

Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee

Item author(s)	Macdonald, Trystan;Zhelev, Zhivko;Liu, Xiaoxuan;Hyde, Christopher;Fajtl, Jiri;Egan, Catherine;Tufail, Adnan;Rudnicka, Alicja;Shinkins, Bethany;Given-Wilson, Rosalind;Dunbar, Kevin;Halligan, Steve;Scanlon, Peter;Mackie, Anne;Taylor-Philips, Sian;Denniston, Alastair;Scanlon, Peter H
Item title	Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee
Citation	Macdonald, T., Zhelev, Z., Liu, X., Hyde, C., Fajtl, J., Egan, C., Tufail, A., Rudnicka, A. R., Shinkins, B., Given-Wilson, R., Dunbar, J. K., Halligan, S., Scanlon, P., Mackie, A., Taylor-Philips, S., & Denniston, A. K. (2025). Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee. <i>The Lancet. Digital health</i> , 7(5), 100840. https://doi.org/10.1016/j.landig.2024.12.004
Link to published version	10.1016/j.landig.2024.12.004
Item License	https://creativecommons.org/licenses/by/4.0/
Date archived on GERR	2025-07-21T11:22:23Z

Generating evidence to support the role of AI in diabetic eye screening: considerations from the UK National Screening Committee



Trystan Macdonald, Zhivko Zhelev, Xiaoxuan Liu, Christopher Hyde, Jiri Fajtl, Catherine Egan, Adnan Tufail, Alicja R Rudnicka, Bethany Shinkins, Rosalind Given-Wilson, J Kevin Dunbar, Steve Halligan, Peter Scanlon, Anne Mackie, Sian Taylor-Philips*, Alastair K Denniston*



Screening for diabetic retinopathy has been shown to reduce the risk of sight loss in people with diabetes, because of early detection and treatment of sight-threatening disease. There is long-standing interest in the possibility of automating parts of this process through artificial intelligence, commonly known as automated retinal imaging analysis software (ARIAS). A number of such products are now on the market. In the UK, Scotland has used a rules-based autograder since 2011, but the diabetic eye screening programmes in the rest of the UK rely solely on human graders. With more sophisticated machine learning-based ARIAS now available and greater challenges in terms of human grader capacity, in 2019 the UK's National Screening Committee (NSC) was asked to consider the modification of diabetic eye screening in England with ARIAS. Following up on a review of ARIAS research highlighting the strengths and limitations of existing evidence, the NSC here sets out their considerations for evaluating evidence to support the introduction of ARIAS into the diabetic eye screening programme.

Introduction

Diabetic retinopathy is a common complication of diabetes¹ and can lead to substantial sight loss.² Diabetes affects the eye through damage to the retinal blood vessels, causing tissue swelling, ischaemia, and the growth of new blood vessels or proliferative retinopathy. Visual loss is due to maculopathy, swelling of the area of retina responsible for fine vision, or complications of proliferative disease including retinal detachment and glaucoma. Patients with advanced diabetic retinopathy are known to have poorer quality of life³⁻⁵ and reduced levels of physical, emotional, and social wellbeing, and to require a high amount of health-care resources.^{6,7}

In the UK, all people aged 12 years and older with types 1 or 2 diabetes are invited to attend annual or biennial diabetic eye screening, as part of a national screening programme. The aim of the programme is to detect retinopathy in its early and asymptomatic stage, facilitating timely diagnosis and intervention before progression to sight-threatening complications when prognosis deteriorates and treatment costs increase. There has been a decrease in sight loss from diabetic retinopathy since the introduction of diabetic eye screening in 2003, particularly in those of working age.^{8,9} The screening programme is, however, resource intensive,⁶ and its costs are expected to rise with projected increases in diabetes prevalence.¹⁰

Artificial intelligence (AI) describes the use of computers to do tasks normally requiring human intelligence. The automated grading of diabetic eye screening photos using automated retinal imaging analysis software (ARIAS) is a particularly promising use-case for AI, and indeed Scottish diabetic eye screening has used an ARIAS for over a decade.^{11,12} This ARIAS belongs to a subset of AI tools called symbolic AI in which the rules by which it should classify an image are manually programmed by humans. The recent

acceleration in AI-enabled diagnostic tests has largely been due to another branch of AI known as machine learning, where software learns patterns from data itself. The advent of machine learning has seen the development of machine learning-based ARIAS (ML-ARIAS) that show better performance than earlier non-ML-ARIAS. This Health Policy review will address only ML-ARIAS as they have unique considerations that must be taken into account in their evaluation and implementation compared with non-ML-ARIAS, namely their requirements for large amounts of data to train, reduced explainability, and potential for adaptation over time. Some ARIAS can use both symbolic and machine learning AI; this Health Policy review will include these as ML-ARIAS owing to these unique considerations.

When considering the introduction of AI to a screening service there is a need to ensure that any change is evidence-based and can be safely implemented. A health technology assessment performed by Tufail and colleagues¹³ suggested an initial approach could be to use ML-ARIAS to screen patients before human grading or to replace the primary human grader (figure 1), both of which are cost saving.¹³ In 2019, a proposal was made to the UK's National Screening Committee (NSC) to consider adopting ML-ARIAS in these roles in England and a review of ARIAS research was commissioned to consider this.¹⁴ The review,¹⁴ published in 2021, concluded that there was sufficient evidence for certain ML-ARIAS to be considered safe and better value for money than current manual grading. However, it recommended further assessment on the social and ethical aspects of ML-ARIAS use.

We seek to outline the UK NSC approach to the critical appraisal of ML-ARIAS evidence, specifically that required to support implementation of these devices in UK diabetic eye screening. The recurrent focus of this article on England (the largest of the UK screening

Lancet Digit Health 2025;7: 100840

Published Online
April 3, 2025
<https://doi.org/10.1016/j.landig.2024.12.004>

*Joint first authors

College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK (T Macdonald MBChB, X Liu PhD, Prof A K Denniston PhD); NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHSFT, Birmingham, UK (T Macdonald, X Liu, Prof A K Denniston); Exeter Test Group, College of Medicine and Health, University of Exeter Medical School, Exeter, UK (Z Zhelev PhD, Prof C Hyde PhD); School of Computer Science and Mathematics, Kingston University London, London, UK (J Fajtl PhD); NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK (C Egan PhD, Prof A Tufail MD); Institute of Ophthalmology (C Egan, Prof A Tufail) and Centre for Medical Imaging, Division of Medicine (Prof S Halligan PhD), University College London, London, UK; Population Health Research Institute, St George's University of London, London, UK (A R Rudnicka PhD); Warwick Medical School, University of Warwick, Coventry, UK (Prof B Shinkins PhD, Prof S Taylor-Philips PhD); St George's University Hospitals NHS Foundation Trust, London, UK (R Given-Wilson FRCR); Vaccination and Screening Directorate, NHS England, London, UK (J K Dunbar MBChB); Gloucestershire Hospitals NHS Foundation Trust, Cheltenham, UK (Prof P Scanlon MD); UK National Screening Committee, Office for Health Improvement and Disparities, Department of Health and Social Care, London, UK (Prof S Taylor-Philips, Prof A Mackie PhD)

Correspondence to: Prof Alastair K Denniston, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK a.denniston@bham.ac.uk
See Online for appendix

programmes) reflects the original nature of the request to the NSC. Furthermore, although set in the context of the English diabetic eye screening programme, these considerations are likely to be relevant to other health-care settings.

ARIAS can be beneficial in two main ways: workload reduction, by triaging no risk or low-risk disease (about 90% of workload)¹⁵ from high-risk disease, leaving diagnostic steps to human graders; or diagnostic capability, classifying all grades of diabetic retinopathy and recommending a clinical outcome. In theory, ARIAS could be positioned at multiple points in the diabetic eye screening pathway, however the two use-cases considered to be most promising in the Tufail and colleagues' health technology assessment were: first, replacing primary graders; and second, as a filter before human screening (a step before primary graders).¹³ The first of these two approaches is shown in figure 1. The approach outlined in this Health Policy review is therefore written with these specific intended uses in mind, although many of the principles of the UK NSC evidence review discussed here would apply to implementations of ARIAS elsewhere in the screening pathway.

accuracy, clinical impact, and cost-effectiveness (appendix p 1). A filter was added to return only references published since the original report's search dates (June 25, 2020, to Dec 31, 2023). References were screened (AKD or TM) using the original inclusion and exclusion criteria as described in Zhelev and colleagues.¹⁴ Articles that met the inclusion criteria were then prioritised (AKD and TM) based on their level of relevance to the primary use-case of evaluating ARIAS' detection of sight-threatening diabetic retinopathy before manual grading in a context analogous to the UK (figure 2). This Health Policy review also draws upon the UK NSC's previous relevant publications in evaluating clinical AI evidence, most notably the UK NSC's approach to reviewing evidence on AI in breast cancer screening. The literature search of methodological publications assessing changes to screening tests which informed that approach is described in full in Taylor-Phillips and colleagues.¹⁶ As with our previous guidance papers,¹⁶ we aim to cite the most relevant sources for each concept, rather than providing an exhaustive list of references for each of the concepts described in this Health Policy review.

UK NSC criteria for appraising population screening programmes

When evaluating changes to an existing screening programme or an entirely new one, the UK NSC evaluates a programme's ability to fulfil 20 criteria.¹⁷ The UK NSC's 2021 review of ARIAS focused on how a

For more on the grading pathway see <https://www.gov.uk/government/publications/diabetic-eye-screening-pathways-patient-grading-referral-surveillance>

Methods

Search strategy and selection criteria

We repeated the search strategy from the NSC's 2021 automated grading in the diabetic eye screening programme report,¹⁴ in MEDLINE, focusing on ARIAS

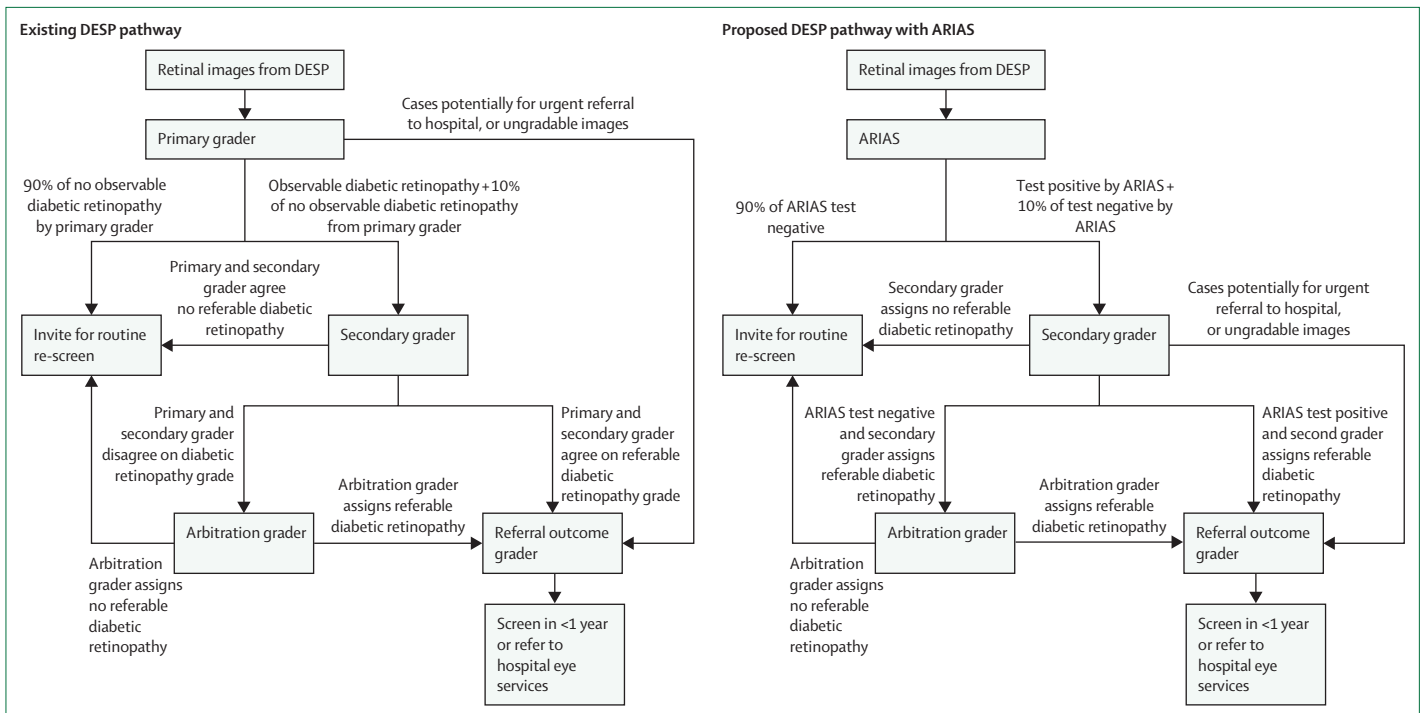


Figure 1: Flow diagram

A simplified representation of the current grading pathway (left), versus replacing human primary grading with ARIAS (right). ARIAS=automated retinal imaging analysis software. DESP=diabetic eye screening programme.

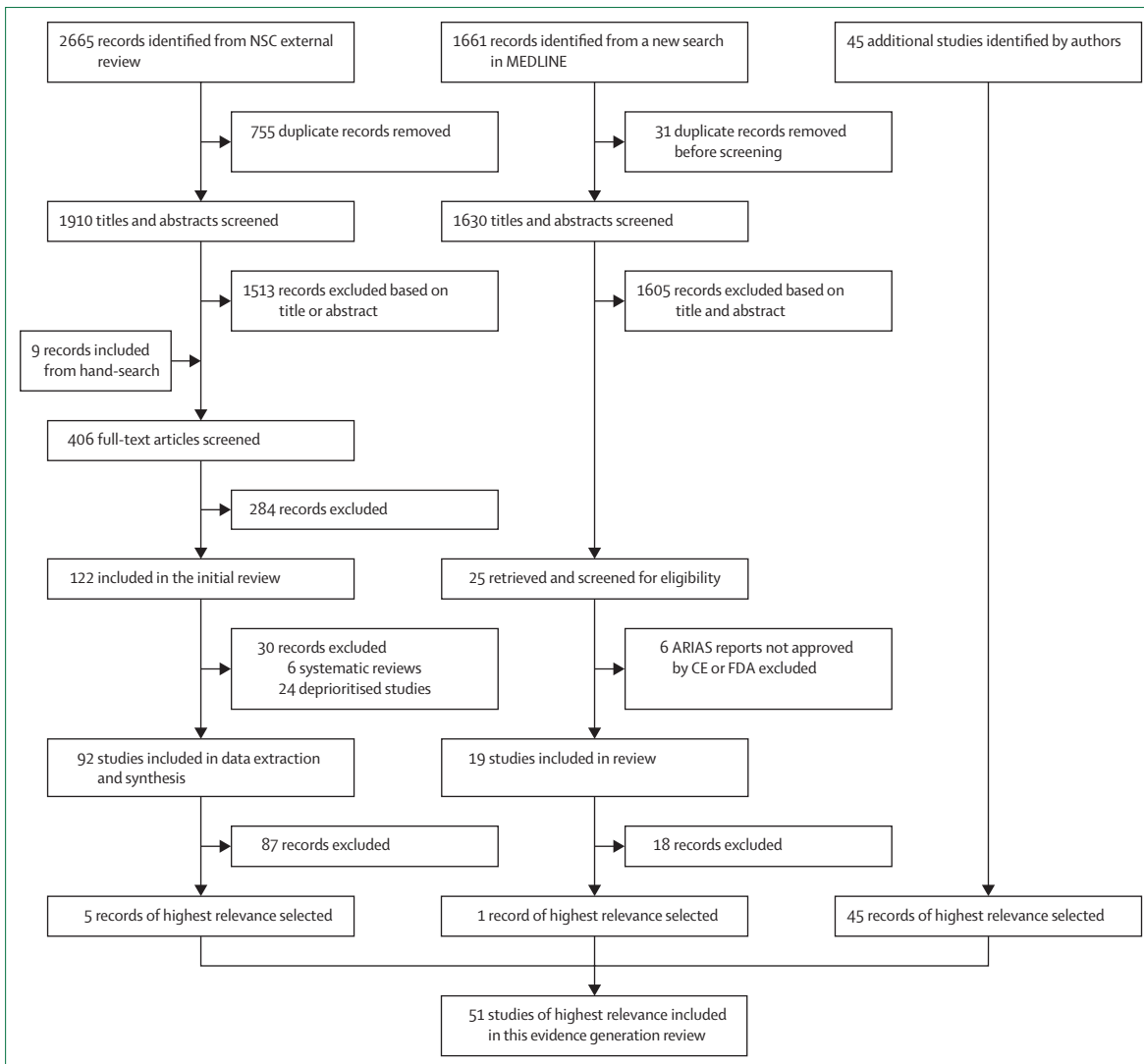


Figure 2: Included studies selection

ARIAS=automated retinal imaging analysis software. CE=Conformité Européenne. FDA=Food and Drug Administration. NSC=National Screening Committee. Left-side flowchart reproduced from Zhelev et al,¹⁴ by permission of the authors.

diabetic eye screening pathway incorporating an ARIAS might fulfil criteria 4, 5, 11, 12, and 14. These broadly cover test accuracy, clinical impact, social and ethical implications, and cost-effectiveness. This Health Policy review will outline how diabetic eye screening pathways incorporating ML-ARIAS might fulfil these as well as criteria 17 and 18, which focus on successful implementation. We will not address all 20 criteria as ML-ARIAS would be incorporated into an existing programme with an established rationale and evidence base addressing the remaining criteria.

Considerations regarding test performance

Test performance metrics

In the context of diabetic eye screening, ARIAS test performance refers to the ability of software to correctly

classify patients either with or without diabetic retinopathy, or with a particular level of diabetic retinopathy in either eye, compared with a reference standard. The definition of a screening test positive or test negative outcome depends on the level of diabetic retinopathy that the ARIAS is intended to detect and the diabetic retinopathy grading system used. English diabetic eye screening uses the National Health Service (NHS) feature-based grading classification.¹⁸ When ARIAS are proposed in the filter role, they are usually expected to classify into disease present (R1, R2, R3, M1, and U) and disease absent (R0M0; panel). Those classified as disease absent are returned to routine screening (ie, 12–24 months), whereas those classified as disease present will have their photos checked by human graders. When deployed to replace a primary grader, ARIAS could

Panel: The context—all grading in the diabetic eye screening programme in England is currently performed by human graders

In England, the standard screening visit involves the capture of two 45-degree colour photographs of the interior surface of the back of the eye or fundus. One photograph is centred on the macula, the central part of the retina; and the other on the optic disc, the point at which the optic nerve enters the eyeball. These are then manually assessed by human graders for diabetic retinopathy using the National Health Service feature-based grading system with patients assigned a grade for retinopathy and maculopathy in each eye. These range from R0 to R3A (no retinopathy to active, proliferative retinopathy), and M0 to M1 (no maculopathy to maculopathy present). Photographs of insufficient quality or clarity to assign a grade are labelled U for ungradeable. Patient outcomes depend on the highest grade in either eye and can range from 24-month review (lowest risk R0M0)* to urgent referral to the hospital eye service (R3AM0 or R3AM1).

Human graders within the multilevel system have varying levels of expertise and grading decisions are escalated according to clinical risk and complexity (figure 1). All fundus photographs are reviewed by primary graders. Blinded to primary graders' decisions, secondary graders review all images graded as R1, R2, R3, M1, and U by primary graders, as well as 10% of those graded as R0 for quality control purposes. If there is agreement between primary and secondary graders no further grading is needed, whereas discrepancies receive a further grade by arbitration graders. Arbitration graders can review both primary and secondary graders' grades. The grade assigned by arbitration graders is considered final. Patients with images determined as R1M1, R2M0, R2M1, R3AM0, R3AM1, or U either undergo additional tests within screening services, or are referred to hospital eye services.

*In the UK, diabetic eye screening was previously undertaken annually, but in 2016 the UK NSC made a recommendation that those with two consecutive screening visits graded as R0M0 could be screened at 24-month intervals. This policy changed in Scotland in 2022 and in England in October, 2023.¹⁹

undertake the same task or instead classify cases into the same set of multiple grades as a human grader (R0-3, M0 or M1, or U; figure 1).

Measures of diagnostic accuracy include sensitivity—ie, the proportion of those with disease, or with specified degree of disease, classified as positive by the ARIAS—and specificity—the proportion of those without disease, or absence of specified degree of disease, classified as negative by ARIAS. Sensitivity and specificity should be considered together since raising one will diminish the other. The receiver operating characteristic curve describes how sensitivity and specificity change across all potential diagnostic thresholds. Although test performance across all thresholds is outlined by a receiver operating characteristic curve, in clinical practice a binary operating point must be chosen that separates positive from negative results. Since the clinical implications of false negative results (ie, missing disease) are rarely equivalent to false positive diagnoses, choosing the optimal operating point is a finely balanced clinical decision, and it is the accuracy at that operating point that will inform decisions around implementation.

Performance in isolation versus performance of pathway as a whole

Most studies consider ARIAS test performance in isolation, but this does not reflect how they might be

implemented in real-life clinical settings. Consideration of ARIAS performance should also recognise any change in the performance of the diabetic eye screening as a whole, which in turn depends on factors such as where the ARIAS sits within the pathway, any change in human performance as a result of implementing the ARIAS, and any pathway redesign that occurs to accommodate the introduction of the ARIAS.

Diabetic eye screening as a whole needs to have high sensitivity and high specificity. High sensitivity minimises the number of false negatives (missed cases) who are at risk of harm through delay in diagnosis and resulting in potential loss of vision. High specificity minimises the number of false positives who are unnecessarily referred to hospital eye services, all of which can have a negative effect on patients (anxiety, time, and cost of attending unnecessary appointments) and the health service (financial and capacity).

If the introduction of ARIAS leads to increased referrals to hospital eye services (an increase in diabetic eye screening sensitivity or a decrease in diabetic eye screening specificity), this will have a negative impact on both patients and the health service. Even small reductions in specificity within a population screening programme can substantially affect receiving services, at a financial and staff time cost, potentially causing patient harm through longer waiting lists and treatment delays.^{15,20,21} This trade-off can be understood in health economic analysis of the data.¹³

Sensitivity in context

When considering ARIAS sensitivity, the nature of any false negatives is also important and should be reported to fully weigh up the potential benefits and harms of deploying the ARIAS. For example, although the misclassification of R0 as R1 is unlikely to result in adverse outcomes, the misclassification of sight-threatening diabetic retinopathy (R2, R3, and M1) as R0 is likely to result in considerable patient harm. Multiclass confusion matrices with confidence intervals around point estimates are therefore most informative in estimating potential benefits and harms.

Specificity in context

Although still important, it is less crucial for an ARIAS deployed in a primary grader role or filter role to have a high specificity when classifying any disease. This is because false positives can be overturned by the human grading system (although this additional human grading should be factored into any calculations with regard to cost savings from ARIAS deployment). However, the presence of anchoring bias (the tendency to rely on a previous piece of information given, in this case, the ARIAS grading) might affect this assumption and should be evaluated. The downstream effects of low ARIAS specificity will depend on how the ARIAS is deployed and its effect on human grading performance. For

example, if human graders are masked to ARIAS decisions, their grading performance might be unaffected, although the anchoring effect of a lower specificity ARIAS could lower the specificity of secondary graders. This further emphasises the importance of considering the sensitivity and specificity of diabetic eye screening as a whole.

Considerations regarding study design

Study design to support evidence of accuracy

Studies to assess accuracy can be retrospective, evaluating ARIAS performance on historical data; or prospective, evaluating performance on data collected after the initiation of the trial. Prospective studies can be further divided into observational, where ARIAS would have no impact on patient care; and interventional, when ARIAS classifications are acted upon. These study types have different advantages and provide different information. For example, a hypothetical portfolio of evidence to support the introduction of an ARIAS into diabetic eye screening might include the following study designs: (1) one or more large-scale retrospective test accuracy studies on previously collected diabetic eye screening data comparing the ARIAS to a human grader, against an acceptable reference standard. This would provide an estimate of diagnostic accuracy of the ARIAS in isolation against the current standard of care. Running multiple ARIAS in the same evaluation on the same dataset allows for direct model comparison.²² This would also provide evidence on safety to support an ARIAS progression to an interventional study; (2) an observational prospective study in which the ARIAS is evaluated within diabetic eye screening in line with its proposed intended use, but without the outputs of the ARIAS being acted upon. This can provide further evidence of diagnostic accuracy and, additionally, might provide evidence of potential implementation challenges. It does not, however, provide any direct evidence of clinical impact since it is non-interventional; (3) an interventional prospective study, enabling the direct comparison of diabetic eye screening with and without the ARIAS. This type of study would provide evidence of the actual effect of introducing the ARIAS on overall diabetic eye screening accuracy, and provides direct evidence of clinical impact and other downstream effects. Interventional designs include test-treat randomised controlled trials, before-and-after studies, and stepped-wedge designs.

Role of retrospective studies

Large-scale comparative test accuracy studies using bio-banked diabetic eye screening data with retinal images provide important evidence on the test performance of the ARIAS compared with current standard of care as the accepted reference standard. Currently, standard of care is a human grader, but it is possible that in the future this will be a combined ARIAS–human grading pathway.

Large datasets of graded images already exist within screening programmes, and can be curated for such studies at relatively low cost.^{23,24} The testing of the ARIAS performance outside the live clinical pathway dramatically reduces the resource and governance requirements to establish the study, and avoids any impact on the patient or the diabetic eye screening itself. Such studies are therefore less complex, less burdensome, less costly, and faster than prospective studies, and can provide an evidence base for further evaluation. They can also be used to compare the performance of different ARIAS on the same dataset.

When evaluating retrospective studies, considerations include: participant selection, the human comparator, the reference standard, subgroup performance, sample size, and (for the UK NSC) relevance to the UK setting.

ARIAS performance must be evaluated in unseen datasets that have not formed part of the ARIAS training data. AI has a tendency to overfit to training datasets, learning noise as well as the signal when making predictions, often then performing poorly when presented with data from new settings.^{25–27}

Studies that include all individuals consecutively screened in a given diabetic eye screening programme over a specific timeframe are most informative. Ideally only individuals excluded from standard pathways²⁸ should be excluded to reduce selection bias and best simulate the environment the ARIAS will encounter after deployment. As prevalence of sight-threatening diabetic retinopathy is low in those screened,^{13,29} large numbers of patients need to be included in such studies to generate tight confidence intervals around the output metrics (including for less common clinical and demographic subgroups). This need for scale can pose challenges, although such studies have taken place in the UK^{13,29} and other countries.³⁰

Datasets enriched with additional positive cases or images that are traditionally difficult to grade (such as patients with a cataract) can be used to evaluate ARIAS performance in detecting rare outcomes or specific scenarios. Enriched datasets can, however, introduce spectrum or selection bias, and are not therefore a substitute for evaluations on large consecutive cohorts to assess the overall performance of ARIAS for use in diabetic eye screening.¹⁶ In addition, enrichment makes evaluation of specificity and cost-effectiveness problematic as the study population no longer reflects the screening population in terms of disease prevalence.

The comparator should be standard of care—ie, human graders operating under standard conditions within diabetic eye screening. Some studies are so-called laboratory studies in which the human grading of the image set is undertaken separately outside a standard clinical environment. Two factors can affect generalisability in this context: first, the human graders might not perform as they would under normal conditions (ie, the laboratory effect);³¹ or the human comparator might

not be representative of graders working within the screening programme in terms of qualifications or experience.

The reference standard is the diagnostic test (or combination of tests) considered the closest approximation of whether an individual does or does not have a given disease in reality. In the context of ARIAS, this can be the final grade assigned by a multilayer grading system comparable to the English diabetic eye screening, using standard two-field digital colour images per eye. Although other imaging modalities or grading systems can be used, such as seven-field standard fundus photography, ultrawide field pseudocolour, or arbitration by experienced medical retina experts or grading centres, these might not be feasible in studies of sufficient size, or offer an improved prediction of true disease status.³² They also represent a different reference standard to that currently accepted in English diabetic eye screening.

There is increasing recognition of the risk of differential performance of health technology across population subgroups, such as those defined by age or ethnicity.^{33,34} This is of particular concern as systematic reviews of publicly available ophthalmic datasets have found these to be unrepresentative of the UK population (if demographic information has been recorded at all),³⁵ and a link between under-representation and disparate AI performance has been observed across multiple scenarios.³⁶ Large-scale retrospective studies across diverse, multi-ethnic groups provide the opportunity to ensure that a satisfactory performance is reached regardless of age, sex, and ethnicity (and other relevant factors such as geography and socioeconomic status), with ARIAS evaluations of this nature currently ongoing in the UK.^{37,38}

Factors related to the health-care setting and the local patient population can affect ARIAS performance substantially.^{27,39} Ideally, datasets used in ARIAS evaluation studies should closely match UK screening populations in terms of age, race, sex, other relevant features, and the prevalence and spectrum of the disease. As there is increasing recognition that there is the potential for differential AI performance between population subgroups, stratified analysis should be performed and reported to reduce the risk of implementing a screening tool which has major algorithmic bias. Furthermore, the testing dataset should comprise images which have been captured in screening programmes similar to the UK in terms of hardware, protocols, pre-processing, and file formats as all can affect AI performance.⁴⁰ For these reasons, well designed test performance evaluation studies done in English diabetic eye screening or in similar health-care settings and populations will be considered more informative than equivalent studies performed in settings or populations that are substantially different to the situation found in the English diabetic eye screening programme.

Role of prospective studies

Prospective studies enable the ARIAS to be evaluated within diabetic eye screening in a live setting. There are potentially a number of variations on both the study design and the level of intervention. A test-treat randomised controlled trial (RCT) might be used, in which patients are randomly assigned to either a standard diabetic eye screening pathway or an ARIAS-enabled one, enabling the direct comparison of the diabetic eye screening as a whole with or without the ARIAS. Randomisation to the intervention reduces allocation bias and can be at an individual or cluster level. Alternatives to the parallel comparator group include various forms of the before-and-after study, such as stepped-wedge designs. Well designed prospective interventional studies undertaken within the English diabetic eye screening programme (or a similar setting and population) provide the strongest estimate of the potential benefits and harms of introducing that ARIAS into the English diabetic eye screening. The effect of incorporating an ARIAS on human grader behaviour can also be assessed.

Alongside interventional prospective studies, there could also be a role for prospective studies in which the ARIAS is run in parallel to the existing human graders but without the findings being acted upon (sometimes called silent trials). Although there are many similarities to a well designed retrospective study (eg, analysis of all consecutive patients over a prespecified period), there is additional opportunity to assess operational factors and costs, including some which might be relevant to assessing accuracy, such as if the performance of the ARIAS declines when embedded in standard workflow systems as opposed to a research server; additionally such studies will be evaluating a more contemporary cohort, whereas biobanked data being older might not reflect current screening populations and protocols. Connectivity and compatibility with diabetic screening and primary databases that guide the patient pathway, trigger human grading, and generate reports can also be assessed. Non-interventional prospective evaluations could also have a role in providing local assurance before deployment to ensure that the level of performance demonstrated elsewhere is replicated in the local population and setting.

In the NSC's evidence review which included studies up to June, 2020, Zhelev and colleagues included ten ARIAS within the narrative synthesis with regard to diagnostic accuracy. They concluded that although a number of these achieved acceptable level of accuracy, only three systems have been evaluated in good quality studies conducted in the UK: EyeArt version 2.1, RetmarkerSR, and iGradingM.¹⁴

Role of vendor-independent studies

Discrepancies have been reported between outcomes from studies undertaken by ML-ARIAS vendors

themselves and those run by independent study groups.³⁰ Lee and colleagues³⁰ recommended the need for external validation where study groups curate image test sets, run the ARIAS, and analyse outputs independent of vendors. These discrepancies between independent and vendor-run studies could be due to a number of factors including how the imaging test set is curated (eg, excluding poor-quality or ungradable images, use of a non-consecutive series); when choosing a dataset, a vendor could be affected by commercial factors (speed and cost of access), rather than being solely focused on the quality, size, or representativeness of that dataset. Running ML-ARIAS independently of vendors also gives insight into deployment issues such as processing time, stalling of processing, and connectivity to other systems and therefore would be the preferred method of evaluation. These independent studies also provide an opportunity to undertake head-to-head studies such as those by Tufail and colleagues¹³ and Lee and colleagues,³⁰ in which independent evaluation of all ARIAS on the same dataset gives a better indication of probable relative performance than multiple separate studies each with their own test dataset.

Considerations for implementation as part of the evaluation process

Integration of ML-ARIAS into existing infrastructure

The current diabetic eye screening programme uses commissioned software from a variety of vendors. These manage the current diabetic eye screening workflow (including human grader review), failsafe systems, contact management, and referral to hospital eye services. Before deployment, it is important that the communication between prospective ML-ARIAS and existing systems are tested. Live implementation studies are currently exploring this.³⁷ Although not essential for the first wave of deployment if there has been appropriate testing between systems, development of a standard ARIAS-application programming interface (API) would allow more flexibility in future deployment as new systems emerge as previously discussed by Tufail and colleagues.¹³

The current National Institute for Health and Care Research (NIHR)-funded ARIAS evaluation integrating commercial state-of-the-art ARIAS in an NHS setting is providing encouraging evidence of its readiness for live deployment.⁴¹ Specifically, all software is already containerised and ready for cloud deployment with no or only minimal cloud configuration necessary. The execution of the ARIAS inference is implemented as a simple data batch processing; a set of images of an encounter or multiple encounters are presented to the ARIAS that returns an estimated diabetic retinopathy outcome, typically within less than 10 s. Although suitable for live deployment, the ARIAS would still require to be integrated with the target diabetic eye screening software, including the communication

functionality verifications, requiring minor additional work by the ARIAS and diabetic eye screening vendors that is already being undertaken.

Benefits of a standardised ARIAS-API

A standardised and preferably open-source ARIAS-API for facilitating communication between existing clinical systems and ML-ARIAS would considerably streamline and lower the cost of evaluating and deploying such algorithms. Vendors who deliver diabetic eye screening software and ML-ARIAS could proactively implement the ARIAS-API, ensuring they are interoperable with each other. This API would expedite the development of an evaluation framework that only needs to be implemented once and could be reused for future evaluations.¹³ Using this approach, ARIAS vendors involved in the process could avoid the need to create an evaluation-specific API. Instead, they could concentrate on a single, more beneficial standardised API that is compatible with both evaluation and live deployment setups.

Test impact—including clinical outcomes and cost-effectiveness

Outcomes

There is agreement between UK NSC and NICE that medical policy decisions should be driven by patient and public health outcomes rather than test performance metrics alone.^{42–44} However, this does not have to be from a single study; instead a linked evidence^{45,46} or analytic framework approach^{47,48} can be used to understand the impact on outcomes through linking different studies together. The purpose of diabetic eye screening is to reduce morbidity from preventable sight loss caused by diabetic retinopathy and although good test accuracy will probably contribute towards this, further evidence linking detection to eye health outcomes needs to be measured to ensure these goals are realised in practice. It would be ideal, therefore, to be able to demonstrate the impact of changing the patient pathway on outcomes such as absolute change or progression to a meaningful threshold such as registerable sight-impairment, rates of progression to a specific threshold of disease (eg R3AM0 or R3AM1), and quality of life measures. However, this is often not practical and requires very large sample sizes and long timescales. Where the link between a test outcome and the clinical outcome is well understood, it might be sufficient to model downstream clinical outcomes based on the diagnostic accuracy and any change in referral behaviour. In the case of the diabetic eye screening programme, the relationship of different grades of retinopathy to progression to sight loss is supported by a wealth of data. If modelling approaches are used, the modelling of clinical outcomes should not consider the ARIAS in isolation, but consider the sensitivity and specificity of the diabetic eye screening programme as a whole (ie, recognising any wider system

effects of the ARIAS that alter overall screening performance).

One consideration sometimes raised is whether the introduction of an ARIAS would have a negative effect on the detection of non-diabetic ocular findings. However, the stated purpose of diabetic eye screening is to detect diabetic eye disease only. Patients are informed they still need to have routine ocular health checks in addition to screening, therefore the detection of other pathologies is out of scope for diabetic eye screening. It is worth noting, however, that ARIAS have not missed serious macular conditions in previous evaluations.¹³

Any decision to introduce an ARIAS must also consider the cost-effectiveness of adopting ARIAS within diabetic eye screening. In their 2016 paper, Tufail and colleagues¹³ evaluated the cost-effectiveness of two different ARIAS compared with manual grading, focusing on the outcome of cost per appropriate screening outcome (defined as a true positive or true negative result). For both use-cases, ARIAS was cost saving compared to manual grading, but less effective. The potential implications of this reduction in effectiveness on patient health was difficult to ascertain due to the aggregate and short-term nature of the outcome selected; both machines were comparable in correctly identifying positive cases, but the proportion of patients receiving an appropriate screening outcome was notably lower due to the relatively high false positive error rate. The inconvenience, additional resource use required, and potential harm to patients associated with false positive results depends on how the ARIAS is implemented, re-emphasising the need to evaluate the effect of ARIAS on diabetic eye screening as a whole, rather than in isolation. Future cost-effectiveness analyses are essential to capture up-to-date, direct and indirect costs associated with adoption of ARIAS, but also to additionally capture and quantify implications in terms of resource use and patient health outcomes, and to incorporate the changing accuracy as ARIAS technology is updated. These types of analyses are ideally conducted alongside high-quality clinical studies, providing the opportunity to collect resource use and cost data in parallel with the evaluation of clinical effectiveness.

Considerations for study design

The clinical impact of introducing ARIAS should be measured through well designed prospective interventional studies in which the ARIAS is evaluated within the diabetic eye screening programme in line with its intended use. The choice of study design needs to balance a number of competing demands, notably the need to provide high-quality evidence by minimising bias while also being safe, acceptable to users, feasible, and affordable. Prospective test-treat RCTs⁴³ provide robust evidence of the effect of the ARIAS on diabetic eye screening programme accuracy (noted earlier), and on wider impacts including patient outcomes, operational issues, and burden on clinical services. Operational

issues that should be actively looked for include reliability of the service, and issues around integrating the ARIAS into existing clinical workflows. Wider service impacts might include: change in resource requirements for diabetic eye screening, change in referral rate to the hospital eye service (costs and resource requirements), or change in behaviour of the human graders due to, for example, alteration in human attention when grading ARIAS test positive encounters.

Prospective intervention studies should only be undertaken when there is sufficient evidence from pre-implementation studies to justify evaluation in a live diabetic eye screening setting. Additional safety measures can be included in these study designs, although resource impact needs to be recognised. Potential safety measures include: increasing the number of screen-negative cases that go to secondary graders for quality assurance (eg, this is currently 10% for human graders but could be increased temporarily or permanently for an ARIAS, analogous to the process used for human graders in training or graders below the quality assurance standard); and undertaking error analysis of ARIAS performance, specifically through ongoing analysis of all disagreements between ARIAS and secondary graders that go to adjudication, looking not only for error rates but the type of errors that ARIAS makes compared with humans.

Although expensive and logistically challenging to perform, prospective comparative trials have been previously carried out in UK screening programmes. Pragmatic adaptations can be made to RCT design to improve feasibility and affordability including through cluster randomisation of screening sites,⁴⁹ or stepped-wedge designs,⁵⁰ provided these are carefully designed.^{51,52} Designing such studies in a manner that they can be directly transitioned into national roll-out if the results are positive both reduces implementation delays of effective technology and optimises the benefits of trial resources.

Cost-effectiveness should also be estimated using a clinical effectiveness outcome that captures the impact of adopting ARIAS within the diabetic eye screening programme on patient outcomes, ideally alongside any prospective interventional studies undertaken. The time horizon of the analysis or model should be sufficiently long to capture all potential downstream cost and health consequences, and a full probabilistic sensitivity analysis should be conducted to characterise the uncertainty in the parameters informing the analysis. Tufail and colleagues provide a template that could be updated with the sensitivities and specificities of newer ML-ARIAS and current direct and indirect costs.¹³

Social and ethical implications

In this context, social and ethical implications refer to the effect of the technology on experiences and outcomes in individuals and population groups within the diabetic eye screening programme, with particular consideration

for under-represented and socioeconomically disadvantaged groups. AI systems are highly sensitive to bias within the data they were trained on, and there is a general risk that AI can exacerbate entrenched health inequalities.^{36,53} In the context of diabetic eye screening, ARIAS could negatively impact inequalities through a variety of mechanisms. These include differential performance in demographic subgroups; the exclusion of specific subgroups from intended-use statements (statements submitted at regulatory approval outlining the patients and situations in which an AI health technology can be used); and issues of perception and trust among subgroups of people with diabetes.⁵⁴

Acceptability and uptake should be assessed, including looking for differences between population groups. There should also be adequate trust in the model's outputs from graders, hospital eye services, and people with diabetes, such that downstream actions specific to ARIAS outputs are predictable and appropriate.

Data governance,⁵⁵ liability,⁵⁶ cybersecurity, and intellectual property generation⁵⁷ lie beyond the scope of test evaluation research, but are all factors that will require careful consideration before ARIAS deployment.

In order to demonstrate acceptable performance across populations, the evidence on test performance and clinical effectiveness described here must present adequate subgroup analyses of relevant populations (such as by ethnicity, sex, and age). To be able to carry out such analyses, the test data supporting such evaluations must include sufficient diversity of populations within the data, adequately labelled to identify these groups and large enough to allow statistical power for the detection of rare events (eg, R3 retinopathy) in subgroups. For example, ARIAS studies currently ongoing in the northeast London diabetic eye screening programme include at least 25 000 screening episodes from each of the key ethnic subgroups of White, Black, and south Asian, and are adequately powered to evaluate ARIAS performance in each.^{37,38}

The NSC's review by Zhelev and colleagues noted that in terms of the social and ethical aspects, there were five primary studies that investigated the impact of AI in the context of screening (including diabetic eye screening programme), with a further 14 primary studies and 38 reviews and opinion papers being considered sufficiently relevant to be included in the evidence map.¹⁴ Ongoing work within English diabetic eye screening settings is now starting to address this evidence gap regarding social and ethical implications as well as equity in ARIAS test performance by population subgroups including age, sex, and ethnicity.³⁷

Conclusion

The UK NSC is committed to ensuring that the diabetic eye screening programmes in the UK continue to deliver improved clinical outcomes to patients, and do so equitably and at good value to the health-care system.

The assessment of evidence for ARIAS integration into English diabetic eye screening programme will therefore prioritise these goals.

There is increasing evidence to support the use of ARIAS with regard to test performance, but there is currently less evidence with regard to their downstream clinical and service outcomes and any wider social impacts or ethical considerations. Some of these issues are currently being explored, such as through the NIHR-funded study focused on acceptability and performance of ARIAS between different ethnic groups.³⁷ Previous health technology assessments have shown that ARIAS can be cost-effective but there would be value in updating these models in line with newer ARIAS coming to market.¹³ Careful consideration should be given to the relative contributions of retrospective and prospective studies, and the elements of study design and wider outcome measurement that are required to provide the evidence that is needed to adequately assess the balance of benefits and harms that the introduction of ARIAS may have. Early engagement between ARIAS companies and the UK NSC is recommended, both to discuss proposed trial designs and to anticipate potential changes in the screening pathway that may be relevant (such as wide-field imaging or non-mydratic approaches). Independent head-to-head evaluations are particularly informative in enabling comparison between ARIAS, and to reduce the variation caused by differences in the underlying datasets. At some point, ARIAS are likely to transform the delivery of diabetic screening; however the introduction and evaluation of ARIAS is complex, and will require a multidisciplinary multi-stakeholder approach to ensure that any introduction of ARIAS is of benefit to patients and can be supported by the wider health service.

Contributors

AKD and XL designed this Health Policy review at the request of the UK National Screening Committee (UK NSC) AI Task Group. AKD and TM undertook the searches, screening, and relevance assessments. This work also drew on the earlier reviews by ZZ, CH, and ST-P. AKD, XL, TM, and ST-P undertook the first draft; all authors contributed to subsequent drafts and approved the final version.

Declaration of interests

XL reports consulting fees from Hardian Health and Conceivable Life Sciences, and a past role as a Health Scientist at Apple. CH is a member of the UK NSC, which funded the Exeter Test Group at University of Exeter to do evidence review of screening for diabetic retinopathy. CE has received a grant or contract and travel support to attend MacTel international clinical study meetings from the Lowy Medical Research Institute; and consulting fees from Heidelberg Engineering, Inozyme Pharma, and Boehringer Ingelheim. AT has received consulting fees from Annexon, Apellis, Bayer, Genentech, Iveric Bio, Novartis, Oxurion, Roche, Heidelberg Engineering, Ocular Therapeutix, Opthea, Oculogics, and Boehringer Ingelheim; payment or honoraria from Apellis; and sits on the Data Safety Monitoring Board or Advisory Board for the Alvotech AVT06 study and J&J 1887 study. RG-W is a member of the UK NSC and Non-Executive Director at Moorfields Eye Hospital. SH is Chair of the UK NSC Artificial Intelligence Task Group, which received institutional funding from the National Institute for Health and Care Research (NIHR) biomedical research centre; is a member of the UK NSC Adult Reference Group and the NIHR health technology assessment prioritisation board; and is a Clinical Professional Advisor for the National Health Service England diabetic eye screening programme; and reports payment for

expert witness work from multiple instructing solicitors. PS reports consulting fees from Boehringer; grants or contracts from Zeiss, Centervue, Optos, and Bayer; speaker fees from Bayer and Topcon; and support for attending meetings from Boehringer and Bayer. ST-P is a member of the UK NSC and Chair of the UK NSC Research and Methodology Group, funded by an NIHR Research Professorship (NIHR302434). All other authors declare no competing interests. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

- 1 Yau JWY, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012; **35**: 556–64.
- 2 Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob Health* 2017; **5**: e1221–34.
- 3 Sharma S, Oliver-Fernandez A, Liu W, Buchholz P, Walt J. The impact of diabetic retinopathy on health-related quality of life. *Curr Opin Ophthalmol* 2005; **16**: 155–59.
- 4 Fenwick EK, Pesudovs K, Rees G, et al. The impact of diabetic retinopathy: understanding the patient's perspective. *Br J Ophthalmol* 2011; **95**: 774–82.
- 5 Cooper OAE, Taylor DJ, Crabb DP, Sim DA, McBain H. Psychological, social and everyday visual impact of diabetic macular oedema and diabetic retinopathy: a systematic review. *Diabet Med* 2020; **37**: 924–33.
- 6 Hex N, Bartlett C, Wright D, Taylor M, Varley D. Estimating the current and future costs of Type 1 and Type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabet Med* 2012; **29**: 855–62.
- 7 Sasongko MB, Wardhana FS, Febryanto GA, et al. The estimated healthcare cost of diabetic retinopathy in Indonesia and its projection for 2025. *Br J Ophthalmol* 2020; **104**: 487–92.
- 8 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.
- 9 Scanlon PH. The contribution of the English NHS Diabetic Eye Screening Programme to reductions in diabetes-related blindness, comparisons within Europe, and future challenges. *Acta Diabetol* 2021; **58**: 521–30.
- 10 Haider S, Thayakaran R, Subramanian A, 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.
- 11 Hipwell JH, Strachan F, Olson JA, McHardy KC, Sharp PF, Forrester JV. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. *Diabet Med* 2000; **17**: 588–94.
- 12 Philip S, Fleming AD, Goatman KA, et al. The efficacy of automated “disease/no disease” grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol* 2007; **91**: 1512–17.
- 13 Tufail A, Kapetanakis VV, Salas-Vega S, et al. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess* 2016; **20**: 1–72.
- 14 Zhelev Z, Peters J, Rogers M, et al. Automated grading in the Diabetic Eye Screening Programme. UK National Screening Committee. 2021. <https://www.gov.uk/government/consultations/automated-grading-in-diabetic-eye-screening-rapid-review-and-evidence-map> (accessed Feb 12, 2024).
- 15 Scanlon PH, Nevill CR, Stratton IM, et al. Prevalence and incidence of diabetic retinopathy (DR) in the UK population of Gloucestershire. *Acta Ophthalmol* 2022; **100**: e560–70.
- 16 Taylor-Phillips S, Seedat F, Kijauskaite G, et al. UK National Screening Committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *Lancet Digit Health* 2022; **4**: e558–65.
- 17 UK Government. Criteria for a population screening programme. 2022. <https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme> (accessed July 5, 2023).
- 18 NHS England. NHS Diabetic Eye Screening Programme: grading definitions for referable disease. 2012. <https://www.gov.uk/government/publications/diabetic-eye-screening-retinal-image-grading-criteria/nhs-diabetic-eye-screening-programme-grading-definitions-for-referable-disease> (accessed Oct 21, 2022).
- 19 Harris M. Diabetic eye screening changes for people at lower risk in England. 2023. <https://nationalscreening.blog.gov.uk/2023/10/03/diabetic-eye-screening-changes-for-people-at-lower-risk-in-england/> (accessed Nov 16, 2023).
- 20 Foot B, MacEwen C. Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. *Eye* 2017; **31**: 771–75.
- 21 Rees S, Hassan H. The hidden waitlist: the growing follow-up backlog. 2023. <https://reform.uk/wp-content/uploads/2024/12/Hospital-of-the-Future-a-framing-paper-1.pdf> (accessed Feb 12, 2024).
- 22 Rajesh AE, Davidson OQ, Lee CS, Lee AY. Artificial Intelligence and Diabetic Retinopathy: AI Framework, Prospective Studies, Head-to-head Validation, and Cost-effectiveness. *Diabetes Care* 2023; **46**: 1728–39.
- 23 Chalkidou A, Shokraneh F, Kijauskaite G, et al. Recommendations for the development and use of imaging test sets to investigate the test performance of artificial intelligence in health screening. *Lancet Digit Health* 2022; **4**: e899–905.
- 24 Kale AU, Mills A, Guggenheim E, et al. A datasheet for the INSIGHT Birmingham, Solihull, and Black Country diabetic retinopathy Screening Dataset. *Ophthalmol Sci* 2023; **3**: 100293.
- 25 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv* 2018; published online July 2. <http://arxiv.org/abs/1807.00431> (preprint).
- 26 Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit Med* 2019; **2**: 31.
- 27 Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 2022; **4**: e351–8.
- 28 UK Government. Diabetic eye screening: cohort management. 2022. <https://www.gov.uk/government/publications/diabetic-eye-screening-cohort-management-overview/diabetic-eye-screening-cohort-management> (accessed Oct 20, 2022).
- 29 Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol* 2021; **105**: 723–28.
- 30 Lee AY, Yanagihara RT, Lee CS, 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.
- 31 Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008; **249**: 47–53.
- 32 Olvera-Barrios A, Heeren TF, Balaskas K, et al. Comparison of true-colour wide-field confocal scanner imaging with standard fundus photography for diabetic retinopathy screening. *Br J Ophthalmol* 2020; **104**: 1579–84.
- 33 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366**: 447–53.
- 34 Alderman JE, Palmer J, Laws E, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *Lancet Digit Health* 2024; **7**: e64–88.
- 35 Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021; **3**: e51–66.
- 36 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27**: 2176–82.

- 37 National Institute for Health and Care Research. Ethnic differences in performance and perceptions of artificial intelligence retinal image analysis systems for the detection of diabetic retinopathy in the NHS diabetic eye screening programme. https://fundingawards.nihr.ac.uk/award/AI_HI200008 (accessed Aug 16, 2023).
- 38 Olvera-Barrios A, Owen CG, Anderson J, et al. Ethnic disparities in progression rates for sight-threatening diabetic retinopathy in diabetic eye screening: a population-based retrospective cohort study. *BMJ Open Diabetes Res Care* 2023; **11**: e003683.
- 39 de Vries CF, Colosimo SJ, Staff RT, et al. Impact of different mammography systems on artificial intelligence performance in breast cancer screening. *Radiol Artif Intell* 2023; **5**: e220146.
- 40 NHS England. Diabetic eye screening: approved cameras and settings. 2014. <https://www.gov.uk/government/publications/diabetic-eye-screening-approved-cameras-and-settings> (accessed Sept 3, 2024).
- 41 Fajtl J, Welikala RA, Barman S, et al. Trustworthy evaluation of clinical AI for analysis of medical images in diverse populations. *NEJM AI* 2024; published online Aug 13. <https://doi.org/10.1056/AIoa2400353>.
- 42 UK Government. UK NSC: evidence review process. <https://www.gov.uk/government/publications/uk-nsc-evidence-review-process/uk-nsc-evidence-review-process> (accessed Oct 21, 2022).
- 43 National Institute for Health and Care Excellence. The guidelines manual. 2012. <https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-pdf-2007970804933> (accessed Feb 12, 2024).
- 44 National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies. 2022. <https://www.nice.org.uk/corporate/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605> (accessed Sept 29, 2022).
- 45 Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care* 2012; **28**: 52–58.
- 46 Lord SJ, Irwig L, Bossuyt PMM. Using the principles of randomized controlled trial design to guide test evaluation. *Med Decis Making* 2009; **29**: e1–12.
- 47 Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001; **20**: 21–35.
- 48 Schünemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies—from test accuracy to patient-important outcomes and recommendations. *J Clin Epidemiol* 2019; **111**: 69–82.
- 49 Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. John Wiley & Sons, 2012.
- 50 Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006; **6**: 54.
- 51 Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Med Res Methodol* 2005; **5**: 10.
- 52 Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ* 2017; **358**: j3064.
- 53 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366**: 447–53.
- 54 Glocker B, Jones C, Bernhardt M, Winzeck S. Risk of bias in chest X-ray foundation models. *arXiv* 2023; published online Sept 7. <http://arxiv.org/abs/2209.02965> (preprint).
- 55 Department of Health and Social Care. Better, broader, safer: using health data for research and analysis. 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf (accessed Feb 12, 2024).
- 56 Smith H, Fotheringham K. Artificial intelligence in clinical decision-making: Rethinking liability. *Med Law Int* 2020; **20**: 131–54.
- 57 Joshi I, Cushman D. A buyer's guide to AI in health and care. NHS. 2020. https://transform.england.nhs.uk/media/documents/NHSX_A_Buyers_Guide_to_AI_in_Health_and_Care.pdf (accessed Feb 12, 2024).

Copyright © 2025 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.