Detailed Project Description: Can automated Diabetic Retinopathy Image Assessment software replace one or more steps of manual imaging grading and is this cost-effective for the English National Screening Programme?

HTA Priority Area: 11/21 - Automated reading of retinal photography in diabetic eye screening

Lead applicant: Adnan Tufail

Co-applicants: Ms Catherine Egan, Dr Alicja Rudnicka, Dr Christopher Owen, Ms Clare Bailey, Dr Caroline Rudisill, Dr Paul Taylor

Project Description Version 3.0

Version History

Project Description Version & date	Comment
Version 1.0	Approved with original submission to HTA
Version 2.0 Dec 2013	Circulated for comment
Version 3.0 March 2014	Approved by Steering committee March 2014, protocol generated directly from this project description Version 3.0 March 2014

Table of Contents

1. Project title: Can automated Diabetic Retinopathy Image Assessment software replace one or more steps of manual imaging grading and is this cost-effective for the English National Screening Programme?

2. How the project has changed since the outline proposal was submitted

Since the outline proposal was submitted, one of our study team (CE) published a manuscript exploring the role of Optical Coherence Tomography (OCT) in screening for Diabetic macular oedema (DMO).[1] DMO is an important cause of vision loss. England has a national systematic photographic retinal screening programme to identify patients with diabetic eye disease. Grading retinal photographs according to this national protocol identifies surrogate markers for DMO, as the colour photographs taken are non-stereo. OCT is an established, well-tolerated, non-invasive, non-contact method of imaging the macula. It is a more accurate and objective method of diagnosing macular oedema than clinical examination, even by experts. [2] Furthermore, OCT imaging is an established clinical trial endpoint for new treatments for diabetes mellitus (DM).

This OCT study showed that, in an unselected screening population of 17, 551 patients with diabetes mellitus (type 1 and type 2) derived from an urban inner city environment (central London, England) 311 were evaluated by standard photographic grading as having mild to moderate non-proliferative diabetic retinopathy (R1) in association with photographic surrogate markers of diabetic maculopathy (M1). Of these, only 38.3% had OCT evidence of macular oedema and required examination by an ophthalmologist to evaluate the need for treatment. We believe that OCT is a useful adjunct to traditional photographic retinal screening and warrants further evaluation, particularly with respect to the role of OCT in screening pathways and the health economics of this introduction. As OCT metrics can be automatically generated by the software attached to the OCT machine we would welcome the opportunity to discuss with the HTA board adding automated OCT analysis to this project. As OCT is beyond the remit of the current proposal we have not included it in the methodology or costing, but feel this could be added into the study at low additional cost without extending the proposed timelines and would potentially future proof the study. In addition our published audit suggest adding an automated OCT pathway would likely enhance the cost-effectiveness of automated screening further by reducing false positive referrals.

3. Planned investigation

The proposed study is an observational study to quantify screening performance, diagnostic accuracy and cost-effectiveness of automated primary grading, in routine English national screening programme for diabetic retinopathy. Four automated software programmes will be assessed. Study phase 1 is retrospective: the automated software packages will grade 24,000 screening episodes, which have been manually graded, from patients attending a diabetic retinopathy (DR) screening programme. Phase 2a is a prospective implementation study of an additional 6,000 screening episodes from the same clinic. Manual grading is based on 2+-field-2/3 step, by human graders who meet English National Screening Programme for Diabetic Retinopathy (ENSPDR) quality assurance standards. Phase 2b is an exploratory analysis based on outcomes of phase one – to explore whether improved outcomes can be obtained by the software if images with metadata tagging of the image location is used.

4. Research objectives

1. What is the estimated screening performance (sensitivity (detection rate), specificity (false positive rate) and likelihood ratios) and estimated precision for different grades of diabetic retinopathy, for each automated software grading system compared with 2+ field 2/3 step manual grading of retinal images?
2. What is the impact on screening performance associated with changing the default cut-offs (if allowable in the software package) used by each automated software grading system for the detection of different grades of retinopathy?

3. What is the clinical and cost effectiveness of implementing automated software grading to identify high-risk groups for manual grading as part of an English national screening programme for diabetic retinopathy?
4. How can automated grading be integrated into existing screening pathways and infrastructure?
5. Does automated grading assign more images as ungradable or a different grade compared with manual graders in particular demographic groups (e.g. ethnic group) or ocular co-morbidities e.g. cataract?

5. Existing research

Diabetes mellitus causes damage to small blood vessels in the retina. This causes vision loss by the following mechanisms: loss of blood supply (ischaemia), leaking blood vessels (macular oedema) and abnormal growth of blood vessels (haemorrhage/retinal detachment). In most cases of diabetic retinopathy (DR), blindness can be prevented if identified early and treated appropriately.

Early signs of the disease can be seen on photographs of the retina and graders are trained to identify these lesions and classify them according to the risk of blindness. Referral pathways to community screening clinics or hospital eye services are based on these grades. If the threshold for referral is set very low, then many patients will suffer needless anxiety and hospital eye services are flooded with patients at a low risk of vision loss, increasing the chance that patients with more severe disease are not treated promptly. However, if the threshold is set too high, then more patients will suffer preventable vision loss. This balance is managed in different ways throughout the world according to local health economics and the availability of specialist doctors.

In England, the National Screening Committee (NSC) launched a programme in 2003 to reduce visual loss due to DR by offering photographic screening to all diabetics. In the United Kingdom, 1.7 million people with diabetes were screened and manually graded for DR in 2007 and 2008.[3] A recent survey [4] reported that many programmes are not resourced to screen all eligible people with diabetes. There is also evidence that the repeatability and sensitivity of manual grading is poor, although much better at higher (sight-threatening) levels of disease severity.

The use of computers to automate at least part of the grading process could be extremely helpful. For example, if low risk patients could be identified automatically, limited human resources could be concentrated on the high risk groups with more severe DR. The Scottish National screening programme allows one type of automated screening software but has a different photography protocol to England and a different retinopathy grade definition (http://www.ndrs.scot.nhs.uk/). Since the introduction of the Scottish programme, other programmes have become commercially viable and international interest in automated diabetic screening has increased due to worldwide rapid increase in the numbers of people diagnosed with diabetes.

Detection programs analyse retinal colour images obtained by fundus cameras from people with diabetes and triage those who have sight threatening DR or other retinal abnormalities from those who can be screened again in one year. The diagnostic accuracy of computer detection programs has been reported to be comparable to that of specialists and expert readers. [5-11]

The external validity of such studies to date is limited because, commonly, images are not publicly available for comparison; non-validated reading protocols were used; and detection programmes were developed on images and populations highly similar to the one they are tested on, including distribution of retinopathy

severity, camera type, image quality,/resolution, field of view and number of images per eye. The Scottish Programme uses one field per eye, whereas 2 fields per eye are used in English Screening Programme. The limited external validity of studies to date may be in part due to the fact that the majority of studies to date 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.

The other problem with studies to date is the low rate of severe levels of DR in the test population. This means that most studies are inadequately powered to show efficacy of the automated software in identification of severe DR. For example, if proliferative DR is rare or absent in the test population, a detection program unable to detect proliferative DR will seem to perform well in that population. But it may miss patients that require immediate treatment, in a different population with a higher number of people with proliferative disease or with a different ethnicity or age profile.

Over the last decade, many computer image analysis methods based on image processing and machine learning have been proposed to interpret digital photographs of the retina to increase the efficiency of early detection of DR. [12-17]Few of these methods have been assessed in a sufficiently large number of patients with diabetes representative of the potential screening population, [5;10] that would also include enough events of the more severe form of retinopathy (R3 or proliferative disease) for evaluation using a fixed algorithm and compared against an established human standard (ENSP NSC standards) using standardised 2 fields images acquired to NSC imaging standards. Only 4 of these programs are commercially available, or have a CE mark or are close to obtaining a CE mark. There is a lack of robust evidence to evaluate the programs independently on the same test set that evaluate whether one or more of these programs meet or exceed the standards set for human graders on imaging compliant with ENSP imaging standards. Even if these programs meet the standard for screening for DR it is important to know how the programs handle poor quality images and non-DR pathology, how cost effective the implementation is and to develop technical standards for implementation into screening programmes.

6. Research methods

6.1 Overview of methods and implementation of outcomes

This study team, who are independent of any commercial links to automated diabetic screening software, will assess screening performance and cost-effectiveness of 4 software packages (CE) marked pending), compared to the English National Screening Committee (NSC) current standard of 2+ field 2/3 step manual grading. The study population will be patients from English diabetic eye screening service compliant with NSC standards: chosen for its demographic/ethnic mix and infrastructure. The study will examine performance and cost-effectiveness of the software packages using the default cut-offs (per CE mark application) and also explore the impact of changing these cut-offs on cost effectiveness. We will assess and develop standards for software integration into existing NSC/NHS IT software systems. Recommendations for the introduction of automated screening will be made, based on assessment of efficacy and cost effectiveness.

This observational study is based on 30,000 screening episodes that have already been graded manually for 24,000 patients attending a national screening programme for diabetic retinopathy. All 30,000 screening episodes will be re-analysed by each automated software. This retrospective re-analysis of screening episodes will not affect patient health outcomes which by design are non-modifiable. The study design is a measurement comparison study based on a health outcome measure that is concerned with the identification of referable diabetic retinopathy. The health economics component of the work will evaluate

the cost-effectiveness of using automated software to either replace one or more steps of manual image grading or to act as a filter to reduce the number of images requiring manual grading.

The prospective phase of the study is an implementation study to develop standard operating procedures, implementation guidelines, and clinical pathways e.g. whether images should be analysed in a store forward role or by real time grading. Our study may find that more than one program is cost effective and equal or superior to human graders (the defined standard for the primary outcome), which may be more financially and technically beneficial for the NHS in the longer term than a single provider in a developing field. We envisage that the information from this study will be provided to the committee of the ENSP NSC for Diabetic Retinopathy to help inform policy decisions on whether one or more software for automated screening should be integrated into the current screening programme and to provide technical standard operating procedures for how this may be achieved.

6.2 Pilot Studies

One of the software packages in the current proposal (iGrading, Medalytix) was used in a pilot study of 1,340 patient screening episodes at St George's Hospital NHS Trust, London, (3 of the applicants were part of this pilot study). Sensitivity was 100% for R2 and for R3 and 91% for M1 comparing this automated software package with manual grading as the reference standard. These data have been used as the basis of the power calculations for this study. A second study is underway, funded by Fight for Sight, to expand the pilot data to test more than one program, and to develop the method to extract images from Digital Healthcare Software (one of the NSC specified software tools for DR screening) and feed them into the servers running the automated programs. This second pilot study will inform the technical aspects of the proposed project and help develop technical standard operating procedures for possible implementation.

6.3 Data Acquisition

24,000 patient screening episodes in total, attending a London PCT screening programme (that meet ENSPDR quality assurance standards) with a wide spectrum of ages and ethnicity will provide the dataset for the evaluation.

Retinal photography

Camera systems used to capture the images must meet minimum NSC requirements. The photographic acquisition protocol is that specified by the NSC. All image capture centres in this study will upload their images to either Orion or Digital Healthcare software (the standard image and workflow software as recommended by the NSC).

Visual acuity

Visual acuity will be used as per the standard measurement acquisition protocol and chart used by each individual screening programme following criteria specified by the NSC; including external quality assurance (EQA) visits this year meet NSC requirements. As visual acuity below a certain level is used as a criterion for referral from a DR screening programme, the impact of this will be evaluated in conjunction with the outputs of the automated softwares.

Health economics

We will collect resource use data for each type of screening strategy (e.g. equipment including overhead and labour-related issues such as time, professional grade and training). Resource use data will then have unit costs attached either through PSSRU unit costs or hospital-specific costs. We will also estimate resource use and then costs for subsequent treatments of patients after they have been screened in our longer term model using cost per QALY as the outcome measure. These will include any procedure, medicines and doctor visit costs after screening.

6.4 Automated grading systems being used

Commercially available software with CE mark, or likely to be CE markable within 6 months of the contractual start of study include: iGrading (Medalytix), Retmarker (Critical-Health), EyeCheck (IDx), EyeArt (EyeNUK). All investigators are independent of commercial software development and have no commercial interest or IP in automated DR screening. Uniquely, we already have access to and support for all 4 software packages for the study, and exclusive access to the IDx software for evaluation in the UK. Any automated software meeting the CE mark standard within 6 months of the contractual start of this study (13 June 2013) that will become available will be considered for inclusion. No additional software will be evaluated thereafter. Additionally, there will be no change in the screening or treatment pathway of individuals.

6.5 Methods to protect against sources of bias

External validity studies of evaluating automating grading of DR to date are limited because, commonly, images are not publicly available for comparison; non-validated reading protocols were used; and detection programmes 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, number of images per eye. With regards to fields of view, the Scottish Programme uses one field per eye, whereas 2 fields per eye are used in English Screening Programme. The limited external validity of studies to date may be in part due to the fact that the majority of studies to date 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. To address these potentials for bias, this study team, who are independent of any commercial links to automated diabetic screening software, will assess screening performance and cost-effectiveness of 4 software packages (CE marked pending), compared to the English National Screening Committee (NSC) current standard of 2+ field 2/3 step manual grading.

Disagreements between the human grader and one or more automated systems will be arbitrated on a subset of images from the 24,000 retrospective screening episodes by an internationally recognised fundus photographic reading centre, masked to the original grading. This gives an independent assessment of the relative performance of manual and automated systems, and provides external validity to the project. Emphasis will be placed on arbitrating discrepancies with manual grades R3, R2, M1 and a proportion of R1 and R0 grades. Technical failure rates for automated and manual grading will be compared.

All four programs being assessed are not learning tools. The cut-offs are fixed which allows for potential FDA approval/CE marking and this fixed cut-off will be used in the primary outcome of disease/no disease. However some of the software have the potential to alter the threshold cut-off (not by altering the algorithm or using a learning set) to potentially separate the R1/M0 grades (that would not require referral) from more severe grade of retinopathy that would potentially make this software implementation more cost effective. This additional analysis is to address point 6 of the question set by the HTA.

6.6 Planned inclusion/exclusion criteria

Inclusion criteria

As defined by the NSC, individuals attending a DR screening programme for the NHS (http://diabeticeye.screening.nhs.uk access date July 2012), which includes

1. Patients aged 12 years and older;
2. With diabetes; and
3. Attending annual screening at the designated programme site and has the appropriate ethics and Caldicott Guardian approvals in place. London site was chosen to encompass a wide age range of

diabetics of different ethnic origins.

Exclusion criteria

1. Consent for research use of images and data not given by the Calidicott Guardian or equivalent or individual permission stored on the Screening Software.
2. Excluding those who do not have perception of light in both eyes.

6.7 Ethical arrangements

Ethics committee and research governance approval will be sought before initiating the study.

Data management and transfer

Patient identifiable information will be removed from each screening episode image and a new unique identifier will be created before data are sent for automated grading or arbitration. Data will be extracted using the batch extraction tool created by Digital Healthcare. Automated grading will take place blind, i.e. without knowledge of the final manual grade. The statistical team at St George's will be responsible for creating a "unique anonymous identifier" for each screening episode which will allow images to be processed by the programs without any patient identifiable information and masked to the original grade allocated by human graders. Data from the programs can then be linked back to the results from the original screening episode at St George's thereby allowing the data from all sources to be merged and compared. Database design and management will be undertaken by the St George's statistical team.

Arbitration images will be sent to Doheny Image Reading Center, Los Angeles, USA. Ethics and Research Governance permission will be sought to be able to send these anonymised images outside of the EU. Images will be transferred to Doheny Reading Center via a secure web based upload service which will be password protected. No data regarding the outcome of manual or automated grading will be sent to the Doheny Image Reading Centre. All data linkage will take place at the 'main centre' where the outcome of manual and automated grading for each screening episode will be linked to the grade obtained from the Doheny Image Reading Centre. Images will be graded according to both NSC and ETDRS grading protocols.

Only patient images and anonymized data will be used in this study and only if appropriate consent has been obtained from the patient, and subject to ethics committee approval. As part of the NSC specified software and process patients are asked at first screening whether they are happy to have their clinical data used for research purposes (see screen shot below) this is recorded and searchable in the database.

6.8 Research Governance

Research activities at each of the participating centres will be carried out in accordance with the DOH research governance framework for Health and Social Care. The project will be registered with Moorfields Eye Hospital NHS Foundation Trust and each appropriate NHS Trust. Records relating to all procedures carried out will be kept for 5 years so that the research process is clearly understandable and repeatable. This study proposal has been reviewed by the Moorfields Eye Hospital Research Governance Panel and they have agreed to act as Sponsors of the Study.

6.9 Proposed sample size

Number of patients and centres

The retrospective study (n=24,000 screening episodes) will look at screening episodes collected 12 months prior to the start of the grant that attended a London PCT diabetic retinopathy screening programme encompassing a wide spectrum of ages and ethnicity (24% South Asian, 17% black African/Carribean and 38% white European). The prospective study to evaluate integration of the systems in clinical use (n=6,000 screening episodes) will begin soon after the grant starts.

Recruitment Rate

As the first part of the study is retrospective and the programme assesses over 24,000 patients annually, recruitment will not be problematic. Patients who agree to research use of their images and data are tagged on the Digital Healthcare screening software, the patients with this research consent can be identified with the Digital Healthcare extraction tool.

Justification of numbers and assumptions underlying the power calculation

For each screening episode, retinopathy will be classified into the following mutually exclusive groups, depending on the highest level of severity observed: R0, R1, M1, R2 or R3. A pilot study of 1,340 patient screening episodes at St George's Hospital NHS Trust, London, (3 of the applicants were part of this pilot study), revealed that the prevalence of R0, R1, M1, R2 and R3 was 68%, 24%, 6.1%, 1.2% and 0.5%, respectively. Additionally, in this pilot study one software (Medalytix) was compared to manual grading as the reference standard. It was found that the sensitivity for R1, M1, R2 and R3 was 82%, 91%, 100% and 100%, respectively, and that 44% of R0 were graded as "disease present". A proportion of disagreements between one or more of the automated systems with the human grade (from all 24,000 screening episodes from the retrospective phase of this study) will be assessed by an independent internationally recognised fundus image reading centre.

The British Diabetic Association (BDA) set standards for the diabetic retinopathy screening programme of at least 80% sensitivity (BDA Report, 1997). The sample size calculation is based on the number of screening episodes to ensure the lower limit of the 95% CI for sensitivity of automated grading does not fall below 97% for retinopathy classified as R3 by human graders. R3 is the rarest and most serious outcome (0.5%) and governs the sample size needed. Under these assumptions, approximately 24,000 screening episodes are required to provide the desired diagnostic accuracy, however, should the sensitivity fall to 90%, the associated binomial exact 95% CI would range from 83% to 95% for R3. Hence, we would be more than 95% certain that the lower limit of this confidence interval would still comply with the BDA standard.

If the sensitivity of all automated systems fell to 90% for image sets graded as R3, R2 or M1 by the human grader, then for each automated software system there would be approximately 200 screening episodes in need of arbitration. A worst case scenario is that for referable retinopathy the 200 disagreements from each automated software are not concordant across all four automated systems. In which case, 4 x 200 (total 800) screening episodes graded referable retinopathy (human grades R3, R2 and M1) would require arbitration. However, it is unlikely that disagreement across automated systems will be mutually exclusive and potentially fewer screening episodes will require arbitration. By assuming independence amongst the 4 different software, a simulation study showed that the number of screening episodes graded as referable retinopathy by the human grader for which at least one software would suggest no referral, would be 660 (95%CI 619 to 701) image sets. It is likely that the image process algorithms of the different software will bear similarities and, therefore, fewer screening episodes will require arbitration. Another point for consideration is that our pilot work was based on just one of the automated programmes. We wish to allow for the fact that the screening performance for the other software may be worse. Allowing for 700 screening episodes graded as *referable* for arbitration should give sufficient coverage. When images are sent for arbitration, the reading centre is assumed to identify the 'true' grade. (See Table 1)

Resources are in place for a total of 1700 screening episodes to be sent for arbitration. To maximise efficiency regarding the cost of arbitration we suggest using the following approach. The majority of screening episodes with disagreement between the human grade and the automated systems is expected to be for retinopathy grades R1 and R0 with no maculopathy; under current guidelines these groups would be recalled for screening annually. Based on data from the pilot study we expect that for each of the automated systems approximately 18% of screening episodes graded R1 by the human grader will be missed by the automated system, and 44% of episodes graded R0 by the human grader will be graded as "disease present" by the automated system. When two or more software disagree with the human grade, a random sample of screening episodes will be sent for arbitration. Table 1 summarises the process of arbitration and our approach for establishing the grade that will serve as the gold standard when evaluating the performance of

each automated system. The percentage of screening episodes falling in cases 3 and 4 (see Table 1) that will be sent for arbitration will be determined by the availability of funds. For example, under the assumption of independence amongst the different software, we would expect 9672 and 868 (total 10,540) screening episodes in cases 3 and 4, respectively. Hence, with 660 episodes for case 2 (as outlined above) we could send an additional 1040 (i.e. 1700-660) screening episodes to the reading centre for arbitration. This represents approximately 10% of screening episodes in cases 3 or 4.

Technical failures are considered as 'referable' in the clinical pathways. Therefore sensitivity and the specificity of each automated system will be determined in two ways, by inclusion and exclusion of technical failures. It will be important to compare technical failure rates for both manual and automated grading.

6.10 Statistical analysis

Primary analyses

Estimating the screening performance of the automated grading systems as compared with human grading in clinical practice

The sensitivity and false positive rate (1-specificity) of each automated grading system will be determined using the final manual grade from clinical records as the reference standard in the first instance. Sensitivity and false positive rate will be provided for referable retinopathy overall and for each grade of retinopathy separately, using data from all 30,000 screening episodes. As 6,000 of the 24,000 individuals in this study will have two screening episodes (a total of 30,000 screening episodes), sensitivity and false positive rate will be estimated using mixed-effects logistic regression models including a random effect for each participant to take account of the clustered nature of the data. The outcome in these models will be the binary result of each software (i.e. disease/no disease) whilst referable retinopathy ascertained from the final human grade will be used as the explanatory variable. Thus for each software these models will estimate the log-odds of an automated system suggesting that a disease is present conditional on the human grading. This will allow the estimation of sensitivity and false positive rate. The analysis will also include the estimation of positive and negative likelihood ratios for each automated software; for a positive test result (referable disease present) the positive likelihood ratio will be estimated in two ways, by inclusion and exclusion of technical failures. Screening episodes with images that cannot be processed by the automated grading systems will be combined with those suggesting that a disease is present, as this is the way technical failures are being treated in the clinical pathway. However, the effect of the exclusion of data from those episodes from the models will be investigated as complementary analysis.

Diagnostic accuracy of sensitivity, false positive rate and likelihood ratios will be further defined by 95% confidence intervals (CI) obtained by using the bootstrap method. Clinical interpretation of these measures will focus in particular on the lower limit of the 95%CI for the detection of proliferative (R3), pre-proliferative (R2) retinopathy grades and maculopathy (M1). This will give an indication of the number of screening episodes requiring clinical intervention that could be missed by the automated systems. The upper confidence limit for false positive rates for retinopathy grades R1 (background) and R0 (no retinopathy) will be important in assessing the additional number of screening episodes requiring further investigation. The level of uncertainty around these estimated lower and upper limits will be investigated using the bootstrap method.

Assessing the predictive value of the automated grading systems

The predictive value of the automated grading systems will be evaluated using data from the 24,000 independent screening episodes by fitting multinomial logistic regression models with the final human grade as outcome and the binary result of each software (i.e. disease/no disease) as the explanatory variable.

Alternative reference standards

Estimating the screening performance of the automated grading systems as compared with human grading in clinical practice refined by an internationally recognized fundus photographic reading centre

The agreement of the automated systems and manual grading against an internationally recognized fundus photographic reading centre; blind to the original classifications from manual and automated systems will be investigated. This will allow independent assessment of screening performance of manual and automated systems, and provide external validity. Arbitration will be undertaken only on a subset of the 24,000 screening episodes from the retrospective element of the study. Emphasis will be placed on arbitrating for all R3, R2, M1 grades and a proportion of R1 and R0 grades; as described in Table 1. Based on the output from the arbitration process we will re-define the reference standard in a subset of images (as outlined in Table 1) and quantify the change in the measures of screening performance defined in our primary analysis above.

Constructing a consensus grade

After completing the primary analysis on the 30,000 images we will be able to identify automated software that achieve an acceptable standard of performance (i.e. at least 80% sensitivity overall and 100% sensitivity for R3). We will evaluate a "consensus grade" approach to defining a reference standard and the impact this has on measures of screening performance. A consensus grade will be formed based on the human grade and/or all automated software outcome that achieve this standard of performance. Consensus will be defined in two ways:

1. A majority classification based on using data from the final manual grade and automated software outcome of disease present/absent.
2. A majority classification based on outcome (disease present/absent) from the automated software only.

We will then compare measures of screening performance from the "consensus grade approach" with that from our primary analysis (where the gold standard is influenced by reading centre results for some screening episodes as outlined in Table 1) using data from all 30,000 screening episodes.

For the 1700 images re-graded by the reader centre we will have a gold standard that is validated and we could evaluate the agreement of the consensus grade with this gold standard. This would constitute a very important component for future work as it would give an estimation of the magnitude and direction of any bias if a consensus grade approach were to be used as the gold standard.

Approaches to technical failure

Comparing the rate of technical failure between manual and automated grading systems

The rate of technical failure for automated and manual grading will be compared using logistic regression. The proportion (with 95%CI) of screening episodes identified by the automated system as disease-negative offset against technical failures will provide an estimate of the potential saving (both time and money) in the number of screening episodes requiring manual grading.

Inclusion of additional metadata

An optional phase 2b will be considered based on the outcomes of Phase 1/2a. The digitial healthcare diabetes screening programme management software (OptoMize), does not attach metadata identifying the image location or whether a non –retinal (lens) image is acquired. Lens shots will be identified as ungradable and therefore will result in making the software tested less cost effective than if lens shots were excluded. Another diabetes screening programme management software (HISVector) has metadata attached to the images and we will repeat the analysis on this alternate dataset looking at a cohort of patient followed up for at least 3 years. The advantage of using an additional NHS Diabetes Screening software is (i) to assess how the software perform on a different system and (ii) allows evaluation of whether use of metadata

attached to the images improves performance while still complying with the directive of the grant to see how these software perform in a current compliant NHS screening programme.

Investigating the influence of other factors (including potential confounders) in screening performance measures

Mixed-effects logistic regression models including patients' ethnicity (South Asians, black African-Caribbean and white Europeans), age and duration of diabetes as explanatory variables will be used to examine the effect of potential confounders on the screening performance of each automated system. In addition, multiple variable logistic regression will be used to explore the influence of ethnicity and age on the frequency of referable retinopathy and technical failure rate.

Health economic analyses

Assessing the impact of different operating points on sensitivity and false positive rate, and their effect on the cost-effectiveness of automated retinopathy screening

The argument for automated retinopathy screening is, ultimately, based on cost-effectiveness. The question is whether the automated screening of retinal images will identify a sufficiently large proportion as not requiring human intervention to justify a) the cost of the software and b) the risk of harm that necessarily follows from automation. In any such calculation there is a trade-off between sensitivity and specificity. Manufacturers of screening software will fix an 'operating point' that provides a trade-off for which regulatory approval can be obtained and a commercial case made for the system's effectiveness. However the chosen operating point may not be optimal for all screening programmes. We will attempt, for each system under study, to identify an optimal operating point for the English setting. Using ROC analysis we will tabulate sensitivity and false positive rate for a range of thresholds of the decision statistics provided by the software companies. We assume, for the purposes of this exercise, that each software package generates a number that reflects the probability of a given grade of retinopathy for that episode. Setting a threshold on this number then determines whether it is classified with that grade or not. We assume that the machine normally operates with a constant threshold and that this determines the sensitivity and false positive rate. This simulation will meet the HTA requirement (point 6 of the question set by the HTA) of exploring the consequences of using the software at different operating points.

The approach will use a Markov Model to simulate the progression of retinopathy through the different stages identified by graders and the identification of disease at screening by the envisaged combination of automated and human reading. Such a model consists of a) a set of states representing the appropriate stages of the disease and the outcomes of investigation and treatment and b) the probabilities of transitions between states. Some of these probabilities will be identified from the literature and some will be identified from our research. If each state can be associated with a value for quality of life and NHS costs, we can then run a simulation of a large screening cohort and measure the costs and benefits that result from a given set of transition probabilities. We can then vary the probabilities to simulate the effects of altering the sensitivity and specificity of screening in order to see what impact this has on the effectiveness of screening. We recognise that modelling disease progression in this way is an approximate business and that reliable and robust estimates of some transition probabilities will not be possible. We will therefore have to augment the analyses described above with some sensitivity analyses to explore the consequences for our conclusions of errors in such estimates.

Model outline

We will model the progress of a population reflecting the demographics of that invited for screening. The states of the model will reflect the different states of the disease. The transitions will model the possible changes of state, including those associated with the different possible outcomes of screening. There will be two versions of the model, one based on the processes involved in manual screening, and one for

automated screening. The manual screening model will reflect the pathway set out in "Public health functions to be exercised by NHS England Service specification No.22: NHS Diabetic Eye Screening Programme". In the automated screening model, automated screening will replace the first primary grading (7) of the two human grading steps. The secondary grader (11) would remain unless the result is R0/M0(8).

Transition probabilities for the model will reflect data on the performance of the automated screening, efficacy of manual screening (as in table 1 in Scotland et al. 2007), likelihood and timing of re-screening, referral rates, prevalence of different levels of retinopathy in the study population, rates of progression and regression of the disease. We will explore the extent to which the rates of disease progression can be estimated using multi-state misclassification model fitted with data from screening programme from which cases for the retrospective analysis were taken. This model takes account of the longitudinal nature of the data, whilst it allows for classification errors with respect to the diabetic retinopathy outcome. Using this statistical framework, we will estimate the probability of transition between the different levels of diabetic retinopathy screening outcomes over time, and we will examine whether these probabilities are being affected by ethnicity, age or duration of diabetes.

Cost data

We will collect resource use data for both screening strategies (e.g. cost of equipment including overhead and labour-related costs - time, professional grade and training). Resource use for subsequent patient treatments after the screening episode will be gathered from existing literature and expert opinion. Resource use data will then be combined with associated costs – taking account of implementation costs - using hospital cost data. Cost effectiveness will be evaluated based on the expected cost per true positive case of retinopathy and cost per QALY. Expected QALYs will not be collected in this study but calculated from the literature for patient outcomes with retinopathy at various stages (e.g. http://www.ajo.com/article/S0002-9394(99)00146-4/abstract). Sensitivity analyses will be performed for key variables.

Assessing software performance on repeated screening episodes on the same patient

In addition to the 24,000 screening episodes we will have accrued approximately 6000+ screening episodes with repeat follow-up data (His Vector & Digital Healthcare acquired images). It would be important to ascertain whether the automated software performs better with repeated screening episodes from the same patient, this additional analysis will only be undertaken if some or all of the softwares take into account the previous visits image in the analysis. We would quantify the screening performance of the automated software firstly by using the human grade as the gold standard and secondly by using a consensus approach informed by the bias we would have quantified above.

7. Publication Policy and Contribution to the Collective Research Effort

Publication decision and content will be by made by the research team named in this protocol and will be undertaken independent of the companies whose software is being tested.

Research findings will be disseminated to the wider research community at both national and international meetings of researchers in ophthalmology, diabetes and health economics. The investigators have considerable experience in presenting and publishing findings in high ranking general medical, and ophthalmology and health economics journals. An important aspect of the work will be the dissemination of the findings to NHS/NSC representatives. One of the stated aims of the English national screening programme is to reduce the anxiety associated with screening by providing timely notification of results. It is likely that further software will be developed to grade retinopathy in the future. The test set of 24,000 patients with manual grades, 4 automated grades as well as arbitration manual grading, will be made available as a standardised set for use by NHS/NIHR investigators looking at future innovations, such as that use of automated software to detect and assess changes in the appearance of the retina between screening visits.

8. Project timetable and milestones

Start Date – within 2 months of receipt of Grant and ethics approval for phase 1 of the study and completion within 6 months of this date. We feel this is achievable as we have undertaken .pilot work funded by a small Fight for Sight grant, and the data has already been gathered in the 12 month retrospective period prior to the start data.

Data analysis and draft manuscript for phase 1 of the study, will be within 4 months of this date, as the discrepant images will need to be evaluated in a reading centre before any analysis can be undertaken.

The prospective part of the study (phase 2) will take 12 months to gather the data and 6 months to analyse, i.e. a total of 18 months, with further publications form the phase 2 data being submitted within 4 months of the end of this date. Within 4 months of the end of the date additional work including developing technical operating procedures for implementation of software(s) into the ENSPDR will be drafted.

9. Expertise

Three members of this research group recently completed a pilot study to assess automated software. The methods in this proposal were implemented in the pilot study and informed this power calculation. We currently hold a small grant (£15,000) from Fight for Sight to expand the pilot study to evaluate 4 different software systems. This study has been initiated. We are therefore in a strong position, to deliver on a definitive study and meet the timelines stated to answer the larger questions required for this commissioned proposal.

The lead applicant, Adnan Tufail, is a disease theme lead at The Moorfields Eye Hospital/UCL NIHR BRC, has published on various retinal disease imaging metrics and endpoints, was previously an ophthalmology lead for diabetic screening and recently led a multicentre RCT which finished on time and on budget. Expertise on diabetic screening: Catherine Egan is lead for diabetic screening Moorfields Eye Hospital group of hospitals, member of the Diabetes Research Network Eye disease group, an elected member of the European Vision Institute Expert Panel for diabetic eye disease, and a Royal College of Ophthalmologists (RCOphth) advisor to the English National Screening Programme for diabetic retinopathy (ENSPDR) and Grading and Assessment subcommittee and a peer reviewer for the ENSPDR EQA team. Clare Bailey is the lead for diabetic screening at Bristol Eye Hospital, is the RCOphth advisor to the ENSPDR, and author on the current RCOphth guidelines for diabetic retinopathy. Health economics expertise: Dr Rudisill is based at the London School of Economics & Political Science and visiting lecturer in health economics at King's College London. She has experience with economic evaluation in the clinical area of diabetes and is supported by Prof Alistair McGuire (LSE) on the advisory panel. Image analysis expertise: Dr Paul Taylor, (Reader, Centre for Health Informatics and Multiprofessional Education, UCL) has expertise in automated image analysis and has previously been awarded a HTA grant in automated mammography screening. Statistical and epidemiological expertise: Dr Rudnicka (Senior Lecturer in Medical Statistics) and Dr Owen (Reader in Epidemiology), both at St George's, University of London. Pertinent to the current proposal, Dr Rudnicka has extensive experience of screening programs, (e.g. antenatal screening for Downs's syndrome), Dr Owen in cardiovascular epidemiology (including the life course determinants of diabetes) and retinal vascular imaging. Owen & Rudnicka have grants related to the epidemiology of diabetes and quantification of retinal vasculature from digital images.

10. The Steering Group

The Steering Group will include Ms I Stratton (Senior Statistician, Gloucester Diabetic Retinopathy Research Group), Mr R. Wormold, (Ophthalmologist, Editor of the Cochrane Eyes and Vision Group, Ophthalmic Epidemiology and evidence based practice), Prof M Jofre-Bonet (City University), Prof I Nazareth (Primary Care & Population Health, UCL, to advise on methodological aspects specific to analyses of data from clinical databases and interpretation of the results and consideration of issues specific to implementation of diabetic retinal screening within the community setting), Mr Steven Aldington (Retinopathy Grading and Screening,

Hon Assc. Prof Univ of Warwick), Prof Simon Jones (Clinical Informatics and Health Care Management, Univ of Surrey).

11. IT Advisory Board

The IT Advisory Board with declared commercial associations for technical support has already been established for the first phase of the study:

Christian Martin (formerly IT expert for the ENSPDR); representatives of NSC manual grading software companies): and each automated software company (D Ramos from Critical Health, Prof M Abramoff from University of Iowa, representative from Medalytix/Digital Healthcare, and Dr K Solanki from EyeNUK). Data extraction and analysis will be independent of this group. Note this board will help draft technical protocols for implementation and connectivity of these programs into screening systems.

12. Service Users

The board from the London Screening Programme, whose screening images are being assessed in this study that include a diabetic patient, a DR grader, a consultant endocrinologist (JA) and a public health physician are being invited advise on the best way to use the software to improve the experience of retinal screening.

13. Justification of support required

The projected total costs of the study of about £369,000 are comprised mainly of purchase and maintenance of computers, software and support costs to link these systems to existing NHS IT system. (About £182,000). The other main costs are that of the arbitration reading centre (up to £80,000) and salary costs (total about £93,000). This is a reduction in the amount requested in the original proposal of about £140,000 which we have achieved without compromising what we intend to do .

IT Costs: We have now shifted our main study site to one with a more modern software system (DH v3) that will allow for data extraction at a much reduced cost. The small Fight for Sight Grant we have to initiate work in this area has allowed us to already purchase some of the required hardware.

The IT budget will pay for SQL Server Licences x 3, High End PC Workstation x 1, Basic Network Servers x 3

Encrypted 500gb Hard-disks x 4, Digital Healthcare Review Station licences x 2, Digital Healthcare Exporter Module x 1, Medalytix Patient Processing Fees, Annual Software Licence fees

Reading Centre costs. These costs are based on the discrepant images that we would expect to have based on pilot data. By altering the structuring of the validation sampling we have reduced arbitration grading costs while maintaining the robustness of the outline proposal design.

Staffing costs- £46,312 to pay for time of the 5 applicants and £46,683 to pay for an IT support technician and statistician.

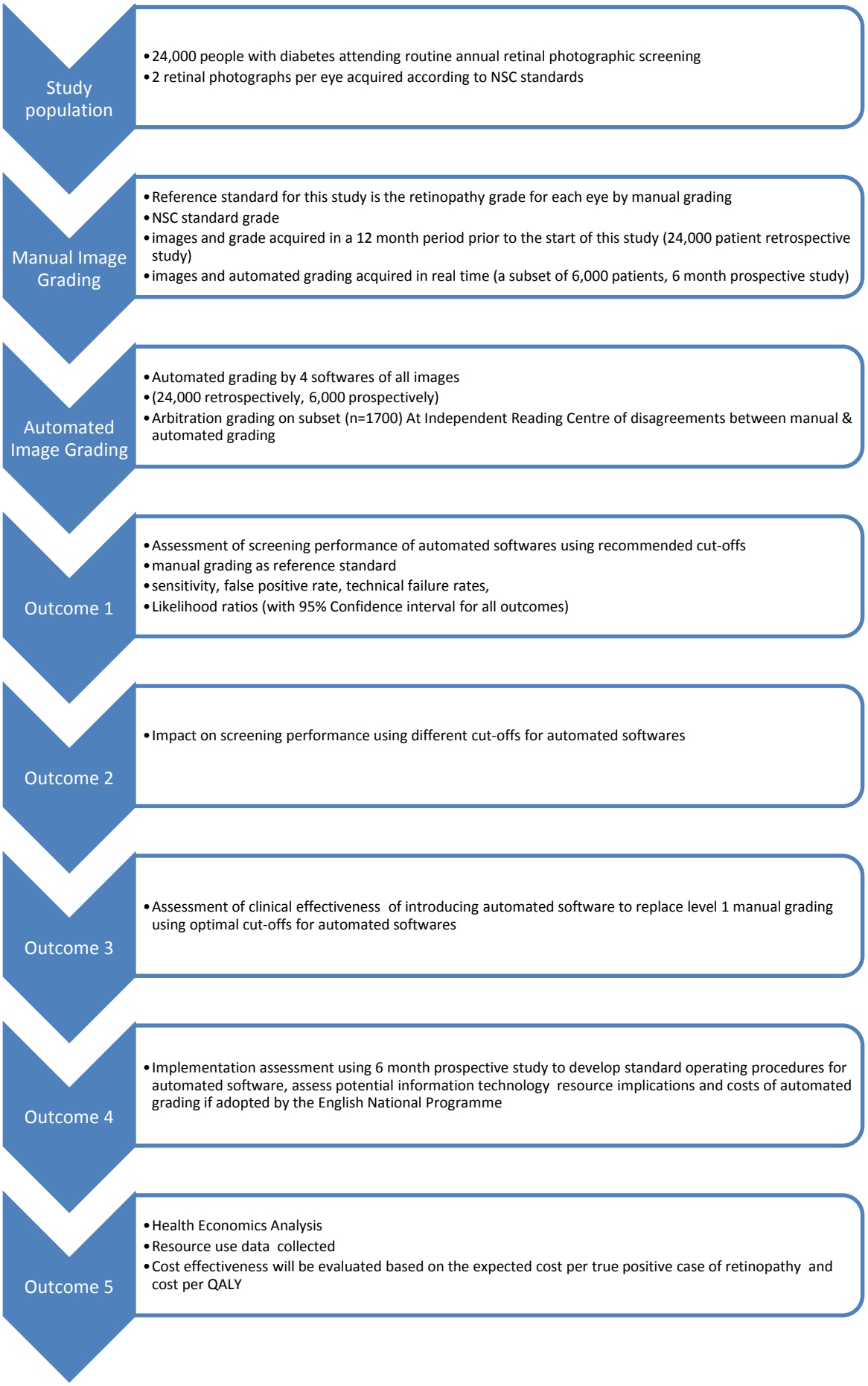
Table 1: Process of selection of discrepancies for external arbitration

Cases		Criteria		N (95%CI)†	Arbitration	Truth
1	IF	Human grade: All software:	M1, R2 or R3 Disease present	1259 (1202 to 1316)	No arbitration	Truth = Human grader
2	IF	Human grade: One or more software:	M1, R2 or R3 Disease not present	660 (619 to 701)	Send all images to reading centre	Truth = Reading centre
3	IF	Human grade: Two or more software:	R0 Disease present	9672 (9542 to 9796)	For each software send a proportion of all screening episodes falling in this case to reading centre (random selection)‡	Truth = Reading centre
4	IF	Human grade: Two or more software:	R1 Disease not present	868 (821 to 918)	For each software send a proportion of all screening episodes falling in this case to reading centre (random selection)‡	Truth = Reading centre
5	IF	Human grade: Only one software:	R0 Disease present	5044 (4940 to 5150)	No arbitration	Truth = Human grader
6	IF	Human grade: Only one software:	R1 Disease not present	2288 (2215 to 2359)	No arbitration	Truth = Human grader
7	IF	Human grade: All software:	R0 Disease not present	1605 (1539 to 1670)	No arbitration	Truth = Human grader
8	IF	Human grade: All software:	R1 Disease present	2603 (2526 to 2684)	No arbitration	Truth = Human grader
9	IF	Human grade: More than one software:	R0, R1, M1, R2 or R3 Technical failure		No arbitration	Truth = Human grader
10	IF	Human grade: Only one software:	R0, R1, M1, R2 or R3 Technical failure		Proceed as in cases 1-8 considering the grades of the remaining software	

† N = Median number of cases (95% CI) based on 1000 simulations, estimated by assuming that the prevalence of R0, R1, M1, R2 and R3 is 68%, 24%, 6.1%, 1.2% and 0.5%, respectively; that all four software are independent with 82% sensitivity for R1, and 90% sensitivity for M1, R2 and R3, and that for each software 44% of R0 will be graded as "disease present".

‡ Appropriate weighting will be used in the random selection process of screening episodes to give increasing preference to episodes for which more than one or all software agree with each other but disagree with the final human grade.

14. Flow diagram



15. References

(1) 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(5):e14811.

(2) Davis MD, Bressler SB, Aiello LP, Bressler NM, Browning DJ, Flaxel CJ et al. Comparison of time-domain OCT and fundus photographic assessments of retinal thickening in eyes with diabetic macular edema. *Invest Ophthalmol Vis Sci* 2008; 49(5):1745-1752.

(3) National Health Service. The English Diabetic Retinopathy Programme Annual Report, 1st April 2007 - 31st March 2008. 2008.

(4) Diabetes UK. State of diabetic care in the UK 2009. 2012.

(5) Abramoff MD, Niemeijer M, Suttorp-Schulten MS, Viergever MA, Russell SR, van GB. 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(2):193-198.

(6) Abramoff MD, Reinhardt JM, Russell SR, Folk JC, Mahajan VB, Niemeijer M et al. Automated early detection of diabetic retinopathy. *Ophthalmology* 2010; 117(6):1147-1154.

(7) 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(6):706-711.

(8) 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(11):1512-1517.

(9) 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(6):712-719.

(10) 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(11):1518-1523.

(11) Niemeijer M, van GB, 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(1):185-195.

(12) Teng T, Lefley M, Claremont D. Progress towards automated diabetic ocular screening: a review of image analysis and intelligent systems for diabetic retinopathy. *Med Biol Eng Comput* 2002; 40(1):2-13.

(13) Cree MJ, Olson JA, McHardy KC, Sharp PF, Forrester JV. A fully automated comparative microaneurysm digital detection system. *Eye (Lond)* 1997; 11 (Pt 5):622-628.

(14) Hipwell JH, Strachan F, Olson JA, McHardy KC, Sharp PF, Forrester JV. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. *Diabet Med* 2000; 17(8):588-594.

(15) Olson JA, Strachan FM, Hipwell JH, Goatman KA, McHardy KC, Forrester JV et al. A comparative evaluation of digital imaging, retinal photography and optometrist examination in screening for diabetic retinopathy. *Diabet Med* 2003; 20(7):528-534.

(16) Larsen N, Godt J, Grunkin M, Lund-Andersen H, Larsen M. Automated detection of diabetic retinopathy in a fundus photographic screening population. *Invest Ophthalmol Vis Sci* 2003; 44(2):767-771.

(17) Walter T, Klein JC, Massin P, Erginay A. A contribution of image processing to the diagnosis of diabetic retinopathy--detection of exudates in color fundus images of the human retina. *IEEE Trans Med Imaging* 2002; 21(10):1236-1243.