV 96.14%, AUC 0.989. Maculopathy – ARIAS: SE 95.7%, SP 97.17%, PPV 93.68%,	Deprioritised: Country, no human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							NPV 98.1%, AUC 0.988.	
Mehra 2022[106], USA	1,052 consecutive adult patients (≥18 yrs) with type 1 or type 2 DM, primary care setting, single centre, telemedicine DR screening; mainly nonmydriatic, dilation if needed.	Retrospective single-gate	IDx-DR (Digital Diagnostics)	Referable only (≥ moderate NPDR [R2+], R3, and/or M1)	No human comparator	Manual overread of fundus photographs by ophthalmologists /optometrists (equivocal cases adjudicated by retina specialist).	mtmDR - IDx-DR: SE 100% (95% CI 90.8-100.00), SP 89.2% (95% CI 87.0-91.1), PPV 27.5% (95% CI 24.0-31.4), NPV 100% (95% CI 99.6-100).	Deprioritised: No human comparator.
Meredith 2023[107], UK	10,000 consecutive people with DM (≥12 years) attending routine appointments in the North West London DESP; mydriatic images.	Retrospective single-gate	RetCAD v.2.1.0 (Thirona, The Netherlands)	Both Any DR (R1 or higher) and Referable DR (R2, R3)	No human comparator	Manual DESP grade in worst affected eye	Any DR – RetCAD (predefined threshold; person-level): SE 69.7% (1,854/2,661), SP 92.2% (6,600/7,156), PPV 76.9% (1,854/2,410). Any DR – RetCAD (optimal threshold; person-level): SE 78.5%, SP 86.5%, AUC 0.885 (95% CI 0.877 – 0.893). Referable DR – RetCAD (predefined threshold; person-level): SE 95.4% (83/87), SP 92.0% (8,947/9,730), PPV 9.6% (83/866), AUC 0.979 (95% CI 0.955–0.993).	Deprioritised: No human comparator
Ming 2021[108], China	193 community-based consecutive DM patients (≥40 years) in Zhengzhou, China, attending a local primary clinic; nonmydriatic images.	Prospective, single-gate	Eye-Wisdom DL-based (Visionary Intelli-	Any DR and Referable DR (RDR, defined as	No human comparator	2 experienced ophthalmologists, with arbitration panel (5 senior ophthalmologists	Eye-wise (n=321 eyes): Referable DR: SE 79.2% (95% CI 57.9–92.9), SP 98.3% (95% CI 96.1–99.5), AUC 0.887	Deprioritised: Country, unclear if ARIAS commercially

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
			gence Ltd, Beijing).	more than mild NPDR).		based on the majority decision)	(95% CI 0.790–0.984). Any DR: SE 83.3% (95% CI 67.2–93.6), SP 97.9% (95% CI 95.5–99.2), AUC 0.906 (95% CI 0.834–0.979). Patient-wise: Referable DR: SE84.6% (95% CI 54.6–98.1), SP 98.0% (95% CI 94.3–99.6), AUC 0.913 (95% CI 0.797–1.000). Any DR: SE 90.0% (95% CI 68.3–98.8), SP 96.6% (95% CI 92.1–98.9), AUC 0.933 (95% CI 0.833–1.000).	available, not CE marked, no human comparator.
Mokhashi 2022[109], USA	260 consecutive DM patients from an internal medicine clinic, ≥18 years, non-mydriatic 2-field retinal imaging.	Retrospective single-gate	EyeArt (Eyenuk)	Referable DR (moderate NPDR or worse, or macular edema features).	No human comparator	Optometrist grading; all disagreements and a subset of 20% (n = 29) of agreements between the optometrist and ARIAS adjudicated by retina specialist.	Referable DR: SE 100% (9/9) SP 77.78% (133/171). PPV 19.15% (9/47). NPV 100% (133/133).	Deprioritised: No human comparator
Musetti 2025[110], Italy	201 consecutive patients with type 1 or type 2 DM (mean age 65±13 years) recruited from the diabetology clinic at University of Genoa; nonmydriatic images with	Prospective single-gate	IDx-DR (IDX, LLC), EyeArt v2.0 (Eyenuk)	Referable DR	No human comparator	2 masked retinal specialists; in case of discordance images were reexamined until	Best-scenario (excluding ungradable cases) Referable DR - EyeArt: SE 100% (95% CI 89.4–100.0), SP 72.9% (95% CI 65.2–79.7).	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	dilation if needed.					a consensus was obtained.	Referable DR - IDx: SE 94.3% (95% CI 80.4–99.3), SP 66.0% (95% CI 57.8–73.5). Worst-scenario (treating ungradable cases as misdiagnoses): Referable DR - EyeArt: SE 86.8% (95% CI 71.9–95.6), SP 69.3% (95% CI 61.6–76.3). Referable DR - IDx: SE 86.8% (95% CI 71.9–95.6 SP 60.7% (95% CI 52.8–60.2).	
Nissen 2023[59], Denmark	Danish National DESP (Steno Diabetes Center, North Jutland); 1,001 mydriatic images, unclear how selected.	Retrospective diagnostic accuracy study	Retina-Lyze (Retina-Lyze A/S)	Any DR: Any disease	Routine grading 2019-2020; 10 ophthalmo-logists' consultants and senior registrars	Majority consensus of 3 independent ophthalmologists	Any DR – ARIAS: Lower threshold (≥1): SE 96.8% (95% CI 93.3 – 98.2), SP 51.7% (95% CI 46.8 – 56.6). Higher threshold (≥3): SE 84.9% (95% CI 81.8 – 87.9), SP 89.9% (95% CI 86.8 – 92.7). Any DR – Human grading: SE 87.0% (95%CI 84.2 - 89.7), SP 85.3% (95%CI 81.8 – 88.6).	Prioritised
Noriega 2021[111], Mexico	Non-mydriatic 45° macula-centred images; inclusion required pupil ≥3 mm (i.e. non-mydriatic). 100 patients from a Mexican ophthalmic hospital: 50% (50/100) non-referable DR and 50% (50/100) referable DR levels.	RCT (2-gate)	ARIA system developed by the study team / MIT & Prosperia Salud (research	Referable DR	17 ophthalmo-logists (retina residents): each evaluated 45 random images (15	3 retina specialists	Mexican hospital image subset: Referable DR – ARIA: AUROC 98.3%, SE 95.2%, SP 90.0%. Referable DR - Human graders (averages across 17 participants): Solo condition:	Deprioritised: Country, ARIAS not CE-marked and not commercially available

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
			ARIA) - DL based		per grading mode). (1) Solo: screen the images by themselves; (2) ARIA answer: screen image after exposure to ARIA classification; (3) ARIA explanation: screen image after exposure to ARIA classification, its level of confidence and heatmap)		SE 87.3%, SP 86.8%. ARIA answer condition: SE 93.3%, SP 89.3%. ARIA explanation condition: SE 91.5%, SP 79.0%.	
Parravano 2025[112], Italy	224 adults with DM (422 eyes) from IRCCS-Fondazione Bietti in Rome, Italy; 91.2% nonmydriatic images acquired with DRSplus retinographer (Centervue SpA); mosaic of partially overlapped retinal images covering 63°of the retina (from central to nasal).	Prospective, unclear how selected	RetCAD (Thirona Retina)	Referable DR (moderate DR or worse)	No human comparator for test accuracy outcomes (only for gradable images rate)	External reading center (InHealth Intelligence, Universal House, Middlewich) using English NHS DESP grading criteria	Referable DR (eye-level): SE 97.3% (95% CI 94.74–99.86), SP 80.4% (95% CI 75.57–85.23), PPV 74.2% (95% CI 68.15–80.33), AUC 0.979 (95% CI 0.0968 – 0.990)	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
Peeters 2023[113], Belgium, France and USA	Private test set and publicly available datasets (EyePACS: USA; n = 16,733 for DR and DME evaluation; Kaggle DR test set: USA; n = 5,000 patients; multiple cameras; Messidor-2: France; n = 874 patients; Topcon NW6; Messidor-2 Iowa reference: France; n = 874 patients; Topcon NW6). Dilated and undilated images.	Retrospective (used existing labelled datasets)	MONA DR DME (MONA. health, Leuven, Belgium), Version 1.0.0.	Referable DR (R2, R3) + DME	No human comparator	Ground truth labels	Referable DR (private test set): AUC 97.28% (95% CI 97.50–97.52), SE 90.67% (95% CI 90.03–91.31), SP 94.62% (95% CI 94.12–95.08). Referable DR (Kaggle test set): AUC 96.91% (95% CI 96.63–97.18), SE 88.45% (95% CI 87.63–89.28), SP 95.16% (95% CI 94.87–95.44). Referable DR (Messidor-2): AUC 97.99% (95% CI 97.08–98.76), SE 93.66% (95% CI 90.62–96.46), SP 92.86% (95% CI 90.73–94.83). Referable DR + DME (Messidor-2 Iowa's reference): SE 97.89% (95% CI 95.65–99.51), SP 84.06% (95% CI 81.30–86.72).	Deprioritised: No human comparator.
Pei 2022[114], China	549 consecutive patients with type 2 DM, ≥18 years, Henan Provincial People's Hospital, China; 1,768 nonmydriatic images.	Prospective single-gate	Eye-Wisdom® DSS and Eye-Wisdom® MCS	Any DR and severity grading (mild NPDR, moderate	No human comparator	Ophthalmologist grading (2 independent graders with adjudication)	Any DR - EyeWisdom®DSS SE 91.0% (95% CI 87.3–93.8), SP 81.3% (95% CI 75.5–86.1). AUC 0.862 (95% CI 0.827–0.897), PPV 87.5% (95% CI 83.4–90.8), NPV 86.3% (95% CI 80.8-90.5).	Deprioritised: Country, unclear if ARIAS CE-marked, no human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
				NPDR, severe NPDR, PDR)			Any DR - EyeWisdom®MCS SE 76.2% (95% CI 71.1–80.7), SP 92.4% (95% CI 87.9–95.4), AUC 0.843 (95% CI 0.809–0.878). PPV 93.5% (95% CI 89.7–96.1), NPV 73.0% (95% CI 67.4-78.0).	
Peris-Martinez 2021[115], Spain	Primary care diabetic retinopathy screening program in Valencia, Spain. 2,680 consecutive patients (after exclusions from the original 3,531) with type 1 and type 2 DM from 43 health centres; nonmydriatic images.	Retrospective single-gate	IDx-DR v2 (IDx Technologies Inc., Coralville, USA)	Referable DR and vision-threatening DR	No human comparator	3 Spanish ophthalmologists independently graded each exam using ICDR scale. If no consensus, anonymous reclassification by retinal specialist from an independent group, with their classification accepted if they agreed with one of the original 3 classifications of the ophthalmologists.	Referable DR – IDx-DR: SE 100.0% (95% CI 96.73-100), SP 81.82% (95% CI 80.27–83.30), PPV 19.20% (95% CI 17.96–20.51), NPV 100%, AUC 0.984 (95% CI 0.97–0.99). Vision-threatening DR – IDx-DR: SE 100.0% (95% CI 94.79-100), SP 94.64% (95% CI 93.70–95.47), PPV 33.01% (95% CI 29.55–36.67), NPV 100%, AUC 0.998 (95% CI 0.997–0.999).	Deprioritised: No human comparator
Piatti 2024[116], Italy	637 consecutive adults with DM (type 1, type 2, or other); screened in 4 diabetes/eye centres in ASLTO5, Turin region; nonmydriatic images.	Prospective single-gate	DAIRET® (comercial Italian version of Retmarker Screening; Retmarker SA,	Both – Any DR and Referable DR (moderate + DR)	No human comparator	Single ophthalmologist grader	Referable DR – DAIRET: SE 100% (35/35), SP 0.80 (399/501) (95% CI ±0.04). Mild DR detection - DAIRET: SE 0.69 (70/101), (95% CI ±0.09).	Deprioritised: No human comparator

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
			Coimbra, Portugal)					
Piatti 2025[117], Italy	506 consecutive adult patients with DM (1,006 eyes) screened at 2 facilities (diabetes centre and eye centre) in Turin, Italy; mydriasis if needed only.	Prospective single-gate	RetCAD (Thirona B.V.) accessed through iCare ILLUME interface, paired with the iCare DRSplus confocal fundus camera (Center-vue).	Both – Any DR and Referable DR (moderate + DR)	No human comparator	Single ophthalmologist (local expert)	Eye-level (analysis set n = 971 eyes where both AI + human provided scores): SE 97.18% (138/142) (95% CI ±2.72), SP 93.73% (777/829) (95% CI ±1.65). Patient-level (analysis set n = 491 patients with paired scores): SE 98.70% (76/77) (95% CI ±2.53), SP 91.06% (377/414) (95% CI ±2.75).	Deprioritised: No human comparator.
Pinto 2025[66], Spain and China	External public test set: random 30% of OIA-DDR dataset (4,105 images); 147 different hospitals using 42 different camera models. No eye drops used. 346 non-gradable images excluded (→ 3,759 images; 1,690 referable DR).	Retrospective (external public test set)	NaIA-RD, a custom in-house algorithm	Referable DR	No human comparator	Labelled OIA-DDR test set (4 professional graders)	Referable DR – NaIA-RD: AUC 0.957 SE 93%, SP 84%.	Deprioritised: ARIAS not commercially available, no human comparator
Poschkamp, 2025[62], Germany	1 specialised diabetes centre (Klinikum Karlsburg); 1,802 consecutive patients with DM; nonmydriatic images.	Single-gate (RetCAD retrospect-tively)	IDx-DR and RetCAD	Referable DR: more-than-mild DR	1 of 6 ophthalmo-logists (indepen-dent from fundoscopy analysis)	Ophthalmologist mydriatic fundoscopy with image analysis (1 of 6 experienced ophthalmologists)	Referable DR - IDx-DR: SE 90.16% (165/183)†, SP 82.95% (798/962)†. Referable DR – RetCAD (manufacturer threshold): SE 70.41% (188/267)†, SP 93.86% (1,268/1,351)†.	Prioritised

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							Referable DR – RetCAD (Youden adjusted threshold): SE 86.89% (232/267)†, SP 82.68% (1,117/1,351)†. Referable DR – RetCAD (customised referral threshold): SE 93.63% (250/267)†, SP 69.06% (933/1,351)†. Referable DR – Human grading: SE 78.23% (151/193)†, SP 99.41% (1,018/1,024)†.	
Rao 2022[119], India	Tertiary eye care centre in south India; 135 consecutive patients (233 eyes) with established DM; mydriatic images	Retrospective single-gate (post-hoc analysis of previously collected dataset)	Medios AI (Medios Technologies)	Any DR + Referable DR	No human comparator	1) Consensus image grading by 2 fellowship-trained vitreo-retina specialists. 2) Clinical examination by a single retina specialist using slit lamp biomicroscopy and indirect ophthalmoscopy.	Any DR (reference standard 1): SE 97.6% (95% CI 95–100), SP 90.9% (95% CI 82.4–99.4), PPV 96.9% (95% CI 93.8-99.9), NPV 93% (95% CI 85.4-100). Any DR (reference standard 2): SE 97.6% (95% CI 94.9-100), SP 88.9% (95% CI 79.7-98.1), PPV 96.1% (95% CI 92.7-99.4), NPV 93.0% (95% CI 85.4-100).	Deprioritised: Country, no human comparator.
Riotto 2024[120], Switzerland	1,350 consecutive DM patients referred to dedicated diabetic retinopathy clinic at Jules Gonin Eye Hospital. 1,141 patients (2,282 images) after excluding 209 patients with insufficient	Retrospective single-gate	IDx-DR	Any DR (the study evaluated no DR, mild DR, moderate DR, and severe DR	No human comparator	Human grading by 2 experienced retinal specialists.	Any DR – IDx-DR: SE 100% (127/127)†, SP 78.4% (795/1,014)†, PPV 36.7% (127/346)†, NPV 100% (795/795)†.	Deprioritised: No human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	quality images; mydriasis unclear.			using ICDR classification)				
Ruamvi-boonsuk 2022[121], Thailand	7,651 consecutive DM patients from 9 primary care centres (8 community hospitals in rural settings, 1 primary care clinic in an urban tertiary care hospital) across 3 regions in Thailand, part of the Thai national DESP; mydriasis if needed only. Stratified sampling for adjudication by retina specialists (1,208 patients).	Prospective single-gate	DLS (Google Health)	Vision-threatening DR	Overreading by 2 regional retina specialists	Panel of 3 US board-certified retina specialists, drawn from a pool of 11 specialists. Only stratified subsample of 1,208 patients received reference standard depending on DLS / overreader agreement: all disagree as to referability, all disagree exact DR grade of referable, all disagreed ungradable/ referable, random 2-5% of other.	Vision-threatening DR – DLS (patient-level): SE 91.4% (95% CI 87.1–95.0), SP 95.4% (95% CI 94.1–96.7). Vision-threatening DR - Regional retina over-readers (patient level): SE 84.8% (95% CI 79.4–90.0) (p = 0.024), SP 95.5% (95% CI 94.1–96.7) (p = 0.98).	Deprioritised: Country.
Ruan 2022[122], China	630 eyes of 315 DM Patients from 3 hospitals (Shanghai General Hospital, West Nanjing Road Community Health Center in Shanghai, and Zhaoqing	Prospective single-gate	Phoebus algorithm (Phoebus Medical, Shanghai).	Referable DR (moderate or worse DR	No human comparator (compares ARIAS performance with	3 masked, experienced ophthalmologists graded image quality and made	Referable DR – Phoebus (table-top): SE 92.7% (95% CI 85.9-99.6), SP 95.9% (95% CI 93.2–98.5), AUC 94.5% (95% CI 91.5-97.5).	Deprioritised: Country, unclear if ARIAS commercially available, no

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	Gaoyao People's Hospital in Guangdong). With or without mydriasis. Only images taken with the table-top camera eligible for our review.			requiring referral)	handheld camera [Aurora, Optomed, Oulu, Finland] to a table-top camera).	diagnoses independently		human comparator.
Rudnicka 2025[61], England	North East London DESP (Homerton Healthcare NHS Foundation Trust; 202,886 consecutive encounters (all screening pathways), 126,365 patients; mydriatic images.	Prospective single-gate (ARIAS grading retrospective)	8 ARIASs: DRISTi 2.0 (Artelus Ltd), EyeArt (EyeNuc Inc), EyeCheck up AI (EyeChec kup), MONA (Mona. health), NEC (NEC Software Solutions), OphtAI 2.3 (Evolucare Technolo-gies SAS), Remidio (Remidio Innovative Solutions Pvt Ltd),	Referable DR, also by DR grade; accuracy overall and by subgroups of age, sex, ethnicity and index of multiple deprivation	NHS DESP primary graders	Final human grade in the worst eye per encounter (up to 3 trained human graders: abnormal or ungradable images and a random 10% of normal images reviewed by secondary graders; disagreements and complex cases are referred to arbitration graders).	Referable DR (excluding Ungradable) – DRISTi 2.0, threshold 1: SE 87.2% (95% CI 86.8 – 87.6), SP 56.0% (95% CI 55.8-56.2)†. EyeArt v3.0.0, threshold 1: SE 98.2% (95% CI 98.1 – 98.4), SP 60.2% (95% CI 59.9-60.4)†. EyeCheckup AI, threshold 1: SE 95.8% (95% CI 95.5 – 96.0), SP 39.9% (95% CI 39.6-40.1)†. MONA, threshold 1: SE 84.2% (95% CI 83.7 – 84.7), SP 86.5% (95% CI 86.3-86.6)†. NEC, threshold 1: SE 98.7% (95% CI 98.5 – 98.8), SP 30.5% (95% CI 30.3-30.7)†. OphtAI 2.3, threshold 1: SE 89.8% (95% CI 89.4 – 90.1), SP 74.5% (95% CI 74.3-74.7)†. Remidio, threshold 1: SE 87.0% (95% CI 86.6 – 87.4), SP 80.9% (95% CI 80.7-81.0)†. Retmarker, threshold 1: SE 83.7% (95% CI 83.2 – 84.1), SP 75.4% (95% CI 75.1-75.6)†. Primary human grader: SE 94.6% (95% CI 94.3 – 94.9),	Prioritised

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
			Retmarker (Retmar-ker SA)				SP NR.	
Sarao 2020[123], Italy	165 consecutive DM patients (type 1 or type 2) attending routine annual visits at an ambulatory surgical centre in Italy. 330 eyes; mydriatic images.	Prospective single-gate	EyeArt v2.1 (Eyenuk Inc., Los Angeles)	Referable DR	No human comparator (compares a conventional flash fundus camera and a white LED confocal scanner).	2 retina specialists masked to each other and to any patient records; involvement of 3rd retina specialist in case of disagreements.	Referable DR – EyeArt (Conventional flash fundus camera): SE 90.8% (95% CI 85.0–94.9), SP 75.3% (95% CI 68.0–81.7), AUC 0.830 (95% CI 0.78–0.87). Referable DR – EyeArt (White LED confocal scanner): SE 94.1% (95% CI 89.1–97.3), SP 86.8% (95% CI 80.7–91.6), AUC 0.905 (95% CI 0.87–0.93). Difference between AUCs: 0.0737 (95% CI 0.0263–0.121), p = 0.0023 (scanner superior).	Deprioritised: No human comparator.
Scheetz 2021[124], Australia	236 consecutive patients with type 1 or type 2 DM attending 2 endocrinology outpatient clinics and 3 Aboriginal Medical Services clinics in Australia. Single 45-degree (macula centred) image without mydriasis.	Prospective single-gate	NR (Custom offline AI system)	Referable DR (pre-proliferative DR or worse)	No human comparator	2 NHS-certified retinal graders using the NHS diabetic eye screening Guidelines.	Referable DR – ARIAS (n=203): AUC 0.92 (95% CI 0.88–0.96), SE 96.9% (95% CI 83.8–99.9), SP 87.7% (95% CI 81.8–92.2), PPV 59.6% (95% CI 45.1–73.0), NPV 99.3% (95% CI 96.4–100).	Deprioritised: ARIAS not CE-marked and not commercially approved, no human comparator.
Sedova 2022[125], Austria	54 consecutive asymptomatic patients with type 1 and type 2 DM without previous diagnosis of DR, recruited from Department of Ophthalmology at Medical University of Vienna. Two 45° colour fundus images from the centre of the macula and the optic disc captured using	Prospective single-gate	IDx-DR v2.2 (Digital Diagno-stics)	Referable DR (moderate / severe)	No human comparator	Human grading of UWF Optomap images: 2 retina specialists graded 1) 7F-mask and 2) full-field on UWF images.	1) Referable DR – IDx-DR (7F-mask grading): SE 100% (95% CI 83–100), SP 47% (95% CI 30–65), PPV 53% (95% CI 36–69), NPV 100% (95% CI 79–100). 2) Referable DR – IDx-DR (UWF full-field grading): SE 95% (95% CI 77–100),	Deprioritised: No human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	the Topcon TRC-NW400 non-mydriatic fundus camera (Topcon Medical Systems, Inc.).						SP 47% (95% CI 29–65), PPV 55% (95% CI 38–71), NPV 94% (95% CI 70–100).	
Sin 2025[60], Czech Republic	4 Czech primary diabetes-care centres, 1,273 consecutive patients; nonmydriatic images.	Prospective single-gate	Aireen (Aireen a.s.)	Any DR: Any DR or ungradable	1. General ophthalmologists – no subspeciality training in retina, 2. Retina specialists	Agreement between the 3 index tests (n = 734); Majority consensus of 3-member high-level expert committee (n = 420)	Any DR – Aireen: SE 92.1% (95% CI 89.3 – 94.9), SP 90.7% (95% CI 88.7 – 92.7). Any DR – General ophthalmologists: SE 87.0% (95% CI 83.6 – 90.4), SP 76.5% (95% CI 73.5 – 79.5). Any DR – Retina specialists: SE 82.9% (95% CI 79.1 – 86.7), SP 81.2% (95% CI 78.5 – 83.9).	Prioritised
Skevas 2024[126], Germany	231 consecutive patients recruited from Ophthalmology Outpatient Department at University Medical Center Hamburg-Eppendorf; with mydriasis.	Prospective single-gate	Integrated AI modules in TeleEye MD platform (TeleMed C).	Referable DR	No human comparator	Human auditor grading (single specialist auditor for this study)	Referable DR – ARIAS: SE 100.00% (95% CI 87.23–100.00), SP 80.10% (95% CI 73.98–85.32), PPV 37.33% (95% CI 25.90–49.91), NPV 100.00% (95% CI 97.79–100.00), AUC 0.90 (95% CI 87–93).	Deprioritised: Unclear if CE-marked, no human comparator
Tan-Torres 2025[127], India	321 consecutive individuals aged 18-45 attending "Diabetes of Young" clinic at AIIMS, New Delhi. 98.8% had type 1 DM, 1.2% had type 2 DM. Mydriatic images.	Prospective single-gate	Deep-learning system (DLS) by Google	Referable DR (moderate-or-worse DR)	No human comparator	3 graders (2 ophthalmologists and 1 optometrist) with the majority opinion taken if consensus not reached within 3 rounds of discussions.	Referable DR – DLS: (eye-level, n = 642): SE 95.1% (95% CI 91.0 – 97.8), SP 95.3% (95% CI 93.2 – 97.2). Subgroup results reported for younger cohort (18 – 25 years, n = 324) and older cohort (26 – 45 years, n = 318).	Deprioritised: Country, unclear if ARIAS CE-marked and commercially available, no human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
Teoh 2023[128], Singapore	Patients with DM attending diabetic retinopathy screening visits at the outpatient ophthalmology clinic at National University Hospital of Singapore. 200 macula-centred fundus photographs randomly selected; unclear mydriasis use.	Retrospective single-gate	Automa-ted DL DR screening software developed by the Common-wealth Scientific and Industrial-Research Organisa-tion (CSIRO); DiaRetDB algorithm	Non-referable DR, non-urgent referable DR, urgent referable DR	No human comparator f or accuracy results	3 retinal specialists: Majority consensus or, if no consensus, reference based on opinion of the most senior retinal specialist.	Non-referable DR – ARIAS: AUC 0.879 (95% CI 0.829–0.929); SE 70.1%, SP 93.5%, PPV 87.1%, NPV 83.3%. Non-urgent referable DR – ARIAS: AUC 0.714 (95% CI 0.643–0.785); SE 90.1%, SP 45.4%, PPV 54.9%, NPV 87.1%. Urgent referable DR – ARIAS: AUC 0.836 (95% CI 0.779–0.893); SE 85.7%, SP 75.9%, PPV 48.6%, NPV 95.2%.	Deprioritised: Country, unclear if ARIAS CE-marked and commercially available, no human comparator.
Tsai 2022[129], Taiwan	DM patients from 3 general practice clinics in Taiwan. 1,407 nonmydriatic colour fundus images taken with 3 different fundus cameras (477 Topcon TRC-NW400, 459 Topcon TRC-NW8 series, and 471 Kowa nonmyd 8 series); unclear how selected.	Retrospective, unclear how images were selected	VeriSee DR (Acer Inc., Taiwan).	Any DR + referable DR	No human comparator (comparison of 3 different fundus cameras)	Majority voting of 3 board-certified ophthalmologists grading images independently	Any DR - VeriSee DR: SE 92.7% (95% CI 87.5–97.9) for TRC-NW400, SE 92.9% (95% CI 88.4–97.4) for the TRC-NW8 series, and SE 95.1% (95% CI 92.0–98.2) for the Kowa nonmyd-8 series. SP 61.7% (95% CI 56.8–66.6) for TRC-NW400, SP 73.9% (95% CI 69.2–78.6) for TRC-NW8, and SP 56.9% (95% CI 51.2–62.7)	Deprioritised: Country, unclear if ARIAS CE-marked and commercially available, no human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
							for the nonmyd-8 series.	
Vaghefi 2023[130], New Zealand	900 consecutive patients recruited from New Zealand DESP from 2 sites: Patients with Type 1 and Type 2 DM, over age 21; mydriasis if needed only.	Prospective single-gate	THEIA™ (Toku Eyes® / University of Auckland group).	Referable DR (R3-R5, M4-M5) (Any DR can be calculated though)	No human comparator	3 independent senior specialist graders (masked); in case of disagreement adjudication by a 4th independent, senior retinal specialist.	Referable DR - THEIA (patient level): SE 100% (95% CI 98.49–100.00), SP 98.33% (95% CI 97.02–99.16), NPV 100%.	Deprioritised: No human comparator.
Van 2024[131], Vietnam	583 consecutive DM patients from hospitals and health centres in Binh Dinh province; 2,332 nonmydriatic images.	Retrospective single-gate	EyeArt system v2.0 (Eyenuk, Inc.)	Any DR + Referable DR – vision-threatening DR	No human comparator	2 eye doctors	Any DR - EyeArt: SE 94.1% (95% CI 90.4–97.8), SP 87.2% (95% CI 83.9–90.2), PPV 69.2% (62.7–75.7), NPV 98.0% (96.5–99.2). Referable DR - EyeArt: SE 96.6% (95% CI 93.3–99.2), SP 90.1% (95% CI 87.3–92.9), PPV 71.4% (64.6–78.3), NPV 99.1% (98.1–99.8).	Deprioritised: Country, no human comparator.
Wang 2024[132], China	Participants with diabetes from a nationwide DR-screening program; 736,083 nonmydriatic images from 237,824 consecutive participants aged 12+ years with DM; random sample for adjudication (10% sample from FN cases and a 5% sample from FP cases)	Retrospective single-gate	Custom DL algorithm	Referable DR	No human comparator	1) Routine human labels from NHS-certificated graders using multi-stage grading workflow. 2) Adjudication by 2 experienced ophthalmologists with any discrepancy sent to a senior	1) Any DR - ARIAS (precorrection, participant level): SE 88.1% (15,236/17,298), SP 84.6% (186,648/220,526). 2) Any DR – ARIAS (postcorrection**, participant level): SE 96.0% (95%CI 95.1-96.8), SP 85.4% (95% CI 85.3 – 85.6). ** When adjudication yielded consistent results with DL instead of human labels, label error rates in the adjudicated FN	Deprioritised: Country, ARIAS not CE-marked and not commercially available, no human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
						ophthalmologist whose grading results were deemed as final.	and FP samples were used to recalculate number of positive and negative cases in the entire data set.	
Whitestone 2024[133], Rwanda	4 diabetes clinics in Rwanda; 827 consecutive participants with type 1 (41.2%) and type 2 (57.4%) DM, ages 18-91 years; nonmydriatic images.	Prospective single-gate	Cybersight AI (Orbis Internatio-nal's Cybersight platform).	Referable DR	No human comparator	Single UK NHS-certified senior retinal image grader (masked) provided human grading; comparisons at patient level (most severe eye)	Referable DR – Cybersight AI: SE 92.0% (95% CI 86.3–96.8). SP 85.0% (95% CI 82.4–87.4). AUC 0.96 (95% CI 0.941–0.975).	Deprioritised: Country, unclear if CE-marked, no human comparator.
Wintergerst 2022,[134] Germany	Primary care/general practice setting; 75 consecutive patients with DM (96% type 2, 4% type 1), mean age 69 years; nonmydriatic images.	Prospective single-gate	EyeArt v2.1.0 (Eyenuk, Inc.).	Any DR + Referable DR	No human comparator	Double manual grading (trained grader + ophthalmologist); third ophthalmologist adjudicated disagreements (graders masked to AI).	Any DR – EyeArt (subset evaluable by the ophthalmologist; n = 63): SE 0.67 (95% CI 0.22–0.96), SP 0.74 (95% CI 0.60–0.84), Any DR – EyeArt (subset evaluable by the algorithm, n = 48): SE 1.00 (95% CI 0.40–1.00), SP 0.98 (95% CI 0.88–1.00), PPV 0.80 (95% CI 0.28–0.99), NPV 1.00 (95% CI 0.92–1.00).	Deprioritised: No human comparator
Wongchai-suwat 2021[135], Thailand	Phase 2: Clinical verification - Patients with DM attending ophthalmology outpatient clinic and diabetes centre at a tertiary care hospital (Bangkok, Thailand); nonmydriatic images taken on Nidek (n = 982) and Eidon (n = 674) cameras.	Phase 2: Prospective single-gate	NR	Referable DR	No human comparator	Retinal examination by a retinal expert (8 retinal experts at Siriraj Hospital) using a biomicroscopic slit	Phase 2: Referable DR – ARIAS: Single photo, all cases: Nidek: SE 82%, SP 92%, PPV 82%, NPV 92%. Eidon (new model):	Deprioritised: Country, ARIAS not CE-marked and not commercially available, no human comparator.

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	First photo macula-centred, remaining 2 photos nasal and temporal overlapping images.					lamp with a non-contact lens in dilated eyes, fundus examinations to zone 3 (i.e., beyond the equatorial zone).	SE 89%, SP 84%, PPV 73%, NPV 94%. 3 photos, all cases: Nidek: SE 97%, SP 30%, PPV 37%, NPV 96%. Eidon (new model): SE 95%, SP 66%, PPV 57%, NPV 97%.	
Xu 2025[136], China	56 consecutive adults (112 eyes) with type 2 DM, aged 39–83 years (mean 60.7), equally male and female, recruited from Shanghai Ninth People's Hospital, China. Macula-centred images captured using an automatic non-mydriatic Topcon TRC-NW400 camera (Topcon, Japan) or RetiCam3100 (SYSEYE New Vision, China).	Multiple reader multiple case (retrospective single-gate)	EvisionAI software	Any DR	21 doctors (4 primary care physicians, 7 ophthalmo-logy residents, 10 attending ophthalmo-logists) graded images first without and then with AI assistance (1-week washout phase)	Gold standard diagnosis and grading established by a senior retinal specialist.	Any DR – AI-assisted grading (all 21 doctors): SE 77.66%, SP 94.55 %, PPV 92.35%, NPV 85.96%. Any DR – Human grading (all 21 doctors): SE 81.67%, SP 79.37%, PPV 78.36%, NPV 87.21%. Also reports accuracy results separately for primary care phy-sicians, ophthalmology residents and attending ophthalmologists.	Deprioritised: Country

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
Yang 2022[137], China	1,001 consecutive adults with DM (type 1 or 2, age ≥18 years) recruited from 3 centres across China (Zhongshan Ophthalmic Center, Peking University People's Hospital, Eye Hospital of Wenzhou Medical University). 2-field images; 2 centres: dilated images; 1 centre: dilation if needed only. Randomly selected 1 eye of each participant.	Prospective single-gate	AIDR-Screening v1.0 (Shen-zhen SiBright Co. Ltd.).	Referable DR	No human comparator	Zhongshan Image Reading Center manual grading performed independently by 2 graders. If discrepancies existed, discussion of the first 2 graders and a third senior grader to reach final conclusive grading.	Referable DR – AIDRScreening: SE 86.72% (95% CI 83.39–90.05), SP 96.09% (95% CI 94.14–97.54), PPV 94.02% (95% CI 91.09–96.22), NPV 91.08% (95% CI 88.49–93.25).	Deprioritised: Country, ARIAS not CE-marked and no FDA approval, no human comparator.
Yao 2024[138], China	474 participants (aged 18-70 years) from permanent residents; stratified multistage cluster sampling from hospital catchment area in China. Mixed population: 279 normal glucose metabolism (58.86%), 117 prediabetes (24.68%), 78 DM (16.46%); 2-field 45° fundus images, dilation if needed only.	Prospective single-gate	Deep-learning convolu-tional network with attention mecha-nism (ensemble / attention-based CNN).	Any DR and Referable DR	No human comparator	Centralised manual reading by 2 professional ophthalmologists, double-blind scoring (two 45° fields per eye), ETDRS.	Any DR – ARIAS (Diabetic population, patient-level): SE 91.6% (95% CI 86.3–95.3), SP 89.0% (95% CI 87.0–90.7), AUC 0.903 (95% CI 0.886- 0.918). Referable DR – ARIAS (Diabetic population, patient-level): SE 98.1% (95% CI 89.7–100.0), SP 82.1% (95% CI 79.9–84.2), AUC 0.901 (95% CI 0.884-0.916).	Deprioritised: Country, ARAIS not CE-marked and not commercially available, no human comparator.
Zhang 2022[139], China	Consecutive DM patients from 3 medical centres in China (Peking Union Medical	Prospective single-gate	Eye-Wisdom V1	Any DR + referable DR = sight-	No human comparator	3 masked senior physicians; in case of	Any DR – EyeWisdom: SE 98.23% (95% CI 96.93–99.08),	Deprioritised: Country, unclear if

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
	College Hospital, Eye Hospital China Academy of Chinese Medical Sciences, and Beijing Friendship Hospital of Capital Medical University). Mean age 56.52 ± 11.13 years; 630 patients with 1,089 non-dilated posterior pole fundus images.		(Visionary Intelli-gence Ltd., Beijing).	threatening DR		disagreements, a 4th grader's (principal investigator) opinion served as final grade	SP 74.45% (95% CI 69.95–78.60), PPV 86.38% (95% CI 83.76–88.72), NPV 96.23% (95% CI 93.50–98.04). Referable DR – EyeWisdom: SE 92.96% (95% CI 90.66–94.84), SP 93.32% (95% CI 90.65–95.42), PPV 94.93% (95% CI 92.89–96.53), NPV 90.78% (95% CI 87.81–93.22).	ARIAS CE-marked and/or FDA approved, no human comparator.
Zhang 2020[140], China	47,269 consecutive patients with DM aged 18+ from 155 diabetes centres (Metabolic Management Centers) across 26 provinces in China. Mean age 54.29 ± 11.60 years, 57.4% male. Study conducted June 2018 to August 2019. One non-mydriatic, macula-centred fundus photograph per eye. 1/3 randomly selected for DL algorithm validation (15,805 patients with 31,498 images).	Prospective single-gate	Voxel-Cloud Retina	Referable DR (moderate NPDR or worse) (Any DR can be calculated though)	Primary human grader	Two-stage human expert panel reading (2 independent human primary graders; senior reviewer adjudicated disagreements)	Referable DR – VoxelCloud: SE 83.3% (95% CI 81.9–84.6), SP 92.5% (95% CI 92.1–92.9), PPV 61.8% (95% CI 60.3–63.3), NPV 97.4% (95% CI 97.2–97.7) Youden index 75.8%. Referable DR – Primary human grader: Not reported but can be calculated from Supplementary Figure 4.	Deprioritised: Country.

AUC – Area Under the ROC Curve, AUROC – Area Under the Receiver Operating Characteristic curve, AUC-PR – Area Under the Precision–Recall Curve, AI – Artificial Intelligence, AMD – Age-related Macular Degeneration, ARIAS – AI Retinal Image Analysis System, CE – Conformité Européenne (CE-marked), CI – Confidence Interval, CNN – Convolutional Neural Network, CSME – Clinically Significant Macular Edema, CSDMO – Clinically Significant Diabetic Macular Oedema, DESP – Diabetic Eye Screening Programme, DL – Deep Learning, DM – Diabetes Mellitus, DMO – Diabetic Macular Oedema, DR – Diabetic Retinopathy, DRPA – Diabetic Retinopathy Progression Assessment, EDESP – English Diabetic Eye Screening Programme, EMR – Electronic Medical Record, ETDRS – Early Treatment Diabetic Retinopathy Study, FFA – Fundus Fluorescein Angiography, FP – False Positive, FN – False Negative, FPRC –

Study and country	Population	Study design	ARIAS	DR grading	Human comparator	Reference standard	Test accuracy	Prioritised/ Deprioritised
Fundus Photograph Reading Centre, ICDR – International Clinical Diabetic Retinopathy (Severity Scale), ICDRS – International Clinical Diabetic Retinopathy and Diabetic Macular Oedema Severity Scale, mtmDR – More-than-mild Diabetic Retinopathy, NPDR – Non-Proliferative Diabetic Retinopathy, NPV – Negative Predictive Value, OCT – Optical Coherence Tomography, PDR – Proliferative Diabetic Retinopathy, PPV – Positive Predictive Value, QA – Quality Assurance, QWK – Quadratic Weighted Kappa, RCT – Randomized Controlled Trial, rDR – referable Diabetic Retinopathy, ROC – Receiver Operating Characteristic, SE – Sensitivity, SP – Specificity, STDR – Sight-Threatening Diabetic Retinopathy, TN – True Negative, TP – True Positive, UWF – Ultra-Widefield (imaging), UK NSC – United Kingdom National Screening Committee, vtDR – vision-threatening Diabetic Retinopathy, WFPRC – Wisconsin Fundus Photograph Reading Center. † Numbers were calculated by reviewers.								

Table 25. Studies meeting inclusion criterion 11 (Question 2), with prioritisation status and rationale

Study & country	Population	Study design	Intervention	Comparator	Wider clinical impact	Prioritised / deprioritised
Chen 2025[65], USA	Adults (≥18 years) with Type 1 or Type 2 DM, no prior DR diagnosis or dilated eye exam in the preceding 12 months, attending 7 affiliated primary care or endocrinology clinics in Northern California/ San Francisco Bay Area (STATUS program). Numbers not reported.	Uncontrolled before-after study	From 04/2021: ARIAS (IDx-DR, now LumineticsCore®) as single grader, with human secondary grading of ARIAS-ungradable cases (Hybrid)	09/2019 to 03/2021 (18 months): Human-based Teleophthalmology, single retina specialists	% DM patients with annual DES Human (03/2021): 67.2% ARIAS (04/2022): 74.8% ARIAS (04/2023): 78.1%.	Prioritised
Dow 2023b[63], USA	Human workflow: 790 patients screened.AI workflow: 2,243 patients screened, of which 664 screened using AI-human hybrid workflow				% with university follow-up appointment within 90 days ARIAS (04/2021 to 09/2022): 35.5% (99/279) (+ 33.7% [94/279] community FU) Hybrid (04/2021 to 09/2022): 11.7% (12/103). Human (09/2019 to 03/2021): 12.0% (14/117).	Prioritised
Chotcom-wongse 2025[83], Thailand	Prospective alternate-week ARIAS vs manual screening at U-Tai District Hospital with referral tracking at Ayutthaya Hospital; single-field 45° macula-centred images on Topcon TRC-NW400; captured by a mid-level practitioner (>10 years DR screening experience); consecutive patients with diabetes - ARIAS: 708 patients; manual: 746 patients; ARIAS: 297 (46.05%) with mydriasis; human: 355 (48.63%) with mydriasis.	Prospective implementa-tion (alternate week); not RCT	ARDA (Automated Retinal Disease Assessment) (Google) with remote overreading once a week by a retina specialist at Ayutthaya Hospital (alternate weeks). For referral positives, the platform automatically sent referral reminders	Retinal photographs graded by trained personnel (alternate weeks). Patients with positive results either referred to Ayutthaya Hospital on the screening day or appointed for in-person retinal examination by a visiting, general ophthalmologist in the monthly eye clinic at U-Tai Hospital to determine referrals.	Referral positives identified at the point-of-screening: ARIAS 31.2% (201/645) vs Manual 24.0% (175/730) (p=0.003). Referral positives with mydriasis: ARIAS 55/297 (18.5%), Manual 55/355 (15.5%) (p=0.304). Referral positives without mydriasis: ARIAS146/348 (42.0%), Manual 120/375 (32.0%) (p=0.006). Referral adherence at Ayutthaya Hospital for TP: ARIAS 115/129 (89.1%),	Deprioritised country

Study & country	Population	Study design	Intervention	Comparator	Wider clinical impact	Prioritised / deprioritised
			to the patients' phones a few days before the referral date.		Manual 17/22 (77.3%) (p=0.158).	
Heuer 2025[67], USA	Adults (>18 years) with DM at 2 rural primary care practices in Maine (WMPC and FHPC) from December 2019 to June 2023. Number not reported.	Controlled before-after study	WMPC: Switch to ARIAS-based DES (EyeArt, Eyenuk) in January 2021	FHPC: No ARIAS-based DES during study period. All diabetics referred to an eye care centre annually.	% DM patients with annual DES WMPC vs FHPC: Before ARIAS: OR = 0.86 (95%CI 0.74 - 0.99) (p=0.041). After ARIAS switch at WMPC: OR = 2.82 (95%CI 2.42 - 3.27) (p<0.001)	Prioritised
Huang 2024[68], USA	Patients with DM who were managed at primary care sites of Johns Hopkins Medicine in the calendar years 2019 and 2021. 2019 (prior ARIAS switch): ARIAS-switched sites: 5,505 patients; Non-ARIAS sites: 12,169 patients. 2021 (post ARIAS switch): ARIAS-switched sites: 5,580 patients; Non-ARIAS sites: 12,010 patients.	Controlled before-after study	Switch to autonomous ARIAS (IDx-DR, now LumineticsCore®) as single grader in 2020	No use of autonomous ARIAS from 2019 to 2021, no other information on graders and grading pathway.	% DM patients with annual DES Change 2019 to 2021 (propensity score weighting analysis): ARIAS-switched: +7.4% Non-ARIAS: -0.3% Difference ARIAS-switched vs Non-ARIAS sites: +7.6% (p<0.001). Also reports change in adherence rates by gender, age and race.	Prioritised
Li 2021[101], Taiwan	Before–after implementation in endocrinology clinic (Taichung Veterans General Hospital): 716 type 2 diabetes patients (June 2019, pre) vs 700 type 2 DM patients (Oct 2019, post); no mydriasis.	Uncontrolled before-after study	ARIAS-assisted grading (October 2019): VeriSee™ developed by Acer Inc. (New Taipei City, Taiwan). Endocrinologists could read	Manual grading (June 2019): Retinal images graded by one of 5 endocrinologists within 3 days after the examination	Monthly referable DR rate: Before ARIAS: 55.1% (258/468), After ARIAS: 42.9% (216/503). Monthly rate of finishing grading on time (within 3 days): Before ARIAS: 66.8% (478/716), After ARIAS: 77.6% (543/700).	Deprioritised country

Study & country	Population	Study design	Intervention	Comparator	Wider clinical impact	Prioritised / deprioritised
			preliminary VeriSee report before making final decision, either confirming the ARIAS grading or revising ARIAS grading			
Mathenge 2022 [101], Rwanda	Rwanda Artificial Intelligence for Diabetic Retinopathy Screening (RAIDERS) study: 823 adults with DM screened at 4 facilities in Rwanda from March 2021 through June 2021; 275 eligible participants with screening results positive for referral by AI randomised 1:1 to grading by AI with immediate feedback (n = 136) vs grading by human with feedback after 3–5 days (control: n = 139).	Investigator-masked, parallel-group RCT	Grading by Cybersight AI (Orbis International) with immediate feedback	Grading by human graders with communication of need for referral after 3-5 days	Referral adherence within 30 days (unadjusted): ARIAS: 51.5% (70/136) vs Human: 39.6% (55/139) (p=0.048). Adjusted for age, sex, and urban or rural residence: OR = 1.74 (95% CI 1.05 – 2.88) (p=0.031). Time to attend (mean ± SD): ARIAS: 6.6 ±1.7 days vs Human: 9.6 ± 5.1 days (p<0.00001)	Deprioritised country
Pinto 2025[66], Spain	University Hospital of Navarre (HUN), ongoing annual DESP. Patients assigned to HUN and diagnosed with Type 2 DM who underwent DES from 2015 to 2023. Before: 19,828 patients screened; After: 22,962 patients screened.	Uncontrolled before-after study	July 2020 to 2023: Custom, in-house AI tool (NaIA-RD) as level 0 grader. Level 1 grader (GPs) review screening proposal of NaIA-RD before assessing the images. Remaining pathway as in 'Comparator'.	2015 to July 2020: 2-level grading: Level 1: Team of 4 primary care GPs with specific training remotely assess images. Referable DR or ungradable images referred to level 2 grading by ophthalmologist.	Referral rate of level 1 grader: 13.3% decrease from 2018/2019 to 2022/2023. % of patients referred to onsite eye examination: 2018/2019: mean 3.08%. 2022/2023: mean 4.65% (1.5 times increase)	Prioritised

Study & country	Population	Study design	Intervention	Comparator	Wider clinical impact	Prioritised / deprioritised
Rahmati 2025[118], USA & Rwanda	Patients with DM without prior DR diagnosis across various settings: Age range: 5-67 years (including paediatric studies). 20,108 patients with DM: 6,476 with AI-based screening and 13,632 with human grading from 6 studies (5 from USA and 1 from Rwanda)	Systematic review and meta-analysis of 2 RCTs and 4 cohort studies	ARIAS-based DR screening	Human-based DR screening	Pooled meta-analysis Follow-up uptake after screening: All 6 studies: OR =1.89 (95% CI 1.78–2.01, p =0.0001) Adult patients (>18 years, 4 studies, 3,261 participants graded using AI and 5,091 participants graded by human graders): OR = 2.75 (95% CI 1.53–4.93, p =0.0007). Adherence to routine DES (3 studies): OR = 13.99 (95% CI 1.93–101.12, p =0.009).	Deprioritised country
Ruamviboon-suk 2022[121], Thailand	7,651 consecutive DM patients from 9 primary care centres (8 community hospitals in rural settings, 1 primary care clinic in an urban tertiary care hospital) across 3 regions in Thailand, part of the Thai national DESP; mydriasis if needed only.	Prospective cohort study	Google Health deep-learning system – DL based	Overreading by 2 regional retina specialists	Number classed as 'Referral': ARIAS: 602 (7.9%), Humans: 525 (6.9%) Number classed as 'Ungradable': ARIAS: 1,184 (15.5%), Humans: 585 (7.6%). Number classed as 'No Referral': ARIAS 5,865 (76.7%), Humans: 6,541 (85.5%).	Deprioritised country
Yang 2022[137], China	1,001 consecutive adults with DM (type 1 or 2, age ≥18 years) recruited from 3 centres across China (Zhongshan Ophthalmic Center, Peking University People's Hospital, Eye Hospital of Wenzhou Medical University). 2-field images; 2 centres:	Prospective cohort study	AIDRScreening system v1.0 - Shenzhen SiBright Co. Ltd., China.	Comprehensive clinical examination, including assessments of visual acuity (Snellen E Chart) and intraocular pressure, slit-lamp biomicroscopy,	% reduction in diagnostic time: ARIAS 24 ± 8 s/eye vs Ophthalmologist 38 ± 32 s/eye.; Average time gain rate -37.32%.	Deprioritised country

Study & country	Population	Study design	Intervention	Comparator	Wider clinical impact	Prioritised / deprioritised
	dilated images; 1 centre: dilation if needed only.			retinal photography, and anterior chamber gonioscopy as determined by an ophthalmologist with >10 years of clinical experience		

AI, Artificial intelligence; ARIAS, Automated retinal image analysis system; CI, Confidence interval; DES, Diabetic eye screening; DESP, Diabetic eye screening programme; DM, Diabetes mellitus; DR, Diabetic retinopathy; FHPC, Franklin Health Primary Care; FU, Follow-up; GP, General practitioner; HUN, University Hospital of Navarre; NR, Not reported; OR, Odds ratio; SD, Standard deviation; STATUS, Stanford Teleophthalmology Autonomous Testing and Universal Screening; WMPC, Western Maine Primary Care.

Appraisal for quality and risk of bias

Quality assessments of the 4 prioritised studies (reported in 5 articles) for question 2 are reported below.

Table 26. Critical appraisal results based on JBI tool for quasi-experimental studies (4 studies, 5 outcomes).

		Internal validity related to:									Statistical conclusion validity
		Domain	Tem-poral prece-dence	Control group?	Partici-pants similar?	Treatment / care otherwise similar?	Multiple measure-ments pre / post?	Same outcome measure-ment?	Reliable outcome measure-ment?	Losses to FU	
		Question No.	1	2	3	4	5	6	7	8	9
Study	**Outcome**	**Result**									
Chen 2025[65]	Screening rates	03/2021 vs 04/2022	yes	no	unclear	no	no	yes	unclear	unclear	no
		02/2021 vs 04/2023								unclear	no
Dow 2023b[63]	FU ap-pointment adherence	Before vs after ARIAS					no	yes	unclear	unclear	no
Heuer 2025[67]	Screening rates	01/2021 vs 06/2023	yes	yes	unclear	unclear	no	yes	unclear	unclear	no
Huang 2024[68]	Screening rates	2019 vs 2021	yes	yes	yes	yes	no	yes	unclear	unclear	yes
Pinto 2024[66]	% with FU ap-pointment referral	2018/2019 vs 2022/2023	yes	no	unclear	yes	no	yes	unclear	unclear	no
	Referral rate of level 1 grader	2018/2019 vs 2022/2023					no	yes	unclear	unclear	no

FU, Follow-up.

Appendix 4 – Supplementary 'Methods' and 'Results Tables and Figures'

Supplementary Methods

Prioritisation of included studies

Studies for questions 1 and 2 were prioritised. Only the highest priority studies were extracted and data synthesised. A similar approach as used in the 2021 UK NSC review[10] was taken to prioritise studies for extraction.

Question 1

The highest priority study would be a comparative UK-based study that evaluates the performance of commercially available, CE-marked and/or FDA approved ARIAS(s) compared to human primary graders to detect 'Any retinopathy' (R1 or higher and/or M1) in dilated eyes (with mydriasis) using an external, geographically separate study population.

In the absence, or minimal volume, of such studies, we prioritised studies as follows:

- Studies from comparable countries (i.e. North-Western European, USA, Canada, Australia);
- Studies assessing ARIAS with pending CE-mark and/or FDA-approval.
- Studies assessing the performance of ARIAS in undilated eyes;
- Studies reporting 'Referable retinopathy' (R2 or higher and/or M1) only;
- Studies using external, temporal validation.

Question 2

The ideal study would be a systematic review of UK studies. In the absence of a suitable systematic review, primary UK-based RCTs or comparative prospective cohort studies were prioritised next.

In the absence, or minimal volume, of such studies, we prioritised studies as follows:

- Studies from comparable countries (i.e. North-Western European, USA, Canada, Australia);
- Retrospective cohort studies or other study designs.

Data extraction

Data of all prioritised studies were extracted by one reviewer, with a random 20% checked by a second reviewer. Disagreements were resolved by discussion and, if necessary, by involving a third reviewer. No formal data extraction was performed for deprioritised studies but study characteristics and main results are reported in **Appendix 3**, **Table 24** (Question 1) and **Table 25** (Question 2).

Methods for analysis/synthesis

Only the highest priority studies were synthesised. Information on their study design, population, setting, the used ARIAS(s), comparator, reference standard and outcome measures were summarised in text and tables.

For question 1, original data extracted from the studies were used to construct 2x2 tables. The resulting pairs of sensitivities and specificities were plotted on a receiver operator characteristic (ROC) curve for ‘Any DR’ and ‘Referable DR’, respectively.

Given the low number of prioritised studies and substantial heterogeneity in photographic protocols, ARIASs, human graders’ speciality and experiences, reference standards and reported test accuracy outcomes, no meta-analysis was carried out, and findings are summarised narratively, with the results being presented in structured tables and plotted in figures where feasible.

Where data permitted, we additionally presented subgroup data by grade of retinopathy, age group and ethnicity.

For question 2, we summarised study characteristics and findings, grouped by study.

Given the low number of identified studies and heterogeneity in study designs, ARIAS implementation in the pathway and reported outcomes, no meta-analysis was carried out, and findings are summarised narratively, with the results being presented in text and table.

Lack of data prevented the presentation of subgroup data and undertaking subgroup analyses by age group and ethnicity, respectively.

For question 3, no eligible UK-based economic evaluations were identified.

Supplementary Tables and Figures (Question 1)

The paired forest plot in **Figure 13** displays individual study sensitivity and specificity to detect 'Referable DR' (as diamond with 95% CIs lines) as well as the corresponding 2x2 table (if available) for ARIAS and human grading (published data only).

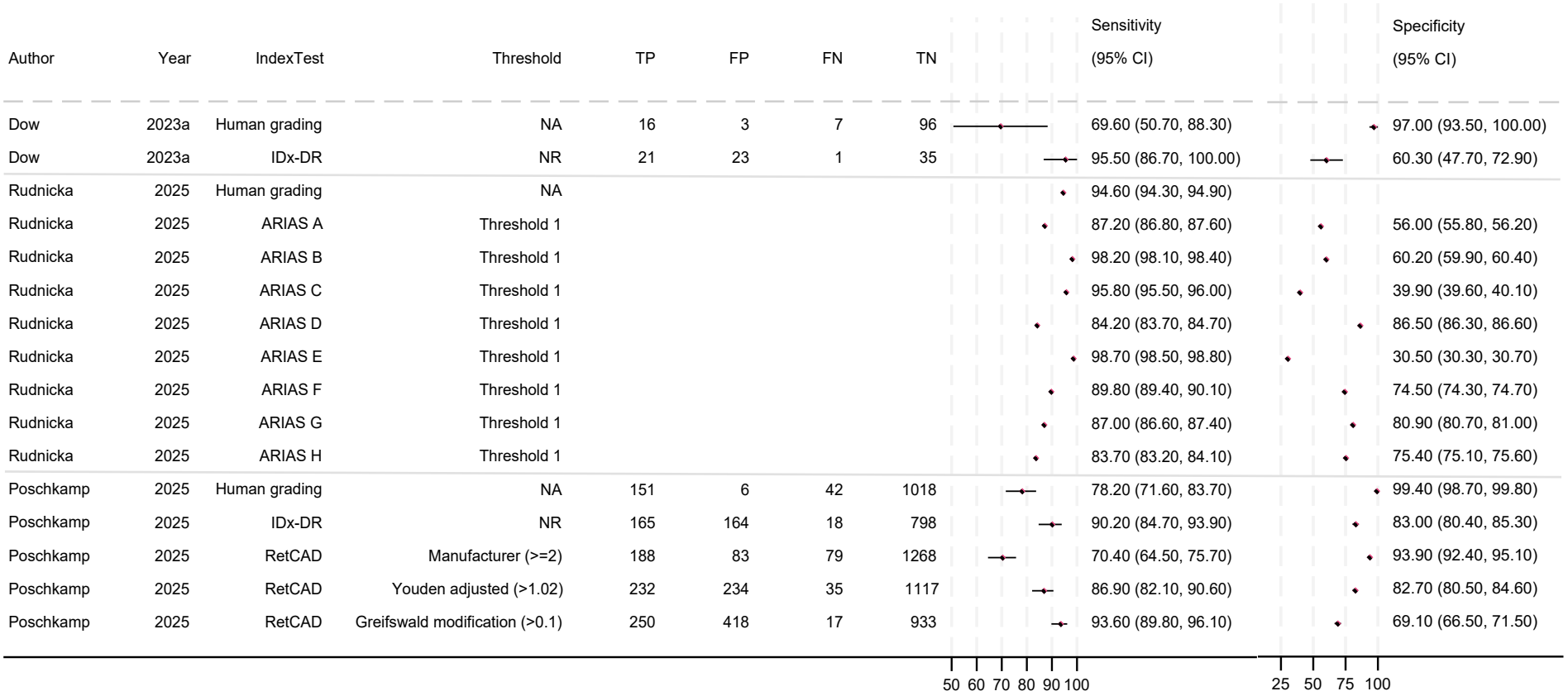


Figure 13. Comparative diagnostic accuracy to detect 'Referable DR' in prioritised studies – ARIAS vs human grading (3 studies)

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; FN, False negative; FP, False positive; NA, Not applicable; NR, Not reported; TN, True negative; TP, True positive.
ARIAS A = DRISTi 2.0; ARIAS B = EyeArt v3.0.0; ARIAS C = EyeCheckup AI; ARIAS D = MONA; ARIAS E = NEC; ARIAS F = OphtAI 2.3; ARIAS G = Remidio; ARIAS H = Retmarker. Threshold 1: ARIAS test positive defined as disease present or ungradable/technical failure.

Table 27. Accuracy to detect referable DR (excluding ungradable images), R2M0 + R2M1 and R3M0 + R3M1, by ethnic group

Index test	Subgroup	Sensitivity, % (95% CI) Referable DR[a]	Sensitivity, % (95% CI) R2M0 + R2M1	Sensitivity, % (95% CI) R3M0 + R3M1	Specificity†, % (95% CI†) R0M0 or R1M0
Primary human grader[b]	Overall	94.6 (94.3 - 94.9)	95.4 (94.7 - 96.1)	95.8 (94.8 - 96.6)	NR
	White	93.5 (92.9 - 94.2)	94.8 (93.3 - 96.0)	94.1 (92.0 - 95.7)	NR
	Black	95.0 (94.4 - 95.6)	95.9 (94.1 - 97.2)	97.6 (95.5 - 98.9)	NR
	South Asian	94.9 (94.4 - 95.3)	95.4 (94.3 - 96.4)	96.1 (94.4 - 97.5)	NR
	Other / missing	95.3 (94.5 - 96.1)	96.2 (94.2 - 97.7)	96.5 (93.4 - 98.4)	NR
DRISTi 2.0, Threshold 1[c]	Overall	87.2 (86.8 - 87.6)	98.0 (97.5 - 98.4)^	98.7 (98.1 - 99.1)^	56.0 (55.8 - 56.2)
	White	91.0 (90.3 - 91.7)	98.8 (98.0 - 99.4)^	99.0 (97.9 - 99.6)^	45.6 (45.2 - 46.0)
	Black	83.9 (82.8 - 84.9)	95.9 (94.1 - 97.2)^	98.1 (96.2 – 99.2)^	62.2 (61.7 - 62.8)
	South Asian	86.4 (85.7 - 87.0)	98.0 (97.2 - 98.7)^	99.0 (97.9 - 99.6)^	61.5 (61.1 - 61.8)
	Other / missing	88.0 (86.8 - 89.1)	98.7 (97.3 - 99.5)^	98.0 (95.5 - 99.4)^	58.2 (57.5 - 58.8)
EyeArt v3.0.0, Threshold 1[c]	Overall	98.2 (98.1 - 98.4)^	99.8 (99.6 - 99.9)^	99.4 (98.9 - 99.7)^	60.2 (59.9 - 60.4)
	White	97.8 (97.4 - 98.1)^	99.9 (99.5 - 100.0)^	98.8 (97.7 - 99.5)^	64.2 (63.8 - 64.6)
	Black	98.5 (98.1 - 98.8)^	99.3 (98.3 - 99.8)^	100.0 (99.0 - 100.0)^	51.9 (51.4 - 52.5)
	South Asian	98.3 (98.0 - 98.5)^	99.9 (99.6 - 100.0)^	99.6 (98.7 - 99.9)^	60.3 (59.9 - 60.6)
	Other / missing	98.6 (98.1 - 99.0)^	99.8 (98.9 - 100.0)^	99.6 (98.7 - 100.0)^	60.0 (59.4 - 60.7)
EyeCheckup AI, Threshold 1[c]	Overall	95.8 (95.5 - 96.0)^	99.6 (99.4 - 99.8)^	97.5 (96.7 - 98.1)^	39.9 (39.6 - 40.1)
	White	96.4 (95.9 - 96.9)^	99.6 (99.1 - 99.9)^	96.6 (94.9 - 97.8)^	37.6 (37.2 - 38.0)
	Black	94.2 (93.5 - 94.9)^	99.0 (97.9 - 99.6)^	97.9 (95.8 - 99.1)^	43.3 (42.8 - 43.9)
	South Asian	96.1 (95.7 - 96.5)^	99.8 (99.4 - 100.0)^	98.4 (97.1 - 99.2)^	39.7 (39.3 - 40.1)
	Other / missing	95.8 (95.0 - 96.5)^	100.0 (99.3 - 100.0)^	96.9 (93.9 - 98.6)^	41.8 (41.1 - 42.4)
MONA, Threshold 1[c]	Overall	82.5 (81.6 - 83.5)	98.4 (98.0 - 98.8)^	96.8 (95.9 - 97.5)^	86.5 (86.3 - 86.6)
	White	84.0 (82.9 - 85.0)	98.3 (97.4 - 99.0)^	95.9 (94.1 - 97.2)^	90.0 (90.3 - 89.8)
	Black	84.6 (83.9 - 85.2)	97.6 (96.2 - 98.6)^	97.9 (95.8 - 99.1)^	83.7 (83.3 - 84.1)
	South Asian	86.6 (85.3 - 87.8)	98.8 (98.1 - 99.3)^	97.3 (95.8 - 98.4)^	84.9 (84.6 - 85.2)
	Other / missing	90.2 (87.2 - 92.7)	98.5 (97.0 - 99.3)^	96.5 (93.4 - 98.4)^	85.5 (85.0 - 86.0)
NEC, Threshold 1[c]	Overall	98.7 (98.5 - 98.8)^	99.8 (99.7 - 99.9)^	99.5 (99.1 - 99.8)^	30.5 (30.3 - 30.7)
	White	99.1 (98.9 - 99.3)^	99.8 (99.3 - 100.0)^	99.4 (98.5 - 99.8)^	29.5 (29.1 - 29.9)
	Black	98.2 (97.8 - 98.5)^	99.7 (98.9 - 100.0)^	100.0 (99.0 - 100.0)^	30.7 (30.1 - 31.2)

Index test	Subgroup	Sensitivity, % (95% CI) Referable DR[a]	Sensitivity, % (95% CI) R2M0 + R2M1	Sensitivity, % (95% CI) R3M0 + R3M1	Specificity†, % (95% CI†) R0M0 or R1M0
	South Asian	98.6 (98.3 - 98.8)^	99.9 (99.5 - 100.0)^	99.6 (98.7 - 99.9)^	31.1 (30.7 - 31.4)
	Other / missing	98.9 (98.4 - 99.2)^	100.0 (99.3 - 100.0)^	99.2 (97.2 - 99.9)^	31.1 (30.5 - 31.7)
OphtAI 2.3, Threshold 1[c]	Overall	89.8 (89.4 - 90.1)	98.5 (98.1 - 98.9)^	95.8 (94.8 - 96.6)^	74.5 (74.3 - 74.7)
	White	92.4 (91.7 - 93.0)^	99.3 (98.6 - 99.7)^	95.9 (94.1 - 97.2)^	74.8 (24.5 - 75.2)
	Black	87.4 (86.4 - 88.3)	96.4 (94.8 - 97.7)^	95.5 (92.8 - 97.3)^	73.3 (72.8 - 73.8)
	South Asian	89.3 (88.7 - 89.8)	98.9 (98.2 - 99.3)^	96.7 (95.1 - 97.9)^	74.5 (74.1 - 74.8)
	Other / missing	90.2 (89.1 - 91.3)	98.5 (97.0 - 99.3)^	93.7 (90.0 - 96.4)^	75.4 (74.8 - 76.0)
Remidio, Threshold 1[c]	Overall	87.0 (86.6 - 87.4)	98.9 (98.5 - 99.2)^	97.1 (96.2 - 97.8)^	80.9 (80.7 - 81.0)
	White	87.8 (86.9 - 88.6)	99.1 (98.3 - 99.6)^	95.7 (93.9 - 97.1)^	82.8 (82.5 - 83.2)
	Black	85.3 (84.3 - 86.3)	97.8 (96.4 - 98.8)^	98.4 (96.5 - 99.4)^	78.2 (77.7 - 78.7)
	South Asian	87.1 (86.5 - 87.7)	99.1 (98.5 - 99.5)^	97.6 (96.2 - 98.6)^	80.3 (80.0 - 80.6)
	Other / missing	87.9 (86.7 - 89.1)	99.4 (98.3 - 99.9)^	97.3 (94.4 - 98.9)^	80.9 (80.4 - 81.4)
Retmarker, Threshold 1[c]	Overall	83.7 (83.2 - 84.1)	96.7 (96.1 - 97.2)^	97.7 (96.9 - 98.3)^	75.4 (75.1 - 75.6)
	White	87.6 (86.7 - 88.4)	98.3 (97.4 - 99.0)^	97.0 (95.5 - 98.2)^	67.5 (67.1 - 67.9)
	Black	80.3 (79.1 - 81.4)	94.4 (92.4 - 96.0)^	98.1 (96.2 - 99.2)^	76.4 (75.9 - 76.9)
	South Asian	82.8 (82.1 - 83.5)	96.5 (95.5 - 97.4)^	98.1 (96.7 - 99.0)^	80.6 (80.3 - 80.9)
	Other / missing	84.7 (83.3 - 85.9)	96.8 (94.9 - 98.1)^	97.6 (94.9 - 99.1)^	78.1 (77.6 - 78.7)

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; DR, Diabetic retinopathy; M0, No observable diabetic maculopathy; M1, Any diabetic maculopathy; NR, Not reported; R0, No observable retinopathy; R1, Mild non-proliferative retinopathy; R2, Moderate-to-severe non-proliferative (pre-proliferative) retinopathy; R3, Proliferative retinopathy.

[a] Referable DR = R2 or higher and/or M1.

[b] Classified as referable diabetic retinopathy by primary human grader.

[c] Threshold 1: ARIAS test positive defined as disease present or ungradable/technical failure.

† ARIAS specificity was calculated by reviewers, i.e. 100% minus the reported false positive rate (%) (for the point estimate and 95% CI, respectively).

^ ARIAS sensitivity same or higher than human grader sensitivity for the particular subgroup.

Table 28. Accuracy to detect referable DR (excluding ungradable images), R2M0 + R2M1 and R3M0 + R3M1, by age group.

Index test	Subgroup	Sensitivity, % (95% CI) Referable DR[a]	Sensitivity, % (95% CI) R2M0 + R2M1	Sensitivity, % (95% CI) R3M0 + R3M1	Specificity†, % (95% CI†) R0M0 or R1M0
Primary human grader[b]	<30 years	95.0 (92.7 - 96.8)	94.7 (88.0 - 98.3)	100.0 (93.3 - 100.0)	NR
	30 to <45 years	95.6 (94.7 - 96.3)	96.4 (94.5 - 97.8)	97.3 (94.6 - 98.9)	NR
	45 to <60 years	95.7 (95.2 - 96.1)	96.3 (95.2 - 97.1)	96.5 (94.8 - 97.8)	NR
	60 to <75 years	94.2 (93.7 - 94.6)	94.6 (93.3 - 95.8)	94.5 (92.6 - 96.0)	NR
	75+ years	92.1 (91.1 - 93.0)	92.7 (89.3 - 95.3)	95.2 (92.0 - 97.4)	NR
DRISTi 2.0, Threshold 1[c]	<30 years	79.4 (75.5 - 82.9)	97.9 (92.5 - 99.7)^	100.0 (93.0 - 100.0)^	72.8 (71.3 - 74.2)
	30 to <45 years	85.9 (84.6 - 87.2)	97.9 (96.3 - 99.0)^	98.9 (96.7 - 99.8)^	74.6 (74.0 - 75.2)
	45 to <60 years	86.8 (86.0 - 87.4)	98.3 (97.5 - 98.9)^	99.2 (98.2 - 99.8)^	66.0 (65.6 - 66.4)
	60 to <75 years	87.8 (87.1 - 88.5)	97.3 (96.3 - 98.1)^	98.4 (97.2 - 99.2)^	50.6 (50.2 - 51.0)
	75+ years	88.9 (87.8 - 90.0)	99.1 (97.3 - 99.8)^	97.8 (95.3 - 99.2)^	33.6 (33.0 - 34.1)
EyeArt v3.0.0, Threshold 1[c]	<30 years	99.2 (97.9 - 99.8)^	100.0 (96.2 - 100.0)^	100.0 (93.3 - 100.0)^	55.4 (53.8 - 57.0)
	30 to <45 years	98.6 (98.0 - 99.0)^	100.0 (99.3 - 100.0)^	99.2 (97.3 - 99.9)^	67.6 (66.9 - 68.3)
	45 to <60 years	98.6 (98.3 - 98.8)^	99.8 (99.5 - 100.0)^	99.7 (98.9 - 100.0)^	67.1 (66.7 - 67.4)
	60 to <75 years	98.0 (97.7 - 98.3)^	99.6 (99.1 - 99.9)^	99.1 (98.1 - 99.6)^	59.4 (59.0 - 59.8)
	75+ years	97.5 (96.9 - 98.0)^	100.0 (98.8 - 100.0)^	99.6 (98.0 - 100.0)^	43.8 (43.2 - 44.4)
EyeCheckup AI, Threshold 1[c]	<30 years	97.3 (95.4 - 98.6)^	98.9 (94.2 - 100.0)^	100.0 (93.3 - 100.0)^	52.4 (50.7 - 54.0)
	30 to <45 years	97.1 (96.4 - 97.7)^	99.8 (99.0 - 100.0)^	98.5 (96.1 - 99.6)^	46.8 (46.1 - 47.5)
	45 to <60 years	96.7 (96.4 - 97.1)^	99.9 (99.6 - 100.0)^	97.9 (96.4 - 98.8)^	42.1 (41.7 - 42.5)
	60 to <75 years	95.0 (94.6 - 95.5)^	99.5 (99.0 - 99.8)^	96.8 (95.2 - 97.9)^	38.6 (38.2 - 39.0)
	75+ years	93.7 (92.8 - 94.5)^	98.7 (96.8 - 99.7)^	97.1 (94.3 - 98.7)^	31.9 (31.4 - 32.4)
MONA, Threshold 1[c]	<30 years	90.2 (87.2 - 92.7)^	97.9 (92.5 - 99.7)^	100.0 (93.3 - 100.0)^	85.6 (84.4 - 86.7)
	30 to <45 years	88.5 (87.2 - 89.6)	99.1 (97.8 - 99.7)^	98.9 (96.7 - 99.8)^	88.6 (88.1 - 89.0)
	45 to <60 years	86.7 (86.0 - 87.4)	98.7 (98.0 - 99.2)^	97.9 (96.4 - 98.8)^	87.7 (87.4 - 87.9)
	60 to <75 years	82.9 (82.1 - 83.7)	98.0 (97.1 - 98.7)^	96.2 (94.6 - 97.5)^	86.3 (86.0 - 86.5)
	75+ years	76.2 (74.7 - 77.7)	97.8 (95.5 - 99.1)^	93.4 (89.7 - 96.0)^	83.1 (82.7 - 83.5)
NEC, Threshold 1[c]	<30 years	99.8 (98.8 - 100.0)^	100.0 (96.2 - 100.0)^	100.0 (93.0 - 100.0)^	16.5 (15.3 - 17.7)
	30 to <45 years	99.0 (98.5 - 99.3)^	100.0 (99.3 - 100.0)^	99.2 (97.3 - 99.9)^	33.7 (33.0 - 34.3)
	45 to <60 years	98.9 (98.7 - 99.1)^	99.8 (99.5 - 100.0)^	99.8 (99.1 - 100.0)^	34.2 (33.8 - 34.6)
	60 to <75 years	98.6 (98.3 - 98.8)^	99.8 (99.3 - 100.0)^	99.5 (98.6 - 99.9)^	30.6 (30.3 - 31.0)

Index test	Subgroup	Sensitivity, % (95% CI) Referable DR[a]	Sensitivity, % (95% CI) R2M0 + R2M1	Sensitivity, % (95% CI) R3M0 + R3M1	Specificity†, % (95% CI†) R0M0 or R1M0
	75+ years	97.9 (97.3 - 98.3)^	100.0 (98.8 - 100.0)^	99.3 (97.4 - 99.9)^	22.5 (22.0 - 23.0)
OphtAI 2.3, Threshold 1[c]	<30 years	97.3 (95.4 - 98.6)^	98.9 (94.2 - 100.0)^	100.0 (93.0 - 100.0)^	75.8 (74.3 - 77.1)
	30 to <45 years	92.6 (91.6 - 93.6)	99.4 (98.4 - 99.9)^	98.1 (95.6 - 99.4)^	79.9 (79.3 - 80.4)
	45 to <60 years	91.4 (90.8 - 92.0)	98.7 (98.0 - 99.2)^	97.5 (96.1 - 98.6)^	78.5 (78.1 - 78.8)
	60 to <75 years	88.5 (87.8 - 89.1)	97.9 (97.0 - 98.6)^	94.2 (92.3 - 95.8)^	73.9 (73.6 - 74.3)
	75+ years	85.2 (83.9 - 86.4)	98.1 (95.9 - 99.3)^	93.0 (89.3 - 95.2)^	64.2 (63.7 - 64.8)
Remidio, Threshold 1[c]	<30 years	92.3 (89.5 - 94.5)^	98.9 (94.2 - 100.0)^	100.0 (93.0 - 100.0)^	79.6 (78.2 - 80.8)
	30 to <45 years	89.7 (88.5 - 90.8)	99.4 (98.4 - 99.9)^	98.5 (96.1 - 99.6)^	86.5 (86.0 - 87.0)
	45 to <60 years	88.4 (87.7 - 89.0)	99.2 (98.6 - 99.6)^	98.5 (96.1 - 99.6)^	84.5 (84.2 - 84.8)
	60 to <75 years	86.3 (85.5 - 87.0)	98.4 (97.5 - 99.0)^	96.5 (94.9 - 97.7)^	80.1 (79.8 - 80.4)
	75+ years	82.2 (80.8 - 83.5)	98.7 (96.8 - 99.7)^	93.4 (89.7 - 96.0)^	71.7 (71.1 - 72.2)
Retmarker, Threshold 1[c]	<30 years	89.8 (86.7 - 92.4)	96.8 (91.0 - 99.3)^	100.0 (93.0 - 100.0)^	64.6 (63.1 - 66.2)
	30 to <45 years	85.7 (84.4 - 87.0)	97.4 (95.6 - 98.6)^	98.9 (96.7 - 99.8)^	82.7 (82.1 - 83.2)
	45 to <60 years	84.6 (83.9 - 85.4)	96.5 (95.5 - 97.4)^	98.0 (96.6 - 98.9)^	82.1 (81.8 - 82.4)
	60 to <75 years	82.4 (81.6 - 83.2)	96.2 (95.0 - 97.1)^	97.4 (96.0 - 98.4)^	74.8 (74.5 - 75.2)
	75+ years	81.9 (80.5 - 83.2)	98.7 (96.8 - 99.7)^	96.0 (92.9 - 98.0)^	59.6 (59.0 - 60.2)

ARIAS, Automated Retinal Image Analysis System; CI, Confidence interval; DR, Diabetic retinopathy; M0, No observable diabetic maculopathy; M1, Any diabetic maculopathy; NR, Not reported; R0, No observable retinopathy; R1, Mild non-proliferative retinopathy; R2, Moderate-to-severe non-proliferative (pre-proliferative) retinopathy; R3, Proliferative retinopathy.

[a] Referable DR = R2 or higher and/or M1.

[b] Classified as referable diabetic retinopathy by primary human grader.

[c] Threshold 1: ARIAS test positive defined as disease present or ungradable/technical failure.

† Specificity was calculated by reviewers, i.e. 100% minus the reported false positive rate (%) (for the point estimate and 95% CI, respectively).

^ ARIAS sensitivity same or higher than human grader sensitivity for the particular subgroup.

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 29**.

Table 29. UK NSC reporting checklist for evidence summaries

Section	Item	Page no.
Title and summaries		
Title Sheet	Identify the review as a UK NSC Evidence summary	Title page
Plain English summary	Plain English description of the executive summary.	5-6
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	7-15
Introduction and Approach		
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 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. Method – briefly outline the rapid review methods used.	16-23 23-24 25
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 a priori	25-27
Appraisal for quality/ risk of bias tool	Details of tool/ checklist used to assess quality, e.g. QUADAS 2, CASP, SIGN, AMSTAR.	27
Search strategy and study selection		
Databases/ sources searched	Give details of all databases searched (including platform/ interface and coverage dates) and date of final search.	28

Section	Item	Page no.
Search strategy and results	Present the full search strategy for at least one database (usually a version of Medline), including limits and search filters if used. 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.	80-89 90
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.	25-27, 91
Study level reporting of results (for each key question)		
Study level reporting, results and risk of bias assessment	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.). Provide a simple summary of key measures, effect estimates and confidence intervals for each study where available. For each study, present the results of any assessment of quality/risk of bias.	Q1: 34-35, 112-144 Q2: 68-70, 145-149 Q1: 38 Q2: 150
Question level synthesis		
Description of the evidence	For each question, give numbers of studies screened, assessed for eligibility, and inclusion in the review, with summary reasons for exclusion	Q1: 32, 92-111 Q2: 66, 92-111 Q3: 75, 92-111
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 compartments should inform the reviewer's judgement on whether the criterion is "met", "not met" or "uncertain": quantity; quality; applicability and consistency.	Q1: 60-62 Q2: 71-73 Q3: 75

Section	Item	Page no.
Summary of findings	Provide a description of the evidence reviewed and included for each question, with reference to their eligibility for inclusion. Summarise the main findings including the quality/ risk of bias issues for each question. Have the criteria addressed been “met”, “not met” or "uncertain”?	Q1: 63 Q2: 73 Q3: 75
Review Summary		
Conclusions and implications for policy	Do findings indicate whether screening should be recommended? IS further work warranted? Are there gaps in the evidence highlighted by the review?	76-77
Limitations	Discuss limitations of the available evidence and of the review methodology if relevant.	78-79

Appendix 6 – Patient and Public Involvement and Engagement (PPIE) (reporting based on GRIPP2† guidelines)

1: Aim of the PPIE

(1) The primary aim was to embed patient and public involvement in the rapid review process, where PPIE contributors and researchers work towards a common goal of ensuring that the research is robust, accountable, relevant and appropriate to end-users. Grounded in the principle 'nothing about us without us', our PPIE group largely consisted of people with lived experience of diabetes and diabetic eye screening.
(2) A second aim was to assess the feasibility of doing PPIE within a rapid review format and within the context of UK NSC processes and timelines. Rapid PPIE is an emerging field that has direct relevance to policy-related work, such as evidence syntheses, which are often conducted within restricted timeframes. We evaluated the process of conducting rapid PPIE, including the perspectives of the PPIE contributors, PPIE facilitator(s), and the review team.

2: Methods

Recruitment:

To facilitate timeliness, recruitment was targeted through known contacts and PPIE networks. Three contributors were recruited through Breakthrough T1D. Informal interviews were conducted with all potential group members to assess suitability, determine any conflict of interest and answer any questions about the project and PPIE expectations. We aimed for diversity in age, gender, ethnicity, socioeconomic background, and experience of diabetes and diabetic eye screening. Eight contributors were recruited.

The group was ethnically diverse and consisted of 4 men and 4 women, with ages ranging from 25 to 76 years. Participants were located in England (north, south, midlands) and Wales. Five had diabetes themselves (types 1 and 2 represented), 2 had close family members with diabetes, and one member of the public had no personal connection to diabetes.

Meetings:

The group met online on 4 occasions between September and November 2025. The meetings covered:

(1) Introduction to the project (including the topic and concept of rapid reviews), introduction to each other and the research team. Discussion of objectives, purpose, research plan, ways of working and expectations of the research team and PPIE group.
(2) Discussion around searching/sifting strategies and data extraction. Emerging findings and interpretations for Questions 2 & 3.

(3) Discussion around integration, clinical impact and practical implications with focus on results for Question 1 (test accuracy).
(4) Summary discussion, including details of reporting and highlighting key messages.

The PPIE group commented on the plain English summary (communication by email) between meetings (3) and (4).

A further meeting (or email correspondence) will provide details of any feedback from the UK NSC and/or stakeholder consultation.

Meetings took a deliberative knowledge space approach, where technical forms of knowledge are integrated through tailoring or adapting information so that it can be explored in an accessible context to enable meaningful PPI contribution. The review team and PPIE facilitator provided the context for the discussion at the start of meetings through short presentations. Comments, questions and discussion from the PPIE group were welcomed and encouraged throughout. The meetings were recorded so that group members who were unable to join meetings were able to catch-up on the recording and provide their feedback.

The level of PPIE involvement varied with the review stage: the group contributed their views and feedback to data analysis and integration discussions, and they directly influenced the communication/dissemination process through the plain English language summary.

Participation was compensated at NIHR rates for involvement. Other costs that might preclude involvement, such as carers/childcare and any additional accessibility costs were also reimbursed as needed.

Evaluation:

Post-meeting surveys were used to evaluate the process. These served to:

(1) help evaluate how to do rapid PPIE and whether it is possible to do effective rapid PPIE in this context;
(2) check in with contributors to ensure they found the experience interesting and worthwhile, and/or whether they had any additional support needs to improve their experience;
(3) give them a further opportunity to provide comments on meeting discussions;
(4) help summarise what information contributors saw as the ‘take-home message’ from meetings and what they considered to be the most important (or surprising) results.

The perspectives of the review team were also gathered, to determine where and how PPIE impacted the review process, what worked well, and where there are areas for further development. Reflections on the PPIE process will also be sought from the UK NSC.

3: Study results

All participants fully engaged with and participated in the process; they freely shared their views and experiences and openly listened to, and respected, the views of others. Participants responded that taking part was a positive experience.

Despite a quick turn-around time, 6 members of the group were able to provide feedback and edits to the plain English summary, improving its accessibility and suitability for a public audience.

Summary of discussion points from the PPIE group:

The group were concerned about the comparability of results in non-UK studies to a UK setting, given the unique economic dynamics of the NHS and the socio-demographic diversity of the UK. It was felt more discussion at outset would have been helpful on what constitutes a country being similar enough to the UK.

The group highlighted the variation between ARIAS (automated retinal imaging analysis software) in a large UK study and, also, that only 8 of the ARIAS vendors invited took part in the study: "*some are very close to human graders and others are not quite there. So what the future looks like, what we would recommend in the future, will very much depend on which vendor you pick*".

Given the lack of UK data (beyond the one large London-based study presented to them) they felt that more information was needed on how ARIAS perform in a real-world situation – specifically in the UK - when compared with human graders. They would also have liked to know more about how the different ARIAS were trained and in which populations.

The group could see the potential for ARIAS, especially given the performance of some of the ARIAS tested in the UK study, but they thought the public would want reassurances that the introduction of this kind of technology was not done primarily for cost-saving (and would not lower their level of care). *"They [the public] are not going to be thinking about the cost effectiveness. I know the policy makers and commissioners will be thinking about cost effectiveness. From patients' point of view is they want to know what is my likelihood of being missed or not. So I think we just need- how we communicate that message is important, but it's clear that actually whichever technology we go with, whichever ARIAS we go with, they are comparable, really*."

Clarity of messaging to the public was re-emphasised as critically important, including being clear on the capabilities of any ARIAS chosen and the risk of false negatives or false positives; particularly as some members of the public may view AI as being less prone to making mistakes "*… if we start saying it's AI, then somehow people expect it to be automation 100% risk free. It's like going on a plane and you know… you're boarding at London Heathrow and you'll get to Glasgow and there won't be an air crash. There is an expectation that it is going to be risk free. So it's how we communicate that message that even with the technology things can go wrong. But we minimise that and this is how we minimise that*".

The proposal to check 10% of negative screening results was welcomed. "*They have to be audited properly, monitored properly ... because at the end of the day it's trust. If you damage the trust with patients and the public, that would be a huge problem".*

Views differed on whether it would be preferable to use ARIAS as an extra level (level 0, prior to the first grader) or replacing the first grader. Some group members were reassured by the high performance levels of some ARIAS in the London study in terms of sensitivity, and the checks that would be put in place. Based on this, they thought that replacing the first grader would be the better option, because of the potential for cost-savings. Others felt that adding an extra level was a safer approach, especially while there are questions on performance in wider settings and between ARIAS.

In summary, PPIE contributors came to similar conclusions to the review team and largely (independently) concurred with the overall conclusions and recommendations. The consensus of the PPIE group was that although the use of ARIAS showed great promise, there were areas where more information is needed.

"*At the moment, based on the data and based on the published results, I really cannot see what endorsement we can offer to the NSC, whether it should be done or not, or it shouldn't be done, other than to say we need more data and we need more time. I think that's the only conclusion based on what has been shared and what we've discussed today*."

The group felt that larger scale trials of ARIAS within the existing NHS screening programme would be one option for consideration. Any implementation (including at pilot level) should include a robust framework for data collection and continuous learning to improve outcomes and ensure patient safety.

Other areas where more information would be beneficial included:

- Understanding the current uptake of diabetic eye screening and the potential impact of socio-economic, geographical, or community-level effects, or people with complex/multiple health conditions on uptake and implementation of ARIAS.
- Other socio-economic or geographical impacts on ARIAS screening outcomes beyond ethnicity – and more in-depth analysis of ethnicity impacts as it was felt these were at a very broad level and there was no clear justification given for doing this.
- Studies on cost effectiveness of implementing ARIAS across socio-demographically different regions.
- More information on the system of regulatory oversight for the use of AI—how does it ensure safety, efficacy, and accountability?
- More exploration into the idea that ARIAS could result in a higher attendance rate for diabetic eye screening or recalls (following a positive screening result) – this needs primary research accounting for (and detailing) the methods for contacting and consenting patients as this will also impact uptake.
- More information on screening outcomes (sensitivity/specificity/acceptability) of using ARIAS in Scotland.

4: Discussion and conclusions

As the PPIE contributors largely agreed with the conclusions of the review team their input was valuable for sense-checking. They also directly contributed to drafting the plain English summary.

The review team also used the opportunity of meeting with the PPIE group to explore particular points from the literature. For example, whether the potential for ARIAS to speed up appointment invites would change uptake behaviour of people invited to screening or invited to follow-up tests following a positive screening test. Given that 5 members of the group had diabetes, they were able to discuss this from their own perspectives. Also, following discussion with representatives from the UK NSC, we asked the group how they felt about adding ARIAS to the screening programme as either a level 1 or level 0 screen, and whether they would feel comfortable (given the information they now had) to have ARIAS as part of their own diabetic eye screening.

Being a rapid review, the PPIE process was condensed into a short space of time. We conducted 4 meetings over approximately 7 weeks. Logistically this presented some challenges, such as:

(1) the point at which the PPIE group were able to begin working with the team. This was after protocol sign-off and the majority of the sifting and final text selection had been completed. Initial discussions around the inclusion criteria and prioritisation of papers suggested that earlier integration of PPIE may have been beneficial. However, the PPIE group did not think that this was a significant issue as they felt they were able to meaningfully contribute from the point at which we began our discussions. The content and level of information provided in the introductory meeting was seen as essential to explaining the purpose of the study and enabling the group to contribute from the start.
(2) PPIE work can involve significant time commitment from contributors, this can be particularly impactful where they have other commitments (including caring roles and/or employment). The impact of this is somewhat mitigated in studies conducted over a longer period, where meetings can be spaced out and planned well in advance. Managing expectations of the group and clear communication were important in this respect. Meetings were recorded so that group members unable to join were able to catch-up on discussions and provide their feedback and perspectives by email, evaluation form, or in a catch-up meeting. The group showed a great deal of flexibility and dedication to the project, which was a significant factor in the success of the PPIE. Several members of the group (including some who have experience of being involved in PPIE projects that span a longer time period) found that the frequent meetings were helpful in keeping the topic fresh in their minds.
(3) Time for preparing documents ahead of meetings was significantly impacted, partly because the work was rapidly evolving and often it was not practical to send information more than one or 2 days ahead of meetings. This also limited the feasibility of spending time to ensure documents and presentations were at a plain English level. The review team were mindful of this, and they provided clear explanations during meetings to help

with accessibility. The PPIE group were encouraged throughout to ask questions to clarify anything that was unclear. The evaluation forms sent after every meeting also gave the group another opportunity to comment on the clarity of the information provided.

(4) Impact on the team. PPIE can be a labour-intensive activity for researchers and PPIE facilitators. From the outset it was decided that the PPIE should not impact on the timeline for the rapid review. As the review team were critical to providing the content for discussion in each meeting, they needed to commit time in their closely planned schedule to prepare for, and attend, the 4 meetings. To minimise impact on the review team, the PPIE work was managed by an experienced PPIE facilitator, who could focus on building and managing relationships with the PPIE contributors, as well as monitoring, evaluating, and supporting the process. Beyond the benefits of PPIE to the research, this study provided valuable insights into the practical implementation of PPIE in this rapid format.

5: Reflections / Critical perspectives

Time pressures were heightened in this rapid PPIE, such as the time commitment of contributors and the research team, shortened time for recruitment and training contributors, and reduced time for document preparation to facilitate creating a deliberative knowledge space. It was not possible to include PPIE perspectives into the early stages of the review.

However, the PPIE group adapted well to the rapid PPIE format and found it to be a rewarding and worthwhile experience. Several of the group were new to PPIE and this experience will hopefully serve as encouragement for them to continue to be involved in research, which is important for capacity-building in PPIE.

Despite the challenges, the group was able to meaningfully contribute. They largely agreed with the conclusions of the review team. They also provided a useful source of information relating to particular issues, based on their lived experience.

Having the involvement of an experienced PPIE facilitator was beneficial, particularly given the timeframe, as they had networks to contact for recruitment, and knowledge of the processes and the points where contributors may need more support. Gathering feedback using an evaluation form throughout the process was vital to assess the success of the process, and any points for adjustment, in reflection with the review team. We explored different techniques for managing some of the time constraints, such as trialling using AI to assist with summarising some of the information presented to the group. Future work with PPIE contributors will continue to refine processes and seek efficiencies to maximise effective support and communication within time constraints.

In summary, the PPIE process was successful from the perspectives of both of our initial aims, by (1) enabling a process whereby contributors could interrogate the review methods, data analysis and conclusions to ensure a robust end-user relevant review product; and (2) by providing valuable information on (and experience of) integrating PPIE within a rapid review for the UK NSC.

† Staniszewska, S., Brett, J., Simera, I. *et al.* GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *Res Involv Engagem* **3**, 13 (2017). https://doi.org/10.1186/s40900-017-0062-2

References

1. Diabetes UK. *About diabetes*: Diabetes UK. URL: https://www.diabetes.org.uk/about-diabetes?gad_campaignid=15219795550&gad_source=1
2. World Health Organization. *Diabetes*: WHO; 2024. URL, https://www.who.int/news-room/fact-sheets/detail/diabetes
3. International Diabetes Federation. *United Kingdom - Diabetes country report 2000 — 2050*. International Diabetes Federation; 2025. https://diabetesatlas.org/data-by-location/country/united-kingdom/
4. Royal National Institute of Blind People. *Understanding diabetes related eye conditions including Diabetic Retinopathy*. RNIB; 2023. https://share.google/BQ2Obp7uL54vMsi3o
5. NHS. *Diabetic retinopathy*; 2021. URL: https://www.nhs.uk/conditions/diabetic-retinopathy/stages/#:~:text=Stage%201%3A%20background%20retinopathy,common%20in%20people%20with%20diabetes.
6. Haider S, Thayakaran R, Subramanian A, Toulis KA, Moore D, Price MJ*, et al.* Disease burden of diabetes, diabetic retinopathy and their future projections in the UK: cross-sectional analyses of a primary care database. BMJ Open 2021;11:e050058. https://doi.org/10.1136/bmjopen-2021-050058
7. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999-2000 with 2009-2010. BMJ Open 2014;4:e004015. https://doi.org/10.1136/bmjopen-2013-004015
8. Amoaku WM, Ghanchi F, Bailey C, Banerjee S, Banerjee S, Downey L*, et al.* Diabetic retinopathy and diabetic macular oedema pathways and management: UK Consensus Working Group. Eye 2020;34:1-51. https://doi.org/10.1038/s41433-020-0961-6
9. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. Acta Diabetol 2017;54:515-25. https://doi.org/10.1007/s00592-017-0974-1
10. Zhelev Z, Peters J, Rogers M, Allen M, Lowe J, Kijauskaite G*, et al. Automated grading in the Diabetic Eye Screening Programme - External review against programme appraisal criteria for the UK National Screening Committee*: UK National Screening Committee; 2021. URL: https://assets.publishing.service.gov.uk/media/619e80088fa8f5037e8ccb2f/Evidence_summary_AI_in_DESP_2021.pdf
11. York and Scarborough Teaching Hospitals - NHS Foundation Trust. *Understanding national grading results and outcomes - Screening for diabetic retinopathy*. 2025. https://www.yorkhospitals.nhs.uk/our-services/organdonation/a-z-of-services/diabetic-eye-screening/understanding-national-grading-results-and-outcomes/
12. Incubator for AI and Digital Healthcare. *Professor Alicja Rudnicka: Are AI-enabled systems in healthcare fit for purpose? Best practice for evaluating AI for screening prior to live implementation*. 2025. https://www.youtube.com/watch?v=KQYl81tcABs. https://www.youtube.com/watch?v=KQYl81tcABs

13. NHS England. *NHS England Diabetic eye screening: Patient grading referral surveillance*: NHS England; 2024. URL, https://assets.publishing.service.gov.uk/media/668423b903469096e184059d/20230725_Diabetic_eye_screening_pathways_-_patient_grading_referral_surveillance_.pdf#:~:text=Screening%20is%20offered%20once%20every%2024%20months%20for,positive%2C%20non-referrable%20result%20or%20a%20single%20negative%20result.

14. Macdonald T, Zhelev Z, Liu X, Hyde C, Fajtl J, Egan C*, et al.* Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee. Lancet Digit Health 2025;7:100840. https://doi.org/10.1016/j.landig.2024.12.004

15. Fleming AD, Mellor J, McGurnaghan SJ, Blackbourn LAK, Goatman KA, Styles C*, et al.* Deep learning detection of diabetic retinopathy in Scotland's diabetic eye screening programme. British Journal of Ophthalmology 2024;108:984-8. https://doi.org/doi:https://dx.doi.org/10.1136/bjo-2023-323395

16. Office for Health Improvement & Disparity. *Corporate report - NHS screening programmes in England: 2020 to 2021*: GOV.UK. URL, https://www.gov.uk/government/publications/nhs-screening-programmes-annual-report/nhs-screening-programmes-in-england-2020-to-2021#nhs-diabetic-eye-screening-des-programme

17. Public Health England. *Diabetes Prevalence Model*; 2016. URL, https://assets.publishing.service.gov.uk/media/5a82c07340f0b6230269c82d/Diabetesprevalencemodelbriefing.pdf

18. HM Government. *National AI Strategy*. London: Office for Artificial Intelligence; 2021. URL: https://www.gov.uk/government/publications/national-ai-strategy

19. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018;286:800-9. https://doi.org/10.1148/radiol.2017171920

20. 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;34:172.

21. 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 https://doi.org/10.1159/000368426

22. 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;19:72. https://doi.org/10.1007/s11892-019-1189-3

23. Macdonald T, Dinnes J, Maniatopoulos G, Solebo A, Hogg H, Alderman J*, et al.* *Development of a Target Product Profile for an Artificial Intelligence for Use in English Diabetic Eye Screening, a Modified Delphi Consensus Study*; 2025. https://doi.org/10.2139/ssrn.5293401

24. Public Health England. *Guidance - UK NSC: evidence review process*. URL, https://www.gov.uk/government/publications/uk-nsc-evidence-review-process/uk-nsc-evidence-review-process

25. Centre for Reviews and Dissemination. *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews*. York (UK): Centre for Reviews and Dissemination; 1995-2025. https://www.crd.york.ac.uk/CRDWeb/. https://www.crd.york.ac.uk/CRDWeb/

26. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB*, et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine 2011;155:529-36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

27. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA*, et al.* A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Transl Vis Sci Technol 2020;9:7. https://doi.org/10.1167/tvst.9.2.7

28. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF*, et al.* QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Annals of Internal Medicine 2021;174:1592-9. https://doi.org/10.7326/M21-2234

29. Barker TH, Habibi N, Aromataris E, Stone JC, Leonardi-Bee J, Sears K*, et al.* The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. JBI Evidence Synthesis 2024;22, https://journals.lww.com/jbisrir/fulltext/2024/03000/the_revised_jbi_critical_appraisal_tool_for_the.4.aspx

30. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J*, et al.* AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008. https://doi.org/10.1136/bmj.j4008

31. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I*, et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. https://doi.org/10.1136/bmj.l4898

32. Joanna Briggs Institute. *Checklist for Cohort Studies - Critical Appraisal tools for use in JBI Systematic Reviews*: JBI; 2020. URL, https://jbi.global/sites/default/files/2021-10/Checklist_for_Cohort_Studies.docx

33. Joanna Briggs Institute. *Checklist for Economic Evaluations*: JBI; 2020. URL, https://jbi.global/sites/default/files/2021-10/Checklist_for_Economic_Evaluations.docx

34. Ong AY, Taribagil P, Sevgi M, Kale AU, Dow ER, Macdonald T*, et al.* A scoping review of artificial intelligence as a medical device for ophthalmic image analysis in Europe, Australia and America. NPJ Digit Med 2025;8:323. https://doi.org/10.1038/s41746-025-01726-8

35. Alqahtani AS, Alshareef WM, Aljadani HT, Hawsawi WO, Shaheen MH. The efficacy of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis. International Journal of Retina and Vitreous 2025;11:48. https://doi.org/10.1186/s40942-025-00670-9

36. Bi Z, Li J, Liu Q, Fang Z. Deep learning-based optical coherence tomography and retinal images for detection of diabetic retinopathy: a systematic and meta analysis. Frontiers in Endocrinology 2025;16:1485311. https://doi.org/https://dx.doi.org/10.3389/fendo.2025.1485311

37. Ciapponi A, Ballivian J, Gentile C, Mejia JR, Ruiz-Baena J, Bardach A. Diagnostic utility of artificial intelligence software through non-mydriatic digital retinography in the screening of diabetic retinopathy: an overview of reviews. Eye 2025;39:2083-9. https://doi.org/10.1038/s41433-025-03809-y

38. Farooq MS, Arooj A, Alroobaea R, Baqasah AM, Jabarulla MY, Singh D*, et al.* Untangling Computer-Aided Diagnostic System for Screening Diabetic Retinopathy Based on Deep Learning Techniques. Sensors 2022;22 https://doi.org/10.3390/s22051803

39. Joseph S, Selvaraj J, Mani I, Kumaragurupari T, Shang X, Mudgil P*, et al.* Diagnostic Accuracy of Artificial Intelligence-Based Automated Diabetic Retinopathy Screening in Real-World Settings: A Systematic Review and Meta-Analysis. American Journal of Ophthalmology 2024;263:214-30. https://doi.org/10.1016/j.ajo.2024.02.012

40. Khan Z, Gaidhane AM, Singh M, Ganesan S, Kaur M, Sharma GC*, et al.* Diagnostic Accuracy of IDX-DR for Detecting Diabetic Retinopathy: A Systematic Review and Meta-Analysis. American Journal of Ophthalmology 2025;273:192-204. https://doi.org/10.1016/j.ajo.2025.02.022

41. Lakshminarayanan V, Kheradfallah H, Sarkar A, Jothi Balaji J. Automated Detection and Diagnosis of Diabetic Retinopathy: A Comprehensive Survey. Journal of Imaging 2021;7 https://doi.org/10.3390/jimaging7090165

42. Moannaei M, Jadidian F, Doustmohammadi T, Kiapasha AM, Bayani R, Rahmani M*, et al.* Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: a meta-analysis. BioMedical Engineering OnLine 2025;24:34. https://doi.org/10.1186/s12938-025-01336-1

43. Parrey MUR, Bhatti MOA, Abdul-Latif MM, Rehman S, Ismail MM, Hamid OA. Meta-analysis of AI algorithm performance in detecting retinal diseases. Anaesthesia, Pain & Intensive Care 2025;29:400-6.

44. Pietris J, Lam A, Bacchi S, Gupta AK, Kovoor JG, Chan WO. Health Economic Implications of Artificial Intelligence Implementation for Ophthalmology in Australia: A Systematic Review. Asia-Pacific Journal of Ophthalmology 2022;11:554-62. https://doi.org/https://doi.org/10.1097/APO.0000000000000565

45. Ramoutar RR. An Economic Analysis for the Use of Artificial Intelligence in Screening for Diabetic Retinopathy in Trinidad and Tobago. Cureus 2024;16:e55745. https://doi.org/10.7759/cureus.55745

46. Sacchini F, Mancin S, Cangelosi G, Palomares SM, Caggianelli G, Gravante F*, et al.* The role of artificial intelligence in diabetic retinopathy screening in type 1 diabetes: A systematic review. Journal of Diabetes and its Complications 2025;39:109139. https://doi.org/https://doi.org/10.1016/j.jdiacomp.2025.109139

47. Tahir HN, Ullah N, Tahir M, Domnic IS, Prabhakar R, Meerasa SS*, et al.* Artificial intelligence versus manual screening for the detection of diabetic retinopathy: a comparative systematic review and meta-analysis. Frontiers in Medicine 2025;Volume 12 - 2025 https://doi.org/10.3389/fmed.2025.1519768

48. Uy H, Fielding C, Hohlfeld A, Ochodo E, Opare A, Mukonda E*, et al.* Diagnostic test accuracy of artificial intelligence in screening for referable diabetic retinopathy in real-world settings: A systematic review and meta-analysis. PLOS Global Public Health 2023;3:e0002160. https://doi.org/10.1371/journal.pgph.0002160

49. Wang Z, Li Z, Li K, Mu S, Zhou X, Di Y. Performance of artificial intelligence in diabetic retinopathy screening: a systematic review and meta-analysis of prospective studies. Frontiers in endocrinology 2023;14:1197783.

50. Wu H, Jin K, Yip CC, Koh V, Ye J. A systematic review of economic evaluation of artificial intelligence-based screening for eye diseases: From possibility to reality. Survey of Ophthalmology 2024;69:499-507. https://doi.org/10.1016/j.survophthal.2024.03.008

51. Zhelev Z, Peters J, Rogers M, Allen M, Kijauskaite G, Seedat F*, et al.* Test accuracy of artificial intelligence-based grading of fundus images in diabetic retinopathy screening: A systematic review. Journal of Medical Screening 2023;30:97-112. https://doi.org/10.1177/09691413221144382

52. Kabunga R, Asasira J, Njuki S, Daniel A, Morley K, Morley M*, et al.* Describing the Performance and the Infrastructure Requirements of the Existing Artificial Intelligence (AI)-Based Diabetic Retinopathy (DR) Screening Algorithms for Diabetic Patients: an Umbrella Review. Journal of medical systems 2025;49:149. https://doi.org/https://dx.doi.org/10.1007/s10916-025-02280-2

53. Teng CW, Patel SD, Barkmeier AJ, Liu TYA, Myung D, Henderer J*, et al.* Autonomous Artificial Intelligence in Diabetic Retinopathy Testing-Lessons Learned on Successful Health System Adoption. Ophthalmology science 2026;6:100935. https://doi.org/https://dx.doi.org/10.1016/j.xops.2025.100935

54. Wang T-W, Luo W-T, Tu Y-K, Chou Y-B, Wu Y-T. Diagnostic Accuracy of EyeArt for Fundus-Based Detection of Diabetic Retinopathy: A Systematic Review and Meta-analysis. American journal of ophthalmology 2025;281:465-79. https://doi.org/https://dx.doi.org/10.1016/j.ajo.2025.09.045

55. Yudistira Y, Hendrawan KA, Andayani A, Suryathi NMA, Ernawati T, Gunawan ACV*, et al.* Diabetic Retinopathy Screening Approaches in Developing Countries: A Systematic Review and Meta-Analysis. Turkish journal of ophthalmology 2025;55:260-75. https://doi.org/https://dx.doi.org/10.4274/tjo.galenos.2025.32916

56. Ran AR, Ding JL, Tang Z, Lam C, Nguyen TX, Zhou J*, et al.* Real-World Prospective Validation and Economic Evaluation of Deep Learning-Based Diabetic Retinopathy Detection From Fundus Photographs: A Systematic Review and Meta-analysis. Diabetes care 2025 https://doi.org/https://dx.doi.org/10.2337/dc25-1493

57. Dow ER, Khan NC, Chen KM, Mishra K, Perera C, Narala R*, et al.* AI-Human Hybrid Workflow Enhances Teleophthalmology for the Detection of Diabetic Retinopathy. Ophthalmology Science 2023;3:100330. https://doi.org/doi:https://dx.doi.org/10.1016/j.xops.2023.100330

58. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE*, et al.* Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. Diabetes care 2021;44:1168-75. https://doi.org/doi:https://dx.doi.org/10.2337/dc20-1877

59. Nissen TPH, Norgaard TL, Schielke KC, Vestergaard P, Nikontovic A, Dawidowicz M*, et al.* Performance of a Support Vector Machine Learning Tool for Diagnosing Diabetic Retinopathy in Clinical Practice. Journal of personalized medicine 2023;13 https://doi.org/doi:https://dx.doi.org/10.3390/jpm13071128

60. Sin M, Zeniskova R, Sliva M, Dvorak K, Valkova J, Bayer J*, et al.* Comparison of the Aireen System with Telemedicine Evaluation by an Ophthalmologist - A Real-World Study. Clinical Ophthalmology 2025;19:957 EP - 64. https://doi.org/doi:https://dx.doi.org/10.2147/OPTH.S511233 PT - Article

61. Rudnicka AR, Shakespeare R, Chambers R, Bolter L, Anderson J, Fajtl J*, et al.* Automated retinal image analysis systems to triage for grading of diabetic retinopathy: a large-scale, open-label, national screening programme in England. The Lancet Digital health 2025:100914. https://doi.org/https://dx.doi.org/10.1016/j.landig.2025.100914

62. Poschkamp B, Kantz L, Augstein P, Tayar A, Kaderali L, Busch M*, et al.* Customizing AI-based screening with real-world data: Practical insights from diabetic retinopathy. Acta ophthalmologica 2025 https://doi.org/https://dx.doi.org/10.1111/aos.17591

63. Dow ER, Chen KM, Zhao CS, Knapp AN, Phadke A, Weng K*, et al.* Artificial Intelligence Improves Patient Follow-Up in a Diabetic Retinopathy Screening Program. Clinical Ophthalmology (Auckland, NZ) 2023;17:3323-30. https://doi.org/doi:https://dx.doi.org/10.2147/OPTH.S422513

64. Public Health England. *The management of grading quality: Good practice in the quality assurance of grading. Version 1.2*; 2016. URL, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/512832/The_Management_of_Grading.pdf

65. Chen KM, Zhao CS, Knapp A, Dow E, Phadke A, Tan M*, et al.* Targeted Interventions Lead to Quality Improvement in Year 2 of an Artificial Intelligence-Based Diabetic Retinopathy Detection Program in Northern California. Retina (Philadelphia, Pa) 2025 https://doi.org/doi:https://dx.doi.org/10.1097/IAE.0000000000004499

66. Pinto I, Olazaran A, Jurio D, De la Osa B, Sainz M, Oscoz A*, et al.* Improving diabetic retinopathy screening using artificial intelligence: design, evaluation and before-and-after study of a custom development. Frontiers in digital health 2025;7:1547045. https://doi.org/doi:https://dx.doi.org/10.3389/fdgth.2025.1547045

67. Heuer R, DayBranch E, Tsomides A, DiBiase JF, Korsen N, Nolan B. AI Diabetic Retinopathy Screening in a Primary Care Setting in Rural Maine. Journal of General Internal Medicine 2025;40:2465 EP - 7. https://doi.org/https://dx.doi.org/10.1007/s11606-024-09319-z

68. Huang JJ, Channa R, Wolf RM, Dong Y, Liang M, Wang J*, et al.* Autonomous artificial intelligence for diabetic eye disease increases access and health equity in underserved populations. npj Digital Medicine 2024;7:196. https://doi.org/doi:https://dx.doi.org/10.1038/s41746-024-01197-3 PT - Article

69. Hailey DM. Health technology assessment in Canada: diversity and evolution. Medical Journal of Australia 2007;187:286-8. https://doi.org/https://doi.org/10.5694/j.1326-5377.2007.tb01245.x

70. Watt A, Cameron A, Sturm L, Lathlean T, Babidge W, Blamey S*, et al.* Rapid reviews versus full systematic reviews: An inventory of current methods and practice in health technology assessment: Corrigendum. International Journal of Technology Assessment in Health Care 2008;24:369-. https://doi.org/10.1017/S0266462308080483

71. Taylor-Phillips S, Geppert J, Stinton C, Freeman K, Johnson S, Fraser H*, et al.* Comparison of a full systematic review versus rapid review approaches to assess a

newborn screening test for tyrosinemia type 1. Research Synthesis Methods 2017;8:475-84. https://doi.org/https://doi.org/10.1002/jrsm.1255

72. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): Protocol for a Qualitative Study. JMIR Res Protoc 2024;13:e58202. https://doi.org/10.2196/58202

73. Tomlinson E, Yang B, Davenport CF, Rutjes AWS, Mallett S, Whiting PF. Piloting QUADAS-3: a revised tool for the quality assessment of diagnostic accuracy studies. Journal of Clinical Epidemiology 2025;188:111983. https://doi.org/https://doi.org/10.1016/j.jclinepi.2025.111983

74. Abreu-Gonzalez R, Susanna-Gonzalez G, Blair JPM, Lasagni Vitar RM, Ciller C, Apostolopoulos S*, et al.* Validation of artificial intelligence algorithm LuxIA for screening of diabetic retinopathy from a single 45degree retinal colour fundus images: the CARDS study. BMJ Open Ophthalmology 2025;10 https://doi.org/doi:https://dx.doi.org/10.1136/bmjophth-2024-002109

75. Al-Turk L, Wawrzynski J, Wang S, Krause P, Saleh GM, Alsawadi H*, et al.* Automated feature-based grading and progression analysis of diabetic retinopathy. Eye 2022;36:524-32. https://doi.org/doi:https://dx.doi.org/10.1038/s41433-021-01415-2

76. Antaki F, Hammana I, Tessier M-C, Boucher A, David Jette ML, Beauchemin C*, et al.* Implementation of Artificial Intelligence-Based Diabetic Retinopathy Screening in a Tertiary Care Hospital in Quebec: Prospective Validation Study. JMIR Diabetes 2024;9:e59867. https://doi.org/doi:https://dx.doi.org/10.2196/59867

77. Arenas-Cavalli JT, Abarca I, Rojas-Contreras M, Bernuy F, Donoso R. Clinical validation of an artificial intelligence-based diabetic retinopathy screening tool for a national health system. Eye 2022;36:78-85. https://doi.org/doi:https://dx.doi.org/10.1038/s41433-020-01366-0

78. Baget-Bernaldiz M, Fontoba-Poveda B, Romero-Aroca P, Navarro-Gil R, Hernando-Comerma A, Bautista-Perez A*, et al.* Artificial Intelligence-Based Screening System for Diabetic Retinopathy in Primary Care. Diagnostics 2024;14 https://doi.org/doi:https://dx.doi.org/10.3390/diagnostics14171992

79. Baget-Bernaldiz M, Pedro R-A, Santos-Blanco E, Navarro-Gil R, Valls A, Moreno A*, et al.* Testing a Deep Learning Algorithm for Detection of Diabetic Retinopathy in a Spanish Diabetic Population and with MESSIDOR Database. Diagnostics 2021;11 https://doi.org/doi:https://dx.doi.org/10.3390/diagnostics11081385

80. Brant A, Singh P, Yin X, Yang L, Nayar J, Jeji D*, et al.* Performance of a Deep Learning Diabetic Retinopathy Algorithm in India. JAMA Network Open 2025;8:e250984. https://doi.org/doi:https://dx.doi.org/10.1001/jamanetworkopen.2025.0984

81. Burlina S, Radin S, Poggiato M, Cioccoloni D, Raimondo D, Romanello G*, et al.* Screening for diabetic retinopathy with artificial intelligence: a real world evaluation. Acta Diabetologica 2024;61:1603-7. https://doi.org/doi:https://dx.doi.org/10.1007/s00592-024-02333-x

82. Chia MA, Hersch F, Sayres R, Bavishi P, Tiwari R, Keane PA*, et al.* Validation of a deep learning system for the detection of diabetic retinopathy in indigenous Australians. British Journal of Ophthalmology 2024;108:268-73. https://doi.org/doi:https://dx.doi.org/10.1136/bjo-2022-322237

83. Chotcomwongse P, Ruamviboonsuk P, Karavapitayakul C, Thongthong K, Amornpetchsathaporn A, Chainakul M*, et al.* Transforming Non-Digital, Clinical Workflows to Detect and Track Vision-Threatening Diabetic Retinopathy via a Digital Platform Integrating Artificial Intelligence: Implementation Research. Ophthalmology and Therapy 2025;14:447-60. https://doi.org/doi:https://dx.doi.org/10.1007/s40123-024-01086-8

84. Curran K, Whitestone N, Zabeen B, Ahmed M, Husain L, Alauddin M*, et al.* CHILDSTAR: CHIldren Living With Diabetes See and Thrive with AI Review. Clinical Medicine Insights Endocrinology and Diabetes 2023;16:11795514231203867. https://doi.org/doi:https://dx.doi.org/10.1177/11795514231203867

85. Dai L, Wu L, Li H, Cai C, Wu Q, Kong H*, et al.* A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications 2021;12:3242. https://doi.org/doi:https://dx.doi.org/10.1038/s41467-021-23458-5

86. Dogan ME, Bilgin AB, Sari R, Bulut M, Akar Y, Aydemir M. Head to head comparison of diagnostic performance of three non-mydriatic cameras for diabetic retinopathy screening with artificial intelligence. Eye (London, England) 2024;38:1694-701. https://doi.org/doi:https://dx.doi.org/10.1038/s41433-024-03000-9

87. Dong X, Du S, Zheng W, Cai C, Liu H, Zou J. Evaluation of an Artificial Intelligence System for the Detection of Diabetic Retinopathy in Chinese Community Healthcare Centers. Frontiers in Medicine 2022;9:883462. https://doi.org/doi:https://dx.doi.org/10.3389/fmed.2022.883462

88. Duggal M, Chauhan A, Gupta V, Kankaria A, Budhija D, Verma P*, et al.* Real-World Evaluation of AI-Driven Diabetic Retinopathy Screening in Public Health Settings: Validation and Implementation Study. JMIR medical informatics 2025;13:e67529. https://doi.org/https://dx.doi.org/10.2196/67529

89. Grzybowski A, Brona P. Analysis and Comparison of Two Artificial Intelligence Diabetic Retinopathy Screening Algorithms in a Pilot Study: IDx-DR and Retinalyze. Journal of Clinical Medicine 2021;10 https://doi.org/doi:https://dx.doi.org/10.3390/jcm10112352

90. Grzybowski A, Jin K, Zhou J, Pan X, Wang M, Ye J*, et al.* Retina Fundus Photograph-Based Artificial Intelligence Algorithms in Medicine: A Systematic Review. Ophthalmology and Therapy 2024;13:2125-49. https://doi.org/https://dx.doi.org/10.1007/s40123-024-00981-4

91. Grzybowski A, Brona P, Krzywicki T, Ruamviboonsuk P. Diagnostic Accuracy of Automated Diabetic Retinopathy Image Assessment Software: IDx-DR and RetCAD. Ophthalmology and Therapy 2025;14:73-84. https://doi.org/doi:https://dx.doi.org/10.1007/s40123-024-01049-z

92. Grzybowski A, Rao DP, Brona P, Negiloni K, Krzywicki T, Savoy FM. Diagnostic Accuracy of Automated Diabetic Retinopathy Image Assessment Softwares: IDx-DR and Medios Artificial Intelligence. Ophthalmic Research 2023;66:1286-92. https://doi.org/doi:https://dx.doi.org/10.1159/000534098

93. Guedes IIG, Saavedra Santana P, Lopez FC, Macias AR, de Miguel AR, Hernandez AG. Evaluation of the degree of agreement in the diagnosis of diabetic retinopathy between ophthalmologists and EyeArt R. International journal of retina and vitreous 2025;11:125. https://doi.org/https://dx.doi.org/10.1186/s40942-025-00748-4

94. Hao S, Liu C, Li N, Wu Y, Li D, Gao Q*, et al.* Clinical evaluation of AI-assisted screening for diabetic retinopathy in rural areas of midwest China. PloS One 2022;17:e0275983. https://doi.org/doi:https://dx.doi.org/10.1371/journal.pone.0275983

95. Hao Z, Xu R, Huang X, Ren X, Li H, Shao H. Application and observation of artificial intelligence in clinical practice of fundus screening for diabetic retinopathy with non-mydriatic fundus photography: a retrospective observational study of T2DM patients in Tianjin, China. Therapeutic Advances in Chronic Disease 2022;13:20406223221097335. https://doi.org/doi:https://dx.doi.org/10.1177/20406223221097335

96. Hu C-L, Wang Y-C, Wu W-F, Xi Y. Evaluation of AI-enhanced non-mydriatic fundus photography for diabetic retinopathy screening. Photodiagnosis and Photodynamic Therapy 2024;49:104331. https://doi.org/doi:https://dx.doi.org/10.1016/j.pdpdt.2024.104331

97. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD*, et al.* Pivotal Evaluation of an Artificial Intelligence System for Autonomous Detection of Referrable and Vision-Threatening Diabetic Retinopathy. JAMA network open 2021;4:e2134254. https://doi.org/doi:https://dx.doi.org/10.1001/jamanetworkopen.2021.34254

98. Karabeg M, Petrovski G, Hertzberg SNW, Erke MG, Fosmark DS, Russell G*, et al.* A pilot cost-analysis study comparing AI-based EyeArt and ophthalmologist assessment of diabetic retinopathy in minority women in Oslo, Norway. International Journal of Retina and Vitreous 2024;10:40. https://doi.org/doi:https://dx.doi.org/10.1186/s40942-024-00547-3

99. Karabeg M, Petrovski G, Holen K, Steffensen Sauesund E, Fosmark DS, Russell G*, et al.* Comparison of Validity and Reliability of Manual Consensus Grading vs. Automated AI Grading for Diabetic Retinopathy Screening in Oslo, Norway: A Cross-Sectional Pilot Study. Journal of clinical medicine 2025;14 https://doi.org/doi:https://dx.doi.org/10.3390/jcm14134810

100. Li N, Ma M, Lai M, Gu L, Kang M, Wang Z*, et al.* A stratified analysis of a deep learning algorithm in the diagnosis of diabetic retinopathy in a real-world study. Journal of diabetes 2022;14:111-20. https://doi.org/doi:https://dx.doi.org/10.1111/1753-0407.13241

101. Li Y-H, Sheu WH-H, Chou C-C, Lin C-H, Cheng Y-S, Wang C-Y*, et al.* The Clinical Influence after Implementation of Convolutional Neural Network-Based Software for Diabetic Retinopathy Detection in the Primary Care Setting. Life (Basel, Switzerland) 2021;11 https://doi.org/doi:https://dx.doi.org/10.3390/life11030200

102. Lim JI, Regillo CD, Sadda SR, Ipp E, Bhaskaranand M, Ramachandra C*, et al.* Artificial Intelligence Detection of Diabetic Retinopathy: Subgroup Comparison of the EyeArt System with Ophthalmologists' Dilated Examinations. Ophthalmology science 2023;3:100228. https://doi.org/doi:https://dx.doi.org/10.1016/j.xops.2022.100228

103. Limwattanayingyong J, Nganthavee V, Seresirikachorn K, Singalavanija T, Soonthornworasiri N, Ruamviboonsuk V*, et al.* Longitudinal Screening for Diabetic Retinopathy in a Nationwide Screening Program: Comparing Deep Learning and Human Graders. Journal of diabetes research 2020;2020:8839376. https://doi.org/doi:https://dx.doi.org/10.1155/2020/8839376

104. Mathenge W, Whitestone N, Nkurikiye J, Patnaik JL, Piyasena P, Uwaliraye P*, et al.* Impact of Artificial Intelligence Assessment of Diabetic Retinopathy on Referral Service

Uptake in a Low-Resource Setting: The RAIDERS Randomized Trial. Ophthalmology science 2022;2:100168. https://doi.org/doi:https://dx.doi.org/10.1016/j.xops.2022.100168

105. Mbaye S, Aw A, Sy EHM, Ka AM, Diagne JP, Diallo HM*, et al.* Assessment of artificial intelligence software for automatic screening of diabetic retinopathy based on fundus photographs in melanoderm subjects. Retina (Philadelphia, Pa) 2025;45:330-4. https://doi.org/doi:https://dx.doi.org/10.1097/IAE.0000000000004310

106. Mehra AA, Softing A, Guner MK, Hodge DO, Barkmeier AJ. Diabetic Retinopathy Telemedicine Outcomes With Artificial Intelligence-Based Image Analysis, Reflex Dilation, and Image Overread. American journal of ophthalmology 2022;244:125-32. https://doi.org/doi:https://dx.doi.org/10.1016/j.ajo.2022.08.008

107. Meredith S, van Grinsven M, Engelberts J, Clarke D, Prior V, Vodrey J*, et al.* Performance of an artificial intelligence automated system for diabetic eye screening in a large English population. Diabetic medicine : a journal of the British Diabetic Association 2023;40:e15055. https://doi.org/doi:https://dx.doi.org/10.1111/dme.15055

108. Ming S, Xie K, Lei X, Yang Y, Zhao Z, Li S*, et al.* Evaluation of a novel artificial intelligence-based screening system for diabetic retinopathy in community of China: a real-world study. International ophthalmology 2021;41:1291-9. https://doi.org/doi:https://dx.doi.org/10.1007/s10792-020-01685-x

109. Mokhashi N, Grachevskaya J, Cheng L, Yu D, Lu X, Zhang Y*, et al.* A Comparison of Artificial Intelligence and Human Diabetic Retinal Image Interpretation in an Urban Health System. Journal of diabetes science and technology 2022;16:1003-7. https://doi.org/doi:https://dx.doi.org/10.1177/1932296821999370

110. Musetti D, Cutolo CA, Bonetto M, Giacomini M, Maggi D, Viviani GL*, et al.* Autonomous artificial intelligence versus teleophthalmology for diabetic retinopathy. European journal of ophthalmology 2025;35:232-8. https://doi.org/doi:https://dx.doi.org/10.1177/11206721241248856

111. Noriega A, Meizner D, Camacho D, Enciso J, Quiroz-Mercado H, Morales-Canton V*, et al.* Screening Diabetic Retinopathy Using an Automated Retinal Image Analysis System in Independent and Assistive Use Cases in Mexico: Randomized Controlled Trial. JMIR formative research 2021;5:e25290. https://doi.org/doi:https://dx.doi.org/10.2196/25290

112. Parravano M, Constanzo E, De Geronimo D, Varano M, Rossi T, Rui CE*, et al.* DRSplus retinographer in patients with diabetes: a comparison study by consensus grading, Reading Center and Artificial Intelligence. Retina (Philadelphia, Pa) 2025 https://doi.org/https://dx.doi.org/10.1097/IAE.0000000000004734

113. Peeters F, Rommes S, Elen B, Gerrits N, Stalmans I, Jacob J*, et al.* Artificial Intelligence Software for Diabetic Eye Screening: Diagnostic Performance and Impact of Stratification. Journal of clinical medicine 2023;12 https://doi.org/doi:https://dx.doi.org/10.3390/jcm12041408

114. Pei X, Yao X, Yang Y, Zhang H, Xia M, Huang R*, et al.* Efficacy of artificial intelligence-based screening for diabetic retinopathy in type 2 diabetes mellitus patients. Diabetes research and clinical practice 2022;184:109190. https://doi.org/doi:https://dx.doi.org/10.1016/j.diabres.2022.109190

115. Peris-Martinez C, Shaha A, Clarida W, Amelon R, Hernaez-Ortega MC, Navea A*, et al.* Use in clinical practice of an automated screening method of diabetic retinopathy that can

be derived using a diagnostic artificial intelligence system. Uso en la practica clinica, de un metodo de cribado automatizado de retinopatia diabetica derivable mediante un sistema de inteligencia artificial de diagnostico 2021;96:117-26. https://doi.org/doi:https://dx.doi.org/10.1016/j.oftal.2020.08.007

116. Piatti A, Romeo F, Manti R, Doglio M, Tartaglino B, Nada E, *et al.* Feasibility and accuracy of the screening for diabetic retinopathy using a fundus camera and an artificial intelligence pre-evaluation application. Acta diabetologica 2024;61:63-8. https://doi.org/doi:https://dx.doi.org/10.1007/s00592-023-02172-2

117. Piatti A, Rui C, Gazzina S, Tartaglino B, Romeo F, Manti R, *et al.* Diabetic retinopathy screening with confocal fundus camera and artificial intelligence - assisted grading. European journal of ophthalmology 2025;35:679-88. https://doi.org/doi:https://dx.doi.org/10.1177/11206721241272229

118. Rahmati M, Smith L, Piyasena MP, Bowen M, Boyer L, Fond G, *et al.* Artificial Intelligence improves follow-up appointment uptake for diabetic retinal assessment : a systematic review and meta-analysis. Eye (London, England) 2025;39:2398-406. https://doi.org/https://dx.doi.org/10.1038/s41433-025-03849-4

119. Rao DP, Sindal MD, Sengupta S, Baskaran P, Venkatesh R, Sivaraman A, *et al.* Towards a Device Agnostic AI for Diabetic Retinopathy Screening: An External Validation Study. Clinical ophthalmology (Auckland, NZ) 2022;16:2659-67. https://doi.org/doi:https://dx.doi.org/10.2147/OPTH.S369675

120. Riotto E, Gasser S, Potic J, Sherif M, Stappler T, Schlingemann R, *et al.* Accuracy of Autonomous Artificial Intelligence-Based Diabetic Retinopathy Screening in Real-Life Clinical Practice. Journal of clinical medicine 2024;13 https://doi.org/doi:https://dx.doi.org/10.3390/jcm13164776

121. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, *et al.* Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. The Lancet Digital health 2022;4:e235-e44. https://doi.org/doi:https://dx.doi.org/10.1016/S2589-7500(22)00017-6

122. Ruan S, Liu Y, Hu W-T, Jia H-X, Wang S-S, Song M-L, *et al.* A new handheld fundus camera combined with visual artificial intelligence facilitates diabetic retinopathy screening. International journal of ophthalmology 2022;15:620-7. https://doi.org/doi:https://dx.doi.org/10.18240/ijo.2022.04.16

123. Sarao V, Veritti D, Lanzetta P. Automated diabetic retinopathy detection with two different retinal imaging devices using artificial intelligence: a comparison study. Graefe's archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie 2020;258:2647-54. https://doi.org/doi:https://dx.doi.org/10.1007/s00417-020-04853-y

124. Scheetz J, Koca D, McGuinness M, Holloway E, Tan Z, Zhu Z, *et al.* Real-world artificial intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and indigenous healthcare settings in Australia. Scientific reports 2021;11:15808. https://doi.org/doi:https://dx.doi.org/10.1038/s41598-021-94178-5

125. Sedova A, Hajdu D, Datlinger F, Steiner I, Neschi M, Aschauer J, *et al.* Comparison of early diabetic retinopathy staging in asymptomatic patients between autonomous AI-based

screening and human-graded ultra-widefield colour fundus images. Eye (London, England) 2022;36:510-6. https://doi.org/doi:https://dx.doi.org/10.1038/s41433-021-01912-4

126. Skevas C, de Olaguer NP, Lleo A, Thiwa D, Schroeter U, Lopes IV, *et al.* Implementing and evaluating a fully functional AI-enabled model for chronic eye disease screening in a real clinical environment. BMC ophthalmology 2024;24:51. https://doi.org/doi:https://dx.doi.org/10.1186/s12886-024-03306-y

127. Tan-Torres A, 3rd, Praveen PA, Jeji D, Brant A, Yin X, Yang L, *et al.* Validation of a Deep Learning Model for Diabetic Retinopathy on Patients with Young-Onset Diabetes. Ophthalmology and Therapy 2025;14:1147-55. https://doi.org/doi:https://dx.doi.org/10.1007/s40123-025-01116-z

128. Teoh CS, Wong KH, Xiao D, Wong HC, Zhao P, Chan HW, *et al.* Variability in Grading Diabetic Retinopathy Using Retinal Photography and Its Comparison with an Automated Deep Learning Diabetic Retinopathy Screening Software. Healthcare (Basel, Switzerland) 2023;11 https://doi.org/doi:https://dx.doi.org/10.3390/healthcare11121697

129. Tsai M-J, Hsieh Y-T, Tsai C-H, Chen M, Hsieh A-T, Tsai C-W, *et al.* Cross-Camera External Validation for Artificial Intelligence Software in Diagnosis of Diabetic Retinopathy. Journal of Diabetes Research 2022;2022:5779276. https://doi.org/doi:https://dx.doi.org/10.1155/2022/5779276

130. Vaghefi E, Yang S, Xie L, Han D, Yap A, Schmeidel O, *et al.* A multi-centre prospective evaluation of THEIATM to detect diabetic retinopathy (DR) and diabetic macular oedema (DMO) in the New Zealand screening program. Eye (Basingstoke) 2023;37:1683 EP - 9. https://doi.org/doi:https://dx.doi.org/10.1038/s41433-022-02217-w

131. Van TN, Thi HLV. Effectiveness of artificial intelligence for diabetic retinopathy screening in community in Binh Dinh Province, Vietnam. Taiwan Journal of Ophthalmology 2024;14:394-402. https://doi.org/doi:https://dx.doi.org/10.4103/tjo.TJO-D-23-00101

132. Wang Y, Han X, Li C, Luo L, Yin Q, Zhang J, *et al.* Impact of Gold-Standard Label Errors on Evaluating Performance of Deep Learning Models in Diabetic Retinopathy Screening: Nationwide Real-World Validation Study. Journal of Medical Internet Research 2024;26:e52506. https://doi.org/doi:https://dx.doi.org/10.2196/52506

133. Whitestone N, Nkurikiye J, Patnaik JL, Jaccard N, Lanouette G, Cherwek DH, *et al.* Feasibility and acceptance of artificial intelligence-based diabetic retinopathy screening in Rwanda. The British Journal of Ophthalmology 2024;108:840-5. https://doi.org/doi:https://dx.doi.org/10.1136/bjo-2022-322683

134. Wintergerst MWM, Bejan V, Hartmann V, Schnorrenberg M, Bleckwenn M, Weckbecker K, *et al.* Telemedical Diabetic Retinopathy Screening in a Primary Care Setting: Quality of Retinal Photographs and Accuracy of Automated Image Analysis. Ophthalmic Epidemiology 2022;29:286-95. https://doi.org/doi:https://dx.doi.org/10.1080/09286586.2021.1939886

135. Wongchaisuwat N, Trinavarat A, Rodanant N, Thoongsuwan S, Phasukkijwatana N, Prakhunhungsit S, *et al.* In-Person Verification of Deep Learning Algorithm for Diabetic Retinopathy Screening Using Different Techniques Across Fundus Image Devices. Translational Vision Science & Technology 2021;10:17. https://doi.org/doi:https://dx.doi.org/10.1167/tvst.10.13.17

136. Xu X, Zhang J, Song X, Liu X, Liu Y, Feng L*, et al.* Enhancing diabetic retinopathy diagnosis and grading: a retrospective study on AI-assisted decision making and cost analysis. The British journal of ophthalmology 2025 https://doi.org/https://dx.doi.org/10.1136/bjo-2025-327442

137. Yang Y, Pan J, Yuan M, Lai K, Xie H, Ma L*, et al.* Performance of the AIDRScreening system in detecting diabetic retinopathy in the fundus photographs of Chinese patients: a prospective, multicenter, clinical study. Annals of Translational Medicine 2022;10:1088. https://doi.org/doi:https://dx.doi.org/10.21037/atm-22-350

138. Yao L, Cao C-Y, Yu G-X, Shu X-P, Fan X-N, Zhang Y-F. Screening and evaluation of diabetic retinopathy via a deep learning network model: A prospective study. World Journal of Diabetes 2024;15:2302-10. https://doi.org/doi:https://dx.doi.org/10.4239/wjd.v15.i12.2302

139. Zhang W-F, Li D-H, Wei Q-J, Ding D-Y, Meng L-H, Wang Y-L*, et al.* The Validation of Deep Learning-Based Grading Model for Diabetic Retinopathy. Frontiers in Medicine 2022;9:839088. https://doi.org/doi:https://dx.doi.org/10.3389/fmed.2022.839088

140. Zhang Y, Shi J, Peng Y, Zhao Z, Zheng Q, Wang Z*, et al.* Artificial intelligence-enabled screening for diabetic retinopathy: a real-world, multicenter and prospective study. BMJ Open Diabetes Research & Care 2020;8 https://doi.org/doi:https://dx.doi.org/10.1136/bmjdrc-2020-001596