UK National Screening Committee

UK National Screening Committee (UK NSC)

Major programme modification: Automated grading in the Diabetic Eye Screening Programme (DESP) – UK NSC review

Date: 25 June 2021

Contents

Aim

1. To ask the UK National Screening Committee (UK NSC) to make a recommendation, based on the evidence presented in this document, whether the proposal to modify the Diabetic Eye Screening Programme (DESP) by implementing an automated retinal image analysis system (ARIAS) should be recommended.

Current Recommendation

2. In England, Scotland and Wales, the Diabetic Eye Screening programme (DESP) has been implemented since 2003, and in Northern Ireland since 2008. The eligible population for these programmes is all people with type 1 and type 2 diabetes aged 12 or over (excluding women who have only gestational diabetes).

3. The programme has a grading pathway to detect individuals with retinopathy and/or maculopathy with 3 levels of human graders. The grading pathways

are similar in all UK nations, except in the Scottish DESP where an ARIAS (iGrading M) has been used as a level 1 grader since 2011.

Proposal for major programme modification

4. In 2019, the UK NSC received a proposal to modify the English DESP with an ARIAS.

Evidence Summary

5. The 2021 evidence summary was undertaken by the University of Exeter.

6. The summary addressed questions on test accuracy [Q1], clinical utility [Q2], cost-effectiveness [Q3] and social and ethical implications [Q4]. In line with the UK NSC evidence review process, questions 1 and 2 were assessed in the form of a rapid review and questions 3 and 4 were assessed in the form of an evidence map.

7. The summary concluded that further evidence is needed to inform the decision on the implementation of ARIAS in the DESPs in England, Wales and Northern Ireland. This is because:

    a. some ARIASs are accurate enough as a single read and 3 ARIASs in particular have been evaluated in good quality studies in the UK. **Criteria 4 and 5 met.**

    b. the effect of ARIAS on patient outcomes is unclear due to the lack of evidence from prospective studies that integrate ARIAS into DESP programmes and compare this pathway with the current one. Consequently, the use case for AI in the pathway for these prospective studies needs to be determined – potential use cases identified in the review were the use of AI as level 1 graders and the use of AI as a pre-screening tool before level 1 graders. **Criterion 11 not met.**

    c. there is some evidence that using AI systems to replace level 1 graders or as a pre-screen provides better value for money than manual grading, but the analyses need updating, for example, to incorporate a 2-year screening interval for low-risk groups and longer than 1 year time horizon to capture clinical impact. Therefore, a rapid review should not be commissioned at this point. **Criterion 14 not met.**

    d. the evidence on social and ethical aspects of using AI systems in screening programmes should be assessed further as the evidence map found a large volume of evidence. **Criterion 12 not met.**

8. Refer to Table A below for criteria.

Consultation

9. A one month consultation (18 May to 21 June 2021) was hosted on the UK NSC website. Direct emails were sent to 59 stakeholders (please note that multiple individuals from the same organisation were invited) (see Appendix A for stakeholders).

10. Comments were received from 6 stakeholders (see Appendix B for comments):

    a. Clinical Lead, Gloucestershire and Oxfordshire Diabetic Eye Screening Programmes and the English NHS DESP

    b. Consultant diabetologist, Association of British Clinical Diabetologists

    c. Senior Medical Statistician in Gloucester, Senior Enterprise Research Fellow in Southampton

    d. A joint response from a group of experts including a Statistical Epidemiologist, Consultant Ophthalmologists, a Clinical Lead in Diabetes and Endocrinology Epidemiology

    e. The Royal College of Ophthalmologists

    f. Retinopathy Research and Professional Development Manager

11. Out of 6 stakeholders, 2 agreed with the recommendation, and remaining stakeholders did not provide a direct statement.

    Several key themes emerged from this consultation: test classification, further work, detection of other eye diseases, appropriateness of expertise and the quality of the review.

    Test classification

    Stakeholders pointed out that ‘low’ and ‘high’ risk groups were not clearly defined in the UK NSC review and wanted clarification on whether ‘low’ risk meant ‘no disease (R0M0)’ or ‘non-referable diabetic retinopathy (R0M0, R1M0)’. Stakeholders unanimously proposed that ARIASs should be used at the ‘no disease’/ ‘disease’ level.

Response: this review did not define ‘low’ and ‘high’ risk groups because there was insufficient evidence to assess the impact of using different decision thresholds, for example (‘no disease’/ ‘disease’ or ‘non-referable’/ ‘referable diabetic retinopathy’). None of the studies assessed the performance of an ARIAS integrated within the DESP pathway; instead, all studies reported performance of ARIASs as a single read. The UK NSC review recommended that future prospective studies and model-based cost-effectiveness evaluations should investigate the effect of different decision thresholds when integrating ARIAS into the DESP pathway.

Suggestions for further research

Below further research suggestions have been summarised *[note: suggestions for further research that are were covered in the review were not included here]*:

- investigate the implications of missing R1 when having 2-year screening intervals for low-risk groups
- focus on ARIASs that are good at detecting any diabetic retinopathy rather than referable or sight threatening diabetic retinopathy

Stakeholders suggested that ARIASs that already show acceptable performance should be further evaluated in staged implementation. It could study the various evidence gaps identified in the review. Stakeholders flagged that they have developed a protocol for this, which was discussed with the Diabetic Eye Screening Programme Research Advisory Committee (RAC).

Response: the committee agreed that further research should take into account the clinical impact of a 2-year interval for low risk groups when evaluating an ARIAS. The recommendations for future research were updated to reflect this. The committee agreed that ARIASs showing acceptable performance should be evaluated in staged implementation. The staged implementation study should have a clear end date after which all gathered evidence should be presented to the UK NSC for recommendation. The recommendations for future research were updated to reflect this. The committee also agreed that decisions on test classification should inform future research including the development of ARIASs.

The committee agreed that it is important to aid further research and thus supported a proposal by the AI Task Group that a guidance document for the UK NSC evidence requirements to inform a recommendation on the use of

ARIAS in the DESP should be developed. This will include discussion of potential use cases of ARIAS in a prospective study, the decision threshold, key outcomes of interest and study design features. The document will be produced in collaboration with AI Task Group members and external experts from relevant fields.

Detection of other eye diseases

There was a disagreement between stakeholders whether the impact of ARIAS on other eye diseases/ pathologies (for example; cataract) should be taken into account in the evaluation of ARIAS process. Some stakeholders suggested that the remit of the current DESP is to detect diabetic retinopathy, and the same remit should apply with the introduction of ARIAS, while others noted that it is important to assess the impact of ARIAS on other eye diseases/ pathologies.

Response: Due to the disagreement, it was proposed that further discussions with DESP managers and clinicians are needed. The committee acknowledged that it would be difficult to assess the impact of an ARIAS on other disease eye diseases / pathologies (for example; incidental findings) in prospective studies. This is because a large sample size would be required due to low prevalence of other eye pathologies. This may also be difficult to address in modelling studies because of a lack of empirical evidence on the accuracy or clinical impact of AI on other eye pathologies. The review has been updated to reflect that the information on incidental findings in the identified studies is limited and difficult to interpret.

Concerns about the authors of this review

A stakeholder was concerned that there was a lack of topic experts among the co-authors.

Response: this review was undertaken in line with the UK NSC evidence review process. The UK NSC reviews are undertaken by external reviewers including consultancy companies and university-based research groups that are experts in review methods and techniques. The review team included an ophthalmologist. In addition, multiple topic experts were consulted while developing this review. For example, the review was discussed at the AI Task Group meeting. The group has a leading expert in ophthalmology. Furthermore, the review and recommendations were discussed at a UK NSC AI in DES workshop held on 17 March, which was hosted specifically to discuss this review, and was attended by the leading experts in

ophthalmology, diabetes, endocrinology, diabetic eye screening programme, artificial intelligence in medical imaging, test methodology and screening.

Concerns about the quality of the UK NSC, PHE, NHSE reviews

A stakeholder was concerned that this UK NSC review and PHE and NHSE document in general are not being quality checked before publishing.

Response: the quality check of the review, including spell-checking and proof-reading, was performed before publishing the final version. As above, the review is quality-checked by the UK NSC evidence team, the ARG reference group, the AI task group and external stakeholders in workshops, meetings, and online consultation.

Recommendation

12. The Committee is asked to approve the following recommendations:

    1. Further research on test accuracy of newer versions of ARIASs, clinical utility and cost-effectiveness in the UK context is needed before a decision on the proposal to modify the Diabetic Eye Screening Programme (DESP) with an automated retinal image analysis system (ARIAS) can be made.

    2. Further work on social and ethical aspects of AI implementation in screening programmes should be commissioned.

    3. The UK NSC evidence team in collaboration with external academics should produce a guidance document on the evidence requirements for the application of AI in DESP.

UK National Screening Committee

Section 1 - Criteria for appraising the viability, effectiveness and appropriateness of a screening programme

This section looks at whether certain UK NSC criteria have been met when reviewing a given screening programme. Only the criteria evaluated by the current review have been included below.

The Test

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

Criterion 4 has been met

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

Criterion 5 has been met

The Screening programme

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" (for example; 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.

- **Criterion 11 has not been met**

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

- **Criterion 12 has not been met**

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 (for example; value for money). Assessment against this criteria should have regard to

evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

- **Criterion 14 has not been met**

UK National Screening Committee

Appendix A: List of organisations and individuals contacted

1. Accelerated Access Collaborative
2. Action for Blind People
3. AGFA
4. Association of British Clinical Diabetologists
5. Association of Optometrists
6. British Association of Retinal Screening
7. College of Optometrists
8. Densitas Health
9. Diabetes UK
10. Faculty of Public Health
11. Faculty.ai
12. Foundation of European Nurses in Diabetes
13. Google Health
14. IBM
15. Imperial College London
16. Institute of Diabetes in Older People
17. Insulin Dependent Diabetes Trust
18. International Diabetes Federation
19. Juvenile Diabetes Research Foundation
20. Kheiron Medical Technologies
21. National Diabetes Information Service
22. NHS
23. NICE
24. Northgateps
25. Optos
26. Primary Care Diabetes Society
27. Researcher with interest in AI (Queen Marry University of London)
28. Researcher with interest in AI (University of Manchester)
29. Royal College of General Practitioners
30. Royal College of Nursing
31. Royal College of Physicians
32. Royal College of Physicians and Surgeons of Glasgow
33. Royal College of Physicians of Edinburgh
34. Royal College of Radiologists

35. Royal National Institute of Blind People (RNIB)
36. Young Diabetologists Forum

UK National Screening Committee

Appendix B: Consultation comments

Note: Personally identifiable information has been redacted from certain comments, where individuals have chosen not to have personal details made public

Automated grading in the Diabetic Eye Screening Programme

Consultation comments

1. Clinical Lead English NHS DESP

Name:	Peter Scanlon	**Email address:**	XXXX XXXX
Organisation (if appropriate):	GHNHSFT and OUHNHSFT		
Role:	**Clinical Lead Gloucestershire and Oxfordshire Diabetic Eye Screening Programmes and the English NHS DESP**		
Do you consent to your name being published on the UK NSC website alongside your response? **Yes**			

Section and / or page number	Text or issue to which comments relate	Comment *Please use a new row for each comment and add extra rows as required.*
1	To triage patients into low and high-risk cases.	Low risk was not adequately defined
31	The systems used in Scotland (iGradingM) and Portugal (RetmarkerSR) are based on Machine Learning (ML) algorithms which extract from the images pre-specified 'hand-crafted' features, such as microaneurysms, and use the information to classify patients into 'high risk' (referral) and 'low risk' (routine recall) categories.	The first definition that I could find of low risk was on page 31 when they talked about low risk being routine referral. I strongly disagree with this because it means that the algorithm is being used at the referral /no referral level. I feel that it should be used at the disease/no disease level to take out those patients with no DR and that the graders should look at all images that have any DR. Future research should then concentrate on algorithms that are good at detecting any DR rather than referable or sight threatening DR

2. Senior Medical Statistician

Name:	xxxx xxxx	**Email address:**	xxxx xxxx
Organisation (if appropriate):	Employed by Gloucester Hospitals NHS FT and University of Southampton		
Role:	Senior Medical Statistician in Gloucester, Senior Enterprise Research Fellow in Southampton		

Do you consent to your name being published on the UK NSC website alongside your response? **No**		
Section and / or page number	**Text or issue to which comments relate**	**Comment** *Please use a new row for each comment and add extra rows as required.*
Front page	List of authors	I have worked in the area of diabetic retinopathy for 33 years and am surprised that I do not recognise any of the names on the list. There is a move to have experts on doing reviews doing most of the work whereas I would go for 50% who know a lot about the subject and 50% who know how to do a systematic review. I was involved in another systematic review, about screening intervals. I left the group because the systematic review professionals did not believe me about some of the papers they cited having mistakes in them making them unusable. So the final report came to the wrong conclusions. Since then the first thing I do when looking at a systematic review is to check all the authors. Sadly the tendency to use the review professionals seems to be unstoppable.

Throughout	Throughout One example, page 6. “With” spelt wrongly – would have been picked up by spell-check. Figure 4 Proposed EDESP modification, replacing level 1 graders wtih an automated system (Tufail 2016 (1)) ..	Spelling not checked. Seems to be the case that neither PHE nor NHSE documents are proof-read or spell-checked. With 9 authors there should not have been any mistakes at all. Very disappointing. Have previously sent list to Goda.
Throughout	Poorly written.	Many sentences are over 6 lines long. This makes it difficult to read and to understand what is going on.
Title	Automated grading to replace level 1 graders	Seems like a narrow definition – should also have covered automatic grading to benchmark graders and to benchmark programmes.
Page 8		Not true that “more experienced graders” do the second grading. It is important that DR is not missed so it is better to use experienced graders as first line graders, not as second line. This is why benchmaking should have been included as a way of using automated analysis.
Page 8	Patients with diabetic retinopathy need to have further assessment and treatment.	WRONG. Those with background retinopathy in either or both eyes (R1) are also referred back for rescreening in one year, they do not need further assessment and certainly not treatment. Few of those with DR need to have treatment. Most people with maculopathy or R2 level DR are kept in in surveillance clinics.

Page 10	(excluding women who have only gestational diabetes) a	"Only" should be removed. Gestational diabetes makes pregnancy more difficult and has impacts on the baby and increases the risk of diabetes for the mother later.
Page 15	All these sources of variation are likely to affect the performance of ARIASs, especially when the system is evaluated away from the setting in which it was developed and initially evaluated.	", especially" should be removed
Page 16	Confusion as to what level 1 graders are doing	They are looking for ANY retinopathy. They may be grading for M1, R2 or R3 but they are also looking for R1. We need to know who has R1 because of risk stratification for extended screening intervals. So looking at sight threatening or referable DR is not correct. We want ANY microanuerysms to be detected.
	There is no direct evidence on the overall impact of ARIAS (including the impact on human graders) but limited evidence from Scotland and Portugal suggest that the risk is low, the performance of ARIAS remains high after implementation, provided robust	We have no independent evidence from Scotland. None of the evaluations looks at the implications of missing R1 in the context of extended screening intervals.

	programme is in place, and the use of ARIAS is likely to increase with time.	Is robust internal and external quality assurance of the ARIAS or of the human graders?
	Should look at outcomes beyond accuracy, such as the actual consequences of false negative and false positive results and the consequences of accidental findings (e.g. missed by ARIAS but referred by human graders).	Any mention anywhere of non-diabetic eye disease? Can include cataract, cancer, melanoma.
Page 24	These estimates will soon be revised in an upcoming systematic review (7).	**This will not give any useful data because the methods employed to measure DR are not comparable.** Those with mydriasis will find more than those without mydriasis. The different imaging methods (number of fields) will probably not be adjusted for. There will be as many teams of graders as there are papers and they will not be quality assured and trained to the same level (i.e. each of them will be using a different assay on samples obtained in different ways on population samples determined in different ways). We don't even know why prevalence differs between programmes in England never mind across the wider world.
Table 1	Error	Wales does not used extended screening intervals.
	The diagnostic test for DMO is Optical Coherence Tomography (OCT).	May also be examination by ophthalmologist.
	These photographs are graded as follows, in order of progression:	Not clear that images are being graded in 2 scales, the R scale and the M scale.

		Possibilities are R0M0 R1M0 R1M1 R2M0 R2M1 R3M0 R3M1 R0M1 is not possible.
Figure 2	Reproduced very poorly	Difficult to read, especially grey on yellow.
Page 27	Patients who are graded R0M0 or R1M0 in the more severely graded eye are invited to return for rescreen after 12 months.	Changing to 24 months. Implications of missing R1 needs to be included in event of 2 year screening being implemented for people with R0.
Figure 3	Reproduced very poorly	
Page 47	The authors estimated that approximately 50% of all screening episodes would require further human grading (which ranged from 47% to 51% across the 3 centres) and will not result in an increased workload	The proportion of image sets with no DR is about 60% in most programmes so 40% need second grading which is less than the 50% here. More work needs to be done here to make the results transparent.

	for the secondary grader while the workload of the tertiary grader (arbitration) is likely to reduce (19).	
Table 8	R1 5.9% (95%CI 84.1–87.5)	Wrong. Should be 85.9% (95%CI 84.1–87.5)

Additional comments received in the email:

The overall conclusion seems to be "More research needed".

If this is to be carried out then I would respectfully suggest that a team should be put together through the Research Advisory Group and include people who have a thorough understanding of EDESP, know something about AI and the implications of sensitivity and specificity as they will be when extended screening intervals are introduced. The sensitivity will not change but the specificity will as there will be fewer people without any DR.

Very disappointed that the document does not seem to have been proof-read or even put through a spell-check programme.

3. **Steve Aldington (Retinopathy Research and Professional Development Manager)**

Feedback received via email:

Good afternoon xxxx xxxx and xxxx xxxx,

apologies xxxx xxxx, as we have never met.

UK National Screening Committee

I am a long-time colleague of xxxx xxxx from years previous when we first worked on xxxx xxxx.

I await the outcome of this Consultation with great interest.

In addition to xxxx xxxx excellent points about the spelling and punctuation errors in the document (xxxx xxxx being renowned for attention to detail), could I also point out that the 16 uses of the term 'EURODIAB criteria' or similar are never once referenced.

As the first author on the EURODIAB methodology paper and effectively the inventor of the criteria used for imaging and assessment in EURODIAB (and other studies), I find it somewhat irksome (to say the least) to have the terms bandied about in an official publication without due recognition of the team who put these original criteria together. The EURODIAB protocol and methodologies provided the cornerstone for much diabetic eye screening around the world.

Equally however, the (Pat) Wilkinson paper reporting the ICDR classification scale was also not referenced, even though these criteria were cited 43 times in this publication. At least the ICDR (and ETDRS) classifications ('severity scales') were included in the list of abbreviations whereas EURODIAB was not.

For EURODIAB not to be recognised and appropriately referenced, when the criteria are used so readily (16 times) as a benchmark in this publication, is a serious omission, in my opinion.

best wishes

Steve

UK National Screening Committee

4. Consultant diabetologist

Name:	Ansu Basu	**Email address:**	xxxx xxxx
Organisation (if appropriate):	Association of British Clinical Diabetologists		
Role:	**Consultant diabetologist** **Involved in the early studies of diabetic retinopathy screening using ANN (artificial neural networks) as a part of MD thesis at Newcastle University 2005**		

Do you consent to your name being published on the UK NSC website alongside your response?

Yes

Section and / or page number	Text or issue to which comments relate	Comment *Please use a new row for each comment and add extra rows as required.*
Question 1	Discussions around question 1	There is some confusion on how you define DL and ML here. DL is a subset of ML. DL uses more hidden layers and DL-based ARIAS are therefore inherently likely to perform better
Page 68	“Depending on whether the system is used to screen out patients with no disease or to differentiate between ‘referable’ and ‘non-referable’ cases, the specificity and the respective	Extremely important and very valid

	workload reduction that could be expected will vary, and is likely to have an impact on the cost-effectiveness of the system."	
Page 78	Review Summary	Completely agree to what has been said

General additional comments

- Significant heterogeneity among clinical studies involving diabetic retinopathy – patient selection, mydriatic vs. non-mydriatic, number of fields imaged, imaging camera used, design of study – comparator, grading protocol used, assessor – ophthalmologist or non-ophthalmologist, the retinal pathology in the cohort etc.
- DL technically is a superior platform as it has a number of hidden layers in the ANN, so my suggestion would be to do a trial using EyeArt 2.1 in the large dataset from EDESP and independent of software company

5. A joint response from a group of experts including Statistical Epidemiologist, Consultant Ophthalmologists, Clinical Lead in Diabetes and Endocrinology, Epidemiologist

Name:	Professor Alicja Rudnicka (1) Professor Adnan Tufail (2) Dr Cathy Egan (2) Dr John Anderson (3)	**Email address:**	XXXX XXXX XXXX XXXX XXXX XXXX XXXX XXXX XXXX XXXX

	Professor Chris Owen (1)		
Organisation (if appropriate):	(1) Population Health Research Institute, St George's, University of London (2) Moorfields Eye Hospital NHS Foundation Trust (3) Homerton University Hospital NHS Foundation Trust		
Role:	**Statistical Epidemiologist (AR)** **Consultant Ophthalmologists (AT & CE)** **Clinical Lead in Diabetes and Endocrinology (JA)** **Epidemiologist (CO)**		

Do you consent to your name being published on the UK NSC website alongside your response?

YES

Section and / or page number	**Text or issue to which comments relate**	**Comment** *Please use a new row for each comment and add extra rows as required.*
P20	Based on the synthesis of evidence against the UK NSC criteria EyeArt v2.1 has consistently high sensitivity, comparable to that of human graders, and could safely be implemented in the EDESP, either as a replacement of level 1 human graders or as a filter before manual grading. It has been shown that the system is cost-effective with either of these	We are in total agreement that there is sufficient evidence for **staged implementation** of selected ARIAS now into the NHS DR screening programme (please see below). A health economic analysis in an English DESP setting already exists,[1 2] which showed considerable cost savings compared to pure manual grading. Since newer versions of

	strategies, although the analyses need updating to reflect the higher performance of the new version; to capture the long-term impact of the system, and to investigate the effect of using different decision thresholds, 'disease/no-disease' vs. 'referable/non-referable' disease. RetmarkerSR (ML-based) also has been shown to have high accuracy (but lower sensitivity than EyeArt v1) and to be cost-effective in the EDESP. Although there is published evidence of its high performance as implemented in the Portuguese DESP, the evidence base is more limited and there is only one UK-based study …	ARIAS algorithms are likely to be more sensitive and specific, it is reasonable to assume therefore that potential cost savings will be even greater.
P21	"….and to investigate the effect of using different decision thresholds, 'disease/no-disease' vs. 'referable/non-referable' disease"	Clarification of disease vs no disease threshold cut-point in evaluation of studies, as opposed to referable vs non-referable disease in studies included in the evaluation: The document is inconsistent in evaluating the different operating thresholds which have clinical and operation importance. A safe initial introduction of ARIAS would be to use the 'disease vs no disease' operating threshold.
P79	This means that out of the 10 ARIASs included in this review, we have applicable high-quality evidence for 3 systems: EyeArt v2.1 (DL), iGradingM (ML) and RetmarkerSR (ML). Evidence from multiple studies show that EyeArt has consistently high sensitivity	We agree that there is sufficient evidence for staged implementation of selected ARIAS now into the NHS DR screening programme, using the disease vs no-disease threshold cut-point. We feel that the conclusions should more

	(~90%) comparable to that of human graders and could be used as an initial screen in the EDESP	strongly recommend staged implementation. The rationale for changing the strength of recommendation are as follows: The evidence review has shown that studies already exist for certain ARIAS algorithms that; • Are of high quality and sufficient size. • Are independently evaluated from the vendors. • Show sufficiently safe sensitivity and specificity for clinical deployment. • Are evaluated on a dataset that reflect real-life NHS implementation.
P79	The HTA conducted by Tufail et al also showed that implementing the system either as a replacement of level 1 graders or as a filter prior to manual grading is cost-effective. Given that the new version of the system evaluated in Heydon 2020 has comparable sensitivity and higher specificity, its implementation could lead to even greater savings than those reported by Tufail et al (1, 19, 38).	We agree. Health economic analysis already exists,[1 2] which showed the potential for considerable cost savings compared to pure manual grading. As mentioned above, it is reasonable to assume that such cost savings will be even greater with newer versions of ARIAS algorithms, which are likely to be more sensitive and specific. Given the demonstrated safety, it would be remiss not to introduce such cost savings now. However, data for an updated health economic analysis can be gathered during a staged implementation phase of ARIAS in EDESP to confirm on-going cost effectiveness (please see below).

P79	Although there is no direct evidence about the overall impact that the implementation of EyeArt could have on the EDESP, Heydon et al estimated that using referable disease as a threshold, approximately 50% of all screening episodes would require further human grading and this will not result in an increased workload for the secondary grader	Heydon[3] directly assessed that 50% of episodes would be test negative and would not require manual grading. Overall impact on EDSEP grading pathways would now be best assessed with staged implementation.
P79 & P81	There is also high quality evidence of the performance of iGradingM in the SDESP including data from an internal quality assurance assessment published in 2017 (15). The latter shows that the system is safe to use in clinical practice with sensitivity of 97%, comparable to that in the published evaluations…… …….We identified a considerable number of surveys looking at the perceptions, attitudes, concerns and educational needs of healthcare professionals and patients with respect the implementation of AI-based technology in screening programmes…. An evidence review of this growing literature will help identify all relevant aspects of the above question, to summarise the existing evidence and identify any gaps that need to be addressed in future research	ARIAS for diabetic screening has been deployed in Scotland for almost 10 years with no major concerns raised by end-users about the “acceptability” of ARIAS vs purely human graders. However, this could be further assessed using a planned stage implementation of ARIAS in EDESP
P81	Future evaluations of ARIAS:	We strongly agree that this is an important process to adhere to for future evaluations of ARIAS.

	Should be done independently from the software developer, in the clinical setting in which the system is meant to be used, under conditions that reflect everyday clinical practice; if possible, they should compare the performance of alternative ARIASs that may have different advantages and disadvantages.	
P81-82	Section entitled: “Future research: Future evaluations of ARIAS”	We agree that further evidence would aid our understanding of optimal implementation of AI in areas such as: • Acceptability of AI in people with diabetes. • Updated Health Economic analysis. • Evaluation of future algorithms and performance on patient subgroups. • Experience and perceptions of healthcare professionals who interact with ARIAS. Further studies on current ARIAS algorithms that already meet accepted standards will not add to the understanding of sensitivity of specificity. However, additional gaps in knowledge would be best evaluated using a staged implementation of sufficiently high performing ARIAS. We have developed a protocol for staged implementation that was proposed and discussed previously with Diabetic Eye Screening Programme Research Advisory Committee (RAC) to PHE, which includes patient group representation (RAC

		meetings from 2017 onwards). We received iterative feedback from RAC, end users and vendors and our template will allow for staged implementation with appropriate safe-guards in place, before consideration of a national rollout.
P81	"..outcomes beyond accuracy, such as the actual consequences of false negative…	Remit of current DESP with purely manual grading is to detect diabetic eye disease, and this same remit should apply with the introduction of ARIAS and not for the detection of other eye disorders (akin to the Scottish system). Other checks within the current DESP pathway would continue (e.g., visual acuity measurements) and therefore the potential to detect other causes of vision loss would be retained.
Pp81-82	AI software is constantly evolving and this is one of its strengths. What is the best way to manage and monitor this, to make sure that the next version is safe, reliable and at least as effective as the previous one? Should the performance of ARIAS, once imple-mented, be monitored in the same way as DESP with human graders is monitored or different process is required? What is the experience with this in Scot-land and Portugal?	As mentioned above we have presented a pipeline/methodology for ongoing assessment of new ARIAS (CE marked or research systems) that would allow for improvement, innovation and competition, but most importantly maintain safety in the future.

References

1. 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(92):1-72. doi: 10.3310/hta20920
2. Tufail A, Rudisill C, Egan C, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology* 2017;124(3):343-51. doi: 10.1016/j.ophtha.2016.11.014
3. 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* 2020 doi: 10.1136/bjophthalmol-2020-316594

6. The Royal College of Ophthalmologists

RCOphth response to NSC on AI in DESP.pd