HEALTH TECHNOLOGY ASSESSMENT

VOLUME 20 ISSUE 92 DECEMBER 2016
ISSN 1366-5278

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, Venediktos V Kapetanakis, Sebastian Salas-Vega, Catherine Egan, Caroline Rudisill, Christopher G Owen, Aaron Lee, Vern Louw, John Anderson, Gerald Liew, Louis Bolter, Clare Bailey, SriniVas Sadda, Paul Taylor and Alicja R Rudnicka

National Institute for Health Research

DOI 10.3310/hta20920

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

This report should be referenced as follows:

Tufail A, Kapetanakis VV, Salas-Vega S, Egan C, Rudisill C, Owen CG, *et al*. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. *Health Technol Assess* 2016;**20**(92).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (SciSearch®) and *Current Contents®*/ Clinical Medicine.

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).

Abstract

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 adnan.tufail@moorfields.nhs.uk

Background: Diabetic retinopathy screening in England involves labour-intensive manual grading of retinal images. Automated retinal image analysis systems (ARIASs) may offer an alternative to manual grading.

Objectives: To determine the screening performance and cost-effectiveness of ARIASs to replace level 1 human graders or pre-screen with ARIASs in the NHS diabetic eye screening programme (DESP). To examine technical issues associated with implementation.

Design: Observational retrospective measurement comparison study with a real-time evaluation of technical issues and a decision-analytic model to evaluate cost-effectiveness.

Setting: A NHS DESP.

Participants: Consecutive diabetic patients who attended a routine annual NHS DESP visit.

Interventions: Retinal images were manually graded and processed by three ARIASs: 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 EyeArt (Eyenuk Inc., Woodland Hills, CA, USA). The final manual grade was used as the reference standard. Arbitration on a subset of discrepancies between manual grading and the use of an ARIAS by a reading centre masked to all grading was used to create a reference standard manual grade modified by arbitration.

Main outcome measures: Screening performance (sensitivity, specificity, false-positive rate and likelihood ratios) and diagnostic accuracy [95% confidence intervals (CIs)] of ARIASs. A secondary analysis explored

the influence of camera type and patients' ethnicity, age and sex on screening performance. Economic analysis estimated the cost per appropriate screening outcome identified.

Results: A total of 20,258 patients with 102,856 images were entered into the study. The sensitivity point estimates of the ARIASs were as follows: EyeArt 94.7% (95% CI 94.2% to 95.2%) for any retinopathy, 93.8% (95% CI 92.9% to 94.6%) for referable retinopathy and 99.6% (95% CI 97.0% to 99.9%) for proliferative retinopathy; and Retmarker 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 proliferative retinopathy. iGradingM classified all images as either 'disease' or 'ungradable', limiting further iGradingM analysis. The sensitivity and false-positive rates for EyeArt were not affected by ethnicity, sex or camera type but sensitivity declined marginally with increasing patient age. The screening performance of Retmarker appeared to vary with patient's age, ethnicity and camera type. Both EyeArt and Retmarker were cost saving relative to manual grading either as a replacement for level 1 human grading or used prior to level 1 human grading, although the latter was less cost-effective. A threshold analysis testing the highest ARIAS cost per patient before which ARIASs became more expensive per appropriate outcome than human grading, when used to replace level 1 grader, was Retmarker £3.82 and EyeArt £2.71 per patient.

Limitations: The non-randomised study design limited the health economic analysis but the same retinal images were processed by all ARIASs in this measurement comparison study.

Conclusions: Retmarker and EyeArt achieved acceptable sensitivity for referable retinopathy and false-positive rates (compared with human graders as reference standard) and appear to be cost-effective alternatives to a purely manual grading approach. Future work is required to develop technical specifications to optimise deployment and address potential governance issues.

Funding: 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.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

ARIAS	automated retinal image analysis system
CE	Conformité Européenne
CI	confidence interval
DESP	diabetic eye screening programme
DICOM	Digital Imaging and Communications in Medicine
GBP	Great Britain Pound
ICER	incremental cost-effectiveness ratio
ID	identification
IT	information technology
JPEG	Joint Photographic Experts Group
M0	no maculopathy
M1	maculopathy
M1a	maculopathy with exudates within 1-disc diameter of the centre of the fovea
M1b	maculopathy with exudates within the macula
OCT	optical coherence tomography
QA	quality assurance
R0	no retinopathy
R1	background retinopathy
R2	pre-proliferative retinopathy
R3	proliferative retinopathy
U	ungradable images
WTE	whole-time equivalent

Plain English summary

Annual diabetic retinopathy screening using digital photographs of the retina assessed by human graders is recognised as the best way to detect vision-threatening disease and reduce visual loss in patients with diabetes mellitus. Vision-threatening disease is referred to hospital eye services for review and possible treatment. With more than 3 million people in England diagnosed with diabetes mellitus, there has been increasing interest in automated systems that detect diabetic retinopathy on digital pictures [so-called automated retinal image analysis systems (ARIASs)] as a way of reducing the need for human graders and the cost of screening. This study compared commercially available ARIASs [IDx-DR (IDx, LLC, Iowa City, IA, USA), 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 UK, Leeds, UK, after conclusion of the study), Retmarker (version 0.8.2, Retmarker Ltd, Coimbra, Portugal) and EyeArt (Eyenuk Inc., Woodland Hills, CA, USA)] with human manual grading on retinal photographs from 20,258 consecutive patients seen in a NHS diabetic eye screening programme. IDx, LLC withdrew from the study, citing commercial reasons. The ability of the remaining three ARIASs to correctly identify patients with diabetic retinopathy was compared against trained human graders. Health-economic analyses were carried out to investigate whether or not it would save money if ARIASs replaced trained human graders in different parts of the screening pathway.

Two ARIASs, Retmarker and EyeArt, achieved an acceptable level of diabetic retinopathy detection in comparison with trained human graders. Retmarker and EyeArt had a modest rate of false alarms, where these ARIASs would identify that there was disease when no disease was actually present. The good detection rate and acceptable false-alarm rate make both ARIASs potentially cost-effective alternatives to human grading in NHS diabetic eye screening programmes.

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.

Chapter 1 Introduction

Background

Diabetes mellitus, particularly type 2 diabetes mellitus, is a major public health problem both in the UK[1] and globally.[2] Diabetic retinopathy is a common complication of diabetes mellitus and is one of the major causes of vision loss in the working-age population.[3] However, in the UK, diabetic retinopathy is no longer the leading cause of vision loss in this age group.[3] This has been attributed to the effectiveness of national screening programmes, including that of the NHS diabetic eye screening programme (DESP),[4] in identifying those in need of treatment, particularly early treatment, which is highly effective in preventing visual loss.[3]

The Airlie House Symposium on the Treatment of Diabetic Retinopathy established the basis for the current photographic method of quantifying the presence and severity of diabetic retinopathy.[5] The characteristic lesions of diabetic retinopathy are presently estimated to affect nearly half of those diagnosed at any given time.[6–9] The advances in the medical management of diabetes mellitus have substantially increased patient survival and life expectancy. However, in doing so people with diabetes mellitus are placed at an increased risk for developing diabetes mellitus-related microvascular complications, the most common of which is diabetic retinopathy.[7,8]

The current treatment recommendations for diabetic retinopathy are highly effective in preventing visual loss.[10–12] Early detection and accurate evaluation of diabetic retinopathy severity, co-ordinated medical care and prompt appropriate treatment represent an effective approach for diabetic eye care.

A total of 2.04 million people were offered diabetic eye screening in 2014/15 in England, among whom the rate of uptake was 83%.[13] This represents a major challenge for the NHS, as current retinal screening programmes are labour intensive (requiring one or more trained human graders), and future numbers will escalate given that both prevalence and incidence of diabetes mellitus are increasing markedly.[14]

The number of individuals with diabetes mellitus is projected to reach 552 million globally by 2030, so technical enhancements addressing image acquisition, automated image analysis algorithms and predictive biomarkers are needed if current diabetic retinopathy screening programmes are to manage this burden successfully. Automated retinal image analysis systems (ARIASs) have been developed. Although such an approach has been implemented in Scotland, little independent evaluation of commercially available or Conformité Européenne (CE)-marked software has been done or been undertaken for the NHS DESP using its image acquisition protocol.

Automated diabetic retinopathy screening

There is increasing interest in systems for the automated detection of diabetic retinopathy. In Scotland, one automated system reported a very high detection rate (100% for proliferative retinopathy and over 97% for maculopathy) in a large, unselected population attending two regional screening programmes.[15] The system differentiates those who have sight-threatening diabetic retinopathy or other retinal abnormalities from those at low risk of progression to sight-threatening retinopathy. However, other systems have since become commercially available and, although the diagnostic accuracy of these computer detection systems has been reported to be comparable with that of experts,[16–22] the independent validity of these systems and their applicability to 'real-life' screening remains unclear. The external validity of such studies to date is limited because, commonly, images are not available for comparison, non-validated reading protocols were used and detection programs were developed on images and populations highly similar to the one they are tested on, including distribution of retinopathy severity, camera type, field of view and number of

images per eye. The Scottish programme uses one field per eye, whereas two fields per eye are used in the NHS DESP. The limited external validity of studies to date may be in part because the majority of studies so far have been designed, run and analysed by individuals linked to the development of the software being tested. Independent expert groups without links to the commercial development of the software are uncommon and the alternative approach of using an independent, internationally recognised fundus photography reading centre is prohibitively expensive. Moreover, these image analysis systems are not currently authorised for use in the NHS DESP. Recent work has also shown that optical coherence tomography (OCT) imaging is a useful adjunct to colour fundus photography in screening for referable diabetic maculopathy and has the potential to significantly reduce over-referral of cases of macula oedema compared with human assessment of colour fundus images.[23] Some hospital eye services are already incorporating OCT assessment within their DESP.[24]

Development and process of automated retinal image analysis systems

The introduction of high-resolution digital retinal imaging systems in the 1990s, combined with growth in computing power, permitted the development of computer algorithms capable of computer-aided detection and diagnosis. Computer-aided detection is the identification of pathological lesions. Computer-aided diagnosis provides a classification that incorporates additional lesion or clinical information to stratify risk or estimate the probability of disease. These recent advances have allowed ARIASs to have potential clinical utility.

Broadly, the approach to ARIASs can be categorised into two components: (1) image quality assessment and (2) image analysis.

Image quality assessment

The NHS DESP currently specifies two-field colour digital photography as the only acceptable method of systematic screening for diabetic eye disease in the NHS and define specific camera systems and minimum resolutions.[25] Currently, retinal digital photography has progressed to a stage at which colour retinal photographs can be obtained using low levels of illumination through an undilated pupil.[26] However, human factors such as movement and positioning and ocular factors such as a cataract and reflections from retinal tissues can produce artefacts. Without pupillary dilatation, artefacts are observed in 3–30% of retinal images to an extent at which they impede human grading.[27] Thus, the importance of good image quality prior to automated image analysis has been recognised and much ancillary research has been conducted in the field of image pre-processing.[28]

Image analysis

The main challenge encountered with processing of colour digital images is the presence of numerous ‘distractors’ within the retinal image (retinal capillaries, underlying choroidal vessels and reflection artefacts), all of which may be confused with diabetic retinopathy lesions. As a result, much research has been focused on the selective identification of diabetic retinopathy features, including microaneurysms, haemorrhages, hard or soft exudates, cotton wool spots and venous beading. These clinical features have been described in great detail in landmark clinical trials, Diabetic Retinopathy Study and Early Treatment Diabetic Retinopathy Study.[10,29]

The main approaches to image analysis of ARIASs can be found in a recent review by Sim *et al.*[30]

Diagnostic accuracy of automated retinal image analysis systems reported to date

Commercial systems that were available at the initiation of this study and invited to participate in the study are described below. ARIASs not available at initiation of the study or that did not meet entry criteria, and that are therefore not evaluated, are summarised in a recent review.[30]

Commercial systems that were available within 6 months of the start of the study include 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 UK, Leeds, UK, after conclusion of the study), IDx-DR (IDx, LLC, Iowa City, IA, USA), Retmarker (version 0.8.2, Retmarker Ltd, Coimbra, Portugal) and EyeArt (Eyenuk Inc., Woodland Hills, CA, USA). A comparison of the published sensitivity and specificity of each ARIAS, as well as a brief description of the published studies supporting each is summarised in a paper by Sim *et al*.[30] Direct head-to-head comparisons between systems to date have proven difficult, mainly because of different photographic protocols, algorithms and patient populations used for validation. A common thread among these automated systems is to identify referable retinopathy, defined as diabetic retinopathy that requires the attention of an ophthalmologist. To date, we have not progressed to a stage at which human intervention can be fully removed from screening programmes. All of the systems described below are semi-automated at some point in the workflow pathway and require the assistance of a human reader/grader.

iGradingM

The iGradingM ARIAS has a class I CE mark and performs 'disease/no disease' grading for diabetic retinopathy.[19] It was developed at the University of Aberdeen, uses previously published algorithms to assess both image quality and disease and has been described in detail elsewhere.[19] A previously trained automated classifier on a set of 35 images containing 198 individually annotated microaneurysm or dot haemorrhages was used in its development. It was first deployed on a large-scale population in the Scottish diabetic retinopathy screening programme in 2010, after being validated using several large screening populations in Scotland,[15,18–21] the largest being 78,601 single-field 45° colour fundus images from 33,535 consecutive patients.[15] In this retrospective study, 6.6% of the cohort had referable retinopathy and iGradingM attained a sensitivity of 97.3% for referable retinopathy. The specificity was not reported in the paper.[15] iGradingM has been further validated on 8271 screening episodes from a south London population in the UK.[31] Here, there was a higher percentage (7.1%) of referable disease, and sensitivity of 97.4–99.1% and specificity of 98.3–99.3% were attained depending on the configuration used.

IDx-DR

The IDx-DR has CE approval as a class IIa medical device for sale in the European Economic Area.[32] It utilises the Iowa Detection Program, which consists of previously published algorithms, including features such as image quality assessment, detection of microaneurysms, haemorrhage, cotton wool spots and a fusion algorithm that combines these analyses to produce the diabetic retinopathy index.[33–35] This index represents a dimensionless number from 0 to 1 and represents the likelihood that the image contains referable disease.

The IDx-DR has been validated in several large screening populations.[16,36] Most recently, the IDx-DR was validated on 1748 eyes with single-field 45° colour fundus images acquired in French primary care clinics.[36] In this study, the proportion of referable diabetic retinopathy was high, at 21.7%, and sensitivity and specificity was reported as 96.8% and 59.4%, respectively.

Retmarker

Retmarker was developed at Coimbra University, Portugal, and has been used in screening programmes in Portugal since 2011. It was launched commercially in 2011 by 'Critical Health', which has since been renamed Retmarker Ltd, and attained CE approval as a class IIa medical device in April 2010. Retmarker

includes an image quality assessment algorithm which has been validated on publicly available data sets,[37] and a co-registration algorithm, which allows comparisons of the same location in the retina to be made between visits.[38]

One study showed that Retmarker can detect referable retinopathy with a sensitivity of 95.8% and a specificity of 63.2%.[39] Retmarker can also perform predictive longitudinal analysis using microaneurysm turnover. Two independent prospective longitudinal studies using the Retmarker on single-field images have demonstrated the relationship of increased microaneurysm turnover rates identified by the Retmarker program and an increased rate for developing clinically significant macular oedema.[40,41]

EyeArt

EyeArt is a class IIa marked ARIAS for sale in the European Economic Area. The screening system is engineered to work on the cloud cluster (Amazon Elastic Cloud; Amazon EC2, Amazon, Seattle, WA, USA).

EyeArt has been validated on diabetic data sets and a recent abstract showed that EyeArt produced a refer/no refer screening recommendation for each patient in the Messidor 2 data set, a large, real-world public data set.[42] This data set comprises images from 874 patients (1748 eyes). EyeArt screening sensitivity was 93.8% [95% confidence interval (CI) 91.0% to 96.6%] and had a specificity of 72.2% (95% CI 68.6% to 75.8%).

Human grader performance and current screening

The effectiveness of different screening modalities has been widely investigated. UK studies show sensitivity levels for the detection of sight-threatening diabetic retinopathy of 48–82% for optometrists and 65% for ophthalmologists. Sensitivities for the detection of referable retinopathy by optometrists have been found in the range of 77–100%, with specificities between 94% and 100%.[43] Photographic methods currently use digital images with subsequent grading by trained individuals. Sensitivities for the detection of sight-threatening diabetic retinopathy have been found to range from 87% to 100% for a variety of trained personnel reading mydriatic 45° retinal photographs, with specificities of 83–96%. Compared with examination by an ophthalmologist, two, 45° field photographs with pupil dilatation has been reported to have a sensitivity of 95% in identifying diabetic patients with sight-threatening retinopathy (specificity 99%) and 83% sensitivity for detecting any retinopathy (specificity 88%).[44] The British Diabetic Association (Diabetes UK) established standard values for diabetic retinopathy screening programmes of at least 80% sensitivity and 95% specificity[45] based on recommendations from a first National Workshop on Mobile Retinal Screening summarised in a paper by Taylor *et al.*[46]

The standard method of grading digital colour photographs in the UK uses trained human graders who meet specific quality standards, with multiple possible levels of grading and quality control checks. One key challenge with all studies that measure diagnostic accuracy is the determination of an appropriate gold standard. It is debatable whether or not a gold standard exists in the detection of diabetic retinopathy, but various methods are thought to have better properties than others. The two methods that might possibly be considered as gold standard are seven-field stereoscopic photography and biomicroscopy carried out by a skilled ophthalmologist through dilated pupils. However, these methods have been compared in a number of studies[47–50] and do not show perfect agreement.[51] Hence, it is clear that one or both allow fairly frequent errors in detecting retinopathy. It is not possible to decide objectively which is in error, but Kinyoun *et al.*[47] did subject disagreements to an expert review, which tended to favour the seven-field photography, with two errors, over the indirect ophthalmoscopy, with 12 errors. Moss *et al.*[49] also examined the disagreement closely and concluded that many involved detection of microaneurysms from photographs that were not detected by ophthalmoscopy. No matter which method was correct, this might suggest that disagreements tend to happen in milder disease states. The direct comparisons of indirect ophthalmoscopy (used by an ophthalmologist) with seven-field photography also raises questions about

the standards of 80% sensitivity and 95% specificity[45] seen as desirable by the British Diabetic Association (Diabetes UK) from The National Workshop on Mobile Retinal Screening in 1994 summarised in a paper by Taylor *et al.*[46] As these standards were not invariably met in comparisons of these two 'gold standard' methods, they may represent an unrealistic target for other methods.

Rationale for the study

Current retinal screening programmes are labour intensive (requiring one or more trained human graders) and with the growing burden of screening there has been increasing interest in systems that allow automated detection of diabetic retinopathy. In Scotland, one automated system is deployed in routine screening. The Scottish Screening Programme utilises one image per eye (field) taken for screening, as opposed to the two-field images used in the NHS DESP. A number of commercially available CE-marked systems can potentially differentiate those who have sight-threatening diabetic retinopathy or other retinal abnormalities from those at low risk of progression to sight-threatening retinopathy. The diagnostic accuracy of these computer detection systems has been reported to be comparable with that of experts; however, the independent validity of these systems and clinical applicability to 'real-life' screening remains unclear. Studies to date have tended to have a low rate of severe levels of diabetic retinopathy in the test population. This means that most studies are inadequately powered to show efficacy of the automated software in identification of severe diabetic retinopathy. For example, if proliferative diabetic retinopathy is rare or absent in the test population, a detection program unable to detect proliferative diabetic retinopathy may seem to perform well. However, in a different population with a higher number of people with proliferative disease or with a different ethnicity or age profile, it may miss patients who require immediate treatment. Moreover, these image analysis systems are not currently authorised for use in the NHS DESP.

There is a need for an independent evaluation of the clinical effectiveness and cost-effectiveness of available ARIAS with CE marking, to inform potential deployment in the NHS DESP. This study addresses this important clinical need in a relevant setting of the NHS DESP. It is based on a large population with sufficient numbers of patients with severe levels of diabetic retinopathy in the test population using the multi-image per eye photographic standard and a spectrum of different ethnicities, age groups and sexes, and a range of cameras typically approved for NHS use. The cost-effectiveness of the intervention specifically for incorporation in different parts of the screening pathway will also be addressed.

Research aim

This report presents the research methods, analysis plan and findings of work commissioned by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme (project number 11/21/02) to evaluate three commercially available automated grading systems against the NHS DESP manual grading in a large population of patients diagnosed with diabetes mellitus attending a real-life diabetic retinopathy screening programme. The health economics of automated versus manual grading provides a key decision-making tool to determine whether or not automated screening should take place, and how it should take place, for diabetic retinopathy in the future.

Research objectives

Primary objectives

This is a measurement comparison study to quantify the screening performance and diagnostic accuracy of commercially available automated grading systems compared with the NHS DESP manual grading. Screening performance was re-evaluated after a subset of images underwent arbitration by an approved Reading Centre to resolve discrepancies between the manual grading and the automated system. Health

economics examined the cost-effectiveness of (1) replacing level 1 manual graders with automated grading and (2) using the automated software prior to manual grading.

Secondary objectives

Alternative reference standards were considered based on a consensus between the ARIASs and the NHS DESP manual grades if performance of ARIASs was sufficiently high for the detection of any retinopathy. Exploratory analyses examined the influence of patients' ethnicity (Asian, black African Caribbean and white European), age, sex and camera on screening performance of ARIASs. Issues related to the implementation of ARIASs to the hospital eye screening service at the Homerton University Hospital Foundation Trust was evaluated prospectively.

Chapter 2 Methods

Study design, population and setting

The study design was an observational retrospective study based on data from consecutive diabetes mellitus patients aged ≥ 12 years who attended the annual Diabetes Eye Screening programme of the Homerton University Hospital Foundation Trust in London between 1 June 2012 and 4 November 2013, which adhered to the NHS DESP guidelines.[52,53] The aim was to obtain around 20,000 unique screening episodes that had been graded manually.

Image acquisition and manual grading

The screening protocols used in the NHS DESP at the time of this study have been published and an updated version of that used is available online.[54] In brief, the protocol requires retinal photography imaging through a dilated pupil to capture four images per patient; for each eye, one image centred on the optic disc and one image centred on the macula. In routine screening practice, additional images are often taken and stored on the screening software to ensure that enough images of sufficient quality for retinal grading are obtained and to document other pathology (e.g. a cataract). Up to three human graders who meet the NHS DESP quality assurance (QA) standards assess the images to determine a disease severity grade and produce a 'final grade' for each eye according to the highest level of severity observed. The grading classifications for the 'eye for which action is most urgently required' in order of increasing severity are no retinopathy (R0), background retinopathy (R1), no maculopathy (M0), ungradable (U), (classification) maculopathy (M1), pre-proliferative retinopathy (R2) and proliferative retinopathy (R3).[52,53] Level 2 grading of images is carried out by more senior graders. Disagreements between level 1 and level 2 graders for episodes that are potentially M1 or R2 are sent to a third grader for arbitration, whose assessment is final. Patients with grades U, M1, R2 and R3 are referred to hospital eye service ophthalmologists, whereas patients with grades R0 and R1 are invited for rescreening in 12 months.

In the screening service for much of the period during which images were retrieved, patients whose photographs were ungradable at their previous screening episodes, for example because of a known cataract that degraded the quality of retinal photography, were 'technically failed' and underwent slit-lamp biomicroscopy by optometrists in a clinic adjacent to the photographic screening clinic. There were 1243 such patients, and they have been omitted from the data set (as they have no gradable images). This reduced the percentage of ungradable images in the set analysed. All manual grades were stored and accessed using an electronic data storage system (Digital Healthcare OptoMize). Patients with no perception of light were excluded as per the standard national guidance.

Reader experience

The community diabetic retinopathy screening programme that was studied had a stable grading team of 18 full-time and part-time optometrists and non-optometrist graders performing well against the national standards and with City and Guilds accreditation relevant to their designation within the grading programme. Performance against national standards is reviewed and reported quarterly at programme board meetings. In addition, the programme was externally quality assured by the national team in 2008, and the 2012 programme performance for the grading team was equivalent to the national average on the six external QA tests reported for that year at the programme board. There was a good level of intergrader agreement within the programme and for arbitration of images and agreement at primary and secondary grader level. All non-optometrist graders had assessed between 1000 and 2000 image sets in the 1-year period (2012–13) and the optometrist graders had assessed > 500 image sets. Hence, the study team is confident that this programme is representative of a diabetic retinal screening service that is performing to the standard expected within the English NHS.

Image management

Images in the main study were extracted using a custom-written extraction tool: Data Export Module for OptoMize version 3 (Digital Healthcare, Cambridge, UK). This software generates a data file containing a number of variables (see *Appendix 1*) including a patient pseudoanonymised identification (ID) code generated by the software, retinopathy grading by each grader at a patient level, visual acuity, camera type, outcome recommended by each grader and final manual grade. A unique image ID was used to link to the images exported by OptoMize Data Export Module as Joint Photographic Experts Group (JPEG) images. The patient pseudoanonymised ID was used to link the images at a patient level. A separate export was undertaken from OptoMize to extract data on patient demographic characteristics (age, sex, ethnicity and camera type) linked back to the patient using the pseudoanonymised image ID code.

A test set of 2500 patient images from the same screening programme (Homerton University Hospital Foundation Trust, London), but not including any images from the main study, were provided to all four ARIAS vendors for the same length of time to develop the importation of the image files and output the final ARIAS classification. Although the NHS DESP stipulated two images per eye at the time of the study image acquisition in 2012–13, in reality additional images were often taken by the screening photographer and additional lens shots were often taken to document the presence of lens opacity. The four ARIAS vendors were advised of these issues to allow them to develop a system to handle non-retinal images and a variable number of images per eye. The vendors were notified that the images were not annotated as to whether or not the image was a retinal image.

The image test sets were exported onto encrypted drives and the corresponding data file, which was stripped of all grading results, was placed on four identical servers to allow linkage at a patient level. Three of the vendors installed their software on the servers [Intel® Xeon® (Intel Corporation, Santa Clara, CA, USA) six core processors, 8 GB random-access memory and Windows Server 2008 R2 (Microsoft Corporation, Redmond, WA, USA)], the fourth vendor (IDx, LLC) supplied a hardware unit to link to the server that processed the test images. Each of the vendors had remote access with administrative privileges to install, test and verify their software. After a predetermined date, remote access ceased and the vendors were not allowed to access installation of their software. In addition, the larger data set for the main study described below was loaded on to each of the servers after this date and processed following the instructions received from each of the companies.

To assess implementation of ARIASs in a real-life screening clinic, the ARIASs on each individual server were connected to the Homerton Univesity Hospital Foundation Trust network connected to the Digital Healthcare system used to store diabetic screening images and associated data. Images were exported on a daily basis using the OptoMize Data Export Module. This attempt to process images in a live, busy screening programme was used in order to understand the technical issues that need to be addressed to allow introduction into the NHS.

Anonymisation of images

Digital Healthcare OptoMize was used to store and extract data from the manual grading process. Data were extracted for 20,258 patients and anonymised to exclude personal identifying data. A second unique identifier was created for patient-screening episodes (and images within an episode) so that data sent for arbitration grading could not be linked with the original anonymised patient episode code.

Automated grading systems

Automated systems for diabetic retinopathy detection with a CE mark obtained or applied for up to 6 months after July 2013 were eligible for evaluation. Three software systems were selected from a literature search and discussion with experts and all three were agreed to participate in the study: iGradingM,[31] Retmarker and IDx-DR.[36] For commercial reasons, IDx, LLC withdrew from the study before the analysis, and in 2013 another software system (EyeArt) asked to join the study, confirming that it

would meet CE-mark eligibility criteria. Three systems (iGradingM, Retmarker and EyeArt) then processed all images from all screening episodes. Permission to extract pseudoanonymised images and process all images by the automated systems was obtained from the Caldicott Guardian. No formal ethics approval was required as all extracted images were anonymised and no change in clinical pathway occurred.

All automated systems are designed to identify cases of diabetic retinopathy of R1 or above. EyeArt is additionally designed to identify cases requiring referral to ophthalmology (diabetic retinopathy of U, M1, R2 or R3). Retmarker and EyeArt process all the images associated with a screening episode and provide a classification per episode. The iGradingM system provides an outcome classification for each image.

Outcome classification from the automated software

To compare screening performance across the automated systems with the manual reference standard, episodes were classified as (1) 'disease present' or 'technical failure' or (2) 'disease absent'. Retmarker automatically classifies screening episodes in this way. For EyeArt, episodes were classified as (1) 'disease', 'definite disease' and 'ungradable' or (2) as 'definite no disease' and 'no disease'. For iGradingM, episodes were classified as 'disease present' or 'ungradable' using the following approach:

1. If the outcome of at least one image in a screening episode is classified as 'disease', then the outcome classification for the episode will be 'disease present'.
2. If the outcome of at least one image in a screening episode is classified as 'ungradable' and 'no disease' is detected in all the other images of the screening episode, then the outcome classification for the episode will be 'ungradable'.
3. If the outcome of all images in a screening episode is 'no disease', then the outcome classification for the episode will be 'disease absent'.

All automated systems also report a numerical value for a decision statistic; values above a certain threshold imply that diabetic retinopathy is present. EyeArt additionally provides a classification for referable versus non-referable disease that is intended to distinguish U, M1, R2 and R3 from R0M0 and R1M0. Screening performance using this outcome classification was also assessed.

Reference standards

Final human grade

Screening performance of each automated system was assessed using the final human manual grade as the reference standard, as well as that modified by arbitration.

Final human grade modified by arbitration

An internationally recognised fundus photographic reading centre (Doheny Image Reading Centre, Los Angeles, CA, USA),[55] masked to the final human grading and ARIAS grading, carried out arbitration on disagreements between the final human grade and the grades assigned by the automated systems. Arbitration was limited to 1700 patient episodes. Emphasis was placed on arbitration of all discrepancies with final manual grades R3, R2 and M1. A random sample of screening episodes when two or more systems disagreed with the final human grade of R1 or R0 were also sent for arbitration. *Table 1* summarises the process of arbitration and how the final reference grade was modified. The percentage of screening episodes falling into cases 3 and 4 (see *Table 1*) that were sent for arbitration was determined by the sensitivity and false-positive rate of the automated systems. The British Diabetic Association (Diabetes UK) sets standards for diabetic retinopathy screening of at least 80% sensitivity for sight-threatening retinopathy[45] and any automated system that did not meet this operating standard was not considered in the process of identifying episodes for arbitration. Anonymised images were sent to the reading centre and diabetic retinopathy feature-based grading was entered into a database provided by the study team (see *Appendix 2*). The SABRE software application (Netsima Ltd, Port Talbot, Wales) was used to convert feature-based grading at an image level to a retinopathy grade at eye and patient level.

TABLE 1 Process of selection of discrepancies between human grade and automated systems for external arbitration

Scenario	Criteria (combined human and software)		Arbitration	Final human grade modified by arbitration
1	Human grade	M1, R2 or R3	No arbitration	Final human grade
	All software	Disease present		
2	Human grade	M1, R2 or R3	All images sent to reading centre	Grade from reading centre
	One or more software	Disease not present		
3	Human grade	R0	For each software a proportion of all screening episodes were sent to the reading centre (random selection)[a]	Grade from reading centre
	Two or more software	Disease present		
4	Human grade	R1	For each software a proportion of all screening episodes were sent to the reading centre (random selection)[a]	Grade from reading centre
	Two or more software	Disease not present		
5	Human grade	R0	No arbitration	Final human grade
	Only one software	Disease present		
6	Human grade	R1	No arbitration	Final human grade
	Only one software	Disease not present		
7	Human grade	R0	No arbitration	Final human grade
	All software	Disease not present		
8	Human grade	R1	No arbitration	Final human grade
	All software	Disease present		
9	Human grade	R0, R1, M1, R2 or R3	No arbitration	Final human grade
	More than one software	Technical failure		
10	Human grade	R0, R1, M1, R2 or R3	Proceed as in cases 1–8 considering the grades of the remaining software	
	Only one software	Technical failure		
11	Human grade	U	No arbitration	Final human grade

M1, maculopathy; R0, no retinopathy; R1, background retinopathy; R2, pre-proliferative retinopathy; R3, proliferative retinopathy; U, ungradable images.

a Appropriate weighting was used in the random selection process of screening episodes to give increasing preference to episodes when more than one or all software agreed with each other but disagreed with the final human grade.

Consensus grade

If the automated systems achieved at least 80% sensitivity in identifying cases of diabetic retinopathy of R1 or above and 100% sensitivity for R3, two 'consensus grades' were considered.

1. a majority classification based on data from the final human grade and the automated system classifications (disease absent/disease present or technical failure)
2. a majority classification based on the automated system classification only (disease absent/disease present or technical failure).

The change in screening performance for each system was compared with the consensus grade.

Statistical methods

Primary analyses

The primary analysis was to assess the screening performance of the automated systems using the final manual grade (with and without reading centre arbitration) as the reference standard. The sensitivity, false-positive rate (specificity) and likelihood ratios of each automated grading system were determined for (1) any retinopathy or ungradable episodes (final human grades R1, U, M1, R2 or R3 vs. R0), (2) episodes with final human grades U, M1, R2 or R3 versus final human grades R1 or R0 and (3) for each grade of retinopathy separately (R0, R1, M1, R2, R3 and U) using data from all 20,258 first screening episodes. The diagnostic accuracy of sensitivity, the false-positive rate and likelihood ratios were defined by 95% CIs obtained using bootstrapping. Clinical interpretation of the lower limit of the 95% CI for the detection of R3, R2 and M1 grades provides an indication of the potential number of screening episodes requiring clinical intervention that could be missed by the automated systems. The upper confidence limit for false-positive rates (and lower limit for specificity) for retinopathy grades R1 and R0 allowed the additional potential number of screening episodes requiring further investigation to be quantified. The level of uncertainty around these estimated lower and upper limits was obtained using bootstrapping.

Secondary analyses

The consensus grade approach outlined above as an alternative reference standard was considered for the evaluation of sensitivity and false-positive rates for the automated systems. Exploratory analyses examined the influence of patient demographic factors on the performance of the automated systems. Sensitivity and false-positive rates were summarised for each automated system for the three main ethnic groups (Asian, black African Caribbean and white European), tertiles of age, sex and camera type, using the final manual grade as modified by arbitration as the reference standard. Multiple variable logistic regression models were used to statistically assess whether or not the odds of the 'disease' versus 'no disease' outcome with the automated system were modified by age, sex, ethnicity or camera type. Patients with the highest manual grade in the worst eye modified by arbitration of R1, M1, U, R2 and R3 were classified as 'cases' and patients with grades R0M0 were classified as non-cases. These cases versus non-case definitions were chosen as the ARIASs are designed to detect retinopathy of R1 or worse. Logistic regression against age (categorised in tertiles), sex, ethnicity and camera type as exploratory variables and interaction terms for each demographic variable with the ARIAS outcome (disease vs. no disease), simultaneously adjusting for the other factors, was explored. An interaction with age was examined first, followed by sex, ethnicity and, finally, camera type. Although the study was not specifically powered to examine interactions, they were examined to identify potential hypotheses requiring further investigation. In view of the number of statistical significance tests for interaction terms, the *p*-value was set a priori to < 0.01 to determine statistical significance.

Sample size

A pilot study of 1340 patient screening episodes at St George's Hospital NHS Trust, London (three of the applicants were part of this pilot study), revealed that the prevalence of R0, R1, M1, R2 and R3 is 68%, 24%, 6.1%, 1.2% and 0.5%, respectively. In addition, in this pilot study one software (iGradingM) was compared with manual grading as the reference standard. The sensitivities for R1, M1, R2 and R3 were 82%, 91%, 100% and 100%, respectively, and 44% of R0 were graded as 'disease present'. The prevalence of different grades of retinopathy and the likely sensitivity of the software guided the sample size required. The sample size calculation was based on the number of screening episodes to ensure that the lower limit of the 95% CI for sensitivity of automated grading did not fall below 97% for retinopathy classified as R3 by human graders. R3 is the rarest and most serious outcome (0.5%) and governed the sample size. Under these assumptions 24,000 screening episodes would be needed. However, should the sensitivity fall to 90%, the associated binomial exact 95% CI for the detection of R3 would range from 83% to 95%. If the prevalence of R3 were instead 1%, only 20,000 episodes would be needed to achieve a similar degree of accuracy for the 95% CI. 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 and, based on pilot data, was expected to provide sufficient R3 events (a more detailed participant flow diagram is given in *Chapter 3*).

Table 2 outlines the prevalence of retinopathy grades reported in this screening programme and the precision with which different levels of sensitivity would be estimated for four potential levels of sensitivity between 80% and 95%. For example, if the automated system detects 85% R3, the 95% CI around this estimate of 85% sensitivity will be from 81% to 89%.

Health economic analyses

Details of the methods used in the health economic analysis, together with the outcomes, are given in *Chapter 4*.

In brief, a health economic analysis was carried out to investigate the economic implications involved if (1) an automated system were to replace level 1 human graders and (2) the automated system were to be used as a filter prior to level 1 grading.

Altered thresholds

The question is whether or not the automated screening of retinal images will identify a sufficiently large proportion of images as not requiring human intervention to justify (1) the cost of the software and (2) the risk of harm that necessarily follows from automation. In any such calculation there is a trade-off between sensitivity and specificity: the larger the proportion of images identified as not requiring human reading, the greater the risk that abnormality will be missed. Manufacturers of ARIASs fix the 'operating point' to provide a trade-off for regulatory approval and a commercial case made for the system's effectiveness. However, the chosen operating point may not be optimal for all screening programmes. We explored the consequences of different operating points.

TABLE 2 Prevalence of retinopathy grade and associated 95% CIs for sensitivity ranging from 80% to 95%

	Screening episodes	95% CI[a] for specified sensitivity of an automated system (%)			
Final human grade	*n* (prevalence, %)	80%	85%	90%	95%
Retinopathy grades					
R0	12,727 (63)	79 to 81	84 to 86	89 to 91	94.6 to 95.4
R1	4749 (23)	79 to 81	84 to 86	89 to 91	94 to 96
M1	1609 (7.9)	78 to 82	83 to 87	89 to 91	94 to 96
R2	637 (3.1)	77 to 83	82 to 88	88 to 92	93 to 97
R3	236 (1.2)	75 to 85	81 to 89	86 to 94	92 to 97
U	300 (1.5)	75 to 84	81 to 89	86 to 93	92 to 97
Combination of grades					
R1, U, M1, R2 or R3	7531 (37)	79 to 81	84 to 86	89 to 91	94.5 to 95.4
U, M1, R2 or R3	2782 (13.7)	79 to 81	84 to 86	89 to 91	94 to 96
Total	20,258 (100)	–	–	–	–

a For R0, the reported 95% CI corresponds to specificity (the true-negative rate).

Chapter 3 Results

Data extraction from the Homerton diabetic eye screening programme

Box 1 shows the number of screening episodes extracted from the Homerton DESP for screening visits between 1 June 2012 and 4 November 2013. *Box 1* shows the degree of data completeness for manual grades. Data from 28,079 screening episodes were obtained (142,018 images) and 20,258 patients were entered into the study. Data from first screening episodes (20,258 patients and 102,856 images) were included in the analysis. The data available for each episode included a unique anonymised patient ID code, episode screening date, retinal image filenames associated with each screening episode, camera type used and manual grades for retinopathy, maculopathy and associated assessment of image quality for each eye from each grader that processed the images. The final manual grade of each episode was obtained by combining information on the quality of the retinal image with retinopathy and maculopathy grades for each eye from each human grader using the highest manual grade, in accordance with the NHS DESP guidelines.[52,53] Individuals with one eye were identifiable in the data set and image quality and manual grades from one eye only were included.

Data on age, sex and ethnicity were available for 20,212 patients (*Table 3*). The median age of patients was 60 years (range 10–98 years), with 0.5% aged < 18 years and 37% aged > 65 years; 41% of patients were white European, 35% were Asian and 20% were black African-Caribbean. White European patients were more likely than Asian or black African-Caribbean patients to have retinopathy graded R0M0, despite being, on average, 7 years older than Asian and black African-Caribbean patients. Conversely, white European patients were less likely to be graded U, R1M1, R2 or R3 (11.2%) than Asian patients (15.0%) or black African-Caribbean patients (16.0%). Black African-Caribbean patients were marginally more likely than white European and Asian patients to have ungradable images (2.0% vs. 1.3% and 1.3%, respectively).

BOX 1 Data extraction of diabetic patients attending the Homerton DESP

Digital Healthcare OptoMize was used to extract data on manual grading for screening episodes between 1 June 2012 and 4 November 2013

A total of 30,529 screening episodes identified from 21,930 patients

A total of 2298 episodes excluded as there were no data on manual grades owing to a data extraction error in Digital Healthcare OptoMize.

A total of 28,231 screening episodes with data on manual grading

A total of 152 episodes excluded:

- 147 episodes incomplete grading pathways as screening process incomplete
- five episodes in which the patient's vision was perception of light vision.

A total of 28,079 screening episodes including 142,018 images

Number of patients with at least one episode: 20,258 (102,856 images).

Number of patients with repeat episodes: 7821 (39,162 images).

TABLE 3 Demographic characteristics and prevalence of retinopathy by ethnicity among patients attending the Homerton DESP

Characteristic	Ethnic group					
	White European	Asian[a]	Black[b]	Mixed[c]	Other[d]	Total
Number of patients, *n* (%)	8358 (41.4)	7018 (34.7)	3957 (19.6)	136 (0.7)	743 (3.7)	20,212 (100)
Median age (years) (IQR)	64.2 (53.9–74.0)	57.5 (48.7–66.4)	56.8 (48.6–69.4)	52.9 (44.7–63.8)	58.5 (50.0–67.2)	60.0 (50.4–70.4)
Percentage male	56	55	49	55	55	54
Manual grade (worst eye)	***Prevalence* n *(%)***					
R0M0	5418 (64.8)	4320 (61.6)	2451 (61.9)	81 (59.6)	437 (58.8)	12,707
R1M0	2002 (24.0)	1648 (23.5)	873 (22.1)	25 (18.4)	190 (25.6)	4738
U	105 (1.3)	90 (1.3)	79 (2.0)	3 (2.2)	16 (25.6)	293
R1M1	504 (6.0)	621 (8.8)	394 (10.0)	17 (12.5)	70 (9.4)	1606
R2	250 (3.0)	247 (3.5)	107 (2.7)	7 (5.1)	25 (3.4)	636
R2M0	94 (1.1)	73 (1.0)	30 (0.8)	3 (2.2)	9 (1.2)	209
R2M1	156 (1.9)	174 (2.5)	77 (1.9)	4 (2.9)	16 (2.2)	427
R3	79 (0.9)	92 (1.3)	53 (1.3)	3 (2.2)	5 (0.7)	232
R3M0	30 (0.4)	25 (0.4)	16 (0.4)	1 (0.7)	1 (0.1)	73
R3M1	49 (0.6)	67 (1.0)	37 (0.9)	2 (1.5)	4 (0.5)	159
Combination of grades						
R0M0, R1M0	7420 (88.8)	5968 (85.0)	3324 (84.0)	106 (77.9)	627 (84.4)	17,445
U, R1M1, R2, R3	938 (11.2)	1050 (15.0)	633 (16.0)	30 (22.1)	116 (15.6)	2767
R1M0, U, R1M1, R2, R3	2940 (35.2)	2698 (38.4)	1506 (38.1)	55 (40.4)	306 (41.2)	7505
Total	8358 (100)	7018 (100.0)	3957 (100.0)	136 (100.0)	743 (100.0)	20,212

IQR, interquartile range.
a Asian includes Asian British, Indian, Pakistani, Bangladeshi and any other Asian background.
b Black includes black British, Caribbean, African and any other black background.
c Mixed includes patients of mixed ethnic origin.
d Other includes all other or unknown ethnic groups.
A total of 46 patients have missing data on demographic characteristics.

Table 31 in *Appendix 3* summarises the screening pathways for all first screening episodes for 20,258 patients; 10,788 episodes (53%) were graded R0M0 by a level 1 grader and did not go on to the level 2 grader. The second largest group comprised 4534 episodes (22%) that were coded as R1M0 by the level 1 grader and were then passed to the level 2 grader, who confirmed the R1M0 grade. Overall, the level 1 grader classified 295 episodes as U (≈1%); this is a relatively low level because patients known to be photographically ungradable were technically failed before image capture and underwent slit-lamp biomicroscopy in the clinic. The Homerton DESP graded M1 as either maculopathy with exudates within 1-disc diameter of centre of the fovea (M1a) or as maculopathy with exudates within the macula (M1b). M1a was graded when visual acuity was 6/6 or better, or 6/9 in the past with no deterioration when the new problem emerged and there were no new symptoms, and 6-month recall was needed. M1b was graded when the level of exudation was more than modest or if the visual acuity is 6/9 or worse, and referral to the hospital eye services was needed.

Screening performance of EyeArt software

Table 4 gives the sensitivity (detection rate) and false-positive rates for the EyeArt software by the highest (worst eye) manual retinopathy grade per episode as modified by the Reading Centre arbitration as the

TABLE 4 Screening performance of EyeArt software compared with manual grade modified by arbitration

	EyeArt outcome, *n* (%[a])		
Manual grade (worst eye)	No disease	Disease	Total, *n* (%[b])
Retinopathy grades			
R0M0	2542 (20)	10,254 (80)	12,796 (63)
R1M0	217 (5)	4401 (95)	4618 (23)
U	98 (23)	329 (77)	427 (2)
R1M1	73 (5)	1485 (95)	1558 (8)
R2	4 (1)	622 (99)	626 (3)
R2M0	3 (2)	190 (98)	193 (1)
R2M1	1 (0)	432 (100)	433 (2)
R3	1 (0)	232 (100)	233 (1)
R3M0	0 (0)	71 (100)	71 (0)
R3M1	1 (1)	161 (99)	162 (1)
Combination of grades			
R0M0, R1M0	2759 (16)	14,655 (84)	17,414 (86)
U, R1M1, R2, R3	176 (6)	2668 (94)	2844 (14)
R1M0, U, R1M1, R2, R3	393 (5)	7069 (95)	7462 (37)
Total	2935 (100)	17,323 (100)	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.

reference standard. The specificity for episodes graded R0M0 (i.e. no retinopathy) was 20%, equivalent to a false-positive rate of 80% (see *Table 4*). For episodes graded manually as R3 the sensitivity was 100%. For any retinopathy (manual grades of R1M0 or higher) the sensitivity was 95% and for grades R1M1 or higher (including ungradable images) the sensitivity was 94%. *Table 5* presents the diagnostic accuracy for the point estimates in *Table 4* and likelihood ratios. For example, the 95% CI for the estimated sensitivity for R3 was 97.0% to 99.9% and the bootstrapped 95% CI around the lower limit was 85.5% to 97.4% and around the upper limit was 99.5% to 99.9%. The likelihood ratios for the software show that patients with manual grades of R1M1, R2 and R3 were all approximately 1.2 times more likely to be classified as 'disease present' than patients with manual grades R0M0. Patients graded U, R1M1, R2, R3 were 1.12 times more likely than patients with manual grades R0M0 and R1M0 to be classified as 'disease present'. The comparable measures of screening performance, diagnostic accuracy and likelihood ratios as obtained prior to arbitration of manual grades by the Reading Centre are presented in *Tables 32* and *33* in *Appendix 3*. Measures of screening performance are very similar when comparing results before and after arbitration of manual grades.

EyeArt provides an alternative classification that attempts to identify cases requiring referral to ophthalmology (grades U, M1, R2 and R3); *Tables 6* and *7* present the findings using this alternative output from EyeArt. This showed a marked effect for patients graded R0M0. In *Table 4*, 20% were classified as 'no disease', compared with 41% classified as 'no refer'. The impact on the other retinopathy grades was less marked, with a reduction in sensitivity for R1M1 from 95% to 91%, and smaller changes in sensitivity for R2 and R3. *Table 7* shows that the precision of detection was not affected but the likelihood ratios were marginally increased. Very similar results were observed prior to arbitration (see *Appendix 3, Tables 34* and *35*).

TABLE 5 Screening performance of EyeArt software compared with manual grade modified by arbitration: 95% confidence limits and likelihood ratios

Manual grade (worst eye)	Proportion classified as disease present or technical failure			Likelihood ratio vs. R0 (95% CI)
	Estimate (95% CI)	Lower[a] (95% CI)	Upper[b] (95% CI)	
Retinopathy grades				
R0M0[c]	0.199 (0.192 to 0.206)	0.192 (0.185 to 0.198)	0.206 (0.199 to 0.212)	–
R1M0	0.953 (0.947 to 0.959)	0.947 (0.941 to 0.953)	0.959 (0.953 to 0.964)	1.189 (1.179 to 1.201)
U	0.770 (0.728 to 0.808)	0.728 (0.691 to 0.773)	0.808 (0.773 to 0.847)	0.961 (0.915 to 1.016)
R1M1	0.953 (0.941 to 0.963)	0.941 (0.930 to 0.951)	0.963 (0.954 to 0.971)	1.189 (1.173 to 1.204)
R2	0.994 (0.983 to 0.998)	0.983 (0.972 to 0.989)	0.998 (0.992 to 1.000)	1.240 (1.227 to 1.251)
R2M0	0.984 (0.953 to 0.995)	0.953 (0.850 to 0.967)	0.995 (0.859 to 0.999)	1.229 (1.000 to 1.247)
R2M1	0.998 (0.984 to 1.000)	0.984 (0.979 to 0.986)	1.000 (0.998 to 1.000)	1.245 (1.000 to 1.255)
R3	0.996 (0.970 to 0.999)	0.970 (0.855 to 0.974)	0.999 (0.995 to 0.999)	1.243 (1.000 to 1.256)
R3M0	1.000	–	–	1.248 (1.237 to 1.258)
R3M1	0.994 (0.958 to 0.999)	0.958 (0.851 to 0.963)	0.999 (0.862 to 0.999)	1.240 (1.000 to 1.254)
Combination of grades				
R0M0, R1M0[c]	0.158 (0.153 to 0.164)	0.153 (0.148 to 0.159)	0.164 (0.159 to 0.170)	–
U, R1M1, R2, R3	0.938 (0.929 to 0.946)	0.929 (0.920 to 0.941)	0.946 (0.939 to 0.957)	1.115 (1.102 to 1.128)[d]
R1M0, U, R1M1, R2, R3	0.947 (0.942 to 0.952)	0.942 (0.937 to 0.949)	0.952 (0.948 to 0.958)	1.182 (1.172 to 1.194)

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion.
c Estimates relate to the proportion classified as 'disease absent' (i.e. the specificity).
d Likelihood ratio estimated compared with R0M0 and R1M0 combined.

Screening performance of Retmarker software

Tables 8 and *9* provide the corresponding results for Retmarker software. The specificity for episodes graded manually as R0M0 was higher for Retmarker than for EyeArt (53% vs. 20%). However, the detection of R2 or R3 retinopathies was slightly lower than that for EyeArt, with sensitivities of 96.5% versus 99.4% for R2 and 97.9% versus 99.6% for R3. *Table 9* gives the measures of diagnostic accuracy around estimates of screening performance in *Table 8*. For example, for R2 the estimated sensitivity was 96.5% with a 95% CI of 94.7% to 97.7%. The bootstrapped 95% CI on the lower bound of this confidence limit ranged from 92.5% to 96.6% for Retmarker compared with 97.2% to 98.9% for EyeArt (see *Table 5*). The precision on the lower bound of the 95% confidence limit for sensitivity for R3 ranged from 92.1% to 96.8% for Retmarker and from 85.5% to 97.4% for EyeArt (see *Table 5*). Focusing on the lower bound for the detection of R3 suggests that the sensitivity was potentially better for Retmarker than for EyeArt, whereas for R2 EyeArt had higher sensitivity. However, when examining combined manual grades of U, R1M1, R2 and R3, the 95% CI on the lower bound for sensitivity ranged from 81.9% to 84.9% for Retmarker and from 92.0% to 94.1% for EyeArt (see *Table 5*). For these combined grades EyeArt had a higher sensitivity than Retmarker when examining the diagnostic accuracy around the lower bound of the 95% CI. The likelihood ratios for Retmarker were generally higher than those for EyeArt, but this appeared to be driven by the higher specificity of Retmarker for grades R0M0 rather than improved sensitivity for manual grades of R1M1 and above. *Tables 36* and *37* in *Appendix 3* give the corresponding results for Retmarker prior to arbitration, which are remarkably similar to *Tables 8* and *9*.

TABLE 6 Screening performance of EyeArt software compared with manual grade in the worst eye modified by arbitration: using EyeArt classification referable vs. non-referable retinopathy

Manual grade (worst eye)	EyeArt outcome, *n* (%[a])		Total, *n* (%[b])
	No refer	Refer	
Retinopathy grades			
R0M0[c]	5212 (41)	7584 (59)	12,796 (63)
R1M0	837 (18)	3781 (82)	4618 (23)
U	145 (34)	282 (66)	427 (2)
R1M1	137 (9)	1421 (91)	1558 (8)
R2	7 (1)	619 (99)	626 (3)
R2M0	5 (3)	188 (97)	193 (1)
R2M1	2 (0)	431 (100)	433 (2)
R3	2 (1)	231 (99)	233 (1)
R3M0	1 (1)	70 (99)	71 (0)
R3M1	1 (1)	161 (99)	162 (1)
Combination of grades			
R0M0, R1M0[c]	6049 (35)	11,365 (65)	17,414 (86)
U, R1M1, R2, R3	291 (10)	2553 (90)	2844 (14)
Total	6340 (100)	13,918 (100)	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.
c Estimates relate to the proportion classified as 'disease absent' (i.e. the specificity).

TABLE 7 Screening performance of EyeArt software compared with manual grade modified by arbitration: 95% confidence limits and likelihood ratios using EyeArt classification referable vs. non-referable retinopathy

Manual grade (worst eye)	Proportion classified as refer or technical failure			Likelihood ratio vs. R0 + R1M0 (95% CI)
	Estimate (95% CI)	Lower[a] (95% CI)	Upper[b] (95% CI)	
Retinopathy grades				
R0M0[c]	0.407 (0.399 to 0.416)	0.399 (0.390 to 0.406)	0.416 (0.407 to 0.423)	–
R1M0	0.819 (0.807 to 0.830)	0.807 (0.797 to 0.819)	0.830 (0.820 to 0.841)	–
U	0.660 (0.614 to 0.704)	0.614 (0.572 to 0.675)	0.704 (0.664 to 0.762)	1.012 (0.950 to 1.102)
R1M1	0.912 (0.897 to 0.925)	0.897 (0.885 to 0.911)	0.925 (0.914 to 0.937)	1.398 (1.375 to 1.421)
R2	0.989 (0.977 to 0.995)	0.977 (0.965 to 0.986)	0.995 (0.988 to 0.999)	1.515 (1.493 to 1.534)
R2M0	0.974 (0.939 to 0.989)	0.939 (0.901 to 0.961)	0.989 (0.969 to 0.998)	1.493 (1.447 to 1.522)
R2M1	0.995 (0.982 to 0.999)	0.982 (0.675 to 0.985)	0.999 (0.688 to 1.000)	1.525 (1.000 to 1.544)
R3	0.991 (0.966 to 0.998)	0.966 (0.677 to 0.973)	0.998 (0.690 to 0.999)	1.519 (1.000 to 1.542)
R3M0	0.986 (0.907 to 0.998)	0.907 (0.868 to 0.928)	0.998 (0.992 to 0.999)	1.511 (1.485 to 1.536)
R3M1	0.994 (0.958 to 0.999)	0.958 (0.681 to 0.963)	0.999 (0.694 to 0.999)	1.523 (1.000 to 1.542)
Combination of grades				
R0M0, R1M0[c]	0.347 (0.340 to 0.354)	0.340 (0.332 to 0.346)	0.354 (0.346 to 0.360)	–
U, R1M1, R2, R3	0.898 (0.886 to 0.908)	0.886 (0.876 to 0.899)	0.908 (0.899 to 0.920)	1.375 (1.354 to 1.396)[d]

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion.
c Estimates relate to the proportion classified as disease absent (i.e. the specificity).
d Likelihood ratio estimated compared with R0M0 and R1M0 combined.

TABLE 8 Screening performance of Retmarker software compared with manual grade modified by arbitration

Manual grade (worst eye)	Retmarker outcome, *n* (%[a])		Total, *n* (%[b])
	No disease	Disease	
Retinopathy grades			
R0M0	6730 (53)	6066 (47)	12796 (63)
R1M0	1585 (34)	3033 (66)	4618 (23)
U	194 (45)	233 (55)	427 (2)
R1M1	207 (13)	1351 (87)	1558 (8)
R2	22 (4)	604 (96)	626 (3)
R2M0	5 (3)	188 (97)	193 (1)
R2M1	17 (4)	416 (96)	433 (2)
R3	5 (2)	228 (98)	233 (1)
R3M0	1 (1)	70 (99)	71 (0)
R3M1	4 (2)	158 (98)	162 (1)
Combination of grades			
R0M0, R1M0	8315 (48)	9099 (52)	17414 (86)
U, R1M1, R2, R3	428 (15)	2416 (85)	2844 (14)
R1M0, U, R1M1, R2, R3	2013 (27)	5449 (73)	7462 (37)
Total	8743 (100)	11,515 (100)	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.

TABLE 9 Screening performance of Retmarker software compared with manual grade modified by arbitration: 95% confidence limits and likelihood ratios

Manual grade (worst eye)	Proportion classified as disease present or technical failure			Likelihood ratio vs. R0 (95% CI)
	Estimate (95% CI)	Lower[a] (95% CI)	Upper[b] (95% CI)	
Retinopathy grades				
R0M0[c]	0.526 (0.517 to 0.535)	0.517 (0.510 to 0.527)	0.535 (0.527 to 0.544)	–
R1M0	0.657 (0.643 to 0.670)	0.643 (0.632 to 0.657)	0.670 (0.659 to 0.684)	1.385 (1.353 to 1.431)
U	0.546 (0.498 to 0.592)	0.498 (0.457 to 0.554)	0.592 (0.552 to 0.647)	1.151 (1.069 to 1.277)
R1M1	0.867 (0.849 to 0.883)	0.849 (0.832 to 0.863)	0.883 (0.868 to 0.896)	1.829 (1.789 to 1.876)
R2	0.965 (0.947 to 0.977)	0.947 (0.925 to 0.966)	0.977 (0.962 to 0.988)	2.035 (1.993 to 2.088)
R2M0	0.974 (0.939 to 0.989)	0.939 (0.898 to 0.964)	0.989 (0.967 to 0.999)	2.055 (1.988 to 2.118)
R2M1	0.961 (0.938 to 0.975)	0.938 (0.913 to 0.961)	0.975 (0.959 to 0.989)	2.027 (1.976 to 2.080)
R3	0.979 (0.949 to 0.991)	0.949 (0.921 to 0.968)	0.991 (0.978 to 0.998)	2.064 (2.015 to 2.109)
R3M0	0.986 (0.907 to 0.998)	0.907 (0.882 to 0.929)	0.998 (0.993 to 0.999)	2.080 (2.037 to 2.125)
R3M1	0.975 (0.936 to 0.991)	0.936 (0.559 to 0.957)	0.991 (0.573 to 0.999)	2.057 (1.990 to 2.106)
Combination of grades				
R0M0, R1M0[c]	0.477 (0.470 to 0.485)	0.470 (0.464 to 0.478)	0.485 (0.478 to 0.493)	–
U, R1M1, R2, R3	0.850 (0.836 to 0.862)	0.836 (0.819 to 0.849)	0.862 (0.846 to 0.874)	1.626 (1.593 to 1.659)[d]
R1M0, U, R1M1, R2, R3	0.730 (0.720 to 0.740)	0.720 (0.709 to 0.731)	0.740 (0.728 to 0.750)	1.540 (1.507 to 1.578)

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion.
c Estimates relate to the proportion classified as 'disease absent' (i.e. the specificity).
d Likelihood ratio compared with R0M0 and R1M0 combined.

Screening performance of iGradingM software

Table 10 shows results for iGradingM using the outcome classification described in *Chapter 2* prior to arbitration by the Reading Centre. None of the screening episodes was classified as 'no disease'. This raised concerns about the iGradingM software processing both disc- and macular-centred images. We examined this directly on the subset of screening episodes that underwent arbitration and we were able to identify that 99% of disc-centred images were classified as 'ungradable' by iGradingM compared with 36% of macular-centred images. Note that *Table 10* is based on the hierarchy for grading outcome at episodes level for iGradingM as defined a priori in *Chapter 2*.[56] If at least one image is classified as 'disease' the outcome at a patient level will be 'disease' even if the disc-centred images are ungradable. However, if at least one image is 'ungradable' and all others are 'no disease' then 'ungradable' is the worst outcome at a patient level.

Arbitration on subset of screening episodes

Table 11 shows the agreement between the manual grade from the Homerton DESP and the Doheny image Reading Centre grade for 1700 screening episodes (8373 images) that underwent arbitration. Some disagreement exists for each grade, but for combinations of grades the agreement was reasonable. For example, the manual grade from the Homerton DESP was R0M0 ($n = 862$) or R1M0 ($n = 362$) in 1224 episodes. The Doheny Image Reading Centre confirmed a grade of R0M0 or R1M0 in 81% [(734 + 22 + 149 + 91)/1224] and an additional 9.8%, although graded by the Homerton DESP, were classified as U [(89 + 32)/1224] and 8.7% were graded R1M1, R2M0, R2M1, R3M0 or R3M1 by the Doheny Image Reading Centre. Episodes with manual grades R1M1, R2M0, R2M1, R3M0 or R3M1 from the Homerton DESP (413 + 32 + 18 + 10 + 3 = 476 episodes) were assigned a grade of at least R1M1 by the Reading Centre in 59% of episodes (279/476 episodes) and an additional 8.4% were classified as U.

TABLE 10 Screening performance of iGradingM software compared with manual grade

	iGradingM outcome, *n* (%[a])			
Manual grade (worst eye)	No disease	Disease	Ungradable	Total, *n* (%[b])
Retinopathy grades				
R0M0	0 (0)	4871 (38)	7856 (62)	12,727 (63)
R1M0	0 (0)	2913 (61)	1836 (39)	4749 (23)
U	0 (0)	118 (39)	182 (61)	300 (1)
R1M1	0 (0)	1136 (71)	473 (29)	1609 (8)
R2	0 (0)	504 (79)	133 (21)	637 (3)
R2M0	0 (0)	169 (80)	41 (20)	210 (1)
R2M1	0 (0)	335 (78)	92 (22)	427 (2)
R3	0 (0)	181 (77)	55 (23)	236 (1)
R3M0	0 (0)	61 (81)	14 (19)	75 (0)
R3M1	0 (0)	120 (75)	41 (25)	161 (1)
Combination of grades				
R0M0, R1M0	0 (0)	7784 (45)	9692 (55)	17,476 (86)
U, R1M1, R2, R3	0 (0)	1939 (70)	843 (30)	2782 (14)
R1M0, U, R1M1, R2, R3	0 (0)	4852 (64)	2679 (36)	7531 (37)
Total	0 (100)	9723 (100)	10,535 (100)	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.

TABLE 11 Results from the arbitration on 1700 screening episodes comparing the manual grade from the Homerton DESP with that from the Doheny Image Reading Centre

Manual grade from Homerton DESP	Arbitration grade from Doheny Image Reading Centre, *n*								
	R0M0	R1M0	U	R1M1	R2M0	R2M1	R3M0	R3M1	Total, *n*
R0M0	734	22	89	15	1	0	1	0	862
R1M0	149	91	32	85	3	0	0	2	362
R1M1	82	60	29	222	0	16	1	3	413
R2M0	7	7	3	9	1	4	0	1	32
R2M1	0	0	4	9	2	2	1	0	18
R3M0	0	1	2	2	0	2	3	0	10
R3M1	0	0	2	0	0	0	0	1	3
Total	972	181	161	342	7	24	6	7	1700

Exploratory analyses of demographic factors on screening performance

Tables 12–17 summarise the screening performance for EyeArt and Retmarker for all manual retinopathy grades as modified by arbitration for three ethnic groups (white European, Asian and black African-Caribbean), three age groups and by sex. There was no evidence to suggest that ethnicity (see *Tables 12* and *13*) or sex (see *Tables 16* and *17*) influences the sensitivity of EyeArt or Retmarker. However, false-positive rates appeared lower in white Europeans for combined grades R0M0 + R1M0 than in Asians or black African-Caribbeans for

TABLE 12 Screening performance for EyeArt (manual grades modified by arbitration) by ethnic group

	EyeArt outcome by ethnic group, *n* (%[a])					
	White European		Asian		Black African-Caribbean	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades						
R0M0	1241 (22.7)	4232 (77.3)	707 (16.3)	3632 (83.7)	480 (19.6)	1963 (80.4)
R1M0	96 (5.0)	1839 (95.0)	60 (3.7)	1551 (96.3)	43 (5.1)	806 (94.9)
U	40 (30.5)	91 (69.5)	23 (18.1)	104 (81.9)	27 (19.4)	112 (80.6)
R1M1	33 (6.7)	460 (93.3)	28 (4.6)	580 (95.4)	9 (2.4)	359 (97.6)
R2	0 (0.0)	246 (100.0)	1 (0.4)	241 (99.6)	3 (2.8)	105 (97.2)
R2M0	0 (0.0)	85 (100.0)	1 (1.4)	68 (98.6)	2 (6.9)	27 (93.1)
R2M1	0 (0.0)	161 (100.0)	0 (0.0)	173 (100.0)	1 (1.3)	78 (98.7)
R3	0 (0.0)	80 (100.0)	0 (0.0)	91 (100.0)	1 (2.0)	49 (98.0)
R3M0	0 (0.0)	29 (100.0)	0 (0.0)	25 (100.0)	0 (0.0)	13 (100.0)
R3M1	0 (0.0)	51 (100.0)	0 (0.0)	66 (100.0)	1 (2.7)	36 (97.3)
Combination of grades						
R0M0, R1M0	1337 (18.0)	6071 (82.0)	767 (12.9)	5183 (87.1)	523 (15.9)	2769 (84.1)
U, R1M1, R2, R3	73 (7.7)	877 (92.3)	52 (4.9)	1016 (95.1)	40 (6.0)	625 (94.0)
R1M0, U, R1M1, R2, R3	169 (5.9)	2716 (94.1)	112 (4.2)	2567 (95.8)	83 (5.5)	1431 (94.5)
Total	1410 (16.9)	6948 (83.1)	819 (11.7)	6199 (88.3)	563 (14.2)	3394 (85.8)

a Percentage per manual grade within each ethnic group.

TABLE 13 Screening performance for Retmarker (manual grades modified by arbitration) by ethnic group

	Retmarker outcome by ethnic group, *n* (%[a])					
	White European		Asian		Black African-Caribbean	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades						
R0M0	3299 (60.3)	2174 (39.7)	2098 (48.4)	2241 (51.6)	1087 (44.5)	1356 (55.5)
R1M0	701 (36.2)	1234 (63.8)	540 (33.5)	1071 (66.5)	267 (31.4)	582 (68.6)
U	69 (52.7)	62 (47.3)	50 (39.4)	77 (60.6)	65 (46.8)	74 (53.2)
R1M1	77 (15.6)	416 (84.4)	83 (13.7)	525 (86.3)	36 (9.8)	332 (90.2)
R2	7 (2.8)	239 (97.2)	5 (2.1)	237 (97.9)	10 (9.3)	98 (90.7)
R2M0	0 (0.0)	85 (100.0)	1 (1.4)	68 (98.6)	4 (13.8)	25 (86.2)
R2M1	7 (4.3)	154 (95.7)	4 (2.3)	169 (97.7)	6 (7.6)	73 (92.4)
R3	2 (2.5)	78 (97.5)	1 (1.1)	90 (98.9)	2 (4.0)	48 (96.0)
R3M0	0 (0.0)	29 (100.0)	1 (4.0)	24 (96.0)	0 (0.0)	13 (100.0)
R3M1	2 (3.9)	49 (96.1)	0 (0.0)	66 (100.0)	2 (5.4)	35 (94.6)
Combination of grades						
R0M0, R1M0	4000 (54.0)	3408 (46.0)	2638 (44.3)	3312 (55.7)	1354 (41.1)	1938 (58.9)
U, R1M1, R2, R3	155 (16.3)	795 (83.7)	139 (13.0)	929 (87.0)	113 (17.0)	552 (83.0)
R1M0, U, R1M1, R2, R3	856 (29.7)	2029 (70.3)	679 (25.3)	2000 (74.7)	380 (25.1)	1134 (74.9)
Total	4153 (49.7)	4154 (49.7)	2777 (39.6)	4241 (60.4)	1467 (37.1)	2490 (62.9)

a Percentage per manual grade within each ethnic group.

TABLE 14 Screening performance for EyeArt (manual grades modified by arbitration) by age group

	EyeArt outcome by age group, *n* (%[a])					
	< 50 years		50 to < 65 years		65 to 98 years	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades						
R0M0	637 (20.7)	2444 (79.3)	962 (19.4)	3997 (80.6)	938 (19.8)	3798 (80.2)
R1M0	73 (6.3)	1078 (93.7)	68 (4.1)	1608 (95.9)	73 (4.1)	1706 (95.9)
U	9 (24.3)	28 (75.7)	26 (23.0)	87 (77.0)	62 (23.0)	208 (77.0)
R1M1	16 (3.9)	391 (96.1)	39 (5.5)	669 (94.5)	18 (4.1)	422 (95.9)
R2	0 (0.0)	162 (100.0)	3 (1.1)	263 (98.9)	1 (0.5)	197 (99.5)
R2M0	0 (0.0)	51 (100.0)	2 (2.6)	75 (97.4)	1 (1.5)	64 (98.5)
R2M1	0 (0.0)	111 (100.0)	1 (0.5)	188 (99.5)	0 (0.0)	133 (100.0)
R3	0 (0.0)	51 (100.0)	0 (0.0)	115 (100.0)	1 (1.6)	62 (98.4)
R3M0	0 (0.0)	16 (100.0)	0 (0.0)	35 (100.0)	0 (0.0)	18 (100.0)
R3M1	0 (0.0)	35 (100.0)	0 (0.0)	80 (100.0)	1 (2.2)	44 (97.8)
Combination of grades						
R0M0, R1M0	710 (16.8)	3522 (83.2)	1030 (15.5)	5605 (84.5)	1011 (15.5)	5504 (84.5)
U, R1M1, R2, R3	25 (3.8)	632 (96.2)	68 (5.7)	1134 (94.3)	82 (8.4)	889 (91.6)
R1M0, U, R1M1, R2, R3	98 (5.4)	1710 (94.6)	136 (4.7)	2742 (95.3)	155 (5.6)	2595 (94.4)
Total	735 (15.0)	4154 (85.0)	1098 (14.0)	6739 (86.0)	1093 (14.6)	6393 (85.4)

a Percentage per manual grade within each age group.

TABLE 15 Screening performance for Retmarker (manual grades modified by arbitration) by age group

	Retmarker outcome by age group, *n* (%[a])					
	< 50 years		50 to < 65 years		65 to 98 years	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades						
R0M0	1231 (40.0)	1850 (60.0)	2924 (59.0)	2035 (41.0)	2567 (54.2)	2169 (45.8)
R1M0	277 (24.1)	874 (75.9)	678 (40.5)	998 (59.5)	624 (35.1)	1155 (64.9)
U	9 (24.3)	28 (75.7)	51 (45.1)	62 (54.9)	133 (49.3)	137 (50.7)
R1M1	33 (8.1)	374 (91.9)	112 (15.8)	596 (84.2)	62 (14.1)	378 (85.9)
R2	1 (0.6)	161 (99.4)	10 (3.8)	256 (96.2)	11 (5.6)	187 (94.4)
R2M0	0 (0.0)	51 (100.0)	3 (3.9)	74 (96.1)	2 (3.1)	63 (96.9)
R2M1	1 (0.9)	110 (99.1)	7 (3.7)	182 (96.3)	9 (6.8)	124 (93.2)
R3	0 (0.0)	51 (100.0)	2 (1.7)	113 (98.3)	3 (4.8)	60 (95.2)
R3M0	0 (0.0)	16 (100.0)	1 (2.9)	34 (97.1)	0 (0.0)	18 (100.0)
R3M1	0 (0.0)	35 (100.0)	1 (1.3)	79 (98.8)	3 (6.7)	42 (93.3)
Combination of grades						
R0M0, R1M0	1508 (35.6)	2724 (64.4)	3602 (54.3)	3033 (45.7)	3191 (49.0)	3324 (51.0)
U, R1M1, R2, R3	43 (6.5)	614 (93.5)	175 (14.6)	1027 (85.4)	209 (21.5)	762 (78.5)
R1M0, U, R1M1, R2, R3	320 (17.7)	1488 (82.3)	853 (29.6)	2025 (70.4)	833 (30.3)	1917 (69.7)
Total	1551 (31.7)	3308 (68.3)	3777 (48.2)	4060 (51.8)	3400 (45.4)	4086 (54.6)

a Percentage per manual grade within each age group.

TABLE 16 Screening performance for EyeArt (manual grades modified by arbitration) in males and females

	EyeArt outcome by sex, *n* (%[a])			
	Male		Female	
Manual grade (worst eye)	No disease	Disease	No disease	Disease
Retinopathy grades				
R0M0	1305 (19.5)	5390 (80.5)	1232 (20.3)	4849 (79.7)
R1M0	121 (4.6)	2521 (95.4)	93 (4.7)	1871 (95.3)
U	46 (22.2)	161 (77.8)	51 (23.9)	162 (76.1)
R1M1	37 (4.1)	864 (95.9)	36 (5.5)	618 (94.5)
R2	2 (0.5)	378 (99.5)	2 (0.8)	244 (99.2)
R2M0	1 (0.8)	117 (99.2)	2 (2.7)	73 (97.3)
R2M1	1 (0.4)	261 (99.6)	0 (0.0)	171 (100.0)
R3	0 (0.0)	152 (100.0)	1 (1.3)	76 (98.7)
R3M0	0 (0.0)	42 (100.0)	0 (0.0)	27 (100.0)
R3M1	0 (0.0)	110 (100.0)	1 (2.0)	49 (98.0)
Combination of grades				
R0M0, R1M0	1426 (15.3)	7911 (84.7)	1325 (16.5)	6720 (83.5)
U, R1M1, R2, R3	85 (5.2)	1555 (94.8)	90 (7.6)	1100 (92.4)
R1M0, U, R1M1, R2, R3	206 (4.8)	4076 (95.2)	183 (5.8)	2971 (94.2)
Total	1511 (13.8)	9466 (86.2)	1415 (15.3)	7820 (84.7)

a Percentage per manual grade within each sex.

TABLE 17 Screening performance for Retmarker (manual grades modified by arbitration) in males and females

	Retmarker outcome by sex, *n* (%[a])			
	Male		Female	
Manual grade (worst eye)	No disease	Disease	No disease	Disease
Retinopathy grades				
R0M0	3597 (53.7)	3098 (53.7)	3125 (51.4)	2956 (51.4)
R1M0	912 (34.5)	1730 (65.5)	667 (34.0)	1297 (66.0)
U	92 (44.4)	115 (55.6)	101 (47.4)	112 (52.6)
R1M1	116 (12.9)	785 (87.1)	91 (13.9)	563 (86.1)
R2	9 (2.4)	371 (97.6)	13 (5.3)	233 (94.7)
R2M0	3 (2.5)	115 (97.5)	2 (2.7)	73 (97.3)
R2M1	6 (2.3)	256 (97.7)	11 (6.4)	160 (93.6)
R3	2 (1.3)	150 (98.7)	3 (3.9)	74 (96.1)
R3M0	1 (2.4)	41 (97.6)	0 (0.0)	27 (100.0)
R3M1	1 (0.9)	109 (99.1)	3 (6.0)	47 (94.0)
Combination of grades				
R0M0, R1M0	4509 (48.3)	4828 (51.7)	3792 (47.1)	4253 (52.9)
U, R1M1, R2, R3	219 (13.4)	1421 (86.6)	208 (17.5)	982 (82.5)
R1M0, U, R1M1, R2, R3	1131 (26.4)	3151 (73.6)	875 (27.7)	2279 (72.3)
Total	4728 (43.1)	6249 (56.9)	4000 (43.3)	5235 (56.7)

a Percentage per manual grade within each sex.

both ARIASs. There was some evidence that as age increased, detection of combined grades U, R1M1, R2 and R3 was reduced for both ARIASs (see *Tables 14* and *15*). Sensitivity falls from 96% in the youngest age group to 92% in the oldest age group for EyeArt and from 93.5% to 78.5% for Retmarker. For Retmarker (but not EyeArt) there was a decline in the false-positive rate with age, from 64% in the youngest age group to 51% in the oldest age group.

Five different cameras were in use at the Homerton DESP: 415 episodes were photographed with the Canon 2CR (Canon, Tokyo, Japan), 7569 episodes with the Canon 2CR-Dgi (Canon), 1230 episodes with the Canon EOS (Canon), 4246 episodes with the Canon EOS2 (Canon) and 3432 episodes with the TOPCONnect (Topcon, Tokyo, Japan). In 3395 episodes the camera type was not recorded. The main reason for the camera type not being recorded was a network interruption or 'processor timeout'. In such situations images are saved to a separate location for uploading later and the camera type code was not retained. Screening performance by camera type is summarised in *Tables 38* and *39* in *Appendix 3*. It appears that both ARIASs performed least well with the Canon EOS2 out of all the cameras.

Multiple variable logistic regression models (cases being defined as those with manual grade refined by arbitration in the worst eye of R1M0, U, M1, R2 or R3 and non-cases patients with a worst eye grade of R0M0) were used to formally test for an interaction between the ARIAS outcome classification of 'disease' versus 'no disease', with age group, sex, ethnicity and camera type as categorical variables.

A multiple variable logistic regression model with adjustment for age, sex, ethnicity and camera type showed that EyeArt cases were 4.42 (95% CI 3.95 to 4.96) times more likely than non-cases to be classified as 'disease' rather than 'no disease'. Inclusion of interaction terms in the model was not statistically significant, suggesting that the performance of EyeArt was not affected by age ($p = 0.51$), sex ($p = 0.20$), ethnicity ($p = 0.58$) or camera type ($p = 0.03$). The corresponding models for Retmarker show that cases were 3.19 (95% CI 2.99 to 3.41) times more likely than non-cases to be classified as 'disease' rather than 'no disease'. However, for Retmarker there were statistically significant interactions at the 1% level for all covariates except sex, suggesting that Retmarker performance was influenced by age, ethnicity and camera type. Odds ratios from the multiple variable logistic regression model including all two-way interactions for Retmarker outcome with age, sex, ethnicity and camera type are given in *Table 18*. The reference category is white European men aged < 50 years photographed with the Canon 2CR DGi camera; cases were 3.73 (95% CI 3.06 to 4.54) times more likely than non-cases to be classified as 'disease' rather than 'no disease, once we take account of all main effects and two-way interactions. This odds ratio increased to 3.98 for the 50 to < 65 years age group but reduced to 2.97 for the oldest age group (65–98 years), suggesting that the discrimination of Retmarker was marginally poorer in the oldest age group. If the man was Asian instead of white European, the odds ratio decreased to 2.63 (95% CI 2.17 to 3.19). A similar odds ratio was observed in black African-Caribbean individuals, implying slightly poorer discrimination of Retmarker in Asian and black African-Caribbean individuals than in white European individuals. The results suggest that the software performs differentially by camera type. For example cases were much more likely than non-cases, to be classified as 'disease' rather than 'no disease' with the Canon EOS than with the Canon 2CR DGi (odds ratios of 3.73 vs. 7.93). The interaction with patient sex, although not significant at the 1% level, has been included for completeness.

TABLE 18 Odds ratios for Retmarker outcome by age group, sex, ethnicity and camera type

Covariate	Odds ratio for Retmarker outcome (disease vs. no disease) (95% CI)	*p*-value for interaction[a]
Age		
< 50 years	3.73 (3.06 to 4.54)	0.006
50 to < 65 years	3.98 (3.38 to 4.67)	
65 to 98 years	2.97 (2.55 to 3.47)	
Sex		
Males	3.73 (3.06 to 4.54)	0.02
Females	3.24 (2.64 to 3.98)	
Ethnic group		
White European	3.73 (3.06 to 4.54)	< 0.0001
Asian	2.63 (2.17 to 3.19)	
Black African-Caribbean	2.52 (2.00 to 3.18)	
Camera		
Canon 2CR-DGi	3.73 (3.06 to 4.54)	< 0.0001
Canon 2CR	3.45 (1.88 to 6.33)	
Canon EOS	7.93 (5.90 to 10.65)	
Canon EOS2	4.38 (3.53 to 5.43)	
TOPCONnect	3.11 (2.45 to 3.95)	

a *p*-value for interaction from the multiple variable logistic regression model (cases were those with manual grade refined by arbitration in the worst eye of R1M0, M1, R2 or R3 and non-cases were patients with worst eye grade of R0M0) including two-way interactions for Retmarker outcome classification with age, sex, ethnicity and camera type.

Altered thresholds

As part of the economic evaluation we explored the consequences for sensitivity and specificity of using different decision thresholds in image classification. In what follows, all episodes given a final grade of R0 or R1 by the human grader, after arbitration, were classified as disease free and all other categories, including unclassified, were considered as disease cases.

Using this as our working definition of 'disease' versus 'no disease' in our sample of 20,258 patients, Retmarker generated values for sensitivity of 0.85 and for specificity of 0.48. It was not possible to achieve this exact level of screening performance by setting a threshold on the decision statistic provided by Retmarker; additional information must be being used in the outcome classification. However, varying the closest obtained threshold to change the sensitivity by an absolute difference of ± 5% gave the following pairs of values for three operating points (*Table 19*).

EyeArt provided two classifications: (1) 'disease versus no disease' based on EyeArt software set to detect 'any retinopathy', that is, R1 or worse, and (2) 'refer versus no refer category' when it is set to detect 'referable retinopathy', that is, M1 or worse. These were implemented as two separate decision statistics rather than two different thresholds on one decision statistic. Varying the threshold used for the 'disease versus no disease' classification to obtain an absolute change in sensitivity of ± 5% gave the pairs of values for three operating points shown in *Table 20*.

Varying the threshold used for the 'refer versus no refer' classification to obtain an absolute change in sensitivity of ± 5% gives the pairs of values shown in *Table 21*.

TABLE 19 The impact on specificity of a change in threshold leading to a 5% change in sensitivity: Retmarker

Threshold	Specificity	Sensitivity
2.12	0.52	0.79
2.50	0.45	0.84
2.90	0.35	0.89

TABLE 20 The impact on specificity of a change in threshold leading to a 5% change in sensitivity: EyeArt's 'disease vs. no disease' classification

Threshold	Specificity	Sensitivity
–1.14	0.03	0.99
–1.00	0.16	0.94
–0.81	0.35	0.89

TABLE 21 The impact on specificity of a change in threshold leading to a 5% change in sensitivity: EyeArt's 'refer vs. no refer' classification

Threshold	Specificity	Sensitivity
–1.14	0.14	0.96
–1.0	0.35	0.91
–0.84	0.54	0.86

The trade-off between sensitivity and specificity was further explored through an analysis of the breakdown of the false-negative results. *Tables 22* and *23* shows how the proportion of false negatives and true negatives increased and sensitivity reduced in order to improve specificity. Sensitivity is largely influenced by episodes graded M1 and U. Both systems, but especially EyeArt, remained relatively successful at detecting the most clinically significant grades of retinopathy, even if sensitivity was relaxed in order to increase specificity, that is, the proportion of disease-free cases correctly identified.

Given the large proportion of ungradable episodes among the false negatives, we looked more carefully at how these were generated. EyeArt, for example, analysed each image and classified it as adequate or inadequate. If more than one image in the episode was classified as adequate, the episode was classified as adequate for image quality. Human graders probably use a very different approach: marking an episode as a technical failure if any area of the retina was not imaged satisfactorily. Of the cases that had a final human grade of U, EyeArt classified 145 as disease free, which reduced its measured sensitivity. In 135 out of those 145 episodes, EyeArt identified one or more of the images as a technical failure but classified the episodes as technically adequate on the strength of the remaining images. A different rule would have provided a classification that performed much better on the metrics used in this evaluation. For example, using the default threshold on the screening decision statistic and taking the final human grade after

TABLE 22 The breakdown by diagnostic categories of the false negatives at different thresholds, shown with the corresponding numbers of true negatives and true and false positives: Retmarker

Threshold	True positives	False negatives	R2	R3	M1	U	True negatives	False positives
2.884	2532	312	17	6	157	132	6230	11,184
2.804	2508	336	19	7	167	143	6569	10,845
2.724	2479	365	24	7	178	156	6881	10,533
2.644	2452	392	30	8	190	164	7231	10,183
2.564	2424	420	33	8	196	183	7562	9852
2.420	2372	472	35	11	221	205	8105	9309
2.340	2347	497	36	12	236	213	8366	9048
2.260	2323	521	41	13	248	219	8636	8778
2.180	2288	556	52	18	260	226	8866	8548
2.100	2230	614	66	22	291	235	9078	8336

TABLE 23 The breakdown by diagnostic categories of the false negatives at different thresholds, shown with the corresponding numbers of true negatives and true and false positives: EyeArt's 'refer vs. no refer' classification

Threshold	True positives	False negatives	R2	R3	M1	U	True negatives	False positives
–1.20	2775	69	1	0	29	39	1056	16,358
–1.15	2720	124	2	0	50	72	2173	15,241
–1.10	2667	177	3	0	80	94	3475	13,939
–1.05	2603	241	4	1	109	127	4769	12,645
–1.00	2553	291	7	2	137	145	6049	11,365
–0.95	2508	336	7	2	163	164	7224	10,190
–0.90	2473	371	8	3	185	175	8275	9139
–0.85	2418	426	9	3	217	197	9239	8175
–0.80	2367	477	11	4	250	212	10,092	7322

arbitration as truth, there are 291 false negatives, the latter group break down into final manual grades after arbitration of 7 R2, 2 R3, 137 M1 and 145 U. Taking the output from the EyeArt image quality assessment and applying an alternative rule that '50% or more of images are technically inadequate, then the episode is technically inadequate' would reclassify 91 of the 291 false negatives as true positives. The result would be a marked improvement in sensitivity, from 90% to 93%. The same rule was applied to all episodes; 450 episodes that were correctly classified as 'disease free' were incorrectly reclassified as technical failures and specificity reduced from 35% to 32%. A tougher rule – 'if more than 50% of images are technically inadequate, then the case is technically inadequate' – reclassified only 47 episodes; hence, sensitivity increased by 1.7% only but at an absolute cost of only a 0.8% reduction in specificity.

Implementation

Image processing by the automated retinal image analysis systems

There were significant problems with some of the ARIASs running the test data. A large number of images were inputted to each of the programs, and two systems (Retmarker and iGradingM) had significant scaling and infrastructure issues. Of note, the number of images given to process in the test set would far exceed the likely daily workload in a large DESP, although this would depend on the implementation model if an ARIAS was used. For example, if batch processing occurred at weekends or overnight from multiple DESPs, then this volume could be encountered. Several attempts to run the software were made and the vendors had to be contacted to help rectify the problems. The limitations of both software packages came from the growth of temporary files and database management.

The software that had the fewest problems executing the test data was EyeArt's cloud-based solution. This infrastructure allows for a scalable, elastic computing that handled the number of test data without any issues.

Implementation of automated retinal image analysis systems into the NHS diabetic eye screening programme

When the software was implanted into the Homerton DESP, there were considerable problems integrating with the OptoMize Data Export Module, which was unable to act in real time. As the study test set relied on the commissioned exporter tool to generate JPEG images and a separate data file, each of the four ARIAS vendors developed importation tools specifically for the data files generated by OptoMize. However, these tools did not allow seamless integration with automated transfer of image files. The only way of handling the images for this study with the OptoMize Data Export Module was to manually transfer them but this would not be optimal if ARIASs were implemented into routine clinical care.

Health economics

The results of the health economic evaluation are given in *Chapter 4*.

Chapter 4 Health economics: methodology and results

Introduction

This chapter covers the economic evaluation of two approaches to incorporate automated screening into the English national screening programme for diabetic retinopathy. Two questions are addressed: (1) is it cost-effective to replace level 1 human graders (manual grading) with automated screening?; and (2) is it cost-effective to use automated screening as a filter prior to level 1 manual grading? Under strategy 1, no images would be seen by level 1 graders and level 2 graders only screen images that have been flagged by the automated system as diseased (see *Figure 1*). In contrast, only a subset of images flagged by automated screening systems would be seen by level 1 graders under strategy 2 (see *Figure 2*).

We considered costs only from the perspective of the NHS. The economic evaluation of these two scenarios did not include costs that would be the same for each screening method. These included costs associated with QA programmes that would be the same under all scenarios and common costs associated with running the screening clinic (e.g. reception, assessment of clinical history, eye drops, fundus photography, rent, overhead and common capital expenditures). We took a time horizon of 1 year, in terms of the patient outcomes evaluated, but costs were amortised over the expected lifetime of the equipment (5–7 years).

Our analysis focused on the outcome of cost per appropriate screening outcome (true positive or true negative correctly identified by the ARIASs). We chose this outcome measure because the objective of this screening programme is to pick up true-positive cases as well true-negative cases, in other words, to maximise the sensitivity and specificity of the screening programme. Both outcomes are important: positive cases should be identified but this should not come at the cost of overly sensitive screening systems that also inconvenience patients with incorrect diagnoses and add further testing costs. The sensitivities and specificities of the alternative screening strategies explored are established in previous chapters, using manual grade as modified by arbitration by the Doheny Image Reader Centre (Doheny Eye Institute, Arcadia, CA, USA) as the reference standard.

We assessed two different automated screening software packages, Retmarker and EyeArt. Although three systems were evaluated as part of this project, the poor efficacy of iGradingM (specified in *Chapter 3*) rendered economic evaluation unnecessary. We undertook identical analyses for Retmarker and EyeArt, altering the input parameters specific to each software. As these systems are not yet used in the English NHS, costing information is tentative and requires extensive sensitivity analysis.

Model

The health economic analysis uses a decision-tree model to calculate the incremental cost-effectiveness of manual grading as it is currently applied within one major provider of diabetic retinopathy screening in England (Homerton University Hospital Foundation Trust, London, UK) versus replacing level 1 human graders with an automated system (strategy 1) or acting as screening mechanism prior to level 1 grading (strategy 2). The decision tree reflected patient screening pathways, the screening levels through which images were processed, and grading outcomes (referral to ophthalmology/hospital eye services or rescreening as part of the annual screening programme).

Figure 1 depicts the processing mechanisms affecting all image sets under strategy 1. All images are first seen by a level 1 manual grader or are run through automated screening software. Patients undergoing manual

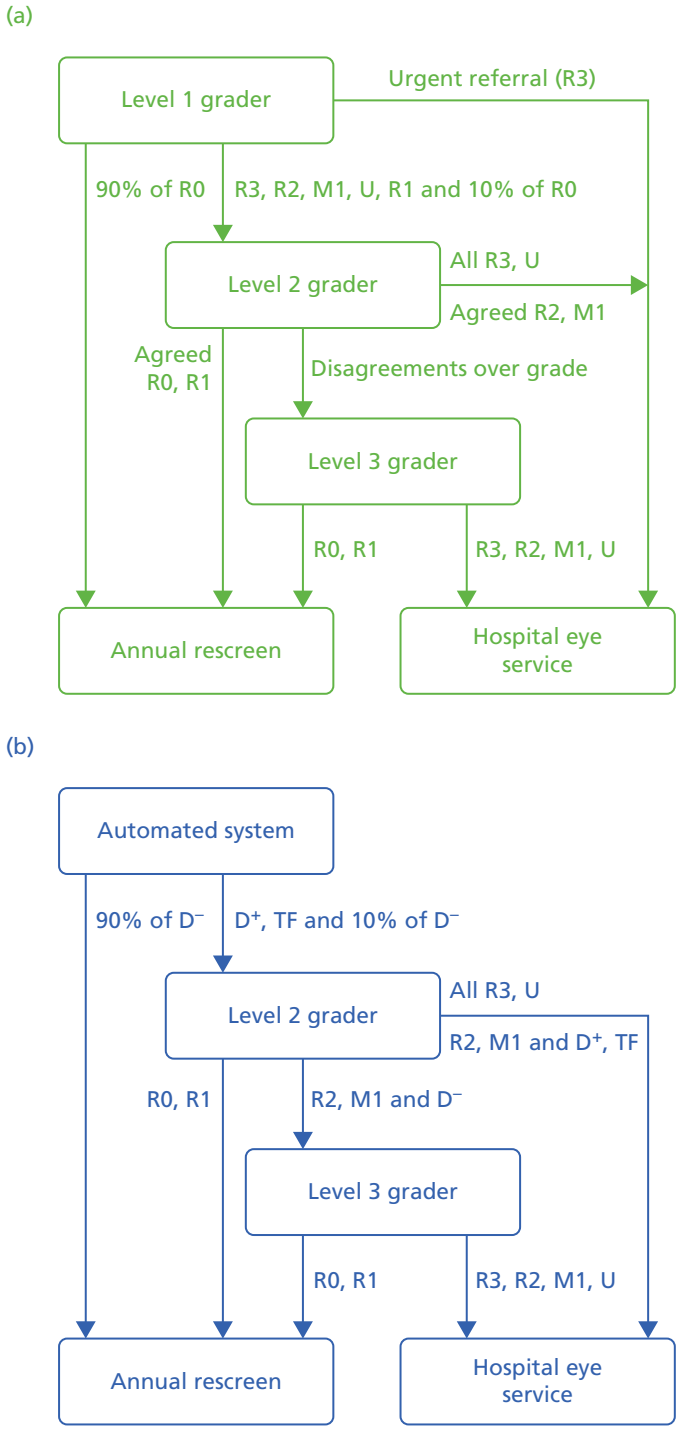


FIGURE 1 Screening pathways (strategy 1). (a) Clinical practice; and (b) replacing level 1 grader by an automated system.

screening immediately fell within one of seven disease categories based on underlying retinopathy status, as per the NHS DESP at the time:[57] R0, R1, R2, R3, M1a, M1b or U (see *Chapter 3*, *Data extraction from the Homerton diabetic eye screening programme* for further details on M1a and M1b).

The screening programme prescribes that 90% of patients receiving a grade of R0 go straight to annual rescreening, whereas other grades and 10% of the R0 grades (for QA purposes) would proceed to level 2 graders. Depending on the agreement or disagreement between level 1 and level 2 graders, a patient would either have their images seen by a level 3 grader (disagreement) or be invited back for annual rescreening (agreed R0 and R1 cases). Anyone receiving a R3, U or agreed R2 or M1a/M1b between the

level 1 and level 2 graders would be referred immediately to the hospital eye service to see an ophthalmologist. If a set of images had been sent to the level 3 grader, this becomes the final outcome grade and any R0 or R1 grades would be invited for annual screening and any R3, R2, M1a or M1b patients would be referred to the hospital eye service.

This decision tree reflected clinical pathways used at our local study site at the time the images were taken and followed through on the basis of existing, manual screening pathways. The manual grading arm consists of patients exactly as they were screened and their images were read at the local site.

The local site had some slight differences at the time of the study to the current national screening pathway.[58] The main difference between our local pathway and the national pathway is that the local pathway splits patients with M1 into M1a and M1b subgroups. The local pathway then rescreens M1a patients within a 6-month interval and refers M1b patients directly to ophthalmology. This subclassification did not affect health economic outcomes. In the absence of any information on diagnoses received during follow-up visits, subsequent treatment costs and clinical outcomes remained the same across patients evaluated at either 6- or 12-month intervals. Therefore, for the present analysis, 6- and 12-month rescreen costs were identical for manual grading and automated screening arms and we called them '12-month rescreen' as they are part of the annual screening programme. If this model had a longer time frame, and incorporating results from the ophthalmological visit, this difference in local pathway would have to be revisited.

Moreover, Doheny arbitration grades did not divide M1 grades into M1a and M1b subgroups. Therefore, to consistently compare screening diagnoses with arbitration grades, both M1a and M1b were assumed to be referable diagnoses. This approach enabled us to refer to arbitration grades to calculate therapeutic effectiveness. It is also consistent with national screening pathways in the UK that call for ophthalmological referral for M1 cases.[58] M1a and M1b patient diagnoses were therefore costed equally. In practical terms, this means that although modelling reflects local practice, our results are nevertheless adaptable to national screening programmes in the UK.

Patients undergoing automated screening fell within three disease categories that are similarly based on underlying retinopathy status: disease present, no disease present or other lesion/ungradable (U). Patient images deemed ungradable result from either technical failures or the capture of ungradable images. Images that proceeded to manual grading after ARIAS would then be given grades as in the manual pathway (R0, R1, R2, R3, M1a, M1b, U) by a level 2 grader.

In strategy 1, when automated screening replaced manual grading for level 1, all episodes with a positive or ungradable disease outcome (as assessed by the software), along with a small proportion (10%) of episodes with a negative disease outcome, would be sent for further human grading (level 2 graders). Given our study's design as a retrospective evaluation, we mapped the results from manual grading data onto the automated screening arms to demonstrate what would have happened to patients if they had undergone ARIAS classification first, followed by manual grading.

Figure 2 shows the screening pathways for strategy 2 in which an automated screening system would act as a screen for all images such that fewer would be seen by a human grader. The manual grading arm stays exactly the same as in strategy 1. Ninety per cent of patients whose images, according to the automated screening system, showed no disease would be invited for annual screening. Those images found to have disease, technical failures and 10% of those images found to have no disease (QA) would be referred to a level 1 grader. The pathway would then work exactly the same as for the manual grading arm.

The health economic model consists of the following data: (1) the probabilities associated with each step of the retinopathy grading pathway, (2) the overall outcome of each screening strategy as being appropriate (true positive and true negative) and (3) bottom-up costing of manual screening strategies and costing from manufacturer data for the ARIAS. It therefore takes into account the screening performance of automated systems, the efficacy of manual screening, the likelihood of rescreening and referral rates to ophthalmologists.

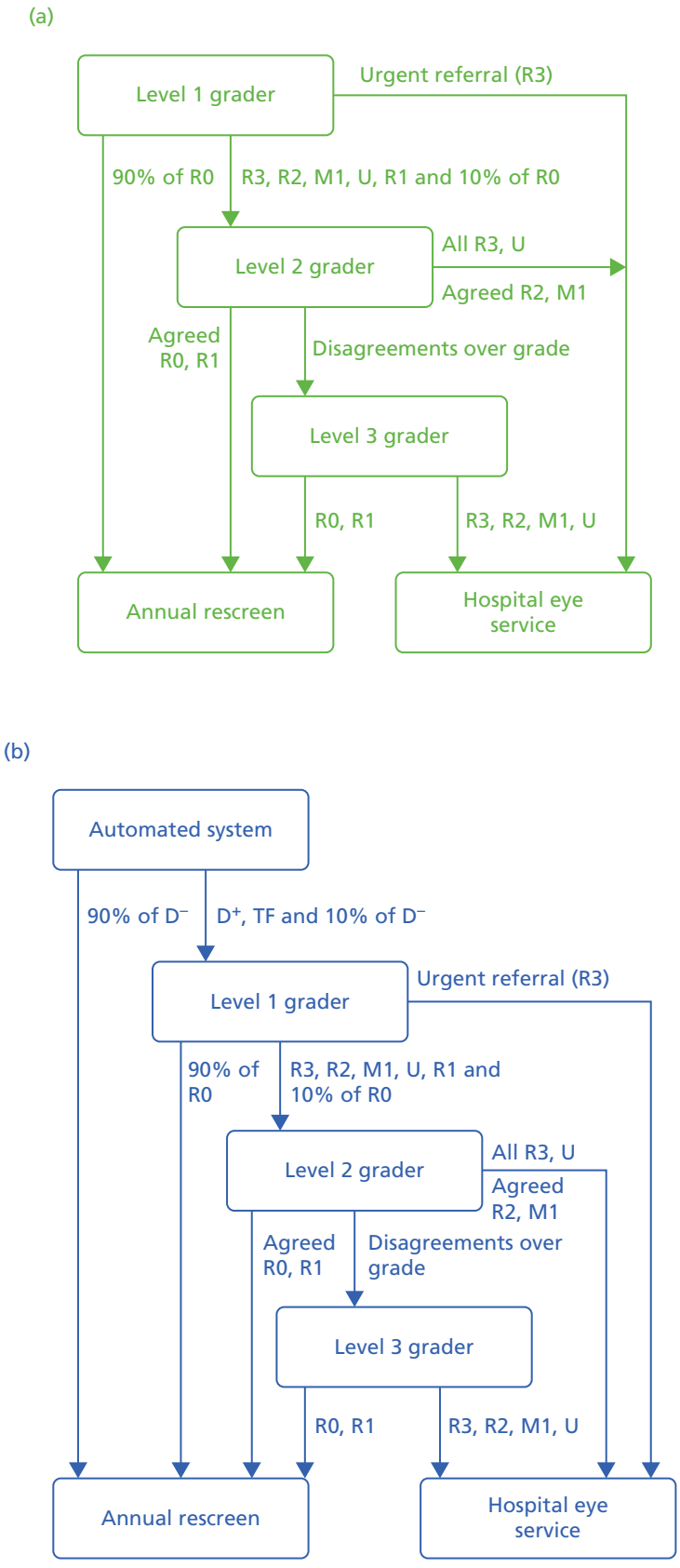


FIGURE 2 Screening pathways (strategy 2). (a) Clinical practice and (b) using an automated system as a filter prior to level 1 grading. D^-, disease absent classification by automated system; D^+, disease present classification by automated system; TF, technical failure.

Data inputs

Efficacy variables

The probability parameters in this model come from patients' experiences on the pathways outlined above. Discounting of clinical effects was not necessary for cost-effectiveness assessments, as our model focused on acute screening outcomes and did not incorporate any time-related element.

Probability parameters for the manual grading arm came from data from the hospital screening programme. They reflect patients' actual journey through screening pathways as they were seen and treated by the Homerton University Hospital Foundation Trust screening programme.

In the automated arm, probability parameters came from a combination of automated screening system results that were mapped on to the tentative protocols for implementing automated screening software in the screening programme and what we know actually happened to patients from their manual grading process (any subsequent manual grades and final outcomes). If automated screening was implemented, 90% of patients found by the system to be disease negative would be referred for annual screening (see *Figure 1*). We therefore took a random sample of the patients noted as disease negative and assumed that they attended annual screening, which means that they would have a variety of final outcomes, as in our patient sample.

Any patient images denoted disease positive or considered technical failures and 10% of those classed as disease negative proceeded to level 2 grading. Again, the 10% was a random sample used for QA purposes. The results of the level 2 grading came from the manual grading records of actual patient practice. When a level 2 grading result was not available (i.e. the patient's image set was seen only by a level 1 grader), we assumed that the level 2 grade would be the same as the level 1 grade; similarly, if necessary, we assumed that the level 3 grade would be the same as the level 2 grade. We then calculated, based on the results from our real patient data, the probabilities that patients would be graded R0 or R1 grade by the level 2 grader and proceed to annual screening, be referred to the hospital eye service or be sent to a level 3 grader. Any level 3 grading that took place was treated in the same way.

Similarly, if a level 3 grader result did not correspond with clinical pathways for automated screening, we modelled the available data to fit within the relevant pathway. For example, if a level 3 grade was available even after level 1 and level 2 manual graders had agreed on a diagnosis under strategy 2, those data were modelled within the pathway that had a 'missing' score for the level 3 grader. The data pertaining to these patient cases were used to calculate system sensitivity and specificity, but the pathway was costed as if it excluded a third-level grader. Finally, if a level 3 grader result was missing owing to early recall or referral that was not reflected in care guidelines, we modelled clinical pathways using the Doheny or arbitration grade as the level 3 grade. In the absence of data from local practice, this approach was necessary to adequately cost and represent tentative screening protocols. In reality, level 3 grades and Doheny arbitration scores frequently agreed when they were both available, giving reason for us to take this this assumption forward.

Costs

Costs were estimated using a number of sources: a cost survey with the local study site, NHS National Tariffs and telephone/e-mail conversations with automated screening system manufacturers. All costs are in Great Britain Pound (GBP) for 2013/14 and, when appropriate, inflated using the 2014 PSSRU hospital and community health services (HCHS) pay and prices index.[59] Recurrent costs (capital costs, periodic charges on technologies) were discounted to reflect opportunity costs over the lifespan of the 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 annualised and incorporated into the model in terms of per patient costs.

Manual screening costs

Fixed and variable screening costs associated with manual grading were obtained directly through a survey of the Homerton University Hospital Foundation Trust's Diabetic Eye Screening Centre, which was also responsible for manual grading of patient cases. The costing information therefore reflects actual resource use for the manual screening pathway presented. This survey obtained equipment, consumables, labour, capital, QA and overhead costs associated with all levels of manual screening.

Labour compensation, working time and productivity (grading rate per hour) were used to derive unit costs per patient case evaluated. Costs associated with QA programmes applying exclusively to either manual or automated screening strategies were included in the appropriate arm of the model as per patient costs. QA costs for each level of grading were calculated as an average per total screened population (35,535 patients).

Manual screening pathways for diabetic retinopathy patients include three sequential levels of screening. Patients enter into the clinical screening pathway through level 1 graders, who are band 4, 5 or 6 employees. If referred on to level 2 graders, patient cases are evaluated by level 2 graders, who are one of two possible bands: bands 5 and 6. Finally, patients who are referred on to arbitration are evaluated by two level 3 graders of two possible bands: bands 6 and 7. The survey was used to obtain salary scales, annual total costs, weeks worked per year, hours worked per week, patient grading rate per hour for each band and level of grader. To reflect labour costs directly associated exclusively with the screening programme, the number of whole-time equivalents (WTEs) of each band employed in the patient screening programme at Homerton University Hospital Foundation Trust (*Table 24*) were used to determine the proportion of salary included. As each band of graders are responsible for their own patient cases, a weighted average per patient cost across employee bands was constructed by applying a patient volume weight to average costs per band reflecting the proportion of total patients evaluated by each band of grader for each grading level (*Table 25*).

Manual graders often work at multiple sites. The Homerton screening programme therefore may employ more or less than one WTE per year for each band of grader. To best reflect screening costs, labour-related expenditures were incorporated as direct costs to the screening programme (weighted by screening WTE), rather than as total yearly salaries from employment in the NHS. Every one WTE was assumed to equal one

TABLE 24 Costing labour inputs across all levels of manual grading

Grading level (band)	Salary per year (£)	Total WTE per year	Total cost per year (£)	Weeks worked per year	Hours worked per week	Patient grading rate (per hour)	Total annual cost per patient screened (£)
Level 1							
Band 6	40,359.00	2	80,718.00	44	56.25	10	3.26
Band 5	32,582.00	0.2	6516.40	44	8	10	1.85
Band 4	25,311.00	0.1	2531.10	44	4	10	1.44
Level 2							
Band 6	40,359.00	0.9	36,323.10	44	30	10	2.75
Band 5	32,582.00	0.1	3258.20	44	4	10	1.85
Level 3							
Band 7	55,391.00	0.2	11,078.20	44	8	10	3.15
Band 6	40,359.00	0.2	8071.80	44	18.75	10	0.98

Note
35,535 patients are screened at Homerton University Hospital Foundation Trust per year. All costs are in 2013/14 GBP.

TABLE 25 Cost allocation per level of manual grading, given by mix of graders

	Total cost per patient screened (£)					
Grading level	Band 7	Band 6	Band 5	Band 4	Aggregated cost per patient	Mean total cost per patient[a]
Level 1	–	3.26	1.85	1.44	6.55	2.99
Level 2	–	2.75	1.85	–	4.60	2.65
Level 3	3.15	0.98	–	–	4.13	1.63

a Volume weighted across bands by proportion of patients observed.
Note
35,535 patients are screened at Homerton University Hospital Foundation Trust per year. All costs are in 2013/14 GBP.

worker; any fraction of one WTE per grader band was assumed to equal part of one worker's time. WTE-based costing was used to derive salary expenditure proportional to screening centre needs and was also used to derive other labour-related expenditures (QA).

Quality assurance costs unique to manual grading were the only QA costs included in the model. This involves all grading staff (levels 1, 2 and arbitration grade/level 3) taking a monthly online test and training course 10 months out of the year that requires 1 hour of their time each month.

Other costs that are universal to both manual grading and automated grading (both strategies) are summarised in *Table 26*. These were not included in our model because they are the same across both screening options.

Table 27 includes all infrastructure and capital costs for the screening programme that are also common to both screening choices (manual grading and automated) and, therefore, not included in our model.

Automated screening costs

As the automated screening systems examined here are not yet available in the English NHS, pre-market costing information was sought directly from individual manufacturers. This yielded system costing for manufacturers that were framed in the same way, cost per patient image set, and included similar components. Therefore, we present models for the two software packages that incorporate cost information gathered from manufacturers using a universal ARIAS cost per image set as a base-case figure. Costing reflects the distribution of actual costs suggested by manufacturers.

The key costing components of automated screening would include software purchase, licensing, user training, server upgrades and software installation and integration.[20,21] We had discussions with each of the automated screening software suppliers to understand what components might be included in cost per image pricing and what would be excluded. We communicated with all three software companies even though only two were subject to economic evaluation, as it gave us more information on how implementations might work in practice.

TABLE 26 Other labour costs

Cost item	Salary per year (£)	Number of employees	Total cost per year (£)	Mean cost per patient per year (£)
Receptionist	24,598.00	3	73,794.00	2.08
Patient history assessment	24,598.00	6	147,588.00	4.15
Fundus photography	24,598.00	6	147,588.00	4.15

Note
35,535 patients are screened at Homerton University Hospital Foundation Trust per year. All costs are in 2013/14 GBP.

TABLE 27 Technological infrastructure and capital costs for screening programme

Cost item	Cost (undiscounted) (£)	Cost (discounted) (£)	Cost per year (discounted) (£)	Cost per patient (discounted) (£)
Server costs	35,000.00	41,569.02	8313.80	0.23
Fundus camera	60,917.00	77,503.62	11,07200	0.31
Eye drops	0.72	0.72	25,585.00	0.72
Fixed capital	125,000.00	1,484,607.88	296,921.58	8.36

Note
35,535 patients are screened at Homerton University Hospital Foundation Trust per year. All costs are in 2013/14 GBP. When noted, costs are discounted over their expected useful lifespan: server costs are discounted over 5 years, whereas fundus camera equipment costs are discounted over 7 years. Fixed capital costs are discounted over 5 years, the expected useful lifespan of automated screening software. Discount rate used is 3.5%.

For all manufacturers, pricing would be contingent on the number of patients for guaranteed contracted volume and this would have major price implications. We discussed procurement of the software for a programme the size of the Homerton DESP at the time of the study (about 35,000 patients), as this would correspond with manual screening information that we captured from this centre. Given the nature of automated screening programmes that could conduct data processing across many screening centres, this may nevertheless reflect a small procurement. Sensitivity analyses were therefore conducted to examine how a range of per patient costs would affect our cost-effectiveness assessment.

Based on our discussions with manufacturers and clinicians regarding system implementation, the assumption taken in this study is that the software could be used with an existing server so there would be no additional costs related to server procurement. The cost per image screened would also include ongoing support and maintenance, and licensing, as well as staff training during initial set-up of the software for a screening programme.

Integration costs were difficult to capture. One manufacturer stated that there would be a 1- to 2-day cost related to set-up (integration costs) in addition to the cost/patient pricing, but that this cost was minimal. However, that manufacturer required that the screening programme have a particular, but common, data management system. Another manufacturer could not commit to the value of integration costs because of uncertainties in the new market. This manufacturer had experience with integration costs being minimal in other markets but this was dependent on factors such as patient numbers and other suppliers involved in the software's deployment. The third manufacturer stated that there would be some integration costs initially but could not estimate the cost.

Two manufacturers provided general cost per image figures, which we used to establish a baseline cost per image (£1 for both manufacturers). An independent bottom-up costing of each software system incorporating the cost components described above indicated that this general figure closely reflected the total cost of implementation of either automated screening programme. However, to address any uncertainty around this general figure, we also undertook extensive deterministic and threshold sensitivity analysis to examine the impact of this figure on results.

Other costs

Table 28 shows the remaining costs used in our model. Rescreening costs were derived as a sum of all applicable variable grading costs per patient. This reflected the maximum possible number of manual graders examining each image set, QA for manual grading and, in the case of automated screening, the cost per image set for use of automated screening software. Rescreening costs do not include capital costs, as they are universal to both manual grading and automated screening options. Rescreening costs per patient are therefore additive of incremental QA, labour and software use. Cost of a referral to the hospital eye service/ophthalmologist came from the NHS national tariffs; this does not include 26.64% market forces factor, or additional imaging costs that many diabetic patients incur at their first visit.[60]

TABLE 28 Additional cost parameters used in health economic model

Cost item	Cost/patient screened (£)
Level 1 (ARIAS)	1.00
Additional QA (manual grading)	0.01
12-month rescreen (manual grading)	7.26
12-month rescreen (ARIAS)	5.27
Ophthalmology referral	104.00[a]

a Excluding market forces factor and imaging cost (2014).[59]
35,535 patients are screened at Homerton University Hospital Foundation Trust per year. All costs are in 2013/14 GBP.
Sources: level 1 (ARIAS), assumption based on discussion with manufacturers, and additional QA and rescreen costs, Homerton Ophthalmology referral PSSRU.

Outcomes

Patient pathways were modelled for all patient episodes that were successfully screened in Homerton University Hospital Foundation Trust (20,258). This figure excluded known repeat admissions, as this could bias clinical pathways selected by the physician or requested by the patient. For each screening episode, the screening outcome from the automated system or level 1 grader was compared with the screening outcome given by the arbitration grader. This comparison was used to identify true positives and true negatives, which were then modelled as the outcome variable for each pathway. These reflected the proportion of cases correctly identified by the automated screening system as disease positive or disease negative, the two outcomes that were common to both automated screening systems. To compare across screening options fairly, R0 and R1 diagnoses from manual grading were considered to indicate a lack of disease, while M1, R2, R3 and technical failures (U) were taken as referable diagnoses. This approach is consistent with the existing English national screening programme for diabetic retinopathy and it applies a common standard for evaluation of both automated and manual screening options.

Analysis of cost per appropriate screening outcome

Results are expressed incrementally as cost per appropriate screening outcome (true positive or true negative correctly identified by the ARIAS) and reflect the two strategies of when to implement automated screening. For the ARIAS, an 'appropriate outcome' was defined as (1) identification of 'disease' present by the ARIAS when the reference human grade indicated prescence of potentially sight-threatening retinopathy or technical failure (including grades M1, R2, R3 and U) and (2) identification of 'no disease' by the ARIAS when the reference human grade indicated absence of retinopathy or background retinopathy only (grades R0 and R1, resulting in annual rescreening).

Deterministic sensitivity analysis

Sensitivity analyses were carried out to quantify the uncertainty and variability inherent within economic valuations on the parameters of key significance in our models. We analysed the impact of ARIAS pricing per patient on the incremental cost-effectiveness ratio (ICER). We also examined a number of other key variables, such as QA rates.

Results

Appropriate screening outcomes

Manual grading exhibited high sensitivity and specificity. It appropriately identified retinopathy status in 97.2% of all patient episodes. Of the images classified as non-disease (R0 or R1) by the level 1 human grader, 98.2% were correctly identified, whereas 90.6% of images classified as disease positive (M1a/b, R2, R3 or U) by the level 1 human grader were correctly classified. Ninety-seven per cent of R3, 97.1% of

R2, 90.6% of M1 and 68.1%% of U cases were correctly identified by level 1 manual graders as disease positive. Of all cases identified by level 1 manual graders as having a classification of M1a/b, R2, R3 or U, 4.0% were in fact R0 and a further 5.4% were R1. Therefore, 9.4% of disease-negative episodes were incorrectly identified by level 1 manual graders as having disease.

Retmarker classified 56.8% of image sets as having disease and identified 53.5% of appropriate outcomes. Of the images classified as non-disease by the software, 95.1% were correctly identified, whereas 21.0% of images classified as disease positive had the correct outcome. Nevertheless, sensitivity was highest at the extremes: 97.9% of R3, 96.4% of R2, 86.7% of M1 and 54.6% of U cases were correctly identified by Retmarker as disease positive. However, the total proportion of cases correctly identified in the disease-positive arm was driven down by a large number of cases that were in fact true negatives: of all cases identified by Retmarker as disease positive, 52.7% were in fact R0 and a further 26.3% were actually R1. Therefore 79.0% of disease-negative episodes were incorrectly identified by Retmarker as having disease. In summary, Retmarker's specificity within the cases that it identifies as disease negative is high; system performance is also high in identifying disease-positive cases when disease is actually present. However, the system appears to have a high type I (false-positive) error rate.

EyeArt classified 85.5% of images as having disease and identified 26.8% of appropriate outcomes overall. Of the images classified as disease negative by the software, 94.0% were correctly identified; 15.4% of images classified as disease positive had the correct outcome. Sensitivity was again highest at the extremes: 99.6% of R3, 99.4% of R2, 95.3% of M1 and 77.0% of U cases were correctly identified by EyeArt as disease positive. However, the total proportion of cases correctly identified within the disease-positive arm was driven down by a large number of cases that were in fact true negatives: of all cases identified by EyeArt as disease positive, 59.2% were in fact R0 and a further 25.4% were actually R1. Therefore, 84.6% of disease-negative episodes were incorrectly identified by EyeArt as having disease. As for the other automated screening system, EyeArt specificity within the cases that it identifies as disease negative is high; system performance is also high in identifying disease-positive cases when disease is actually present. However, EyeArt also appears to have a high type I (false-positive) error rate.

Cost per appropriate screening outcomes

Table 29 shows the costs of screening patients in our sample under either strategy 1 or strategy 2 and using either EyeArt or Retmarker. Under both strategies, the results for both software systems are directionally the same in that the ARIASs are both cheaper but also less effective than the current manual grading system. Because the 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. A lower ICER means that the intervention is less preferred.[61] In addition, for both Retmarker and EyeArt, strategy 1 provides more cost savings per appropriate outcomes missed than strategy 2.

For strategy 1 with Retmarker, the ICER can be interpreted as £11.81 in savings per appropriate outcome not identified but missed by Retmarker. If this figure were to fall, then an appropriate outcome would be missed but for a lower amount of savings.

For strategy 2 with Retmarker, ICER results still lie in the south-west quadrant. However, in comparison with strategy 1, there would be lower cost savings per appropriate outcome missed at £9.71. The effectiveness of strategies 1 and 2 for the same software system are almost identical. This probably 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 are higher than level 2 graders in our study. The average difference in cost between strategy 1 and strategy 2 for Retmarker is £0.24 per patient in the no-disease arm and £1.47 in the disease arm. Therefore, the biggest cost difference comes in those patients who were more likely to see a higher number of human graders because the automated screening system here acts as a filter rather than a replacement.

TABLE 29 Base-case results for 20,258 patients

Screening strategy and software	Total cost of grading (£)	Incremental cost (£)	Appropriate outcomes	Incremental appropriate outcomes	Reduced cost per appropriate outcome missed (ICER)[a] (£)
EyeArt					
Strategy 1[b]					
Manual grading	502,544.03	–	19,684.17	–	–
ARIAS	438,193.71	64,350.32	5427.00	14,257.20	4.51
Strategy 2[c]					
Manual grading	502,544.03	–	19,684.17	–	–
ARIAS	426,687.64	39,856.39	5428.00	14,256.17	2.80
Retmarker					
Strategy 1[b]					
Manual grading	502,544.03	–	19,684.17	–	–
ARIAS	396,841.49	105,702.54	10,731.00	8953.17	11.81
Strategy 2[c]					
Manual grading	502,544.03	–	19,684.17	–	–
ARIAS	415,863.95	86,680.11	10,760.55	8923.62	9.71

a If the intervention (ARIAS) was 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 but for a lower amount of savings.
b Strategy 1 replaces level 1 grader with an automated system.
c Strategy 2 is when an automated system acts as a filter prior to level 1 grader.

For strategy 1 with EyeArt, the ICER is £4.51 in savings relative to manual grading per appropriate outcome missed, whereas for strategy 2 savings are lower relative to manual grading, at £2.80. As is the case with Retmarker, cost implications drive the difference in savings and not effectiveness, as effectiveness is the same.

Deterministic sensitivity analysis

Automated screening costs

We undertook one-way sensitivity analysis to check the robustness of our findings to 50% changes in ARIAS pricing. *Table 30* shows the results of varying the cost of ARIASs per patient to demonstrate that, as expected, as the cost of ARIASs rises, using ARIASs is less cost saving per appropriate outcome missed (for all strategies and both software systems).

TABLE 30 Impact of variations in automated screening pricing

	ICER[a] (£)				
Screening strategy and software	ARIAS = £0.50 per patient	ARIAS = £0.75 per patient	ARIAS = £1.00 per patient[b]	ARIAS = £1.25 per patient	ARIAS = £1.50 per patient
EyeArt					
Strategy 1[c]	5.83	5.17	4.51	3.85	3.19
Strategy 2[d]	4.12	3.46	2.80	2.13	1.47
Retmarker					
Strategy 1[c]	13.90	12.86	11.81	10.76	9.71
Strategy 2[d]	11.85	10.78	9.71	8.65	7.58

a ICER is reduced cost per appropriate outcome missed.
b Base case.
c Strategy 1 replaces level 1 grader with an automated system.
d Strategy 2 is when an automated system acts as a filter prior to level 1 grader.

The results of the threshold analysis testing the highest ARIAS cost per patient before becomes more expensive per appropriate outcome than manual grading are given in *Figure 3*. This analysis demonstrates that for strategy 1 and Retmarker this figure is £3.82; therefore, if the ARIAS cost £3.82 more per patient, it would be costlier than manual grading. In strategy 2 for Retmarker this figure is £3.28. Given that the other components of the pathway come out more expensive in strategy 2, the lower figure for the ARIAS pricing would be expected.

For EyeArt for strategy 1, if the ARIAS cost more than £2.71 per patient, then it would be costlier than manual grading. For strategy 2, this figure is £2.05. Again, cost components in the rest of the pathway for

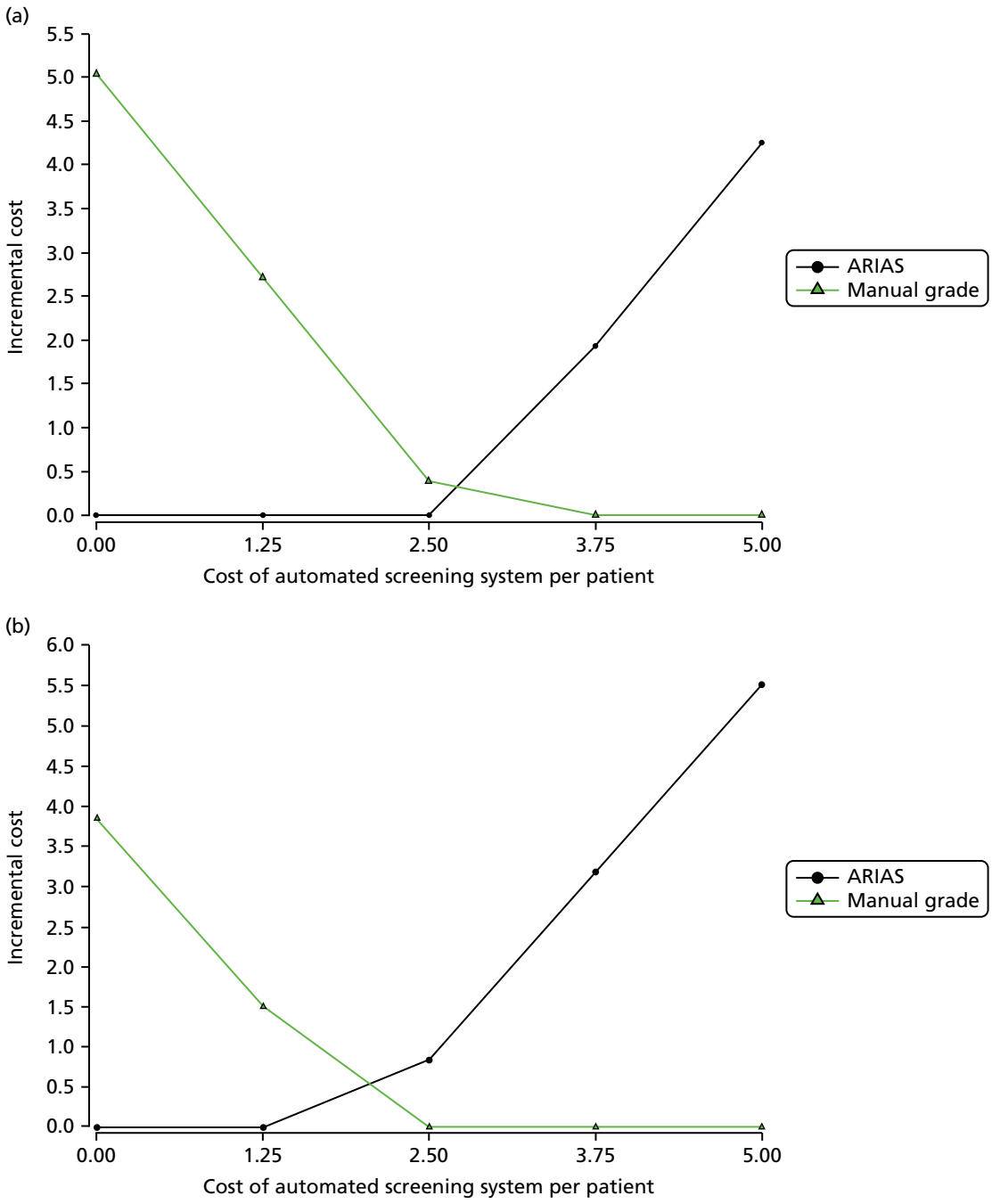


FIGURE 3 Deterministic sensitivity analysis of variations in automated screening pricing. (a) Strategy 1 EyeArt; (b) strategy 2 EyeArt; (c) strategy 1 Retmarker; and (d) strategy 2 Retmarker. (*continued*)

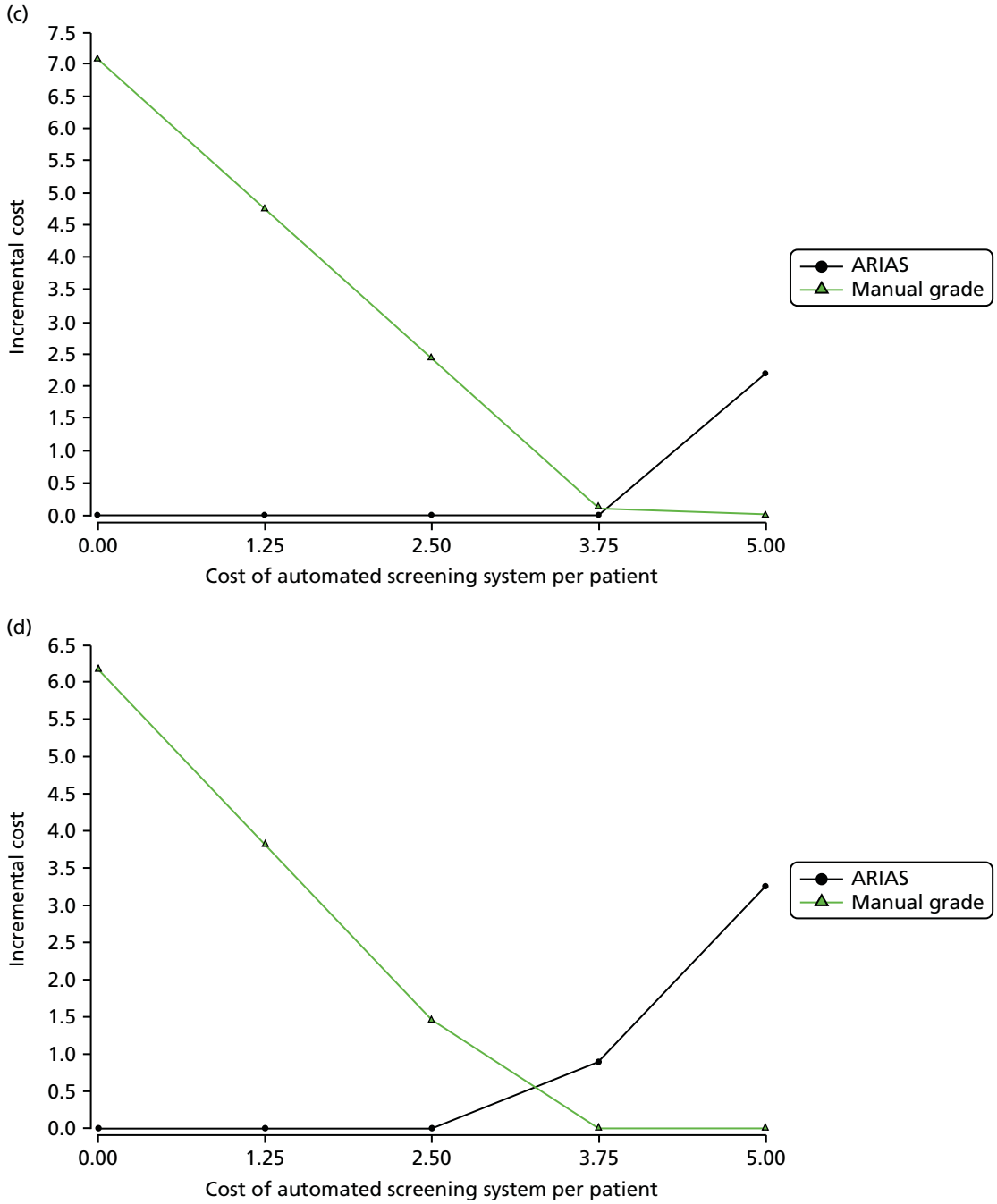


FIGURE 3 Deterministic sensitivity analysis of variations in automated screening pricing. (a) Strategy 1 EyeArt; (b) strategy 2 EyeArt; (c) strategy 1 Retmarker; and (d) strategy 2 Retmarker.

strategy 2, namely that level 1 graders have the highest cost per patient image screened compared with other level graders, are behind this result.

Manual grading costs

In terms of other costs, the cost of level 1, 2 and 3 graders would be expected to influence the appropriate choice of ARIAS or a manual grader acting as a level 1 grader. If one was maximising cost savings and willing to accept lower effect levels, then a 50% change in all three graders' costing levels (in both directions simultaneously) would not affect the choice of automated screening, as automated screening would still be cheaper than manual grading (strategies 1 and 2 Retmarker). Examining the cost of level 1 graders in particular, as the cost of level 1 graders increases, automated screening would present

additional cost savings if used as in both strategies for Retmarker. Level 3 grader cost changes have limited effect on cost savings in both strategies but level 1 changes make a major difference. For example, in strategy 2, the ICER varies by £2.64 savings per appropriate outcome missed when the cost of level 1 graders varies by 50%. For level 2 graders the figure is £1.09, and for level 3 graders the figure is £0.01.

The same can be said for EyeArt; varying level 1, 2 or 3 grader costs by 50% does not change the cost-saving nature of the software and the costs of level 1 have the greatest impact on cost savings, level 2 having the next greatest impact and level 3 having minimal impact (strategies 1 and 2).

Quality assurance protocols

We also examined how changing the percentage of patient image sets being sent for QA by a manual grader affects cost-effectiveness. QA was modelled to occur when the ARIAS states that no disease is present. Rather than allowing for immediate 12-month patient recall, 10% of these image sets are sent to a level 2 grader under strategy 1 and 10% are sent to a level 1 grader anyway under strategy 2. This made no difference in where the ICER lies on the cost-effectiveness quadrant (south-west corner) when varying the figure between the extremes of 0% and 100%. As expected, a lower proportion of image sets requiring QA was associated with the automated systems offering more cost savings per patient (strategies 1 and 2 for Retmarker and EyeArt).

Probability of automated screening system identifying disease

For Retmarker, 56.8% of the image sets were labelled as 'disease'. This figure would have to increase to 74.8% for strategy 1 or 2 to become costlier than manual grading.

For EyeArt, 85.5% of image sets were labelled as 'disease'. This figure would have to increase to 98.5% for strategy 1 to become costlier than manual grading.

Summary of findings

The findings of this model suggest that both software options offer cost-effective alternatives to a purely manual grading approach to diabetic retinopathy screening. Although they are less effective in picking up appropriate outcomes overall than manual grading, they are less expensive per patient, with these cost results being robust to significant variations in automated system pricing.

The automated systems are comparable with manual grading in picking up true negatives. However, both systems are overly sensitive in classifying images as disease positive. In cases when disease is actually present, both machines are comparable with human graders in correctly identifying positive cases, but aggregate performance of appropriate outcomes is driven down for both automated systems by a relatively high false-positive error rate.

Nevertheless, the automated screening systems tested here could reduce the requirement for manual grading. Both systems offer cost savings per appropriate outcome missed and further technical advances are likely to make the software increasingly cost-effective.

Chapter 5 Discussion

Primary analysis

In this observational retrospective study based on data from consecutive diabetic patients who attended an annual DESP, our primary analysis was to quantify the screening performance and diagnostic accuracy of all three ARIASs compared with manual grading and its cost-effectiveness at an acceptable sensitivity for detecting any retinopathy, referable retinopathy and sight-threatening retinopathy.[45] The point estimates for sensitivity using the manual grade modified by arbitration as the reference standard were as follows: EyeArt 94.7% (95% CI 94.2% to 95.2%) for any retinopathy, 93.8% (95% CI 92.9% to 94.6%) for referable retinopathy and 99.6% (95% CI 97.0% to 99.9%) for proliferative retinopathy. 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 proliferative retinopathy. For EyeArt, the 95% CI for the lower bound of the confidence limits for sensitivity was 93.7% to 94.9% for any retinopathy, 92% to 94% for referable retinopathy and 85.5% to 97.4% for R3. For Retmarker, the corresponding lower bound confidence limits were 70.9% to 73.1% for any retinopathy, 81.9% to 84.9% for referable retinopathy and 92.1% to 96.8% for proliferative retinopathy. The diagnostic accuracy of both EyeArt and Retmarker also falls within sensitivity levels set by the British Diabetic Association, achieving good sensitivity when compared with human graders for referable retinopathy. The false-positive rates were moderate for R0M0, at 80% and 47% for EyeArt and Retmarker, respectively. For iGradingM, none of the screening episodes was classified as 'no disease'. This raised concerns about how the iGradingM software processed both disc- and macular-centred images. We examined this directly on the subset of screening episodes that underwent arbitration and found that 99% of disc-centred images were classified as 'ungradable' by iGradingM, in addition to 36% of macular-centred images. iGradingM does not function on disc-centred images, but it can pick up some disease on macular-centred images. Therefore, iGradingM is of no use as a screening tool because it classified all encounters as containing at least one image with disease or ungradable.

Owing to the very poor performance of iGradingM, health economic analyses were undertaken only for EyeArt and Retmarker ARIASs. This study examined the cost-effectiveness of EyeArt and Retmarker ARIASs using two different strategies (strategies 1 and 2) versus manual grading, as currently performed at the Homerton DESP. We found that, when used as a replacement for level 1 grading (strategy 1), both automated screening systems were cost saving relative to manual grading but also less effective. When used as a filter prior to level 1 grading (strategy 2), thus reducing the number of level 2 grading episodes, both automated screening systems are less cost saving than if used as a replacement for level 1 graders.

The issue with both ARIASs appears to be not cost but effectiveness. Both systems are overly sensitive and performed very well in picking up cases of disease, but both had moderately high false-positive rates (low specificity). Hence, cost-effectiveness suffers because too many disease-free episodes are classified as disease positive, necessitating further investigation. The decision of whether or not to fund and implement a system into routine diabetic retinopathy screening programmes, and which system to use, would be based on the tolerability of the level of misclassified screening episodes balanced against cost. If the ARIASs were to be used without additional input from manual graders, a positive ARIAS classification, as modelled in strategies 1 and 2, will have both positive and negative consequences for patients. If the ARIAS outcome was reported to the patient, false positives may cause patients undue stress. On the other hand, overprovision of services via additional referrals to higher-level graders may provide a safety check for clinical diagnoses and other concurrent non-diabetic eye disease. From this perspective, cost-effectiveness figures derived in this economic evaluation provide a conservative estimate of the gains from the use of ARIASs in screening programmes and justify further cost–utility analyses of technical issues affecting adoption. One outstanding issue is that this study examines the cost-effectiveness of using ARIASs in two different places along the screening pathway for a set number of patients as per our study site. In reality, however, the scalability of

ARIASs is quite different in that, unlike manual graders, they do not take time off and can work 24 hours a day, 7 days a week, and therefore can grade many more images than manual graders over the same time period . This overall ability to scale up the screening programme by implementing an ARIAS should be considered for implementation.

Assumptions about patient numbers are key to this decision process. The Homerton DESP is a relatively small screening programme and an automated screening system would probably be procured for a much larger patient population.

Modes of implementation other than the pathways modelled here might handle some of these issues related to the realities of automated screening and better reflect gains to society. For example, under strategy 1, even if the automated screening system identified disease but a level 2 grader gave the conservative diagnosis of R0, implementation policy could be designed so that for this patient, the automated system reading could be ignored and no level 3 grading would be needed. This would allow the automated system to capture true-positive and true-negative cases while lowering costs associated with episodes identified by automated screening programmes as false positives.

Gold standard issues: missing proliferative disease

None of the systems had a 100% detection rate for proliferative disease. The study was not designed to look at the accuracy of human grading. No study has specifically looked at the sensitivity of human graders to detect R3 disease. It is clear that human graders at different levels differed in classifying R3, demonstrating that human graders themselves are not 100% accurate in detecting R3. It is striking that, of the 13 episodes classified as R3 by the Homerton DESP graders and sent to arbitration, only four were classified as R3 by the Doheny Image Reading Centre (three were classified as R1 or M1 and four were classified as U).

Secondary analyses

The colour of the fundus is determined by the choroidal blood supply and the amount of pigmentation in the choroid and overlying retinal pigment epithelium. Different ethnicities may have different levels of fundus pigmentation and fundus colour, which may influence the performance of an ARIAS. However, in this study there was no strong evidence to suggest that the sensitivity of the EyeArt or Retmarker ARIASs automated systems varies by ethnicity group but false-positive rates for combined grades R0M0 and R1M0 appeared to be lower in white European patients than in Asian and black African-Caribbean patients. There was some evidence that as patient age increased the ability of both the EyeArt and Retmarker ARIASs to detect combined grades U, R1M1, R2 and R3 fell. Whether or not this is an effect of poorer image quality with an ageing lens remains to be established. The performance of both ARIASs, but partularly Retmarker, differed according to camera type. The largest discrepancy for the detection of proliferative disease ranged from 100% for two out of the five cameras to 80% with another commonly used camera for Retmarker and from 100% to 97.4% for EyeArt. EyeArt appeared robust to variations in age, ethnicity and camera type, whereas Retmarker seemed to be less robust to variations in these patient-level factors. There is a need to understand why one ARIAS may be more sensitive to differences in camera type than another.

Detection of non-diabetic pathology

Although the study was not designed to evaluate the performance of the ARIASs on non-diabetic eye disease, we evaluated this in the subset of patient images sent for arbitration. A total of 467 images from 211 patients were coded with other pathologies by the Doheny Image Reading Centre. No episodes with at least one image coded with presence of age-related macular degeneration (11 patients), myopic degeneration (seven patients) or central retinal vein occlusion (two patients) were classified as ‘no disease’

by either software. Many more images were coded as having intermediate-sized drusen (298 images from 188 patients) or large-sized drusen (133 images from 70 patients). For intermediate-sized drusen 21 out of 133 episodes were graded as 'no disease' by EyeArt compared with 48 out of 133 for Retmarker. For large-sized drusen 10 out of 70 episodes were graded as 'no disease' by EyeArt, compared with 33 out of 70 for Retmarker. It is clear that a beneficial consequence of ARIASs is that other eye diseases may be detected.

Comparison of performance to previous studies

Sensitivity and specificity

In this study, the iGradingM ARIAS had a very poor specificity (0%). In a previous retrospective Scottish study[15] that used a single image per eye (in contrast to two images per eye in this study) iGradingM attained a sensitivity of 97.3% for referable retinopathy. However, the specificity was not reported. iGradingM has been further validated on a south London population with two fields per eye,[31] which reported sensitivity of 97.4–99.1% and specificity of 98.3–99.3%. These studies were not run completely independently of the software developers/vendor and it is unclear whether or not there was a different importation tool or pre-processing step undertaken. iGradingM, as configured for this study, did not recognise disc-centred images and classified them as ungradable, meaning that in this study the vast majority of episodes were ungradable because at least one image per eye was ungradable. The Scottish diabetic retinopathy screening programme, in contrast to the English programme, uses one macular-centred image per eye. It is possible that the iGradingM software used was not programmed to handle disc-centred images. However, all vendors were made aware of the two fields per eye standard and had equal access to a test set of representative images from the screening programme from which the test images were taken at initiation of this study. EyeArt and Retmarker in this study produced similar sensitivity to that previously reported on different data sets.[39,42]

Health economics comparison with previous literature

A number of previous studies have examined the cost-effectiveness of diabetic retinopathy screening (e.g. James *et al.*,[62] Javitt *et al.*[63]) and there is little doubt that these screening programmes are cost-effective. The use of automated screening systems to replace a grading step(s) and/or act as an image filter has been less well researched. As technology moves on and new systems become available, these studies need to be updated. Scotland *et al.*[21] examined the cost-effectiveness of introducing automated screening into national screening for diabetic retinopathy in Scotland. As in our study, their research used a decision-tree model and assessed a variant of our strategy 1 (replacing level 1 graders with automated screening). They also found automated grading to be less costly and largely less effective than manual grading. In their probabilistic sensitivity analysis, 18% of cases showed automated screening to be both less costly and more effective than manual grading. On balance, their recommendation was that automated screening should be chosen for Scotland, as the small reduction in finding referable cases when using automated screening is worth the cost saving. Sharp *et al.*[64] found automated screening was not cost-effective; however, the image processing technology has changed considerably since both Scotland *et al.*[21] and Sharp *et al.*,[64] making direct comparability of cost-effectiveness difficult. Issues around implementation and costing automated screening remain, however, as these systems have not been purchased and implemented in England.

Prescott *et al.*[65] examined the cost-effectiveness of a number of strategies in screening for diabetic macular oedema, including automated screening, but this did not cover retinopathy screening.

Altered thresholds

In any assessment of screening, automated or otherwise, there is a trade-off between sensitivity and specificity. Manufacturers of ARIASs fix an 'operating point' at which a commercial and clinical case can be made for the system's effectiveness. As part of the economic evaluation we explored different operating points. With Retmarker, varying the closest obtained threshold to alter the sensitivity by an absolute

difference of ± 5%, generated values of specificity between 35% and 52%. EyeArt provided two classifications, one with the software set to detect 'any retinopathy' and one when it is set to detect 'referable retinopathy'. These were implemented as two separate decision statistics rather than two different thresholds on one decision statistic. It was the former decision statistic that was used in the economic evaluation as it seemed to offer the closest comparison with the Retmarker. However, the latter statistic perhaps offered the better basis for cost-effective screening. At thresholds set to deliver sensitivity of between 86% and 96%, it could deliver specificities of between 14% and 54%.

A breakdown of the false-negative results further illuminated the trade-off between sensitivity and specificity. Sensitivity was largely influenced by episodes graded M1 and U. Both systems, but especially EyeArt, remained relatively successful at detecting the most clinically significant grades of retinopathy, even if overall sensitivity was relaxed in favour of specificity. At a threshold set to attain a specificity of > 40%, 98.9% of R2 and 99.2% of R3 cases were correctly identified.

In assessing the metric used in the economic evaluation, 'cost per appropriate outcome', it is worth considering that many of the outcomes we have had to regard as 'inappropriate' would not have had any adverse clinical outcome. For example, we regarded an episode with a final human classification of 'U' as a false negative if classed as being of adequate quality and disease free by the software. Given the large proportion of these ungradable episodes within the false negatives, we looked more carefully at how these were generated. EyeArt, for example, analysed each image and if more than one image in the episode was classified as adequate, the episode was classified as adequate for image quality. Requiring that an episode be classified as U if 50% or more of images were technically inadequate would improve sensitivity, on our metric, from 90% to 93%, albeit at the cost of slightly reduced specificity (from 35% to 32%).

Conformité Européenne marking

The two ARIASs that underwent health economic evaluation in this study, Retmarker and EyeArt, have class IIa CE marking, allowing them to be legally placed on the UK market. Of note, since the initiation of this study, new versions of both software packages have been released but were not tested in this study.

Implementation in real life and generalisability

In contrast to previous studies of ARIASs, multiple systems were compared on the same data set of consecutive patient screening episodes from a relatively large NHS diabetic retinopathy screening programme, using multiple different camera types and on populations of patients from different ethnic groups spanning a wide age range. All steps of the study were undertaken by study personnel independent of ARIAS vendors. The ARIAS vendors were locked out of the systems after the testing phase of their software. We are confident that the results are generalisable to NHS screening programmes. We noted a variation in screening performance by camera type and it may be useful to run test sets of previously graded images for each camera before implementation of a particular ARIAS in a new programme.

Technical issues for implementation

There were significant problems with some of the software vendors in running the test data. Given the large number of images, we noted that two ARIASs (Retmarker and iGradingM) had significant scaling and infrastructure issues. Several unsuccessful attempts to run the software were made, and the vendors had to be contacted to help resolve the issues. The limitations of both systems came from the growth of temporary files and a need for considerable database management. This may not be an issue with the likely number of patient episodes processed daily even in a large DESP.

The software that had the fewest problems executing the test data was Eyenuk Inc.'s cloud-based solution, EyeArr. This infrastructure allows for a scalable, elastic computing that handled the test data volume without difficulty. However, there are significant concerns with using cloud-based software, the foremost of which is the security implications of allowing imaging data to leave the hospital information technology (IT) infrastructure. In the evaluation phase of the study, significant effort was taken to ensure that all imaging data fed into the programs had all patient-identifiable information removed as part of the study protocol. In a real implementation there are security risks associated with patient information stored in metadata or in the Digital Imaging and Communications in Medicine (DICOM) headers format containing unencrypted identifiable information.

In addition to security risks, there are concerns of network bandwidth. Eyenuk Inc.'s software runs within the hospital system and downsamples the images prior to uploading the images to the cloud. Although this would minimise the number of data transferred, there are concerns over the impact of the software on the network load on the hospital. In a real implementation scenario there could be a bottleneck created by the number of images processed and the connection bandwidth of the hospital to the internet and the cloud computing infrastructure.

Many of the software vendors have trained their software to expect only retinal images, and thus the inclusion of non-retinal images might have affected the ability of the software to detect retinopathy at an episode level. In future implementations, images should be annotated to specify whether or not the image was a retinal image. The lens images could then be precluded from being inputted into the software.

Limitations of our study

This study is limited by its time horizon. We mapped an acute episode in the diabetic retinopathy screening programme, as in Scotland *et al.*,[21] but not the ongoing activities of the programme. Importantly, missed episodes of retinopathy could possibly have implications for patient sight. Recent work provides the transition probability data necessary to undertake this analysis (Catherine Egan, Moorfields Eye Hospital, 2016, personal communication) and should be an area of future examination.

Although a randomised controlled trial of ARIAS versus manual grading may seem desirable, the problem with a randomised controlled trial in the context of unproven software is that, for safety/ethical reasons, the investigative pathway would need to incorporate a safety check to ensure that cases of sight-threatening diabetic retinopathy were not missed by the ARIAS. This would likely lead to discrepancies that could result in an intervention in the screening pathway altering the outcome of missed retinopathy, which in turn would alter the referral pathway for a patient. Therefore, there would be no advantage of having a prospective randomised trial, as the intervention would obviate the ability to judge the implications of missed retinopathy over time. However, as there were three ARIASs tested, the advantage of a non-randomised retrospective study design is that all three ARIASs can be tested on exactly the image set. This created limitations in our analysis because of the need to map ARIAS findings onto manual grading results. The reference standard based on manual grading varies, in terms of if a screening episode received grading by level 1, 2 or 3 graders. According to the pathway protocol for strategy 1 (replacement of level 1 grader with ARIAS) if the ARIAS classified the episode as 'disease' the episode should receive a level 2 grade. However, some patients did not receive a level 2 grade in real life if the level 1 grader determined that no disease was present. In order to populate the health economics model we needed to assign level 2 and 3 grades to patients with automated screening outcomes that would have gone on to the level 2 grader. For these cases, we assumed a level 1 grader would act as a level 2 grader and a level 2 grader would act as a level 3 grader. If level 1 graders are, in reality, less experienced than level 2 graders, this may have underestimated the efficacy of automated screening programmes under strategy 1. However, comparison of results from both strategy 1 and 2 suggests that even if such a bias existed, its effect on final cost-effectiveness assessments was nominal.

In practice, level 2 graders are blind to grades received from level 1 graders, but level 3 graders are not blind to previous grades and would not be blind to automated system classifications. We assumed that level 1 graders are independent of later grades. In reality, however, it is possible that level 2 and 3 graders may be influenced by the disease/no disease classification from the automated system or from diagnoses given from the level 1 manual grader. Given the greater precision of manual grading diagnoses, one would expect level 3 grading scores to be more sensitive to level 1 manual grading scores than ARIAS classifications. This may mean that our cost-effectiveness assessment would vary slightly if ARIASs are implemented. Nevertheless, this ex ante modelling approach was necessary because ARIASs have yet to be adopted in the NHS.

Our findings are also limited by the effectiveness variable chosen. We chose 'appropriate outcome' because this allows us to analyse the screening goals of capturing true positives and true negatives together. It has also been used in other existing papers on this topic.[20,21] However, the end result for all patients either having automated screening or manual grading will be referral or annual rescreening. Pathway-specific safety checks by way of secondary evaluations of patient episodes by level 1, level 2 and level 3 manual graders after automated screening may avert inappropriate patient referrals or rescreens from automated screening scores. In reality, it is likely that the cost-effectiveness assessment produced here overestimates welfare losses occurring as a result of high false-positive rates in ARIASs. Although these systems do perform poorly under some circumstances, namely when disease is identified, tentative implementation protocols do not allow for direct referral from ARIAS disease classification. At the extremes, ARIASs are observed to demonstrate high rates of sensitivity and specificity, that is, true-negative and true-positive detection rates are high in ARIASs and comparable with level 1 manual graders. However, the impact on patient referability in the case of false positives is likely to be mitigated in practice through the existence of safety checks by way of level 2 and level 3 graders. Observational trials evaluating ARIAS effectiveness by way of correct patient referral or rescreen rates are needed to validate this hypothesis.

Our interpretation of results was also limited by the fact that the automated screening systems had differing ways of treating technical failures. Retmarker had two output options: disease and no disease. Technical failures were classified as disease positive to reflect manual grading protocols. Although this does not affect the main results, it did not allow us to differentiate the reasons behind 'disease' classifications for Retmarker. For EyeArt, technical failures were also included in the disease-positive arm of our decision tree, although this software does offer a technical failure differentiation. We treated technical failures as the same for comparability purposes but further analyses may wish to differentiate technical failures and disease-positive classifications to model the cost-effectiveness of EyeArt. Given the high rate of false positives associated with both automated screening systems, disaggregating disease-positive classifications into disease positive and technical failure may allow researchers to examine the inherent effectiveness of ARIASs after controlling for potentially confounding latent variables (e.g. dust on image).

Should these technologies be adopted, future experience with implemented automated screening systems in the NHS for diabetic retinopathy may allow for a better understanding of the costs associated with automated screening systems. Although we used sensitivity analysis to examine this point, the practicalities of implementation could be better quantified on actual experience.

Chapter 6 Conclusions and future research

Two ARIASs, Retmarker and EyeArt, met acceptable sensitivity levels for referable disease[45] and achieved sufficient levels of specificity to make them potentially cost-effective for deployment in DESPs.

These two ARIASs (Retmarker and EyeArt) offer cost-effective alternatives to a purely manual grading approach to diabetic retinopathy screening. Although they are less effective in picking up appropriate outcomes overall than manual grading, they are less expensive per patient, with these cost results being robust to significant variations in automated system pricing. Although this may seem counterintuitive to recommend ARIASs that are less effective at picking up appropriate outcomes (i.e. correctly classifying true positives and true negatives), this is because they are overly sensitive and so are still safe to potentially deploy and then, pragmatically, the reason to deploy would be cost saving.

The two ARIASs tested offer promise to diabetic retinopathy screening programmes as an option to reduce the extent of manual grading required. Both systems offer cost savings per appropriate outcome missed and further delineation of false positives will only improve effectiveness. In the light of the screening programme protocols evaluated, even if an automated screening software is overly sensitive, the patient will, in practice, probably still 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 a filter strategy is chosen for implementation of the automated screening systems. The advantages and disadvantages of the ARIASs, Retmarker and EyeArt, might usefully be evaluated by Public Health England for consideration of introduction into screening pathways when additional technical and governance issues raised have been addressed.

Prioritised research recommendations

The following is a list of research recommendations from this study:

1. Planning for an ARIAS-based programme: we propose technical standards for image acquisition and recommendations for service provision. Testing on previously acquired images from a particular programme before implementation will be important in the development of a screening programme protocol.
2. Messaging development: develop pathways of communication for current and future ARIASs to exchange information and images between the primary screening software and ARIASs.
3. Image repository test set: set up an image repository of the test images used in this study to allow for rapid evaluation of newer software versions of currently available ARIASs, as well as new CE-marked ARIASs that have become available since this study. The data set should include images that were not read by arbitration graders to provide a resource representative of 'real life'.
4. Image repository training set: set up an imaging repository of graded images that are not part of the definitive evaluation set, which will help accelerate the development of more effective ARIASs.
5. Image metadata: an early priority will be to develop quality standards for retinal imaging that may make the potential effectiveness of ARIASs even greater.
6. New ARIAS testing: it is important to establish a protocol that allows standardised evaluation of updated/novel ARIASs to be undertaken, optimising rapid uptake in screening practice.
7. Screen interval: it will be important to assess if annual or biennial screening after two no disease/disease cut-off points is appropriate for a future programme using ARIASs.
8. Health economics transition: probabilities of progression between grades of retinopathy have just become available. A Markov Model for evaluating ARIAS cost-effectiveness in screening for diabetic retinopathy could now be developed using these recently available data.
9. Wide field imaging modalities: if wide-field imaging becomes available for diabetic retinopathy screening, a modified training and validation set should be developed to help validate ARIAS use in this context.

Future prospects

In our opinion, a technical panel should be set up with sufficient expertise to advise Public Health England on technical and integration standards that would help create a market for ARIAS. We suggest that the panel should be independent of all commercial vendors but may seek advice from vendor representatives. A standard 'real-life' data set should be set up (e.g. the data set from this study) and an adjudicating panel that can evaluate the test performance of new/updated ARIASs and consider implementation into the NHS, independently of the software vendors.

Export of the images to the ARIASs in this study relied on a tool commissioned to transfer anonymised images from the Digital Healthcare screening database to a location for the ARIASs to analyse. This system was not automated. Standards to allow automated data interchange between ARIASs and the primary screening systems are needed for effective implementation. All of the software tested relied on manual instructions to initiate processing of the images. Future work should not only include DICOM compatibility but also a standardised method for automating the initiation and processing of output from the grading programmes, without the need for human intervention.

The workflow of the output from the ARIAS should be optimised to link back to the screening system, to set up a grading queue and randomly sample a proportion of images for arbitration.

The ARIASs may be even more cost-effective if fully integrated into the screening system to allow for real-time grading. This may allow for selective OCT to be done at one patient visit.

We noted a variation in grading test performance by camera type. This may be because of defects in the camera's image quality (e.g. dust on the lens). Developing quality standards for the retinal imaging process would be helpful.

Diabetic retinopathy screening as currently implemented with manual graders is not designed to pick up other non-diabetic eye disease but may detect other pathology as a by-product of implementation. This study was not powered or designed to look at non-diabetic retinopathy eye disease, such as a cataract. It may be that a combination of a visual acuity filter and image quality assessment will help pick non-diabetic retinopathy eye disease. This should be considered in the pathways during implementation.

Acknowledgements

We would like to thank:

Moorfields Eye Hospital NHS Trust IT department.

Vikash Chudasama (IT Systems Manager, Moorfields Eye Hospital NHS Trust) for maintenance and set-up of study servers.

Homerton DESP.

Ryan Chambers (Diabetes Retinal Screening Data Manager, Homerton University Hospital Foundation Trust) for help with extraction and merging of patient demographic and medical history data.

Homerton University Hospital Foundation Trust IT department.

Robin McNamara (IT Systems Administrator, Homerton University Hospital Foundation Trust) for maintenance and set-up of study servers.

Doheny Image Reading Centre.

Sowmya Sriniva (Postdoctoral Research Fellow/Senior Grader, Doheny Image Reading and Research Laboratory).

Muneeswar Gupta Nittala (Senior Research Associate, Doheny Image Reading and Research Laboratory).

Srinivas Sadda (Professor of Ophthalmology, Doheny Eye Institute, and Medical Director, Doheny Image Reading and Research Laboratory).

We wish to thank Ms Sarah Kerry for her assistance in developing the electronic database and data entry form used by the Doheny Image Reading Centre for feature-based grading of images.

Steering Committee

We would like to thank the members of the Steering Committee for their time and invaluable advice.

Steven Aldington: independent member.

Mireia Jofre-Bonet: non-independent member.

Simon Jones: independent member.

Irwin Nazareth: independent member.

Adnan Tufail: chief investigator member.

Irene Stratton (Senior Statistician Gloucestershire Retinal Research Group): chairperson, independent member and statistician member.

Gillian Vafidis: independent member.

Richard Wormald (Sponsor Representative): non-independent member.

Contributions of authors

Adnan Tufail (Consultant Ophthalmologist, Moorfields Eye Hospital, and Professor of Ophthalmology Institute of Ophthalmology, University College London) was chief investigator of the study. He was involved in the conception and design of study, collection of data, interpretation of data and writing the final report.

Venediktos V Kapetanakis (previously Senior Research Fellow in Medical Statistics, St George's, University of London): was involved in the statistical analysis, acquisition of data, database management, interpretation of data and writing the final report.

Sebastian Salas-Vega (Doctoral student in Health Economics, Department of Social Policy and LSE Health, London School of Economics and Political Science): health economic analysis, interpretation of data and writing the final report.

Catherine Egan (Consultant Ophthalmologist, Moorfields Eye Hospital NHS Trust, and Honorary Senior Lecturer Institute of Ophthalmology, University College London) was co-investigator of the study. She was involved in the conception and design of the study, interpretation of data and writing the final report.

Caroline Rudisill (Associate Professor in Health Economics, Department of Social Policy and LSE Health, London School of Economics and Political Science) was involved in the conception and design of the study, acquisition of data, health economic analysis, interpretation of data and writing the final report.

Christopher G Owen (Professor of Epidemiology, St George's, University of London) was involved in the conception and design of the study, interpretation of data and writing the final report.

Aaron Lee (previously Retina Fellow, Moorfields Eye Hospital NHS Trust) was involved in the set up and activation of ARIASs, database management and editing of the final report.

Vern Louw (Data Warehouse Developer, IT Department, Moorfields Eye Hospital NHS Trust) was involved in the set up and activation of ARIASs and database management.

John Anderson (Consultant Physician and Endocrinologist, Homerton University Hospital Foundation Trust) was involved in the conception and design of the study, interpretation of human grading data and costs and review of the final report.

Gerald Liew (previously Retina Fellow, Moorfields Eye Hospital NHS Trust) was involved in data management, operational issues, interpretation of data and writing of the final report.

Louis Bolter (Divisional Service Manager, Homerton University Hospital Foundation Trust): was involved in the interpretation of human grading data and costs, support for the real-time study and review of the final report.

Clare Bailey (Consultant Ophthalmologist, Bristol Eye Hospital) was involved in the conception and design of the study and writing the final report.

SriniVas Sadda (Professor of Ophthalmology, Doheny Eye Institute, and Medical Director, Doheny Image Reading and Research Laboratory) was involved in the arbitration of study images and editing of the final report.

Paul Taylor (Reader in Health Informatics, Centre for Health Informatics & Multiprofessional Education (CHIME), Institute of Health Informatics, University College London) was involved in the conception and design of the study, analysis of altered thresholds, interpretation of data and writing the final report.

Alicja R Rudnicka (Reader in Medical Statistics, St George's, University of London) was involved in the conception and design of the study, acquisition of data, statistical analysis, interpretation of data and writing the final report.

Publications

Kapetanakis VV, Rudnicka AR, Liew G, Owen CG, Lee A, Louw V, Bolter L, Anderson J, Egan C, Salas-Vega S, Rudisill C, Taylor P, Tufail A. 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**:112–18.

Owen CG, Rudnicka AR, Kapetanakis V. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared to human graders. *Ophthalmology* 2016; in press.

Data sharing statement

We shall make data available to the scientific community with as few restrictions as feasible, while retaining exclusive use until the publication of major outputs. Anonymised demographic data can be obtained by contacting the corresponding author. The corresponding author welcomes proposals for collaborative projects using the image data set.

References

1. Diabetes UK. *State of the Nation*. 2013. URL: www.diabetes.org.uk/Documents/About%20Us/What%20we%20say/0160b-state-nation-2013-england-1213.pdf (accessed 20 August 2014).
2. Neubauer AS, Kernt M, Haritoglou C, Priglinger SG, Kampik A, Ulbig MW. Nonmydriatic screening for diabetic retinopathy by ultra-widefield scanning laser ophthalmoscopy (Optomap). *Graefes Arch Clin Exp Ophthalmol* 2008;**246**:229–35. http://dx.doi.org/10.1007/s00417-007-0631-4
3. Liew G, Michaelides M, Bunce C. A comparison of the causes of blindness certifications in England and Wales in working age adults (16-64 years), 1999–2000 with 2009-2010. *BMJ Open* 2014;**4**:e004015. http://dx.doi.org/10.1136/bmjopen-2013-004015
4. García M, López MI, Alvarez D, Hornero R. Assessment of four neural network based classifiers to automatically detect red lesions in retinal images. *Med Eng Phys* 2010;**32**:1085–93. http://dx.doi.org/10.1016/j.medengphy.2010.07.014
5. Davis MD, Fine SL, Goldberg MF, McMeel JW, Norton EWD, Okun E, Wetzig P. O'Hare Classification of Diabetic Retinopathy. In Goldberg MF, Fine SL, editors. *Symposium on the Treatment of Diabetic Retinopathy*. Public Health Service Publication No. 1890. Washington, DC: US Government Printing Office; 1969. pp. xxi–xxiv.
6. Kempen JH, O'Colmain BJ, Leske MC, Haffner SM, Klein R, Moss SE, *et al.* The prevalence of diabetic retinopathy among adults in the United States. *Arch Ophthalmol* 2004;**122**:552–63. http://dx.doi.org/10.1001/archopht.122.4.552
7. Klein R, Klein BE, Moss SE, Davis MD, DeMets DL. The Wisconsin epidemiologic study of diabetic retinopathy. III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch Ophthalmol* 1984;**102**:527–32. http://dx.doi.org/10.1001/archopht.1984.01040030405011
8. Klein R, Klein BE, Moss SE, Davis MD, DeMets DL. The Wisconsin epidemiologic study of diabetic retinopathy. II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Arch Ophthalmol* 1984;**102**:520–6. http://dx.doi.org/10.1001/archopht.1984.01040030398010
9. Roy MS, Klein R, O'Colmain BJ, Klein BE, Moss SE, Kempen JH. The prevalence of diabetic retinopathy among adult type 1 diabetic persons in the United States. *Arch Ophthalmol* 2004;**122**:546–51. http://dx.doi.org/10.1001/archopht.122.4.546
10. Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy. *Opthalmology* 1991;**98**(Suppl. 5):766–85. http://dx.doi.org/10.1016/S0161-6420(13)38011-7
11. Korobelnik JF, Do DV, Schmidt-Erfurth U, Boyer DS, Holz FG, Heier JS, *et al.* Intravitreal aflibercept for diabetic macular edema. *Ophthalmology* 2014;**121**:2247–54. http://dx.doi.org/10.1016/j.ophtha.2014.05.006
12. Brown DM, Nguyen QD, Marcus DM, Boyer DS, Patel S, Feiner L, *et al.* Long-term outcomes of ranibizumab therapy for diabetic macular edema: the 36-month results from two phase III trials: RISE and RIDE. *Ophthalmology* 2013;**120**:2013–22. http://dx.doi.org/10.1016/j.ophtha.2013.02.034
13. Public Health England. *NHS Screening Programmes in England 2014 to 2015*. URL: www.gov.uk/government/uploads/system/uploads/attachment_data/file/480968/Screening_in_England_2014-15_online_version.pdf (accessed 27 August 2016).
14. Shaw JE, Sicree RA, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res Clin Pract* 2010;**87**:4–14. http://dx.doi.org/10.1016/j.diabres.2009.10.007

15. Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *Br J Ophthalmol* 2010;**94**:1606–10. http://dx.doi.org/10.1136/bjo.2009.176784

16. Abràmoff MD, Niemeijer M, Suttorp-Schulten MS, Viergever MA, Russell SR, van Ginneken B. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. *Diabetes Care* 2008;**31**:193–8. http://dx.doi.org/10.2337/dc07-1312

17. Abràmoff MD, Reinhardt JM, Russell SR, Folk JC, Mahajan VB, Niemeijer M, Quellec G. Automated early detection of diabetic retinopathy. *Ophthalmology* 2010;**117**:1147–54. http://dx.doi.org/10.1016/j.ophtha.2010.03.046

18. Fleming AD, Goatman KA, Philip S, Williams GJ, Prescott GJ, Scotland GS, *et al.* The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol* 2010;**94**:706–11. http://dx.doi.org/10.1136/bjo.2008.149807

19. Philip S, Fleming AD, Goatman KA, Fonseca S, McNamee P, Scotland GS, *et al.* The efficacy of automated 'disease/no disease' grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol* 2007;**91**:1512–17. http://dx.doi.org/10.1136/bjo.2007.119453

20. Scotland GS, McNamee P, Fleming AD, Goatman KA, Philip S, Prescott GJ, *et al.* Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *Br J Ophthalmol* 2010;**94**:712–19. http://dx.doi.org/10.1136/bjo.2008.151126

21. Scotland GS, McNamee P, Philip S, Fleming AD, Goatman KA, Prescott GJ, *et al.* Cost-effectiveness of implementing automated grading within the national screening programme for diabetic retinopathy in Scotland. *Br J Ophthalmol* 2007;**91**:1518–23. http://dx.doi.org/10.1136/bjo.2007.120972

22. Niemeijer M, van Ginneken B, Cree MJ, Mizutani A, Quellec G, Sanchez CI, *et al.* Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE Trans Med Imaging* 2010;**29**:185–95. http://dx.doi.org/10.1109/TMI.2009.2033909

23. Mackenzie S, Schmermer C, Charnley A, Sim D, Vikas T, Dumskyj M, *et al.* SDOCT imaging to identify macular pathology in patients diagnosed with diabetic maculopathy by a digital photographic retinal screening programme. *PLOS ONE* 2011;**6**:e14811. http://dx.doi.org/10.1371/journal.pone.0014811

24. Public Health Policy and Strategy Unit, Department of Health. *NHS Public Health Functions Agreement 2014–15: Public Health Functions to be Exercised by NHS England Service*. 2013. URL: www.gov.uk/government/uploads/system/uploads/attachment_data/file/256502/nhs_public_health_functions_agreement_2014-15.pdf (accessed 6 September 2015).

25. Public Health England. *Diabetic Eye Screening: Guidance on Camera Approval.* 2015. URL: www.gov.uk/government/publications/diabetic-eye-screening-approved-cameras-and-settings/diabetic-eye-screening-guidance-on-camera-approval (accessed 1 August 2015).

26. Ahmed J, Ward TP, Bursell SE, Aiello LM, Cavallerano JD, Vigersky RA. The sensitivity and specificity of nonmydriatic digital stereoscopic retinal imaging in detecting diabetic retinopathy. *Diabetes Care* 2006;**29**:2205–9. http://dx.doi.org/10.2337/dc06-0295

27. Scanlon PH, Foy C, Malhotra R, Aldington SJ. The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. *Diabetes Care* 2005;**28**:2448–53. http://dx.doi.org/10.2337/diacare.28.10.2448

28. Fleming AD, Philip S, Goatman KA, Olson JA, Sharp PF. Automated assessment of diabetic retinal image quality based on clarity and field definition. *Invest Ophthalmol Vis Sci* 2006;**47**:1120–5. http://dx.doi.org/10.1167/iovs.05-1155

29. Early Treatment Diabetic Retinopathy Study research group. Photocoagulation for diabetic macular edema. Early Treatment Diabetic Retinopathy Study report number 1. *Arch Opthalmol* 1985;**103**:1796–806. http://dx.doi.org/10.1001/archopht.1985.01050120030015

30. Sim DA, Keane PA, Tufail A, Egan CA, Aiello LP, Silva PS. Automated retinal image analysis for diabetic retinopathy in telemedicine. *Curr Diab Rep* 2015;**15**:14. http://dx.doi.org/10.1007/s11892-015-0577-6

31. 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**:e27524. http://dx.doi.org/10.1371/journal.pone.0027524

32. Goatman KA. A reference standard for the measurement of macular oedema. *Br J Ophthalmol* 2006;**90**:1197–202. http://dx.doi.org/10.1136/bjo.2006.095885

33. Niemeijer M, van Ginneken B, Staal J, Suttorp-Schulten MS, Abràmoff MD. Automatic detection of red lesions in digital color fundus photographs. *IEEE Trans Med Imaging* 2005;**24**:584–92. http://dx.doi.org/10.1109/TMI.2005.843738

34. Niemeijer M, Abràmoff MD, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. *Med Image Anal* 2006;**10**:888–98. http://dx.doi.org/10.1016/j.media.2006.09.006

35. Tang L, Niemeijer M, Reinhardt JM, Garvin MK, Abràmoff MD. Splat feature classification with application to retinal hemorrhage detection in fundus images. *IEEE Trans Med Imaging* 2013;**32**:364–75. http://dx.doi.org/10.1109/TMI.2012.2227119

36. Abràmoff MD, Folk JC, Han DP, Walker JD, Williams DF, Russell SR, *et al.* Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* 2013;**131**:351–7. http://dx.doi.org/10.1001/jamaophthalmol.2013.1743

37. Pires Dias JM, Oliveira CM, da Silva Cruz LA. Retinal image quality assessment using generic image quality indicators. *Inf Fusion* 2014;**19**:73–90. http://dx.doi.org/10.1016/j.inffus.2012.08.001

38. Nunes S, Pires I, Rosa A, Duarte L, Bernardes R, Cunha-Vaz J. Microaneurysm turnover is a biomarker for diabetic retinopathy progression to clinically significant macular edema: findings for type 2 diabetics with nonproliferative retinopathy. *Ophthalmologica* 2009;**223**:292–7. http://dx.doi.org/10.1159/000213639

39. Oliveira CM, Cristovao LM, Ribeiro ML, Abreu JR. Improved automated screening of diabetic retinopathy. *Ophthalmologica* 2011;**226**:191–7. http://dx.doi.org/10.1159/000330285

40. Haritoglou C, Kernt M, Neubauer A, Gerss J, Oliveira CM, Kampik A, Ulbig M. Microaneurysm formation rate as a predictive marker for progression to clinically significant macular edema in nonproliferative diabetic retinopathy. *Retina* 2014;**34**:157–64. http://dx.doi.org/10.1097/IAE.0b013e318295f6de

41. Nunes RP, Gregori G, Yehoshua Z, Stetson PF, Feuer W, Moshfeghi AA, Rosenfeld PJ. Predicting the progression of geographic atrophy in age-related macular degeneration with SD-OCT en face imaging of the outer retina. *Ophthalmic Surg Lasers Imaging Retina* 2013;**44**:344–59. http://dx.doi.org/10.3928/23258160-20130715-06

42. Solanki K, Ramachandra C, Bhat S, Bhaskaranand M, Nittala MG, Sadda SR, editors. *EyeArt: Automated, High-throughput, Image Analysis for Diabetic Retinopathy Screening*. Denver, CO: The Association for Research in Vision and Ophthalmology; 2015.

43. Facey K, Cummins E, Macpherson K, Morris A, Reay L, Slattery J. *Organisation of Services for Diabetic Retinopathy Screening. Health Technology Assessment Report 1*. Glasgow: Health Technology Board for Scotland; 2002.

44. 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**:2086–90. http://dx.doi.org/10.2337/diacare.24.12.2086

45. British Diabetic Association. *Retinal Photography Screening for Diabetic Eye Disease: A British Diabetic Association Report*. London: British Diabetic Association; 1997.

46. Taylor R, Broadbent DM, Greenwood R, Hepburn D, Owens DR, Simpson H. Mobile retinal screening in Britain. *Diabet Med* 1998;**15**:344–7. http://dx.doi.org/10.1002/(SICI)1096-9136(199804)15:4%3C344::AID-DIA588%3E3.0.CO;2-O

47. Kinyoun JL, Martin DC, Fujimoto WY, Leonetti DL. Ophthalmoscopy versus fundus photographs for detecting and grading diabetic retinopathy. *Invest Ophthalmol Vis Sci* 1992;**33**:1888–93.

48. Pugh JA, Jacobson JM, Van Heuven WA, Watters JA, Tuley MR, Lairson DR, *et al.* Screening for diabetic retinopathy. The wide-angle retinal camera. *Diabetes Care* 1993;**16**:889–95. http://dx.doi.org/10.2337/diacare.16.6.889

49. Moss SE, Klein R, Kessler SD, Richie KA. Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy. *Ophthalmology* 1985;**92**:62–7. http://dx.doi.org/10.1016/S0161-6420(85)34082-4

50. Schachat AP, Hyman L, Leske MC, Connell AM, Hiner C, Javornik N, Alexander J. Comparison of diabetic retinopathy detection by clinical examinations and photograph gradings. Barbados (West Indies) Eye Study Group. *Arch Ophthalmol* 1993;**111**:1064–70. http://dx.doi.org/10.1001/archopht.1993.01090080060019

51. Kinyoun J, Barton F, Fisher M, Hubbard L, Aiello L, Ferris F. Detection of diabetic macular edema. Ophthalmoscopy versus photography – Early Treatment Diabetic Retinopathy Study Report Number 5. The ETDRS Research Group. *Opthalmology* 1989;**96**:746–50. http://dx.doi.org/10.1016/S0161-6420(89)32814-4

52. Core National Diabetic Eye Screening Programme team. *Feature Based Grading Forms, Version 1.4*. 2012. URL: www.gov.uk/government/uploads/system/uploads/attachment_data/file/402295/Feature_Based_Grading_Forms_V1_4_1Nov12_SSG.pdf (accessed 2 August 2016).

53. Taylor D. *Diabetic Eye Screening Revised Grading Definitions*. 2012. URL: www.gov.uk/government/uploads/system/uploads/attachment_data/file/402294/Revised_Grading_Definitions_V1_3_1Nov12_SSG.pdf (accessed 2 August 2016).

54. NHS Diabetic Screening Programme. *Operational Guidance*. 2013. URL http://diabeticeye.screening.nhs.uk/operational-guidance (accessed 10 July 2014).

55. Doheny Image Reading Center. *Doheny Image Reading Center*. 2013. URL: http://dirc.doheny.org/Default.aspx (accessed 10 July 2014).

56. Kapetanakis VV, Rudnicka AR, Liew G, Owen CG, Lee A, Louw V, *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**:112–18. http://dx.doi.org/10.1177/0969141315571953

57. UK National Screening Committee. *Essential Elements in Developing a Diabetic Retinopathy Screening Pathway. National Screening Programme for Diabetic Retinopathy Workbook 4.3*. 2009. URL http://rcophth-website.www.premierithosting.com/docs/publications/published-guidelines/ENSPDR_Workbook_2009.pdf (accessed August 2015).

58. Public Health England. *Diabetic Eye Screening: Commission and Provide*. 2015. URL: www.gov.uk/government/collections/diabetic-eye-screening-commission-and-provide (accessed August 2015).

59. Curtis L. *Unit Costs of Health and Social* Care. 2014. URL: www.pssru.ac.uk/project-pages/unit-costs/2014/ (accessed 19 June 2015).

60. Monitor and NHS England. *National Tariff Payment System 2014/15*. 2013. URL: www.gov.uk/government/publications/national-tariff-payment-system-2014-to-2015 (accessed 19 June 2015).

61. Gray AM, Clarke PM, Wolstenholme J, Wordsworth S. *Applied Methods of Cost-effectiveness Analysis in Healthcare*. Oxford: Oxford University Press; 2011.

62. James M, Turner DA, Broadbent DM, Vora J, Harding SP. Cost effectiveness analysis of screening for sight threatening diabetic eye disease. *BMJ* 2000;**320**:1627–31. http://dx.doi.org/10.1136/bmj.320.7250.1627

63. Javitt JC, Aiello LP, Chiang Y, Ferris FL, Canner JK, Greenfield S. Preventive eye care in people with diabetes is cost-saving to the federal government. Implications for health-care reform. *Diabetes Care* 1994;**17**:909–17. http://dx.doi.org/10.2337/diacare.17.8.909

64. Sharp PF, Olson J, Strachan F, Hipwell J, Ludbrook A, O'Donnell M, *et al.* The value of digital imaging in diabetic retinopathy. *Health Technol Assess* 2003;**7**(30). http://dx.doi.org/10.3310/hta7300

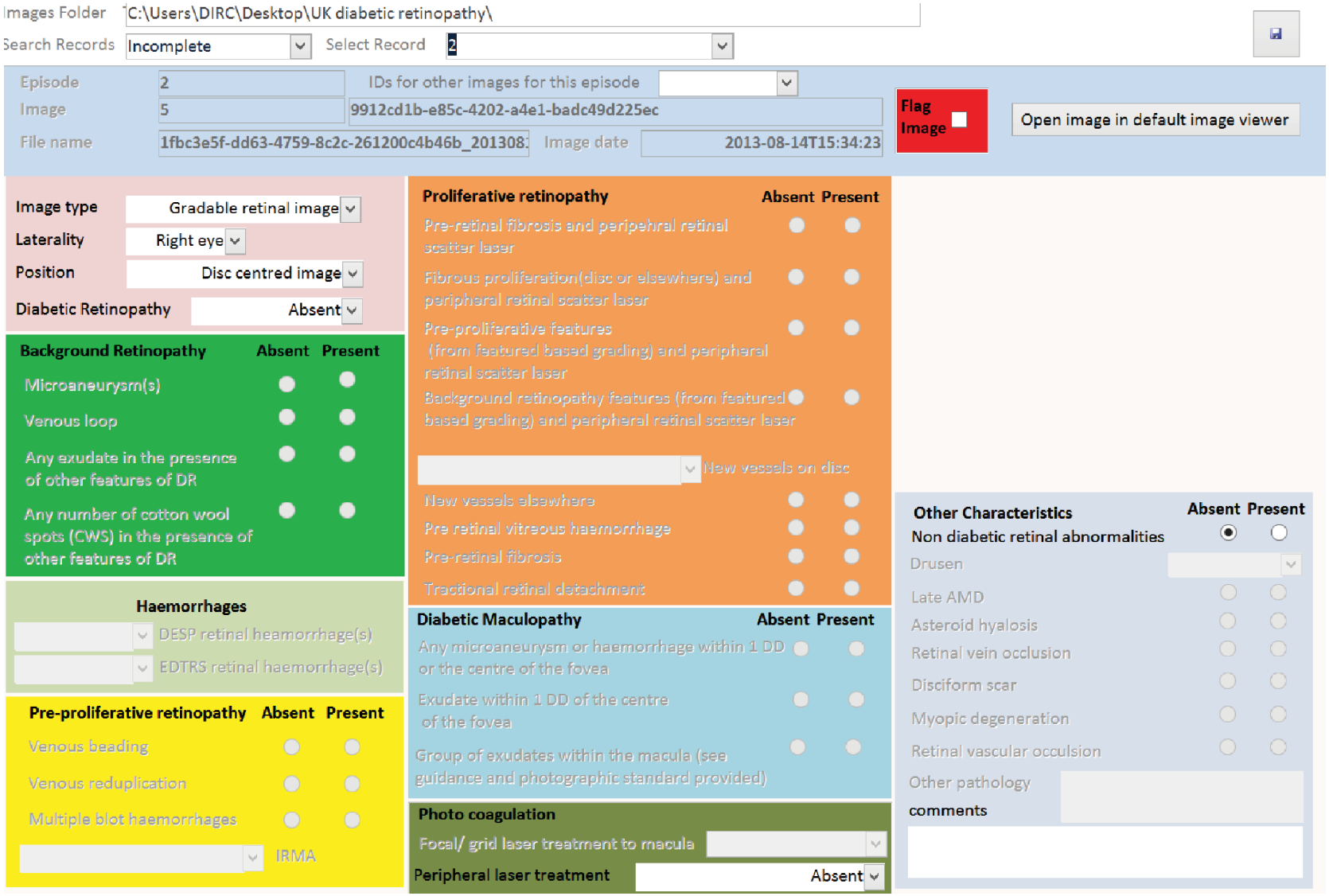
65. Prescott G, Sharp P, Goatman K, Scotland G, Fleming A, Philip S, *et al.* Improving the cost-effectiveness of photographic screening for diabetic macular oedema: a prospective, multi-centre, UK study. *Br J Ophthalmol* 2014;**98**:1042–9. http://dx.doi.org/10.1136/bjophthalmol-2013-304338

Appendix 1 Variables exported from OptoMize Data Export Module

Variables exported from Data Export Module for OptoMize version 3 database. Grading data were removed and the data file was used to link images to patient episode grading.

- ImageFileName
- ImageUniqueId
- ImageDate
- Laterality
- Camera
- BestVARight
- BestVALeft
- PrimaryImageQualityRight
- PrimaryImageQualityLeft
- PrimaryRetinopathyRight
- PrimaryRetinopathyLeft
- PrimaryMaculopathyRight
- PrimaryMaculopathyLeft
- PrimaryOutcome
- SecondaryImageQualityRight
- SecondaryImageQualityLeft
- SecondaryRetinopathyRight
- SecondaryRetinopathyLeft
- SecondaryMaculopathyRight
- SecondaryMaculopathyLeft
- SecondaryOutcome
- ArbitrationImageQualityRight
- ArbitrationImageQualityLeft
- ArbitrationRetinopathyRight
- ArbitrationRetinopathyLeft
- ArbitrationMaculopathyRight
- ArbitrationMaculopathyLeft
- ArbitrationOutcome
- FinalKind
- FinalImageQualityRight
- FinalImageQualityLeft
- FinalRetinopathyRight
- FinalRetinopathyLeft
- FinalMaculopathyRight
- FinalMaculopathyLeft
- FinalOutcome
- PatientID
- EncounterID

Appendix 2 Reading centre grading form



Appendix 3 Additional tables referred to in *Chapter 3*

TABLE 31 Grading pathways for first screening episodes from 20,258 patients attending the Homerton DESP 1 June 2012 and 4 November 2013

Level 1 grader	Level 2 grader	Arbitration grader	Final grade	Total number of patients
R0	Missing[a]	Missing[a]	R0	10,788
R0	R0	Missing[a]	R0	1658
R0	R0	R0	R0	1
R0	R1	Missing[a]	R1	69
R0	U	Missing[a]	U	20
R0	U	R0	R0	1
R0	M1a	Missing[a]	M1a	1
R0	M1a	R0	R0	3
R0	M1a	M1a	M1a	1
R0	M1b	Missing[a]	M1b	1
R0	R2	R2	R2	3
R0	R3	Missing[a]	R3	1
R1	R0	Missing[a]	R0	224
R1	R1	Missing[a]	R1	4534
R1	U	Missing[a]	U	21
R1	U	R1	R1	2
R1	M1a	Missing[a]	M1a	3
R1	M1a	R1	R1	18
R1	M1a	M1a	M1a	51
R1	M1a	M1b	M1b	1
R1	M1b	Missing[a]	M1b	33
R1	R2	Missing[a]	R2	6
R1	R2	R1	R1	6
R1	R2	R2	R2	7
R1	R2	R3	R3	1
R1	R3	Missing[a]	R3	4
U	Missing[a]	Missing[a]	U	10
U	R0	Missing[a]	R0	3
U	R0	R0	R0	15
U	R0	U	U	1
U	R1	Missing[a]	R1	4

continued

TABLE 31 Grading pathways for first screening episodes from 20,258 patients attending the Homerton DESP 1 June 2012 and 4 November 2013 (*continued*)

Level 1 grader	Level 2 grader	Arbitration grader	Final grade	Total number of patients
U	R1	R0	R0	1
U	R1	R1	R1	10
U	R1	U	U	1
U	U	Missing[a]	U	239
U	M1b	Missing[a]	M1b	3
U	R2	Missing[a]	R2	5
U	R3	Missing[a]	R3	1
M1a	R0	Missing[a]	R0	2
M1a	R0	R0	R0	17
M1a	R0	R1	R1	3
M1a	R0	M1a	M1a	2
M1a	R0	M1b	M1b	1
M1a	R1	Missing[a]	R1	2
M1a	R1	R0	R0	2
M1a	R1	R1	R1	55
M1a	R1	M1a	M1a	12
M1a	U	Missing[a]	U	2
M1a	M1a	Missing[a]	M1a	1054
M1a	M1a	M1a	M1a	23
M1a	M1a	M1b	M1b	1
M1a	M1b	Missing[a]	M1b	99
M1a	M1b	M1a	M1a	1
M1a	R2	Missing[a]	R2	16
M1a	R2	M1a	M1a	5
M1a	R2	M1b	M1b	1
M1a	R2	R2	R2	12
M1a	R3	Missing[a]	R3	3
M1b	R0	Missing[a]	R0	3
M1b	R0	R0	R0	2
M1b	R0	M1a	M1a	1
M1b	R1	Missing[a]	R1	6
M1b	R1	R1	R1	7
M1b	R1	M1a	M1a	1
M1b	U	Missing[a]	U	2
M1b	M1a	Missing[a]	M1a	2
M1b	M1a	R1	R1	2
M1b	M1a	M1a	M1a	13

TABLE 31 Grading pathways for first screening episodes from 20,258 patients attending the Homerton DESP 1 June 2012 and 4 November 2013 (*continued*)

Level 1 grader	Level 2 grader	Arbitration grader	Final grade	Total number of patients
M1b	M1a	M1b	M1b	7
M1b	M1a	R2	R2	3
M1b	M1b	Missing[a]	M1b	260
M1b	R2	Missing[a]	R2	12
M1b	R2	R2	R2	1
M1b	R3	Missing[a]	R3	9
R2	R0	Missing[a]	R0	2
R2	R0	R0	R0	5
R2	R1	Missing[a]	R1	2
R2	R1	R1	R1	27
R2	R1	R2	R2	3
R2	U	Missing[a]	U	2
R2	M1a	M1a	M1a	8
R2	M1a	M1b	M1b	1
R2	M1a	R2	R2	3
R2	M1b	Missing[a]	M1b	18
R2	R2	Missing[a]	R2	525
R2	R2	R1	R1	1
R2	R2	R2	R2	21
R2	R2	R3	R3	1
R2	R3	Missing[a]	R3	27
R2	R3	R2	R2	1
R3	R1	R1	R1	1
R3	U	Missing[a]	U	2
R3	M1a	Missing[a]	M1a	1
R3	M1a	R3	R3	1
R3	M1b	Missing[a]	M1b	4
R3	R2	Missing[a]	R2	17
R3	R2	R2	R2	2
R3	R2	R3	R3	1
R3	R3	Missing[a]	R3	187
Total	–	–	–	20,258

a Missing in the context of this table means grading was not performed as not required by the pathway. For each grader the highest retinopathy grade for the screening episode is given.

TABLE 32 Screening performance of EyeArt software compared with manual grade prior to arbitration

Manual grade (worst eye)	EyeArt outcome, *n* (%[a])		Total, *n* (%[b])
	No disease	Disease	
Retinopathy grades			
R0M0	2426 (19)	10,301 (81)	12,727 (63)
R1M0	388 (8)	4361 (92)	4749 (23)
U	70 (23)	230 (77)	300 (1)
R1M1	47 (3)	1562 (97)	1609 (8)
R2	2 (0)	635 (100)	637 (3)
R2M0	2 (1)	208 (99)	210 (1)
R2M1	0 (0)	427 (100)	427 (2)
R3	2 (1)	234 (99)	236 (1)
R3M0	2 (3)	73 (97)	75 (0)
R3M1	0 (0)	161 (100)	161 (1)
Combination of grades			
R0M0, R1M0	2814 (16)	14,662 (84)	17,476 (86)
U, R1M1, R2, R3	121 (4)	2661 (96)	2782 (14)
R1M0, U, R1M1, R2, R3	509 (7)	7022 (93)	7531 (37)
Total	2935	17,323	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.

TABLE 33 Screening performance of EyeArt software compared with manual grade prior to arbitration: 95% confidence limits and likelihood ratios

Manual grade (worst eye)	Proportion classified as disease present or technical failure			Likelihood ratio vs. R0 (95% CI)
	Estimate (95% CI)	Lower[a] (95%CI)	Upper[b] (95%CI)	
Retinopathy grades				
R0M0[c]	0.191 (0.184 to 0.198)	0.184 (0.177 to 0.190)	0.198 (0.191 to 0.204)	–
R1M0	0.918 (0.910 to 0.926)	0.910 (0.903 to 0.920)	0.926 (0.920 to 0.934)	1.135 (1.122 to 1.146)
U	0.767 (0.715 to 0.811)	0.715 (0.675 to 0.783)	0.811 (0.777 to 0.864)	0.947 (0.900 to 1.018)
R1M1	0.971 (0.961 to 0.978)	0.961 (0.949 to 0.969)	0.978 (0.968 to 0.984)	1.199 (1.185 to 1.211)
R2	0.997 (0.988 to 0.999)	0.988 (0.845 to 0.989)	0.999 (0.855 to 1.000)	1.232 (1.000 to 1.244)
R2M0	0.990 (0.963 to 0.998)	0.963 (0.849 to 0.971)	0.998 (0.858 to 0.999)	1.224 (1.000 to 1.236)
R2M1	1.000	–	–	1.236 (1.225 to 1.245)
R3	0.992 (0.967 to 0.998)	0.967 (0.851 to 0.974)	0.998 (0.860 to 0.999)	1.225 (1.000 to 1.238)
R3M0	0.973 (0.900 to 0.993)	0.900 (0.851 to 0.932)	0.993 (0.861 to 0.999)	1.203 (1.000 to 1.228)
R3M1	1.000	–	–	1.236 (1.225 to 1.245)
Combination of grades				
R0M0, R1M0[c]	0.161 (0.156 to 0.167)	0.156 (0.150 to 0.161)	0.167 (0.160 to 0.172)	–
U, R1M1, R2, R3	0.957 (0.948 to 0.963)	0.948 (0.940 to 0.956)	0.963 (0.957 to 0.970)	1.140 (1.128 to 1.150)[d]
R1M0, U, R1M1, R2, R3	0.932 (0.927 to 0.938)	0.927 (0.921 to 0.934)	0.938 (0.933 to 0.944)	1.152 (1.142 to 1.164)

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion
c Estimates relate to the proportion classified as disease absent (i.e. the specificity).
d The likelihood ratio here is estimated compared with R0 and R1M0 combined.

TABLE 34 Screening performance of EyeArt software compared with manual grade in the worst eye prior to arbitration using EyeArt classification referable vs. non referable retinopathy

Manual grade (worst eye)	EyeArt outcome, *n* (%[a])		Total, *n* (%[b])
	No refer	Refer	
Retinopathy grades			
R0M0[c]	5098 (40)	7629 (60)	12,727 (63)
R1M0	1002 (21)	3747 (79)	4749 (23)
U	106 (35)	194 (65)	300 (1)
R1M1	126 (8)	1483 (92)	1609 (8)
R2	5 (1)	632 (99)	637 (3)
R2M0	4 (2)	206 (98)	210 (1)
R2M1	1 (0)	426 (100)	427 (2)
R3	3 (1)	233 (99)	236 (1)
R3M0	3 (4)	72 (96)	75 (0)
R3M1	0 (0)	161 (100)	161 (1)
Combination of grades			
R0M0, R1M0[c]	6100 (35)	11,376 (65)	17,476 (86)
U, R1M1, R2, R3	240 (9)	2542 (91)	2782 (14)
Total	6340	13,918	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.
c Estimates relate to the proportion classified as disease absent (i.e. the specificity).

TABLE 35 Screening performance of EyeArt software compared with manual grade in the worst eye prior to arbitration: 95% confidence limits and likelihood ratios using EyeArt classification referable vs. non referable retinopathy

Manual grade (worst eye)	Proportion classified as refer or technical failure			Likelihood ratio vs. R0 + R1M0 (95% CI)
	Estimate (95% CI)	Lower[a] (95% CI)	Upper[b] (95% CI)	
Retinopathy grades				
R0M0[c]	0.401 (0.392 to 0.409)	0.392 (0.383 to 0.399)	0.409 (0.400 to 0.416)	–
R1M0	0.789 (0.777 to 0.800)	0.777 (0.767 to 0.790)	0.800 (0.790 to 0.812)	–
U	0.647 (0.591 to 0.699)	0.591 (0.544 to 0.672)	0.699 (0.653 to 0.763)	0.993 (0.929 to 1.101)
R1M1	0.922 (0.908 to 0.934)	0.908 (0.896 to 0.922)	0.934 (0.923 to 0.946)	1.416 (1.390 to 1.441)
R2	0.992 (0.981 to 0.997)	0.981 (0.967 to 0.988)	0.997 (0.990 to 0.999)	1.524 (1.502 to 1.541)
R2M0	0.981 (0.950 to 0.993)	0.950 (0.913 to 0.967)	0.993 (0.974 to 0.999)	1.507 (1.459 to 1.533)
R2M1	0.998 (0.984 to 1.000)	0.984 (0.979 to 0.985)	1.000 (0.999 to 1.000)	1.533 (1.000 to 1.555)
R3	0.987 (0.961 to 0.996)	0.961 (0.683 to 0.973)	0.996 (0.696 to 0.999)	1.517 (1.000 to 1.542)
R3M0	0.960 (0.883 to 0.987)	0.883 (0.684 to 0.925)	0.987 (0.932 to 0.998)	1.475 (1.000 to 1.524)
R3M1	1.000	–	–	1.536 (1.518 to 1.549)
Combination of grades				
R0M0, R1M0[c]	0.349 (0.342 to 0.356)	0.342 (0.334 to 0.347)	0.356 (0.348 to 0.362)	–
U, R1M1, R2, R3	0.914 (0.903 to 0.924)	0.903 (0.891 to 0.915)	0.924 (0.913 to 0.934)	1.404 (1.383 to 1.424)

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion.
c Estimates relate to the proportion classified as disease absent (i.e. the specificity).

TABLE 36 Screening performance of Retmarker software compared with manual grade prior to arbitration

Manual grade (worst eye)	Retmarker outcome, *n* (%[a])		
	No disease	Disease	Total, *n* (%[b])
Retinopathy grades			
R0M0	6584 (52)	6143 (48)	12,727 (63)
R1M0	1733 (36)	3016 (64)	4749 (23)
U	131 (44)	169 (56)	300 (1)
R1M1	255 (16)	1354 (84)	1609 (8)
R2	31 (5)	606 (95)	637 (3)
R2M0	20 (10)	190 (90)	210 (1)
R2M1	11 (3)	416 (97)	427 (2)
R3	9 (4)	227 (96)	236 (1)
R3M0	6 (8)	69 (92)	75 (0)
R3M1	3 (2)	158 (98)	161 (1)
Combination of grades			
R0M0, R1M0	8317 (48)	9159 (52)	17,476 (86)
U, R1M1, R2, R3	426 (15)	2356 (85)	2782 (14)
R1M0, U, R1M1, R2, R3	2159 (29)	5372 (71)	7531 (37)
Total	8743	11,515	20,258 (100)

a Percentage within each manual grade.
b Percentage of the total number screened.

TABLE 37 Screening performance of Retmarker software compared with manual grade prior to arbitration: 95% confidence limits and likelihood ratios

Manual grade (worst eye)	Proportion classified as disease present or technical failure			Likelihood ratio vs. R0 (95% CI)
	Estimate (95% CI)	Lower[a] (95% CI)	Upper[b] (95% CI)	
Retinopathy grades				
R0M0[c]	0.517 (0.509 to 0.526)	0.509 (0.501 to 0.519)	0.526 (0.519 to 0.536)	–
R1M0	0.635 (0.621 to 0.649)	0.621 (0.609 to 0.635)	0.649 (0.637 to 0.663)	1.316 (1.285 to 1.358)
U	0.563 (0.507 to 0.618)	0.507 (0.449 to 0.567)	0.618 (0.567 to 0.678)	1.167 (1.051 to 1.299)
R1M1	0.842 (0.823 to 0.859)	0.823 (0.804 to 0.840)	0.859 (0.842 to 0.874)	1.743 (1.698 to 1.791)
R2	0.951 (0.932 to 0.966)	0.932 (0.914 to 0.958)	0.966 (0.953 to 0.983)	1.971 (1.925 to 2.027)
R2M0	0.905 (0.857 to 0.938)	0.857 (0.815 to 0.908)	0.938 (0.909 to 0.975)	1.874 (1.799 to 1.984)
R2M1	0.974 (0.954 to 0.986)	0.954 (0.934 to 0.972)	0.986 (0.974 to 0.995)	2.018 (1.969 to 2.067)
R3	0.962 (0.928 to 0.980)	0.928 (0.889 to 0.960)	0.980 (0.957 to 0.996)	1.993 (1.927 to 2.056)
R3M0	0.920 (0.833 to 0.964)	0.833 (0.724 to 0.891)	0.964 (0.902 to 0.993)	1.906 (1.740 to 2.026)
R3M1	0.981 (0.944 to 0.994)	0.944 (0.911 to 0.961)	0.994 (0.977 to 0.999)	2.033 (1.970 to 2.092)
Combination of grades				
R0M0, R1M0[c]	0.476 (0.469 to 0.483)	0.469 (0.462 to 0.477)	0.483 (0.477 to 0.492)	–
U, R1M1, R2, R3	0.847 (0.833 to 0.860)	0.833 (0.816 to 0.846)	0.860 (0.844 to 0.872)	1.616 (1.581 to 1.651)[d]
R1M0, U, R1M1, R2, R3	0.713 (0.703 to 0.723)	0.703 (0.692 to 0.715)	0.723 (0.712 to 0.735)	1.478 (1.442 to 1.516)

a Lower limit of 95% CI of estimated proportion.
b Upper limit of 95% CI of estimated proportion.
c Estimates relate to the proportion classified as disease absent (i.e. the specificity).
d The likelihood ratio here is estimated compared with R0M0 and R1M0 combined.

TABLE 38 Screening performance for EyeArt (manual grades modified by arbitration) by camera type

	EyeArt outcome, *n* (%[a])											
	Canon 2CR		Canon 2CR-DGi		Canon EOS		Canon EOS2		TOPCONnect		Unknown	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades												
R0M0	6 (2.6)	224 (97.4)	482 (10.1)	4274 (89.9)	168 (22.3)	584 (77.7)	1029 (36.6)	1783 (63.4)	245 (12.3)	1749 (87.7)	612 (27.2)	1639 (72.8)
R1M0	2 (2.2)	89 (97.8)	39 (2.1)	1833 (97.9)	11 (3.7)	285 (96.3)	106 (12.0)	778 (88.0)	24 (3.1)	745 (96.9)	35 (5.0)	671 (95.0)
U	0 (0.0)	14 (100.0)	17 (14.5)	100 (85.5)	6 (28.6)	15 (71.4)	36 (28.6)	90 (71.4)	23 (23.5)	75 (76.5)	16 (31.4)	35 (68.6)
R1M1	1 (2.7)	36 (97.3)	8 (1.5)	525 (98.5)	2 (2.5)	77 (97.5)	32 (11.9)	236 (88.1)	8 (2.1)	374 (97.9)	22 (8.5)	237 (91.5)
R2	0 (0.0)	33 (100.0)	1 (0.5)	215 (99.5)	0 (0.0)	54 (100.0)	3 (3.4)	85 (96.6)	0 (0.0)	141 (100.0)	0 (0.0)	94 (100.0)
R2M0	0 (0.0)	11 (100.0)	1 (1.3)	74 (98.7)	0 (0.0)	10 (100.0)	2 (6.9)	27 (93.1)	0 (0.0)	44 (100.0)	0 (0.0)	24 (100.0)
R2M1	0 (0.0)	22 (100.0)	0 (0.0)	141 (100.0)	0 (0.0)	44 (100.0)	1 (1.7)	58 (98.3)	0 (0.0)	97 (100.0)	0 (0.0)	70 (100.0)
R3	0 (0.0)	10 (100.0)	0 (0.0)	75 (100.0)	0 (0.0)	28 (100.0)	1 (2.6)	37 (97.4)	0 (0.0)	48 (100.0)	0 (0.0)	34 (100.0)
R3M0	0 (0.0)	4 (100.0)	0 (0.0)	21 (100.0)	0 (0.0)	12 (100.0)	0 (0.0)	9 (100.0)	0 (0.0)	14 (100.0)	0 (0.0)	11 (100.0)
R3M1	0 (0.0)	6 (100.0)	0 (0.0)	54 (100.0)	0 (0.0)	16 (100.0)	1 (3.4)	28 (96.6)	0 (0.0)	34 (100.0)	0 (0.0)	23 (100.0)
Combination of grades												
R0M0, R1M0	8 (2.5)	313 (97.5)	521 (7.9)	6107 (92.1)	179 (17.1)	869 (82.9)	1135 (30.7)	2561 (69.3)	269 (9.7)	2494 (90.3)	647 (21.9)	2310 (78.1)
U, R1M1, R2, R3	1 (1.1)	93 (98.9)	26 (2.8)	915 (97.2)	8 (4.4)	174 (95.6)	72 (13.8)	448 (86.2)	31 (4.6)	638 (95.4)	38 (8.7)	400 (91.3)
R1M0, U, R1M1, R2, R3	3 (1.6)	182 (98.4)	65 (2.3)	2748 (97.7)	19 (4.0)	459 (96.0)	178 (12.7)	1226 (87.3)	55 (3.8)	1383 (96.2)	73 (6.4)	1071 (93.6)
Total	9 (2.2)	406 (97.8)	547 (7.2)	7022 (92.8)	187 (15.2)	1043 (84.8)	1207 (28.6)	3009 (71.4)	300 (8.7)	3132 (91.3)	685 (20.2)	2710 (79.8)

a Percentage per manual grade within camera type.

TABLE 39 Screening performance for Retmarker (manual grades modified by arbitration) by camera type

	Retmarker outcome, *n* (%[a])											
	Canon 2CR		Canon 2CR-DGi		Canon EOS		Canon EOS2		TOPCONnect		Unknown	
Manual grade (worst eye)	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease	No disease	Disease
Retinopathy grades												
R0M0	48 (20.9)	182 (79.1)	2085 (43.8)	2671 (56.2)	579 (77.0)	173 (23.0)	1736 (61.7)	1076 (38.3)	529 (26.5)	1465 (73.5)	1753 (77.9)	498 (22.1)
R1M0	15 (16.5)	76 (83.5)	538 (28.7)	1334 (71.3)	148 (50.0)	148 (50.0)	346 (39.1)	538 (60.9)	139 (18.1)	630 (81.9)	399 (56.5)	307 (43.5)
U	2 (14.3)	12 (85.7)	41 (35.0)	76 (65.0)	12 (57.1)	9 (42.9)	79 (62.7)	47 (37.3)	29 (29.6)	69 (70.4)	31 (60.8)	20 (39.2)
R1M1	1 (2.7)	36 (97.3)	46 (8.6)	487 (91.4)	13 (16.5)	66 (83.5)	63 (23.5)	205 (76.5)	23 (6.0)	359 (94.0)	61 (23.6)	198 (76.4)
R2	0 (5.0)	33 (95.0)	7 (8.5)	209 (91.5)	2 (0.0)	52 (100.0)	6 (0.0)	82 (100.0)	5 (16.5)	136 (83.5)	2 (2.7)	92 (97.3)
R2M0	0 (0.0)	11 (100.0)	2 (2.7)	73 (97.3)	0 (0.0)	10 (100.0)	3 (10.3)	26 (89.7)	0 (0.0)	44 (100.0)	0 (0.0)	24 (100.0)
R2M1	0 (0.0)	22 (100.0)	5 (3.5)	136 (96.5)	2 (4.5)	42 (95.5)	3 (5.1)	56 (94.9)	5 (5.2)	92 (94.8)	2 (2.9)	68 (97.1)
R3	0 (31.4)	10 (68.6)	1 (0.0)	74 (100.0)	0 (0.0)	28 (100.0)	1 (20.9)	37 (79.1)	2 (14.3)	46 (85.7)	1 (0.0)	33 (100.0)
R3M0	0 (0.0)	4 (100.0)	0 (0.0)	21 (100.0)	0 (0.0)	12 (100.0)	0 (0.0)	9 (100.0)	0 (0.0)	14 (100.0)	1 (9.1)	10 (90.9)
R3M1	0 (0.0)	6 (100.0)	1 (1.9)	53 (98.1)	0 (0.0)	16 (100.0)	1 (3.4)	28 (96.6)	2 (5.9)	32 (94.1)	0 (0.0)	23 (100.0)
Combination of grades												
R0M0, R1M0	63 (19.6)	258 (80.4)	2623 (39.6)	4005 (60.4)	727 (69.4)	321 (30.6)	2082 (56.3)	1614 (43.7)	668 (24.2)	2095 (75.8)	2152 (72.8)	805 (27.2)
U, R1M1, R2, R3	3 (3.2)	91 (96.8)	95 (10.1)	846 (89.9)	27 (14.8)	155 (85.2)	149 (28.7)	371 (71.3)	59 (8.8)	610 (91.2)	95 (21.7)	343 (78.3)
R1M0, U, R1M1, R2, R3	18 (9.7)	167 (90.3)	633 (22.5)	2180 (77.5)	175 (36.6)	303 (63.4)	495 (35.3)	909 (64.7)	198 (13.8)	1240 (86.2)	494 (43.2)	650 (56.8)
Total	66 (15.9)	349 (84.1)	2718 (35.9)	4851 (64.1)	754 (61.3)	476 (38.7)	2231 (52.9)	1985 (47.1)	727 (21.2)	2705 (78.8)	2247 (66.2)	1148 (33.8)

a Percentage per manual grade within camera type.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library