**Title: Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared to human graders**

**Authors:** Adnan Tufail FRCOphth,1 Caroline Rudisill PhD,2 Catherine Egan FRANZCO,1 Venediktos V Kapetanakis PhD,3 Sebastian Salas-Vega MSc,2 Christopher G Owen PhD,3 Aaron Lee MD,1,4 Vern Louw,1 John Anderson FRCP,5 Gerald Liew FRANZCO,1, Louis Bolter,5 Sowmya Srinivas MBBS,6 Muneeswar Nittala MPhil,6 SriniVas Sadda MD,6 Paul Taylor PhD,7 Alicja R Rudnicka PhD.3

1. Moorfields BRC, Moorfields Eye Hospital, London, EC1V 2PD, United Kingdom

2. Department of Social Policy, LSE Health, London School of Economics and Political Science, London, WC2A 2AE, United Kingdom

3. Population Health Research Institute, St George’s, University of London, Cranmer Terrace, London, SW17 0RE, United Kingdom

4. University of Washington, Department of Ophthalmology, Seattle, Washington, USA

5. Homerton University Hospital, Homerton Row, E9 6SR, London, United Kingdom

6. Doheny Eye Institute, Los Angeles, CA, 90033, USA

7. CHIME, Institute of Health Informatics, University College London, London, NW1 2HE, United Kingdom

**Word Count (excluding title page, abstract, 36 references, 3 figures and 3 tables):** 3875 words

**Keywords:** Diabetes mellitus, Diabetic retinopathy, digital image, screening, validation, automatic classification, sensitivity, specificity, detection, health economics, cost effectiveness

**Financial Support:** This project was funded by the National Institute for Health Research HTA programme (project no. 11/21/02); a Fight for Sight Grant (Hirsch grant award); and the Department of Health’s NIHR Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and UCL Institute of Ophthalmology. The views expressed are those of the authors, not necessarily those of the Department of Health. The funder had no role in study design, data collection, analysis, or interpretation, or the writing of the report.

**Conflicts of interest:** Adnan Tufail has received funding from Novartis and is on the Advisory Board of Heidelberg Engineering and Optovue. Sadda Srinivas has received personal fees from Optos, Carl Zeiss Meditec, Alcon, Allergan, Genentech, Regeneron and Novartis.

**Running head:** Automated diabetic retinopathy image assessment software performance

**Address for reprints:** Mr Adnan Tufail, Moorfields Eye Hospital NHS Trust, 162 City Road, London, EC1V 2PD, United Kingdom. Email:- Adnan.tufail@moorfields.nhs.uk

Telephone:- 0207 253 3411

## Abstract

**Objective:** With increasing prevalence of diabetes, annual screening for diabetic retinopathy (DR) by expert human grading of retinal images is challenging. Automated DR image assessment systems (ARIAS) of retinal images may provide clinically- and cost-effective detection of retinopathy. We aimed to determine if available Automated DR image assessment systems (ARIAS) can be safely introduced into DR screening pathways and replace human graders.

**Design:** Observational measurement comparison study of human graders following a national screening program for DR versus ARIAS.

**Participants:** Retinal images from 20,258 consecutive patients attending routine annual diabetic eye screening between 1st June 2012 and 4th November 2013.

**Methods:** Retinal images were manually graded following a standard national protocol for DR screening and were processed by three ARIAS: iGradingM, Retmarker, and EyeArt. Discrepancies between manual grades and ARIAS were sent for arbitration to a reading center.

**Main outcomes:** Screening performance (sensitivity, false positive rate), and diagnostic accuracy (95% confidence intervals of screening performance measures) were determined. Economic analysis estimated the cost per appropriate screening outcome.

**Results:** Sensitivity point estimates (95% confidence interval) of the ARIAS were as follows: EyeArt 94.7% (94.2 to 95.2) for any retinopathy, 93.8% (92.9 to 94.6) for referable retinopathy (human graded as either ungradable, maculopathy, pre-proliferative or proliferative), 99.6% (97.0 to 99.9) for proliferative retinopathy; Retmarker 73.0% (72.0 to 74.0) for any retinopathy, 85.0% (83.6 to 86.2) for referable retinopathy, 97.9% (94.9 to 99.1) for proliferative retinopathy. iGradingM classified all images as either having disease or being ungradeable. EyeArt and Retmarker were cost saving compared to manual grading both as a replacement for initial human grading, or as a filter prior to primary human grading, although the latter approach was less cost-effective.

**Conclusions:** Retmarker and EyeArt achieved acceptable sensitivity for referable retinopathy when compared with human graders and had sufficient specificity to make them cost-effective alternatives to manual grading alone. ARIAS have the potential to reduce costs in developed world healthcare economies and to aid delivery of DR screening in developing or remote healthcare settings.

**Introduction**

Patients with diabetes are at risk of developing retinal microvascular complications that can cause vision loss, and indeed diabetes is the leading cause of incident blindness among the working age population. Early detection through regular surveillance by clinical examination, or grading of retinal photographs is essential if sight-threatening retinopathy is to be identified in time to prevent visual loss.1-4 Annual screening of the retina is recommended but presents a huge challenge, given that the global prevalence of diabetes is estimated to be 9% among adults in 2014.5 The delivery of diabetic screening will become more problematic as the number of people with diabetic retinopathy (DR) is expected to increase 3 fold in the USA by 2050,6;7 and double in the developing world by 2030, particularly in Asia, the Middle East, and Latin America.8

National screening programmes for DR, including that of the UK National Health Service Diabetic Eye Screening Programme (NHS DESP),9 are effective, however, they are also labor and capital intensive, requiring trained human graders. Similar teleretinal imaging programs have been initiated in the USA, including the Veterans Health Administration and elsewhere.10;11

Computer processing of medical images, including ophthalmic images, has benefited from advances in processing power, the availability of large datasets and new image processing techniques, which means many hitherto challenges associated with their wider application are now tractable. For instance, Automated Retinal Image Analysis Systems (ARIAS) allow the detection of DR without the need for a human grader. A number of groups have reported success in the use of their ARIAS for the detection of diabetic retinopathy.12-14 These systems triage those who have sight-threatening DR or other retinal abnormalities, from those at low risk of progression to sight-threatening retinopathy. However, while the diagnostic accuracy of some of these computer detection systems has been reported to be comparable to that of expert graders, the independent validity of ARIAS, and clinical applicability of different commercially available ARIAS to ‘real life’ screening has not been evaluated.

These image analysis systems are not currently authorized for use in NHS DESP and their cost-effectiveness is not known. Moreover, their applicability to US health settings is yet to be realized. There is a need for independent validation of one or more of the ARIAS to meet the global challenge of diabetic retinopathy screening.

This study examines the screening performance of ARIAS and the health economic implications of replacing human graders with ARIAS the UK’s National Health Service, or using ARIAS as a filter prior to manual grading.15

**Methods**

*Study design and participants:* The main aim of the study was to quantify the screening performance and diagnostic accuracy of ARIAS using NHS DESP manual grading as the reference standard.15 The study design has been previously described,15 and the protocol was published online.16

In brief, retinal images were obtained from consecutive patients with a diagnosis of diabetes mellitus, who attended their annual visit at the Diabetes Eye Screening programme of the Homerton University Hospital, London, between 1st June 2012 and 4th November 2013.17;18 Two photographic image fields were taken on each eye, one centered on the optic disc and the other on the macula in accordance with NHS DESP protocol.17 During the delivery of the screening service, patients previously screened at the Homerton University Hospital and known to be photographically ungradable underwent slit-lamp biomicroscopy in the clinic. This was part of the routine screening pathway as set by the Homerton University Hospital. Since these patients have no photographic images they could not be included in our study. Otherwise, all other patients that underwent routine retinal photography as part of the screening programme, even if images were or poor quality or classified as ‘ungradable’ by the human graders, were included in the dataset.

Research Governance approval was obtained. Images were pseudonymized, and no change in the clinical pathway occurred.

*Automated Retinal Image Analyses Systems (ARIAS):-* Automated systems for DR detection with a CE (Conformité Européenne) mark obtained or applied for within 6 months of the start of this study (July 2013) were eligible for evaluation. Three software systems were identified from a literature search and discussions with experts in the field, and all three met the CE mark standards: iGradingM (version 1.1 by Medalytix/EMIS Health, Leeds, UK),19 Retmarker (version 0.8.2. 2014/02/10 by Critical-Health, Coimbra, Portugal), IDx-DR (by IDx, Iowa City, Iowa, USA).14 IDx, Medalytix and Critical-Health agreed to participate in the study. IDx later withdrew, citing commercial reasons. An additional company, Eyenuk Inc, (Woodland Hills, California, USA) with software EyeArt, contacted us in 2013 to join the study and undertook to meet the CE mark eligibility criterion.

All the automated systems are designed to identify cases of DR of mild non-proliferative (R1) or above. EyeArt is additionally designed to identify cases requiring referral to ophthalmology (DR of ‘ungradable’ or above). A test set of 2500 images also from the Homerton screening programme (but not the same patients) was provided to the vendors to optimize their file handling processes, to address the fact that in practice, screening programmes often capture more than the 2 requisite image fields per eye, and include non-retinal images (e.g., images of crystalline lens/ cataracts) that need to be identified. During the study period ARIAS vendors had no access to their systems and all processing was undertaken by the research team.

*Reference Standards:-* All screening episodes were manually graded following NHS DESP guidelines. Each ARIAS processed all screening episodes. The study was not designed to establish the screening performance of human graders,20-22 but to compare the automated systems with outcomes from clinical practice. Screening performance of each automated system was assessed using a reference standard consisting of the final human grade modified by arbitration, by an internationally recognized fundus photographic reading center (Doheny Image Reading Center; Los Angeles, USA). Arbitration was carried out on a subset of disagreements between the final manual grade and the grades assigned by the ARIAS, without knowledge of the assigned grade. All discrepancies with final human grades for proliferative retinopathy (R3), pre-proliferative retinopathy (R2) or maculopathy (M1) were sent for arbitration to the reading center. A random sample of 1224 screening episodes (including 6000 images) where two or more systems disagreed with the final human grade of mild non-proliferative (R1) or no retinopathy (R0) were also sent for arbitration.

*Reader experience:* The Homerton Diabetes eye screening programme had a stable grading team of 18 full and part-time optometrists and non-optometrist graders holding appropriate accreditation for their designation within the programme. Performance against national standards is reviewed and reported quarterly at board meetings. In addition, the programme had been quality assured externally by the national team. Primary and secondary graders both meet minimum requisite standards to grade retinopathy and are continuously monitored to maintain quality assurance.23 In the current screening pathway,24 all retinal images are reviewed by a primary grader (level 1 grader) and any patients with mild or worse retinopathy or maculopathy are reviewed by an additional grader (secondary grader; level 2 grader) with discrepancies between primary and secondary grader reviewed by an arbitration grader (level 3 grader).

*Sample size calculations:* A pilot study of 1,340 patient screening episodes revealed that the prevalence of no retinopathy (R0), mild non-proliferative (R1, approximately equal to ETDRS level >=20 to <=43), maculopathy (M1), pre-proliferative (R2, ETDRS level > 43) and proliferative retinopathy (R3) 18 was 68%, 24%, 6.1%, 1.2% and 0.5%, respectively. One of the ARIAS (iGradingM) was compared to manual grading as the reference standard. The sensitivity for mild non-proliferative (R1), maculopathy (M1), pre-proliferative (R2) and proliferative (R3) was 82%, 91%, 100% and 100%, respectively, and 44% of R0 were graded as "disease present". The number of unique patient screening episodes (not repeat screens) undertaken in a 12-month period at the Homerton University Hospital, was 20,258. The pilot data suggested that this would provide sufficient R3 events to estimate sensitivity with an acceptable level of precision of 95% confidence intervals (CI) for sensitivity ranging from 80% to 95% for each grade (and combination of grades) of retinopathy.15 All manual grades of screened patients were stored and accessed using the Digital Health Care system version 3.6.

*Statistical Analysis:* Screening performance (sensitivity, false positive rates) and diagnostic accuracy of ARIAS (95% CI of screening performance measures) were quantified using the final manual grade with arbitration by the reading centre as the reference standard for each grade of retinopathy, as well as combinations of grades. Diagnostic accuracy of all screening performance measures was defined by 95% CI obtained by bootstrapping. Secondary analyses used multi variable logistic regression to explore whether camera type and patients’ age, gender, ethnicity influenced the ARIAS output.

*Health economic analysis:-* A decision tree model was used to calculate the incremental cost-effectiveness of replacing initial grading undertaken by human graders (level 1 graders) with ARIAS (Strategy 1 – Figure 1) and of using ARIAS prior to manual grading (Strategy 2 - Figure 2). The decision tree was designed to reflect patient screening pathways shown in Figures 1 and 2,25 and incorporated the screening levels through which images were processed (Levels 1, 2 and 3 human graders), as well as grading outcomes (referral to ophthalmology/hospital eye services or re-screening as part of the annual screening programme).

The health economic model used the following data: (i) the probabilities associated with the likelihood of a patient image continuing down each step of the retinopathy grading pathway shown in Figures 1 and 2, (ii) the overall likelihood of correct outcome classification of each screening strategy (true positives and true negatives correctly identified) and (iii) bottom-up costing of manual screening strategies and costing analysis of ARIAS via interviews and analysis estimates. It therefore took into account screening performance of automated systems (sensitivity and false positive rates), efficacy of manual screening, likelihood of re-screening and referral rates to ophthalmologists. For the ARIAS an ‘appropriate outcome’ was defined as (i) identification of ‘disease’ present by the ARIAS when the reference human grade indicated presence of potentially sight threatening retinopathy or technical failure (including grades M1, R2, R3 and U); (ii) identification of ‘no disease’ by the ARIAS when the reference human grade indicated absence of retinopathy or background retinopathy only (grades R0,R1; resulting in annual rescreening).

The model focused on assessing the relative performance of potential screening strategies and did not incorporate quality- or time-related elements. Probability parameters were modelled on the basis of Homerton hospital screening data for manual grading performance. ARIAS performance was mapped onto tentative implementation protocols for automated screening software in the NHS screening programme for diabetic retinopathy (Figures 1 and 2). Fixed and variable screening cost data were obtained through a survey of the local study centre, NHS National Tariffs, hospital cost data, phone/email conversations with automated screening system manufacturers, existing literature and expert opinion. All costs were standardized to UK Pounds Sterling for 2013/14 and, where appropriate, inflated using the 2014 Personal Social Services Research Unit costs, hospital and community health services pay and prices index.26 Screening center full time equivalent staff costs and productivity (grading rate per hour) were used to derive unit costs per screened patient across the entire screened population. Recurrent costs (capital costs, periodic charges on technologies) were discounted to reflect opportunity costs over the lifespan of investment. Medical capital equipment and hospital capital charges, including overhead charges for utilities and floor space, were discounted at 3.5% per annum over the expected lifespan of the equipment or the ARIAS. All discounted charges were annualized and incorporated into the model in terms of per patient costs. Costing results were converted into US dollar equivalents using yearly average exchange rates for 2014 from the Internal Revenue Service.27

Costing information regarding technological adoption was sought directly from manufacturers as the systems are not yet available on the English National Health Service. This yielded system costing for manufacturers which were framed as an estimated cost for screening per patient image set and included similar components in this estimate. Pricing would be contingent on the number of patients for a given guaranteed contracted volume, which has major price implications. Hence, the base case estimates used reflect the size of the screening programme for which we have manual screening data. We present models for EyeArt and Retmarker that incorporate cost information gathered from manufacturers using a universal ARIAS cost per image set as a base case figure. Costing elements of automated screening included software purchase, licensing, user training, server upgrades, and software installation and integration.28;29 We undertook extensive deterministic and threshold sensitivity analysis to examine the impact of these pricing figures on results, since there are many uncertainties related to costing a system which have not yet been implemented in the health service.

**Results**

Figure 3 shows the degree of data completeness for manual grades. Data from 20,258 consecutive screening episodes (102,856 images) were included in the analysis. Data available for each episode included a unique anonymized patient identifier; episode screening date, age, gender, ethnicity, image filenames associated with each screening episode, camera type used, retinopathy grade, maculopathy grade and associated assessment of image quality for each eye from the grader who assessed the image. The median age was 60 years (range 10 to 98 years), with 37% of patients over 65 years of age. The main ethnic groups were White (41%), Asian (35%) and Black (20%). Table 1 shows the ARIAS outcomes classifications for EyeArt and Retmarker, using the worst eye manual retinopathy grade refined by arbitration as the reference standard. The sensitivity (detection rates) point estimates (95% CI) of the ARIAS are presented in Table 2. For EyeArt sensitivity for any retinopathy (defined as manual grades mild non-proliferative [R1], pre-proliferative [R2], proliferative [R3], maculopathy [M1] and ungradable [U] combined) was 94.7% (95% CI 94.2-95.2%), 93.8% (95% CI 92.9-94.6%) for referable retinopathy (defined as manual grades, pre-proliferative [R2], proliferative [R3], maculopathy [M1] and ungradable combined), and 99.6% (95% CI 97.0-99.9%) for proliferative disease (R3). The corresponding results for Retmarker (Table 2) were 73.0% (95% CI 72.0-74.0) for any retinopathy, 85.0% (95% CI 83.6-86.2) for referable retinopathy, and 97.9% (95% CI 94.9-99.1) for proliferative retinopathy (R3). This means that per 100 screening episodes with referable retinopathy, 94 would be correctly classified as ‘disease’ by EyeArt and 6 would be incorrectly classified as ‘no disease’ (false negatives), whereas for Retmaker 85 would be correctly classified as ‘disease’ and 15 would be incorrectly classified as ‘no disease’. The false positive rate for EyeArt was 80% for retinopathy graded R0M0, meaning that out of 100 screening episodes without any retinopathy 80 would be incorrectly classified as ‘disease’ and the remaining 20 would be correctly classified as ‘no disease’ (specificity of 20%). The corresponding false positive rate for Retmarker is lower at 47.7% (specificity of 52.3%).

Unfortunately iGradingM classified all screening episodes as “disease” or “ungradable”, hence although the sensitivities were 100% the false positive rate was also 100%. Examination of a subset of images showed that the software was unable to process disc centred images. Sensitivity and false positive rates for EyeArt were not affected by ethnicity, gender or camera type, but there was weak evidence of a marginal decline in sensitivity with increasing patient’s age. Retmarker performance appeared to be marginally influenced by patient’s age, ethnicity and camera type.

Due to the performance of the iGradingM ARIA, health economic analysis was undertaken for EyeArt and Retmarker only. This study explored the cost-effectiveness of EyeArt and Retmarker ARIAS using two different strategies versus manual grading, including replacing initial manual grading (level 1 graders) with ARIAS (Strategy 1), or using ARIAS as a filter prior to manual grading (Strategy 2).

Table 3 shows the costs of screening patients in our sample under either Strategy 1 or Strategy 2 and using either EyeART or Retmarker. The results for both software systems were similar, in that the ARIAS were both cheaper, but also less likely to correctly identify the presence or absence of disease than the current manual grading system. Although the misclassification of R0 and R1 as ‘disease’ was relatively high for the ARIAS (see Tables 1 and 2), the proportion of potentially sight threatening retinopathy correctly identified was 93.8% for EyeArt and 85% for Retmarker. In this sample of 20,258 patients screened with 2844 cases of potentially sight threatening retinopathy, 2668 cases were correctly classified by EyeArt and 2416 cases by Retmarker. Hence the proportion of these 2844 cases missed was therefore 6% (176 cases) for EyeArt and 15% (428 cases) for Retmarker. Reassuringly, for the most severe retinopathy grade (R3 – proliferative retinopathy) all cases received the appropriate outcome for EyeArt and 98.6% for Retmarker. Because the incremental cost effectiveness ratio (ICER) lies in the south-west quadrant of a cost-effectiveness plane (intervention being less costly and less effective than the status quo), we have to think carefully about interpretation. Here, a lower ICER means that the intervention is less cost-effective.30 For both Retmarker and EyeART, Strategy 1 provides more cost savings per appropriate outcomes missed than Strategy 2. With Strategy 2, ICER results for Retmarker still lie in the south-west quadrant. However, in comparison to Strategy 1, there would be lower cost savings per appropriate outcome missed at $15.36. The effectiveness of Strategies 1 and 2 for the same software system were nearly identical. This likely reflects the fact that the presence of a Level 1 grader has no bearing on the disease classification given to patient episodes from automated screening systems. The cost implications emerge because patients are more likely to see more graders in Strategy 2, and Level 1 grader costs per patient are higher than Level 2 graders, reflecting a proportionally larger share of whole time equivalents dedicated to the screening clinic. The average difference in cost in the no disease arm between Strategy 1 and Strategy 2 for Retmarker was $0.38 per patient and in the disease arm $2.33. Therefore, the biggest cost difference comes in those patients who were more likely to see a higher number of human graders when the automated screening system acts as a filter rather than a replacement.

Of key importance to our findings was the cost of automated screening. We undertook one-way sensitivity analysis to check the robustness of our findings to 50% changes in ARIAS pricing. When used as a replacement for Level 1 grading (Strategy 1), both ARIAS were cost saving relative to manual grading, but offer lower effectiveness (appropriate identification of disease status in patient episodes). However, although both ARIAS are deemed ‘less effective’ overall than human graders, this was due to over sensitivity and the ARIAS very rarely missed any pre-proliferative/proliferative retinopathy or maculopathy with mild grades of retinopathy. When used as a filter prior to Level 1 grading (Strategy 2), thus reducing the volume of Level 2 grading episodes, both ARIAS were less cost saving than if used as a replacement for Level 1 graders. Threshold analysis testing was used to identify the highest ARIAS cost per patient before which they become more expensive per appropriate outcome than human grading. For Retmarker, this figure was $6.04 under Strategy 1, and $5.19 for Strategy 2. For EyeArt, ARIAS pricing above $4.29 and $3.24 per patient would make the system more expensive than manual grading under Strategy 1 and Strategy 2, respectively.

**Discussion**

The detection of diabetic retinopathy is a complex image interpretation task, and a key step to any successful screening programme. We have shown that Retmarker and EyeArt achieved acceptable sensitivity for referable retinopathy when compared with human graders at a level of specificity that makes them cost effective alternatives to a purely manual grading of diabetic retinopathy. While these two ARIAS systems have good sensitivity their low specificity makes them less effective in detecting appropriate outcomes overall than manual grading, but they are less expensive per patient, with these cost results being robust to significant variations in automated system pricing. It is important to note that although both ARIAS are deemed ‘less effective’ overall than human graders due to excessive sensitivity, they rarely missed any pre-proliferative/proliferative retinopathy or maculopathy with mild grades of retinopathy (e.g. EyeArt picked up 95% of maculopathy with mild retinopathy – (R1M1)). In light of the screening programme protocols evaluated, even if an automated screening software is overly sensitive, the patient is likely to achieve the appropriate outcome at the end of his/her acute episode. This is expected to come at a total grading cost that is cheaper regardless of whether a replacement or filter strategy is chosen for implementation of the automated screening systems. For implementation into screening pathways, some additional technical issues have to be addressed, including system integration, which this study showed was a problem in a real screening environment.

This study was not designed to look at the accuracy of human graders. In the Scottish Diabetic Retinopathy Screening Programme using similar feature based grading with one field photography, and a reference standard defined as a consensus grade from the top level graders, found the sensitivity for referable retinopathy for human graders to be 91.1% on average. Sensitivity varied by centre from 81.9% (75.2-87.1%) to 95.0% (91.5-97.1). The inter-grader agreement for referable retinopathy across all grading episodes was 88.7% (95%CI 88.0, 89.4%).31 A recent modelling study on a diabetic retinopathy screening dataset from the UK showed an estimated 11% of cases would have sight threatening retinopathy missed by human graders.32 These findings suggest that similar screening programmes using trained human graders, have comparable test performance to the ARIAS used.

This study, in keeping with the remit of established DR screening programmes such as the NHS DESP, was not designed to diagnose non-DR eye disease. However, Retmarker and EyeArt did not miss any vision threatening non-DR retinal conditions from the subset of images that went to the reading centre for arbitration.

As one of the ARIAS processes images using cloud-based technology, governance issues associated with this form of data storage need to be addressed before implementation. Health economic models may be used to evaluate the cost-effectiveness of ARIA systems under different circumstances, including in developing country settings. Additional studies are also required to shed light on the sensitivity of ARIA software to non-DR eye disease.

The ARIAS shown effective in this study have the potential to support the impending challenge of diabetic retinopathy screening in developed, as well as developing, countries. China, for instance, faces the challenge of currently having an estimated 92 million patients with diabetes,33 of which at least 50% receive no retinopathy check.34;35 In India, even though diabetes is projected to affect over 100 million people by 2035,36 it may be problematic to deliver screening even with low labor costs. Introducing ARIAS in these settings, could help scale eye screening delivery programmes, while also reducing the number of images that are manually read by up to 200 million per year in each country, assuming all patients were screened. If properly implemented, ARIAS may offer the opportunity to widen provision of a needed health service, while also freeing resources for other areas in health. The use of ARIAS, in conjunction with availability of low cost retinal digital cameras and IT infrastructure, may therefore help make the prevention of diabetic-associated blindness a tractable problem.

**Study registration**

This study protocol was registered with the HTA study number 11/21/02 and protocol published online.16

**Acknowledgements**

Our thanks to Vikash Chudasama (IT Systems Manager, Moorfields Eye Hospital) and Robin McNamara (IT Systems Administrator, Homerton University Hospital) for maintenance and setup of study servers, Ryan Chambers (Diabetes Retinal Screening Data Manager, Homerton University Hospital) for help with extraction and merging of patient demographic and medical history data. We are also grateful to out Steering Committee chaired by Irene Stratton (Senior Statistician Gloucestershire Retinal Research Group), with Steven Aldington (Independent member), Mireia Jofre-Bonet (non-Independent member), Simon Jones (Independent member), Irwin Nazareth (Independent member), Gillian Vafidis (Independent member), and Richard Wormald (Sponsor representative – non-independent member).

**Details of contributors**

All Authors read and contributed to this manuscript. AT,CE, ARR, CR, CGO,PT designed the study and raised funding. AL, VL, JA, AT, CE collected data for the study and undertook data management. VVK, ARR, SSV, CR,PT analysed the data. SS, MN carried out image grading. AT wrote the first draft of the report, to which all authors contributed. AT, VVK, ARR, SSV, CR, PT are responsible for data integrity.

Table 1: Outcome classification of EyeArt and Retmarker ARIAS compared to manual grade modified by arbitration

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Manual grade**  **(worse eye)** | **Number of Screening** | **EyeArt outcome (row %)** | | **Retmarker (row %)** | |
| **Episodes (column %)** | **No disease** | **Disease** | **No disease** | **Disease** |
| Retinopathy grade |  |  |  |  |  |
| R0M0 | 12796 (63%) | 2542 (20%) | 10254 (80%) | 6730 (53%) | 6066 (47%) |
| R1M0 | 4618 (23%) | 217 (5%) | 4401 (95%) | 1585 (34%) | 3033 (66%) |
| U | 427 (2%) | 98 (23%) | 329 (77%) | 194 (45%) | 233 (55%) |
| R1M1 | 1558 (8%) | 73 (5%) | 1485 (95%) | 207 (13%) | 1351 (87%) |
| R2 | 626 (3%) | 4 (1%) | 622 (99%) | 22 (4%) | 604 (96%) |
| R2M0 | 193 (1%) | 3 (2%) | 190 (98%) | 5 (3%) | 188 (97%) |
| R2M1 | 433 (2%) | 1 (0%) | 432 (100%) | 17 (4%) | 416 (96%) |
| R3 | 233 (1%) | 1 (0%) | 232 (100%) | 5 (2%) | 228 (98%) |
| R3M0 | 71 (0.4%) | 0 (0%) | 71 (100%) | 1 (1%) | 70 (99%) |
| R3M1 | 162 (1%) | 1 (1%) | 161 (99%) | 4 (2%) | 158 (98%) |
| Combination of grades |  |  |  |  |  |
| R0M0, R1M0 | 17414 (86%) | 2759 (16%) | 14655 (84%) | 8315 (48%) | 9099 (52%) |
| U, R1M1, R2, R3 | 2844 (14%) | 176 (6%) | 2668 (94%) | 428 (15%) | 2416 (85%) |
| R1M0, U, R1M1, R2, R3 | 7462 (37%) | 393 (5%) | 7069 (95%) | 2013 (27%) | 5449 (73%) |
| **Total** | **20258 (100%)** | **2,935** | **17,323** | 8,743 | 11,515 |

*No retinopathy* (R0), no maculopathy (M0), *background retinopathy* (R1), *ungradable images* (U), *maculopathy* (M1), *pre-proliferative retinopathy*

(R2) and *proliferative retinopathy* (R3) 17;18

Table 2: Sensitivity and false positive rates (%) for EyeArt and Retmaker ARIAS compared to manual grade modified by arbitration

|  |  |  |
| --- | --- | --- |
| **Manual grade**  **(worse eye)** | **% classified by ARIAS as disease present**  **(95% confidence interval)** | |
| **EyeArt** | **Retmarker** |
| Retinopathy grade |  |  |
| R0M0 | 80.1 (79.4 to 80.8) | 47.7 (46.5 to 48.3) |
| R1M0 | 95.3 (94.7 to 95.9) | 65.7 (64.3 to 67.0) |
| U | 77.0 (72.8 to 80.8) | 54.6 (49.8 to 59.2) |
| R1M1 | 95.3 (94.1 to 96.3) | 86.7 (84.9 to 88.3) |
| R2 | 99.4 (98.3 to 99.8) | 96.5 (94.7 to 97.7) |
| R2M0 | 98.4 (95.3 to 99.5) | 97.4 (93.9 to 98.9) |
| R2M1 | 99.8(98.4 to 100) | 96.1 (93.8 to 97.5) |
| R3 | 99.6 (97.0 to 99.9) | 97.9 (94.9 to 99.1) |
| R3M0 | 100 | 98.6 (90.7 to 99.8) |
| R3M1 | 99.4 (95.8 to 99.9) | 97.5 (93.6 to 99.1) |
| Combination of grades |  |  |
| R0M0, R1M0 | 84.2 (83.6 to 84.7) | 52.2 (51.5 to 53.0) |
| U, R1M1, R2, R3 | 93.8 (92.9 to 94.6) | 85.0 (83.6 to 86.2) |
| R1M0, U, R1M1, R2, R3 | 94.7 (94.2 to 95.2) | 73.0 (72.0 to 74.0) |

*No retinopathy* (R0), no maculopathy (M0), *background retinopathy* (R1), *ungradable images* (U), *maculopathy* (M1), *pre-proliferative retinopathy* (R2) and *proliferative retinopathy* (R3) 17;18

For manual grades R0M0 classified as “disease present” by the ARIAS the percentages correspond to false positive rates.

Table 3: Base case results for 20,258 patients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Screening strategy and ARIAS | Total cost of grading | Incremental cost | Appropriate outcomes | Incremental appropriate outcomes | Cost reduction per appropriate outcome missed (ICER)\* |
| **EyeArt** |  |  |  |  |  |
| *Strategy 1* |  |  |  |  |  |
| MG | $795,164.60 | - | 19,684 | - | - |
| ARIAS | $693,344.48 | $(101,820.13) | 5,427 | (14,257) | $7.14 |
| *Strategy 2* |  |  |  |  |  |
| MG | $795,164.60 | - | 19,684 | - | - |
| ARIAS | $675,138.67 | $(63,063.91) | 5,428 | (14,256) | $4.43 |
| **Retmarker** |  |  |  |  |  |
| *Strategy 1* |  |  |  |  |  |
| MG | $795,164.60 |  | 19,684 |  |  |
| ARIAS | $627,913.75 | $(167,250.85) | 10,731 | (8,953) | $18.69 |
| *Strategy 2* |  |  |  |  |  |
| MG | $795,164.60 | - | 19,684 | - | - |
| ARIAS | $658,012.58 | $(137,152.07) | 10,760 | (8,923) | $15.36 |

Key – MG - manual grader, ARIAS - Automated diabetic retinopathy image assessment systems, ICER - incremental cost effectiveness ratio

Strategy 1 replaces the initial grading (Level 1 Grader) with ARIAS.  
Strategy 2 ARIAS is used as a filter prior to manual grading by Level 1 Grader.  
\*If the ARIAS were more costly and more effective the ICER would be stated in terms of cost/appropriate outcome.

ICER can also be interpreted as cost savings per appropriate outcome missed.

Figure 1: Decision tree model used to calculate the incremental cost-effectiveness of manual grading versus replacing initial grading undertaken by human graders (level 1 graders) with ARIAS.



Figure 2: Decision tree model used to calculate the incremental cost-effectiveness of manual grading versus replacing initial grading undertaken by human graders (level 1 graders) with ARIAS prior to manual grading.



Figure 3: Data extraction of diabetic patients attending the Homerton Diabetic Eye Screening Programme.



**References**

(1) Photocoagulation for diabetic macular edema. Early Treatment Diabetic Retinopathy Study report number 1. Early Treatment Diabetic Retinopathy Study research group. *Arch Ophthalmol* 1985; 103(12):1796-1806.

(2) Cheung N, Wong IY, Wong TY. Ocular anti-VEGF therapy for diabetic retinopathy: overview of clinical efficacy and evolving applications. *Diabetes Care* 2014; 37(4):900-905.

(3) Aiello LP, Gardner TW, King GL, Blankenship G et al. Diabetic retinopathy. *Diabetes Care* 1998; 21(1):143-156.

(4) Mohamed Q, Gillies MC, Wong TY. Management of diabetic retinopathy: a systematic review. *JAMA* 2007; 298(8):902-916.

(5) World Health Organization. Global status report on noncommunicable diseases 2014. http://www.who.int/nmh/publications/ncd-status-report-2014/en/ [Accessed June 2016]; 2014.

(6) Congdon N, O'Colmain B, Klaver CC, Klein R et al. Causes and prevalence of visual impairment among adults in the United States. *Arch Ophthalmol* 2004; 122(4):477-485.

(7) Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: US. https://www.cdc.gov/diabetes/pubs/pdf/ndfs\_2011.pdf [Accessed June 2016]; 2011.

(8) Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010; 87(1):4-14.

(9) Public Health England. Diabetic eye screening: programme overview. https://www.gov.uk/guidance/diabetic-eye-screening-programme-overview [Accessed June 2016]; 2014.

(10) Murchison AP, Friedman DS, Gower EW, Haller JA et al. A Multi-Center Diabetes Eye Screening Study in Community Settings: Study Design and Methodology. *Ophthalmic Epidemiol* 2016; 23(2):109-115.

(11) Kirkizlar E, Serban N, Sisson JA, Swann JL et al. Evaluation of telemedicine for screening of diabetic retinopathy in the Veterans Health Administration. *Ophthalmology* 2013; 120(12):2604-2610.

(12) Abramoff MD, Folk JC, Han DP, Walker JD et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* 2013; 131(3):351-357.

(13) Soto-Pedre E, Navea A, Millan S, Hernaez-Ortega MC et al. Evaluation of automated image analysis software for the detection of diabetic retinopathy to reduce the ophthalmologists' workload. *Acta Ophthalmol* 2015; 93(1):e52-e56.

(14) Goatman K, Charnley A, Webster L, Nussey S. Assessment of automated disease detection in diabetic retinopathy screening using two-field photography. *PLoS One* 2011; 6(12):e27524.

(15) Kapetanakis VV, Rudnicka AR, Liew G, Owen CG et al. A study of whether automated Diabetic Retinopathy Image Assessment could replace manual grading steps in the English National Screening Programme. *J Med Screen* 2015; 22(3):112-118.

(16) Tufail A, Egan C, Rudnicka A, Owen C, Bailey C, Rudisill C et al. Detailed project description: Can automated diabetic retinopathy image assessment software replace one or more steps of manual imaging grading and is this cost-effective for the English National Screening Programme? http://www.nets.nihr.ac.uk/\_\_data/assets/pdf\_file/0019/81154/PRO-11-21-02.pdf [Accessed June 2016]; 2014.

(17) NHS Diabetic Eye Screening Programme (NDESP). Diabetic eye screening feature based grading forms: Guidance on standard feature based grading forms to be used in the NHS Diabetic Eye Screening Programme. https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/402295/Feature\_Based\_Grading\_Forms\_V1\_4\_1Nov12\_SSG.pdf [Accessed June 2016]; 2012.

(18) Taylor D. Diabetic eye screening revised grading definitions: To provide guidance on revised grading definitions for the NHS Diabetic Eye Screening Programme. https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/402294/Revised\_Grading\_Definitions\_V1\_3\_1Nov12\_SSG.pdf [Accessed June 2016]; 2012.

(19) Philip S, Fleming AD, Goatman KA, Fonseca S et al. The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol* 2007; 91(11):1512-1517.

(20) Stellingwerf C, Hardus PL, Hooymans JM. Two-field photography can identify patients with vision-threatening diabetic retinopathy: a screening approach in the primary care setting. *Diabetes Care* 2001; 24(12):2086-2090.

(21) Taylor R, Broadbent DM, Greenwood R, Hepburn D et al. Mobile retinal screening in Britain. *Diabet Med* 1998; 15(4):344-347.

(22) Kinyoun JL, Martin DC, Fujimoto WY, Leonetti DL. Ophthalmoscopy versus fundus photographs for detecting and grading diabetic retinopathy. *Invest Ophthalmol Vis Sci* 1992; 33(6):1888-1893.

(23) Taylor D, Widdowson S. The management of grading quality: Good practice in the quality assurance of grading. https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/512832/The\_Management\_of\_Grading.pdf [Accessed June 2016]; 2016.

(24) NHS Diabetic Eye Screening Programme (NDESP). Diabetic eye screening pathway overviews: Overview diagrams for patient pathway, grading pathway, surveillance pathways and referral pathways. https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/403074/Pathway\_Diagrams\_V1\_2\_29Oct12\_SSG\_\_1\_.pdf [Accessed June 2016]; 2012.

(25) UK National Screening Committee. Essential elements in developing a diabetic retinopathy screening programme. Workbook 4.3. http://rcophth-website.www.premierithosting.com/docs/publications/published-guidelines/ENSPDR\_Workbook\_2009.pdf [Accessed June 2016]; 2009.

(26) Personal Social Services Research Unit. Unit costs of health and social care 2014. http://www.pssru.ac.uk/project-pages/unit-costs/2014/index.php [Accessed June 2016]; 2016.

(27) Internal Revenue Service (IRS). Yearly Average Currency Exchange Rates Translating foreign currency into U.S. dollars. Available from https://www.irs.gov/individuals/international-taxpayers/yearly-average-currency-exchange-rates [Acessed June 2016]; 2016.

(28) Scotland GS, McNamee P, Fleming AD, Goatman KA et al. Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol* 2010; 94(6):712-719.

(29) Scotland GS, McNamee P, Philip S, Fleming AD et al. Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *Br J Ophthalmol* 2007; 91(11):1518-1523.

(30) Gray AM, Clarke PM, Wolstenholme JL, Wordsworth S. Applied methods of cost-effectiveness analysis in healthcare. Oxford University Press; 2011.

(31) Goatman KA, Philip S, Fleming AD, Harvey RD et al. External quality assurance for image grading in the Scottish Diabetic Retinopathy Screening Programme. *Diabet Med* 2012; 29(6):776-783.

(32) Oke JL, Stratton IM, Aldington SJ, Stevens RJ et al. The use of statistical methodology to determine the accuracy of grading within a diabetic retinopathy screening programme. *Diabet Med* 2016; 33(7):896-903.

(33) Yang W, Lu J, Weng J, Jia W et al. Prevalence of diabetes among men and women in China. *N Engl J Med* 2010; 362(12):1090-1101.

(34) Peng J, Zou H, Wang W, Fu J et al. Implementation and first-year screening results of an ocular telehealth system for diabetic retinopathy in China. *BMC Health Serv Res* 2011; 11:250.

(35) Wu B, Li J, Wu H. Strategies to Screen for Diabetic Retinopathy in Chinese Patients with Newly Diagnosed Type 2 Diabetes: A Cost-Effectiveness Analysis. *Medicine (Baltimore)* 2015; 94(45):e1989.

(36) Guariguata L, Whiting DR, Hambleton I, Beagley J et al. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract* 2014; 103(2):137-149.