OPEN ACCESS

Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients

Peter Heydon ,[1] Catherine Egan,[1,2] Louis Bolter,[3] Ryan Chambers,[3] John Anderson,[3] Steve Aldington,[4] Irene M Stratton,[4] Peter Henry Scanlon ,[4] Laura Webster,[5] Samantha Mann,[5] Alan du Chemin,[5] Christopher G Owen ,[6] Adnan Tufail,[1,2] Alicja Regina Rudnicka [6]

For numbered affiliations see end of article.

Correspondence to
Alicja Regina Rudnicka, Population Health Research Institute, St George's, University of London, London, UK; arudnick@sgul.ac.uk

Received 28 April 2020
Accepted 28 May 2020
Revised 13 May 2020

To cite: Heydon P, Egan C, Bolter L, *et al. Br J Ophthalmol* Epub ahead of print: [*please include* Day Month Year].doi:10.1136/bjophthalmol-2020-316594

ABSTRACT

Background/aims Human grading of digital images from diabetic retinopathy (DR) screening programmes represents a significant challenge, due to the increasing prevalence of diabetes. We evaluate the performance of an automated artificial intelligence (AI) algorithm to triage retinal images from the English Diabetic Eye Screening Programme (DESP) into test-positive/technical failure versus test-negative, using human grading following a standard national protocol as the reference standard.

Methods Retinal images from 30 405 consecutive screening episodes from three English DESPs were manually graded following a standard national protocol and by an automated process with machine learning enabled software, EyeArt v2.1. Screening performance (sensitivity, specificity) and diagnostic accuracy (95% CIs) were determined using human grades as the reference standard.

Results Sensitivity (95% CIs) of EyeArt was 95.7% (94.8% to 96.5%) for referable retinopathy (human graded ungradable, referable maculopathy, moderate-to-severe non-proliferative or proliferative). This comprises sensitivities of 98.3% (97.3% to 98.9%) for mild-to-moderate non-proliferative retinopathy with referable maculopathy, 100% (98.7%,100%) for moderate-to-severe non-proliferative retinopathy and 100% (97.9%,100%) for proliferative disease. EyeArt agreed with the human grade of no retinopathy (specificity) in 68% (67% to 69%), with a specificity of 54.0% (53.4% to 54.5%) when combined with non-referable retinopathy.

Conclusion The algorithm demonstrated safe levels of sensitivity for high-risk retinopathy in a real-world screening service, with specificity that could halve the workload for human graders. AI machine learning and deep learning algorithms such as this can provide clinically equivalent, rapid detection of retinopathy, particularly in settings where a trained workforce is unavailable or where large-scale and rapid results are needed.

Diabetic eye disease is a microvascular complication of Type 1 and Type 2 diabetes, and is a leading cause of incident blindness in people of working age. Early detection through screening programmes for diabetic retinopathy (DR) with digital retinal imaging can reduce the impact of this condition.[1] National screening programmes for DR, including the National Health Service (NHS) Diabetic Eye Screening Programme (DESP), represent a major challenge to healthcare providers. The incidence and prevalence of diabetes mellitus are increasing; 425 million adults were estimated by the International Diabetes Federation to be living with diabetes in 2017, a prevalence that has doubled since 1980 and is projected to rise to 629 million by 2045.[2] Screening for DR is generally based on human grading, which is labour intensive, requiring human graders, who should be trained, undergo regular quality assurance and be retained. Applying the current UK annual screening protocol globally would require 2.2 billion retinal images to be graded in 2030. Emerging automated retinal image analysis systems (ARIAS) are artificial intelligence (AI) (machine learning) algorithms[3 4] that may provide cost-effective alternatives to human grading, designed to have a high sensitivity for detection of sight-threatening retinopathy in need of clinical intervention. These systems could be used to triage those who have sight-threatening DR or other retinal abnormalities, from those at low risk of sight-threatening retinopathy.

This study expands on our previous study published in 2016–2017[5 6] that quantified the screening performance and diagnostic accuracy of three ARIAS using NHS DESP human grading as a reference standard. Two of the ARIAS achieved acceptable sensitivity when compared with human graders and had specificities that made them cost-effective alternatives to human grading alone. We reported how such software could be incorporated into pre-existing screening pathways[6] and that replacement of a primary human grader was the most cost-effective strategy, compared with a strategy that used the ARIAS as a filter prior to primary human grader (figure 1 shows the pathway on which cost-effectiveness analyses were performed.[6])

This study provides a large prospective evaluation using an updated version of one of the ARIAS in three current DESPs conducted in North East London (NEL), South East London (SEL) and Gloucestershire (GS) during a different time period

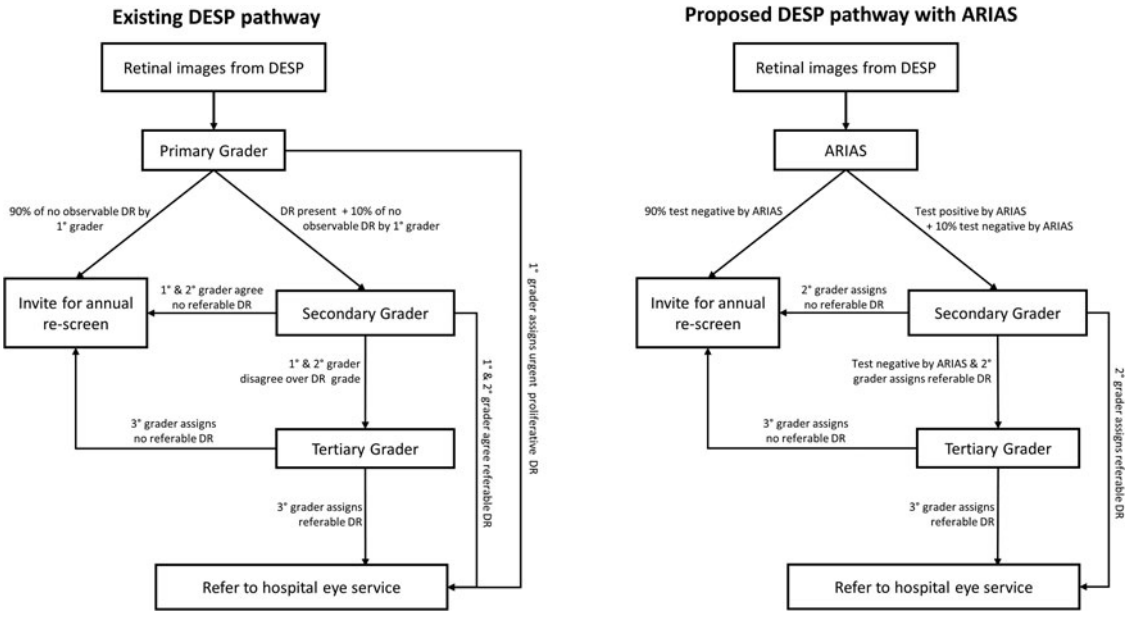

Figure 1 Decision tree model used to calculate the incremental cost-effectiveness of human grading versus replacing initial grading undertaken by human graders (primary graders) with automated retinal image analysis systems (ARIAS).[5 6] The clinical practice pathway as illustrated reflects the pathway and grading[23] that was used in previous cost-effectiveness analyses.[5 6] Non-referable retinopathy refers to human grades of no observable retinopathy (R0) and mild-to-moderate non-proliferative retinopathy (R1) and non-referable maculopathy (M0). Referable retinopathy refers to referable maculopathy (M1), moderate-to-severe non-proliferative retinopathy (R2) and proliferative retinopathy (R3) and ungradable images (U).[8 9] DESP, Diabetic Eye Screening Programme; DR, diabetic retinopathy.

to the first study. Therefore, different populations with diabetes, different human grading teams and different cameras are contributing to this evaluation of a single algorithm. In its current form, the ARIAS is not considered precise enough to be a diagnostic test capable of replacing human grading altogether. The updated version of the ARIAS evaluated in this study has now received CE marking (Conformité Européene marking indicates that it is compliant with European Union product legislation and meets safety, health or enviromental requirements) and is in its final stages of Food and Drug Administration evaluation. However, no large independent evaluation on a UK or other screening population has yet been undertaken. The aim of this study is to evaluate whether the ARIAS is sensitive and specific enough to successfully triage patients into low- and medium-/high-risk DR cases, thereby potentially reducing the need for all screening episodes to receive human grading, allowing resources to be focused on high-risk screening episodes. Hence, the aim is to evaluate ARIAS in combination with manual grading, rather than to replace all manual grading.

METHODS

Three screening programmes contributed to the study, DESPs of NEL, GS and SEL. Programme activity and patient demographic characteristics for the 2017–2018 period are presented in table 1. The rate of uptake of the screening invitation ranged from 77% to 83%. Sex differences were similar, with more men being screened in each centre reflecting the predominance of males with diabetes. There was a higher proportion of Asian and black patients screened within the London centres, which reflects the local population with diabetes.

In the current screening pathway, all retinal images are reviewed by a primary grader, and any patients with mild or worse retinopathy or maculopathy (in addition to the 10% graded 'no retinopathy') are reviewed by a secondary grader, with discrepancies between the primary and secondary grader reviewed by an arbitration grader (tertiary grader).[7] A sample size commensurate with our previous work[5 6] was sought. Hence, our target was to obtain retinal images from approximately 10 000 consecutive screening episodes at each centre that had been graded for retinopathy level by the local team of graders as part of standard care. At least two digital image fields were taken of each eye, one centred on the optic disc and the other on the macula, in accordance with NHS

Table 1 Activity and patient characteristics across three screening centres for the 2017–2018 period

	NEL DESP	GS DESP	SEL DESP
Number of patients invited	106 257	30 658	77 086
Number of patients examined (participation rate %)	88 519 (83%)	23 697 (77%)	62 738 (81%)
Ethnicity of patients (%)			
White	35.8	66.9	50.8
Asian	43.0	2.5	13.2
Black	15.4	1	26.3
Mixed	1.2	0.5	2.1
Not-specified	4.6	29.3	7.6
Gender			
Male (%)	54	57	54.4
Female (%)	46	43	45.6
Not specified (%)	<0.1		<0.1
Age			
Mean (range) in years	60 (12–107)	66 (12–104)	61 (12–107)
50 years or older (%)	79.5	87.1	83.3

DESP, Diabetic Eye Screening Programme; GS, Gloucestershire; NEL, North East London; SEL, South East London.

DESP protocol.[8] Images that were of poor quality or classified as *ungradable* by human graders were included in the data set. The data set included all images captured for each eye for each episode, including partial retinal images, poor quality images and non-retinal images such as cataracts. Hence, this study analysed 30 000 real live episodes with images captured as part of routine clinical care without any editing or selection of images prior to processing with ARIAS, EyeArt. Research Governance approval for the study was obtained. Images were pseudonymised, and no change in the clinical pathway occurred.

The CE-marked software complied with relevant European legislation, EyeArt v2.1.0 (Eyenuk, Woodland Hills, CA), was used to assess retinal images for each person for the presence of DR. The software was installed on 3 January 2017 on two servers within a secure computer facility at the Homerton Hospital. Pseudo-anonymised image data were brought to the secure computer facility from the three sites using encrypted disks and processed. Images from NEL were processed on 14 January 2017, from SEL 6 February 2017 and from GS 25 April 2017. The patient clinical pathway was not affected by the processing of images by the machine learning algorithm as this process was undertaken in parallel with usual care. The human grade for each image was not available while the images were being processed by the EyeArt software. EyeArt classification for the presence of retinopathy for each person was either *test-positive or test-negative*. Episodes with images found to be un-assessable, that is, technical failure by the software, were allocated as *test-positive*, on the basis that these cases would need further human assessment.

The human grading was undertaken in each programme using NHS DESP guidelines.[8 9] The final human grade in the worst eye was used as the reference standard for all analyses presented. We have previously shown that the final human grade is a stable reference standard since measures of screening performance and accuracy of performance metrics were not materially altered when the reference standard was further refined by arbitration by an expert reading centre.[5 6] Human grading classifications were, no observable retinopathy (R0), mild-to-moderate non-proliferative retinopathy (R1), non-referable maculopathy (M0), ungradable images (U), referable maculopathy (M1), moderate-to-severe non-proliferative retinopathy (R2) and proliferative retinopathy (R3). The commensurate ETDRS retinopathy grade scores are as follows[9 10]: R0 equivalent to 'no apparent retinopathy'; R1 ETDRS scores 20–35 inclusive; R2 ETDRS scores 43–53 inclusive; R3 ETDRS scores 61+. A more detailed description of NHS DESP grades alongside ETDRS final retinopathy grade is available.[9] In the English DESP, retinopathy grades R0M0, R1M0 are non-referable retinopathy and grades M1, R2 and R3 are referable retinopathy. Patients with ungradable images are referred for slit-lamp biomicroscopy within the DESP. Patients with referable retinopathy are then referred to the hospital eye service. Patients with non-referable retinopathy receive an invitation for re-screening within 1 year.

Statistical analysis

All analyses were performed using STATA 15.0 IC (STATA Corps LP, College Station, TX USA). Estimates of screening performance and diagnostic accuracy were determined (sensitivity/detection rate, false-positive rates, specificity) and the corresponding 95% CIs for each centre and overall were also determined. Screening performance estimates are given for each grade separately as well for referable versus non-referable retinopathy. Logit-transformed 95% CIs were determined or binomial exact CIs in the presence of detection rates of 100%. Systematic differences between centres in the likelihood of screening outcome being classified as test-positive versus test-negative, conditional on human retinopathy grade, were examined using a χ^2 test.

RESULTS

The number of consecutive screening episodes with complete human grading are given in table 2 for each centre and overall. Each centre contributed at least 10 000 consecutive screening episodes. The prevalence of different retinopathy grades was similar in NEL and SEL with marginally lower prevalence of retinopathy grades R1M1, R2 and R3 (and consequently a higher proportion of R1M0 and R0 M0) in GS compared with the other two centres. The proportion of screening episodes that were ungradable ranged from 1.9% in GS DESP up to 3% at the NEL DESP.

Table 3 presents the EyeArt classification of test-negative and test-positive for each screening episode by centre and overall. For episodes manually graded as R0M0 (no retinopathy), the specificity (EyeArt classification of test-negative) overall is about two-thirds, ranging from 65% in SEL to 71% in GS. A high proportion (from 84% to 93%) of those with mild-to-moderate non-proliferative retinopathy and no-referable maculopathy (human grade R1M0) were classified as test-positive by EyeArt software. All cases of moderate-to-severe non-proliferative retinopathy and proliferative retinopathy (human grades R2 and R3, with or without macular involvement) were detected by the software as test-positive or 'technical failure', that is, a sensitivity of 100%.

Table 4 provides measures of diagnostic accuracy in terms of the 95% confidence limits (CI) around the estimates for each centre and overall for episodes classified as test-positive (including technical failures) by the software.

The sensitivity (95% CIs) of EyeArt was 95.7% (94.8% to 96.5%) for referable retinopathy (human graded ungradable, referable maculopathy, moderate-to-severe non-proliferative or proliferative: grades U, M1, R2 and R3). This figure is composed of sensitivities of 98.3% (97.3% to 98.9%) for mild-to-moderate non-proliferative retinopathy with referable maculopathy (R1M1), 100% (98.7% to 100%) for moderate-to-severe non-

Table 2 Prevalence of retinopathy based on the final human grade in the worst eye for each centre and for all three centres combined

Final human retinopathy grade	Number of episodes in each DESP (column %)			Combined (column %)
	NEL	GS	SEL	
R0M0	7031 (69.4%)	6867 (68.1%)	7414 (72.9%)	21 312 (70.1%)
R1M0	2252 (22.2%)	2645 (26.2%)	1994 (19.6%)	6891 (22.7%)
R1M1	368 (3.6%)	287 (2.8%)	346 (3.4%)	1001 (3.3%)
R2	127 (1.3%)	50 (0.5%)	113 (1.1%)	290 (1%)
R2M0	38 (0.4%)	24 (0.2%)	31 (0.3%)	93 (0.3%)
R2M1	89 (0.9%)	26 (0.3%)	82 (0.8%)	197 (0.6%)
R3	57 (0.6%)	49 (0.5%)	66 (0.6%)	172 (0.6%)
R3M0	28 (0.3%)	23 (0.2%)	20 (0.2%)	71 (0.2%)
R3M1	29 (0.3%)	26 (0.3%)	46 (0.5%)	101 (0.3%)
Ungradable	302 (3%)	193 (1.9%)	244 (2.4%)	739 (2.4%)
Total	10 137	10 091	10 177	30 405

Retinopathy grades: No retinopathy (R0); mild-to-moderate non-proliferative retinopathy (R1); non-referable maculopathy (M0); ungradable images (U); referable maculopathy (M1), moderate-to-severe non-proliferative retinopathy (R2) and proliferative retinopathy (R3).[8 9]
DESP, Diabetic Eye Screening Programme; GS, Gloucestershire; NEL, North East London; SEL, South East London.

Table 3 Screening performance of EyeArt software compared with the final human grade in the worst eye for each centre and for all three centres combined

	EyeART classification (row % within each centre)							
	NEL		GS		SEL		Combined (row %)	
Final human grade	Test-negative	Test-positive	Test-negative	Test-positive	Test-negative	Test-positive	Test-negative	Test-positive
Retinopathy grades								
R0M0	4787 (68.1%)	2244 (31.9%)	4891 (71.2%)	1976 (28.8%)	4793 (64.6%)	2621 (35.4%)	14 471 (67.9%)	6841 (32.1%)
R1M0	184 (8.2%)	2068 (91.8%)	431 (16.3%)	2214 (83.7%)	133 (6.7%)	1861 (93.3%)	748 (10.9%)	6143 (89.1%)
R1M1	4 (1.1%)	364 (98.9%)	8 (2.8%)	279 (97.2%)	5 (1.4%)	341 (98.6%)	17 (1.7%)	984 (98.3%)
R2	0 (0%)	127 (100%)	0 (0%)	50 (100%)	0 (0%)	113 (100%)	0 (0%)	290 (100%)
R2M0	0 (0%)	38 (100%)	0 (0%)	24 (100%)	0 (0%)	31 (100%)	0 (0%)	93 (100%)
R2M1	0 (0%)	89 (100%)	0 (0%)	26 (100%)	0 (0%)	82 (100%)	0 (0%)	197 (100%)
R3	0 (0%)	57 (100%)	0 (0%)	49 (100%)	0 (0%)	66 (100%)	0 (0%)	172 (100%)
R3M0	0 (0%)	28 (100%)	0 (0%)	23 (100%)	0 (0%)	20 (100%)	0 (0%)	71 (100%)
R3M1	0 (0%)	29 (100%)	0 (0%)	26 (100%)	0 (0%)	46 (100%)	0 (0%)	101 (100%)
U	29 (9.6%)	273 (90.4%)	22 (11.4%)	171 (88.6%)	27 (11.1%)	217 (88.9%)	78 (10.6%)	661 (89.4%)
Total	5004	5133	5352	4739	4958	5219	15 314	15 091

Retinopathy grades: No retinopathy (R0); mild-to-moderate non-proliferative retinopathy (R1); non-referable maculopathy (M0); ungradable images (U); referable maculopathy (M1), moderate-to-severe non-proliferative retinopathy (R2) and proliferative retinopathy (R3).[8 9]
GS, Gloucestershire; NEL, North East London; SEL, South East London.

Table 4 Screening performance measures along with 95% confidence limits for each centre and for all three centres combined

	Percentage classified as test-positive (including technical failure) by EyeArt (95% CI)*			
Final human grade	NEL	GS	SEL	Combined
R0M0	31.9 (30.8, 33.0)	28.8 (27.7, 29.9)	35.4 (34.3, 36.4)	32.1 (31.5, 32.7)
R1M0	91.8 (90.6, 92.9)	83.7 (82.2, 85.1)	93.3 (92.1, 94.3)	89.1 (88.4, 89.9)
R1M1	98.9 (97.1, 99.6)	97.2 (94.5, 98.6)	98.6 (96.6, 99.4)	98.3 (97.3, 98.9)
R2*	100.0 (97.1, 100.0)	100.0 (92.9, 100.0)	100.0 (96.8, 100.0)	100.0 (98.7, 100.0)
R2M0	100.0 (90.7, 100.0)	100.0 (85.8, 100.0)	100.0 (88.8, 100.0)	100.0 (96.1, 100.0)
R2M1	100.0 (95.9, 100.0)	100.0 (86.8, 100.0)	100.0 (95.6, 100.0)	100.0 (98.1, 100.0)
R3*	100.0 (93.7, 100.0)	100.0 (92.7, 100.0)	100.0 (94.6, 100.0)	100.0 (97.9, 100.0)
R3M0	100.0 (87.7, 100.0)	100.0 (85.2, 100.0)	100.0 (83.2, 100.0)	100.0 (94.9, 100.0)
R3M1	100.0 (88.1, 100.0)	100.0 (86.8, 100.0)	100.0 (92.3, 100.0)	100.0 (96.4, 100.0)
U	90.4 (86.5, 93.3)	88.6 (83.2, 92.4)	88.9 (84.3, 92.3)	89.4 (87.0, 91.5)

*95% CIs are binomial exact.
Retinopathy grades: No retinopathy (R0); mild-to-moderate non-proliferative retinopathy (R1); non-referable maculopathy (M0); ungradable images (U); referable maculopathy (M1), moderate-to-severe non-proliferative retinopathy (R2) and proliferative retinopathy (R3).[8 9]
GS, Gloucestershire; NEL, North East London; SEL, South East London.

proliferative retinopathy (R2), and 100% (97.9 to, 100%) for proliferative disease (R3) (table 4).

For non-referable retinopathy (human grade R1M0), 89.1% (88.4% to 89.9%) were classified as test-positive (this is the sensitivity/detection rate for non-referable retinopathy).

EyeArt agreed with the human grade of no retinopathy (R0M0) in 68% (67% to 69%). Hence, the specificity for no retinopathy (test-negative) is 68%; the equivalent expressed as a false-positive rate is 32.1% (31.5% to 32.7%) (table 4). If this group is considered in combination with non-referable retinopathy (R1M0), the specificity is 54% (53.4% to 54.5%) (equivalent to 46.0% (45.5% to 46.6%) expressed as a false-positive rate), which is the specificity corresponding to the detection rate of 95.7% for referable retinopathy given above.

Approximately 50% (15 091/30 405) of all screening episodes would require further human grading after EyeArt classification and this percentage ranged from 47% to 51% across the three centres.

The EyeArt likelihood ratio for a test-positive result for referable disease was stable across centres. There was some evidence of variation across the three centres in the proportion of episodes classified as test-negative for screening episodes with human retinopathy grades no retinopathy (R0M0) or mild-to-moderate non-proliferative retinopathy with no-referable maculopathy (R1M0) (interaction $p<0.001$). However, the maximum absolute difference in specificity between centres was only 3.6 percentage points.

DISCUSSION

This prospective study has demonstrated the high sensitivity of an AI-enabled algorithm to detect referable DR across three different UK screening centres, with a diverse ethnic mix of individuals with diabetes. Among 30 405 screening episodes, all 462 cases of moderate-to-severe non-proliferative retinopathy (human grade R2) and proliferative DR (human grade R3) were classified by EyeArt as test-positive (including technical failures) and would

have been sent for human grading. Using the EyeArt system to triage screening episodes (rather than replace manual grading altogether) could halve the workload for human graders. This report used a methodology framework for independent performance evaluation[5] that could be applied to future evaluations of other algorithms developed for this purpose.

Strengths of the current study include the large sample size, based on patients from real-world screening environments within three current DESPs, and evaluated independently of any commercial partner. Although there was a small difference in the performance of the EyeArt software across the three centres, mainly in terms of specificity for non-referable retinopathy, this is not surprising given there will be systematic differences between the three DESPs in terms of the patient- and centre-specific characteristics, such as age, ethnicity profiles (fundus pigmentation may influence software performance)[5], quality of image capture and variation in human grading (ie, random error or systematic grader-bias), which is more common with lower grades of retinopathy.[11] Although our previous work has shown that screening performance of two ARIAS appeared to be robust to variations in age, ethnicity and camera type, this was only within one DESP, which is likely to show less variation than across different DESPs.[5 6] It is noteworthy that detection rates were high and likelihood ratios for referable retinopathy were stable across centres. The sensitivity of EyeArt is high for detection of sight-threating retinopathy and exceeds any published AI algorithm to date.[12] Importantly, the specificity is sufficient to make an even greater cost-saving to the NHS than previously described.[5]

Automated screening software, including recent developments using machine learning, has been available for some time,[13] but independent, large-scale validation of commercially available licences[14] has been limited until recently.[5 6] Population screening programmes have not routinely used automation for retinopathy detection, with the exception of the Scottish national programme, which uses a computer program to triage macula-centred images (one per eye) into presence or absence of disease.[15] However, in our previous work, we showed that this system could not operate on disc-centred images as part of the English DESP.[5]

The methods previously described[5] have been adopted by other groups, when validating new approaches to automated assessment of retinal images.[16] This latter work tested the algorithm on a data set of over 10 000 retinal images. Our current work tests the software on over 120 000 retinal images (30 000 screening episodes with a minimum of four retinal images per episode and on average five images per patient).

A recent evaluation of this software on the EyePACS telescreening programme by the software developer achieved lower levels of sensitivity for referable DR of 91% vs 95.7% in the current study, but with higher levels of specificity of 91% for 'non-referable retinopathy'.[17] Differences in EyeArt performance between this study and ours are likely to result from differing software thresholding cut-offs being implemented as well as differences in image capture systems, image quality and human grading. Hence, there is an ongoing need to evaluate such systems locally to evaluate the impact on estimates of screening performance of any ARIAS.

Diabetes is recognised as a health challenge for every country. Good glycaemic and blood pressure control reduce the risk of incident diabetic eye disease, but regular screening for DR is necessary to detect sight-threatening diabetic eye disease so that appropriate treatment can be given to prevent vision loss. The UK has one of the largest, systematic DESPs in the world, with well-documented procedures to train and quality assure human graders. However, there are recognised variations in the ability of different countries and healthcare models to provide screening and treatment for sight-threatening complications of DR.

The majority of those with diabetes will have a very low risk of vision loss (either no retinopathy or mild non-proliferative retinopathy). Performance of human graders can vary at this end of the spectrum, without harming the patient. This is because an image with no retinopathy or an image with a single microaneurysm (mild retinopathy) will have the same outcome for the patient, that is, a routine review in 1 year (in the UK), or even less frequently in some countries with longer screening intervals for low-risk disease. Nonetheless, vast numbers of grading hours are devoted to this repetitive task, reducing the time available for grading high-risk images. Machines can address the problem of repeated grading. We have shown one such machine-learning software could halve the amount of DR screening images requiring human grading without missing sight-threatening disease.

Two other measures of screening performance worth mentioning are the positive predictive value and negative predictive value. Among all test-positive results from EyeArt, the probability of any retinopathy was 55% and 14% had referable retinopathy according to human grading. Among all test-negative results from EyeArt when compared with human grader, 94.5% did not have any observable retinopathy, 99% did not have referable retinopathy.

A previous version of the software (v1) with poorer specificity was shown to considerably reduce the cost of screening.[5 6] Figure 1 demonstrates the clinical screening pathways that were in use at that time, and the associated cost-effectiveness analysis evaluated a potential strategy of implementing ARIAS into the screening pathway by replacing the primary grader. Hence, all test-positive results from ARIAS (~50% of all screened) would immediately pass to the secondary grader. This is unlikely to result in an increased workload for the secondary grader, because in the existing fully manual grading pathways we have observed, between 40% and 50% of all screening episodes pass from the primary to the secondary grader across the three NHS DESPs included in this study. In addition, about 10% pass to the tertiary grader for arbitration. Under the proposed pathway, only screening episodes classified at test-negative by the ARIAS but as referable retinopathy by the secondary grader would pass to tertiary grader for arbitration. In reality, this will be a small proportion, since only 10% of test-negatives pass to the secondary grader, of which its estimated that 1.7% (tables 3, 4) would include potentially referable disease. This would equate to approximately 50 cases per 30 000 screening episodes. Hence, the workload for the tertiary (arbitration) grader would reduce. The implementation of ARIAS as outlined in the figure could potentially save £0.5 million per 100 000 screening episodes, offering huge savings in relation to over 2.2 million screening episodes per year (2017–2018) in England alone, with numbers increasing.[18] We also demonstrated that this was more cost-effective than using the ARIAS as a filter prior to human grading by level 1 graders.[5 6] The performance of the software (v2.1.0) in this current study has improved since the original report and we expect, therefore, that the cost savings will also have improved.

The number of people diagnosed with diabetes globally is estimated at 625 million by 2045,[2] which would generate between 2 and 3 billion retinal images per year following an English screening programme approach. Current evaluations predict that diabetes-associated blindness is likely to rise dramatically in the developing world.[19] Given the numbers of people with diabetes, costs and quality of eye care will become even more important. The use of AI

represents a new avenue for DR screening,[13 16–20] and the use of neural networks has also demonstrated promise in staging DR and triaging other retinal conditions.[21 22] This technology could be extended to screening programmes in developed and developing nations. The potential for nearly instantaneous triage at the point of image capture has not yet been fully explored but is another area where this technology could enhance diabetes eye care. Integration of ARIAS with OCT (for detection of macular oedema) at image capture is another potential pathway. All systems should undergo rigorous independent evaluation before changes are made to existing DR screening pathways.

Author affiliations
[1]Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, UK
[2]Institute of Ophthalmology, UCL, London, UK
[3]Homerton University Hospital NHS Trust, London, UK
[4]Gloucestershire Hospitals NHS Foundation Trust, Cheltenham, UK
[5]Guy's and Saint Thomas' NHS Foundation Trust, London, UK
[6]Population Health Research Institute, St George's, University of London, London, UK

Acknowledgements Thanks to Steve Chave for preparing the Gloucestershire data set.

Contributors CE, CGO, AT and ARR designed the study and raised funding. PH, CE, LB, RC, JA, SA, IS, PS, LW, SM, AdC and AT were involved in collection of data for the study. ARR undertook data management and analysed the data; the writing committee (PH, CE, CGO, AT and ARR) wrote the first draft of the report, which was critically appraised by all authors. The final draft was approved by all authors. ARR is responsible for data integrity.

Funding This research has received a proportion of its funding from the Department of Health's NIHR Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and UCL Institute of Ophthalmology. The views expressed in the publication are those of the authors and not necessarily those of the Department of Health. Diabetes prevention research at St George's, University of London, is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) (grant reference NIHR200152).

Competing interests None to declare.

Ethics statement According to the UK Governance Arrangements for Research Ethics Committees (GAfREC), ethical review is not required for anonymised data. This was confirmed by the Caldicott Guardian, who is responsible for protecting the confidentiality of patients seen in the National Health Service (NHS).

Provenance and peer review Not commissioned; internally peer reviewed.

Data sharing statement For general data sharing inquiries, contact Professor Alicja Rudnicka arudnick@sgul.ac.uk.

ORCID iDs
Peter Heydon http://orcid.org/0000-0001-7029-4188
Peter Henry Scanlon http://orcid.org/0000-0001-8513-710X
Christopher G Owen http://orcid.org/0000-0003-1135-5977
Alicja Regina Rudnicka http://orcid.org/0000-0003-0369-8574

REFERENCES

1 Mohamed Q, Gillies MC, Wong TY. Management of diabetic retinopathy: a systematic review. *JAMA* 2007;298:902–16.
2 International Diabetes Federation. *IDF diabetes atlas*. 8th Edn edn. Brussels, Belgium: International Diabetes Federation, 2017.
3 Lee A, Taylor P, Kalpathy-Cramer J, *et al.* Machine learning has arrived! *Ophthalmology* 2017;124:1726–8.
4 Bhaskaranand M, Ramachandra C, Bhat S, *et al.* Automated diabetic retinopathy screening and monitoring using retinal fundus image analysis. *J Diabetes Sci Technol* 2016;10:254–61.
5 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.
6 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;124:343–51.
7 Public Health England. NHS diabetic eye screening programme overview of patient pathway, grading pathway, surveillance pathways and referral pathways, 2017 March.
8 Core NDESP team. Guidance on standard feature based grading forms to be used in the NHS diabetic eye screening programme. Diabetic Eye Screening Feature Based Grading Forms, 2013 November1.
9 Public Health England. NHS Diabetic Eye Screening Programme grading definitions for referable disease, 2017 January.
10 Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. *Ophthalmology* 1991;98:823–33.
11 Oke JL, Stratton IM, Aldington SJ, *et al.* The use of statistical methodology to determine the accuracy of grading within a diabetic retinopathy screening programme. *Diabet Med* 2016;33:896–903.
12 Bellemo V, Lim G, Rim TH, *et al.* Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep* 2019;19:72.
13 Norgaard MF, Grauslund J. Automated screening for diabetic retinopathy - a systematic review. *Ophthalmic Res* 2018;60:9–17.
14 U.S. Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems 2018 April 11. Available https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye (accessed 9 Jun 2019).
15 Zachariah S, Wykes W, Yorston D. The Scottish Diabetic Retinopathy Screening programme. *Community Eye Health* 2015;28:s22–3.
16 Gulshan V, Peng L, Coram M, *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
17 Bhaskaranand M, Ramachandra C, Bhat S, *et al.* The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther* 2019;21:635–43.
18 Public Health England. NHS screening programmes in England 2017 to 2018: PHE publications gateway number: GW-243, March 2019.
19 Roglic G. *Global report on diabetes*. Geneva, Switzerland: World Health Organization, 2016.
20 Abramoff MD, Lavin PT, Birch M, *et al.* Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digit Med* 2018;1. UNSP 39.
21 Lam C, Yi D, Guo M, *et al.* Automated detection of diabetic retinopathy using deep learning. *AMIA Jt Summits Transl Sci Proc* 2018;2017:147–55.
22 De Fauw J, Ledsam JR, Romera-Paredes B, *et al.* Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
23 Taylor D. Diabetic eye screening revised grading definitions. 2012 November 1.