# 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

Adnan Tufail,[1]* Venediktos V Kapetanakis,[2] Sebastian Salas-Vega,[3] Catherine Egan,[1] Caroline Rudisill,[3] Christopher G Owen,[2] Aaron Lee,[1] Vern Louw,[1] John Anderson,[4] Gerald Liew,[1] Louis Bolter,[4] Clare Bailey,[5] SriniVas Sadda,[6] Paul Taylor[7] and Alicja R Rudnicka[2]

[1]National Institute for Health Research Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, UK
[2]Population Health Research Institute, St George's, University of London, London, UK
[3]Department of Social Policy, LSE Health, London School of Economics and Political Science, London, UK
[4]Homerton University Hospital Foundation Trust, London, UK
[5]Bristol Eye Hospital, Bristol, UK
[6]Doheny Eye Institute, Los Angeles, CA, USA
[7]Centre for Health Informatics & Multiprofessional Education (CHIME), Institute of Health Informatics, University College London, London, UK

*Corresponding author

**Declared competing interests of authors:** SriniVas Sadda received grants and personal fees from Optos and Carl Zeiss Meditec during the duration of the study.

Published December 2016
DOI: 10.3310/hta20920

# Scientific summary

## Automated diabetic retinopathy image softwares

Health Technology Assessment 2016; Vol. 20: No. 92
DOI: 10.3310/hta20920

NIHR Journals Library www.journalslibrary.nihr.ac.uk

# Scientific summary

## Background

National screening programmes for diabetic retinopathy, including the NHS diabetic eye screening programme (DESP), have been effective in identifying those in need of treatment and so preventing visual loss. However, at present there are > 2.5 million people (aged ≥ 12 years) in England living with diabetes mellitus.

This represents a major challenge for the NHS, as current retinal screening programmes are labour intensive and require trained human graders, who are hard to find and retain. In addition, it is projected that future numbers to be screened for retinopathy will escalate given that both the prevalence and incidence of diabetes mellitus are increasing markedly.

There is increasing interest in systems to detect diabetic retinopathy automatically. These systems differentiate those who have sight-threatening diabetic retinopathy or other retinal abnormalities from those at low risk of progression to sight-threatening retinopathy. However, while the accuracy of these computer systems has been reported to be comparable with that of expert graders, the independent validity of these systems and clinical applicability to 'real-life' screening within the NHS DESP protocol of two retinal field photographs per eye (macular- and optic disc-centred images) remains unclear. Moreover, these image analysis systems are not currently authorised for use in DESPs and their cost-effectiveness is not known.

## Objectives

1. To quantify the screening performance and diagnostic accuracy of automated retinal image analysis systems (ARIASs) using NHS DESP manual grading as the reference standard.
2. To re-evaluate screening performance after a subset of images were regraded by an approved reading centre for discrepancies found between manual grades and ARIASs.
3. To estimate cost-effectiveness of (1) replacing level 1 manual graders with automated retinal image analysis and (2) using the automated retinal image analysis prior to manual grading.
4. To examine the influence of camera type and patients' ethnicity, age and sex on screening performance.
5. To examine issues of implementing ARIASs in a real screening environment at the Homerton DESP.

## Study population

The study was conducted on patients attending the annual DESP at Homerton University Hospital Foundation Trust, London. A consecutive series of diabetic patients aged ≥ 12 years who attended the hospital between 1 June 2012 and 4 November 2013 and had macular- and disc-centred retinal images taken in accordance with imaging standards laid out by the NHS DESP were included. As the patient data and retinal images were anonymised, Caldicott Guardian and Research Governance approval was obtained but full ethics committee approval was not required, given that the study did not change the clinical pathway for the patients and the data were anonymised.

## Intervention

Automated systems for diabetic retinopathy detection with a Conformité Européenne (CE) mark obtained or applied for up to 6 months after the contractual 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. All met the CE mark standards: iGradingM (version 1.1, originally Medalytix Group Ltd, Manchester, UK, but purchased by Digital Healthcare, Cambridge, UK, at the initiation of the study, purchased in turn by EMIS Health, Leeds, UK, after conclusion of the study), Retmarker (version 0.8.2, Retmarker Ltd, Coimbra, Portugal) and IDx-DR (IDx, LLC, Iowa City, IA, USA). Medalytix Group Ltd, IDx, LLC and Retmarker Ltd agreed to participate in the study in 2012. An additional company, Eyenuk Inc., contacted us in 2013 to join the study within the 6-month time limit and stated that its system, EyeArt (Eyenuk Inc., Woodland Hills, CA, USA), would meet the CE mark eligibility criterion.

All automated systems are designed to identify cases of diabetic retinopathy of background retinopathy (R1) or above. EyeArt is additionally designed to identify cases requiring referral to ophthalmology [diabetic retinopathy of ungradable images (U), maculopathy (M1), pre-proliferative retinopathy (R2) and pre-proliferative retinopathy (R3) or above]. A set of 2500 images from the same screening programme (but not the same patients) was provided to the four ARIAS vendors to optimise their file-handling processes. This optimisation step allowed vendors to address the fact that, in practice, screening programmes often capture more than the two requisite image fields per eye and often take additional non-retinal images, for example photographs of cataracts that would have to be filtered out of the retinal grading process in an automated system, but provide useful information to a manual grading team. During the study period, ARIAS vendors had no access to their systems and all processing was undertaken by the research team. One of the software vendors, IDx, LLC, withdrew from the study after processing the test set, citing commercial reasons.

## Methods

### *Sample size and inclusion criteria*

A pilot study of 1340 patient screening episodes at a London NHS screening programme revealed that the prevalence of no retinopathy (R0), R1, M1, R2 and R3 was 68%, 24%, 6.1%, 1.2% and 0.5%, respectively. In this pilot study, one software (iGradingM) was compared with manual grading as the reference standard. The sensitivity for R1, M1, R2 and R3 was 82%, 91%, 100% and 100%, respectively, and 44% of R0 images were graded as 'disease present'. The sample size calculation was based on the number of screening episodes required to ensure that the lower limit of the 95% confidence interval (CI) for sensitivity of automated grading did not fall below 97% for the most severe level of retinopathy, classified as R3 by human graders. If the prevalence of R3 was 0.5% and ARIASs detected all R3 images (as in the pilot study), the required sample size would be 24,000 episodes. The number of unique patient screening episodes (not repeat screens) undertaken in a 12-month period at the Homerton University Hospital Foundation Trust was 20,258. This would provide sufficient R3 events based on pilot data. All manual grades of screened patients were stored and accessed using the Digital Healthcare OptoMize version 3.6 (Digital Healthcare, Cambridge, UK).

### *Reference standards*

All screening episodes were manually graded following NHS DESP guidelines. Each ARIAS processed all screening episodes. Screening performance of each automated system was assessed using two reference standards: (1) the final manual grade and (2) the final manual grade modified by arbitration. An internationally recognised fundus photographic reading centre (Doheny Image Reading Centre, Los Angeles, CA, USA), masked to all grading, arbitrated on disagreements between the final human grade and the grades assigned by the ARIASs. Arbitration was limited by the available financial resources to no more than 1700 episodes. Emphasis was placed on arbitration of all discrepancies with final manual grades R3, R2, M1, that is patients at risk of vision loss with more severe grades of diabetic retinopathy. A random sample of screening episodes when two or more systems disagreed with the final manual grade of R1 or R0 were also sent for arbitration.

### *Statistical analysis*

The screening performance (sensitivity, false-positive rates and likelihood ratios) and diagnostic accuracy (95% CI of screening performance measures) were quantified using the final manual grade (with and without reading centre arbitration) as the reference standard for each grade of retinopathy, as well as combinations of grades. The diagnostic accuracy of all screening performance measures was defined by 95% CI obtained by bootstrapping. Secondary analyses used multiple variable logistic regression analyses to explore whether or not camera type and patient characteristics, including age, sex and ethnicity, influenced the outcome classification of ARIASs.

### *Health economics*

Analyses were carried out to investigate the economic implications of (1) if an automated system were to replace level 1 human graders and (2) if the automated system were to be used as a filter prior to level 1 graders. Cost data were obtained from Personal Social Services Research Unit, hospital cost data, the existing literature and expert opinion.

A prospective study was undertaken in which three ARIASs were integrated with the systems used in the screening service at the Homerton University Hospital Foundation Trust: the Digital Healthcare OptoMize. An exporter tool developed for this study [Data Export Module for OptoMize version 3 database (Digital Healthcare)] was used daily to transfer images to each of the three ARIAS servers to be processed. The purpose of this element of the study was to identify technical issues that may arise if ARIASs were used in routine NHS diabetic eye screening.

## Results

A total of 142,018 images from 28,079 screening episodes involving 20,258 patients were included in the study. Only data from 20,258 primary patient episodes were analysed (102,856 images) as none of the ARIASs altered their performance by knowledge of a previous patient episodes grade. Data on age, sex and ethnicity were available for 20,212 patients. The median age was 60 years (range 10–98 years) and 41% of patients were white European, 35% were Asian and 19.6% were black African Caribbean. The sensitivity point estimates of the ARIASs were as follows: EyeArt 94.7% (95% CI 94.2% to 95.2%) for any retinopathy (manual grades R1, U, M1, R2 and R3 as refined by arbitration), 93.8% (95% CI 92.9% to 94.6%) for referable retinopathy (manual grades U, M1, R2 and R3 as refined by arbitration), 99.6% (95% CI 97.0% to 99.9%) for R3 proliferative disease; corresponding sensitivities for Retmarker were 73.0% (95% CI 72.0% to 74.0%) for any retinopathy, 85.0% (95% CI 83.6% to 86.2%) for referable retinopathy and 97.9% (95% CI 94.9% to 99.1%) for R3 proliferative retinopathy. For manual grades R0 and no maculopathy (M0), specificity was 20% (95% CI 19% to 21%) for EyeArt and 53% (95% CI 52% to 54%) for Retmarker. The iGradingM outcome at the episode level was either ‘disease’ or ‘ungradable’. An examination of the subset of images from arbitration grading showed that this software was unable to process disc-centred images. Sensitivity and false-positive rates for EyeArt were not affected by ethnicity, sex or camera type, but sensitivity was marginally lower with increasing patient age. The screening performance of Retmarker appeared to vary according to the patient’s age, ethnicity and camera type. We did not systematically assess images for non-diabetic eye disease; however, in a subset of images that were arbitrated, no late age-related macular degeneration, central retinal vein occlusion eyes or myopic degeneration eyes were categorised as ‘no disease’ by the ARIASs among the arbitration episodes.

Owing to the very poor performance of the iGradingM ARIAS in a two-field per eye image acquisition protocol, health economic analysis was undertaken for EyeArt and Retmarker only. This study explored the cost-effectiveness of EyeArt and Retmarker ARIASs using two different strategies compared with manual grading as currently performed at the Homerton screening service. When used as a replacement for level 1 grading (strategy 1), both automated screening systems were cost saving relative to manual grading but offered lower clinical effectiveness (appropriate identification of disease status in patient episodes).

When used as a filter prior to level 1 grading (strategy 2), thus reducing the number of level 1 and level 2 grading episodes, both automated screening systems were less cost saving than with strategy 1. Threshold analysis testing the highest ARIASs cost per patient before which ARIASs become more expensive per appropriate outcome than human grading demonstrated that, for strategy 1 with Retmarker, this figure was £3.82. In strategy 2 for Retmarker this figure was £3.28. For EyeArt, it would be more expensive than manual grading if the ARIAS was priced above £2.71 per patient for strategy 1 and £2.05 for strategy 2.

## Conclusions

Retmarker and EyeArt achieved acceptable sensitivity and false-positive rates for referable retinopathy, when compared with manual grades as a reference standard, to make them cost-effective alternatives to a purely manual grading approach. The economic costs appear robust to significant variations in automated system pricing. Even if an automated screening software is overly sensitive, the patient is likely to achieve the appropriate outcome at the end of his or her acute episode. This is expected to come at a total grading cost that is cheaper regardless of whether ARIASs replace level 1 graders or are used as filter prior to level 1 manual grading. Retmarker and EyeArt should therefore be considered for screening pathways when additional technical issues have been addressed.

This future technical work would involve Digital Imaging and Communications in Medicine (DICOM) compatibility and standardised methods for automated preparation of image processing by ARIASs. Output from the ARIAS should link back to the screening system to set up a grading queue and randomly sample a proportion of images that were classified as not having any disease for manual grading. As one of the ARIASs processes images using the cloud, governance issues associated with this need to be addressed before implementation.

## Study registration

This study protocol was registered with the HTA study number 11/21/02 and protocol published online on 23 May 2013 [www.nets.nihr.ac.uk/__data/assets/pdf_file/0019/81154/PRO-11-21-02.pdf (accessed 23 May 2013)].

## Funding

This study was funded by the National Institute for Health Research (NIHR) Health Technology Assessment programme, 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 the University College London Institute of Ophthalmology.

# Health Technology Assessment

HTA/HTA TAR

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 4.058

*Health Technology Assessment* is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the ISI Science Citation Index.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

### Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

### HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hta

### This report

The research reported in this issue of the journal was funded by the HTA programme as project number 11/21/02. The contractual start date was in December 2012. The draft report began editorial review in November 2015 and was accepted for publication in July 2016. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health.



Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).