# Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy

G S Scotland,[1] P McNamee,[1] A D Fleming,[2] K A Goatman,[2] S Philip,[3] G J Prescott,[4] P F Sharp,[2] G J Williams,[5] W Wykes,[6] G P Leese,[7] J A Olson[3], on behalf of the Scottish Diabetic Retinopathy Clinical Research Network

► Supplementary material is published online only. To view these files please visit the journal online (http://bjo.bmj.com).

[1]Health Economics Research Unit, University of Aberdeen, Foresterhill, Aberdeen, UK
[2]Biomedical Physics, University of Aberdeen, Foresterhill, Aberdeen, UK
[3]Diabetes Retinal Screening Service, David Anderson Building, Foresterhill, Aberdeen, UK
[4]Section of Population Health, University of Aberdeen, Foresterhill, Aberdeen, UK
[5]Department of Ophthalmology, MacQuaker Building, Glasgow, Lanarkshire, UK
[6]Department of Ophthalmology, Southern General Hospital, Glasgow, UK
[7]Department of Diabetes, Ninewells Hospital, Dundee, UK

**Correspondence to**
Mr G S Scotland, Health Economics Research Unit, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK; g.scotland@abdn.ac.uk

Scottish Diabetic Retinopathy Clinical Research Network: NHS Grampian/University of Aberdeen (Alan D Fleming, Keith A Goatman, John A Olson, Sam Philip, Gordon Prescott, Paul McNamee, Graham S Scotland, Peter F Sharp, Graeme J Williams); NHS Greater Glasgow (William Wykes); NHS Lanarkshire (Meena Virdi); NHS Lothian (Ken Swa); NHS Tayside (Graeme Leese).
Guarantor: Graham Scotland.

Accepted 2 October 2009

## ABSTRACT

**Aims** To assess the cost-effectiveness of an improved automated grading algorithm for diabetic retinopathy against a previously described algorithm, and in comparison with manual grading.

**Methods** Efficacy of the alternative algorithms was assessed using a reference graded set of images from three screening centres in Scotland (1253 cases with observable/referable retinopathy and 6333 individuals with mild or no retinopathy). Screening outcomes and grading and diagnosis costs were modelled for a cohort of 180 000 people, with prevalence of referable retinopathy at 4%. Algorithm (b), which combines image quality assessment with detection algorithms for microaneurysms (MA), blot haemorrhages and exudates, was compared with a simpler algorithm (a) (using image quality assessment and MA/dot haemorrhage (DH) detection), and the current practice of manual grading.

**Results** Compared with algorithm (a), algorithm (b) would identify an additional 113 cases of referable retinopathy for an incremental cost of £68 per additional case. Compared with manual grading, automated grading would be expected to identify between 54 and 123 fewer referable cases, for a grading cost saving between £3834 and £1727 per case missed. Extrapolation modelling over a 20-year time horizon suggests manual grading would cost between £25 676 and £267 115 per additional quality adjusted life year gained.

**Conclusions** Algorithm (b) is more cost-effective than the algorithm based on quality assessment and MA/DH detection. With respect to the value of introducing automated detection systems into screening programmes, automated grading operates within the recommended national standards in Scotland and is likely to be considered a cost-effective alternative to manual disease/no disease grading.

## INTRODUCTION

Systematic screening for diabetic retinopathy has been identified as a cost-effective use of health service resources,[1–4] with national screening programmes based on digital photography being implemented across Europe. Given the increasing prevalence of diabetes, national screening programmes may struggle to meet the ever increasing demand using manual grading alone. For example, the number of people recorded on Scottish diabetes registers rose from 162 000 to 197 000 between 2004 and 2006.[5]

We recently reported a study assessing the efficacy and cost-effectiveness of an automated grading system, using digital images from a single grading centre in Scotland.[6 7] Software algorithms were developed to perform the task of image quality (clarity and field definition) assessment and disease/no disease decision making based on the detection of microaneurysms (MA) and dot haemorrhages (DH).[8–10] The automated system was found to be comparable with manual grading in terms of efficacy, and it was estimated that it would result in a cost saving to the Scottish health service of approximately £200 000 per year if implemented within the national screening programme.

Since the publication of the above study, our research group has developed a new algorithm incorporating macular exudate and blot haemorrhage detection, both signs that may require referral to an ophthalmology clinic.[11] The efficacy (sensitivity/specificity) of this new algorithm was compared with the above previously developed algorithm using a test set of reference graded images from multiple screening centres.[12] Here we apply these new efficacy data in a decision tree model to assess the relative cost-effectiveness of using the alternative algorithms within the national screening programme in Scotland. In addition, we model the cost-effectiveness of implementing the new algorithm compared with the current practice in Scotland, which relies on manual grading alone.

## METHODS

### Cost-effectiveness of algorithm (b) versus algorithm (a)

Cost-effectiveness of the new automated algorithm (b) was assessed relative to the previously developed algorithm (a). Algorithm (a) used image quality assessment and MA/DH detection and was identical to that used in our previous studies.[6] Algorithm (b) combined quality assessment and MA/DH detection algorithms with new algorithms for detecting blot haemorrhages and macular exudates. The development and efficacy evaluation of algorithm (b) is described in detail elsewhere.[12–14]

Briefly, the sensitivities and specificities of algorithms (a) and (b), for the main categories of retinopathy defined in the Scottish grading scheme (no retinopathy; mild retinopathy; observable retinopathy/maculopathy; referable retinopathy/maculopathy; and image quality failures), were assessed relative to a reference standard grading using a test

set which included 1253 cases with observable/referable disease and 6333 individuals with mild or no retinopathy. The performance of level-two (full-disease) and level-three (ophthalmologist) manual graders from participating centres was assessed relative to the same reference standard. The reference standard grading was based on review of all photographic images by a clinical research fellow (SP or GW). Disagreements between the reference grading and the final screening programme grade, concerning status of observable/referable retinopathy, were arbitrated by the lead clinician (JAO), again by review of photographic images.

The efficacy estimates (table 1) were incorporated into our previously developed decision tree model,[7] enabling us to assess the cost-effectiveness of implementing the alternative automated algorithms as the first level of grading (disease/no disease) within the national screening programme in Scotland (figure 1),[15] where manual grading is performed using the Scottish Diabetic Retinopathy Grading scheme.[11]

The cost per patient estimate for automated level-one grading is described elsewhere[7] and is based on the assumption that the software will run on a central server covering the whole of Scotland. Costs for level-two and level-three manual grading were based on a survey of staffing arrangements and self reported grading rates at five Scottish grading centres. The cost of level-two and level-three grading were estimated for each centre, and the average used in the base case analysis. The cost of slit-lamp grading was derived from a previous survey of the Grampian retinal screening programme.[7] Finally, we included a cost of £65 for all referrals to ophthalmology (to confirm or refute the screening outcome) (see http://www.isdscotland.org/isd/files/Costs_R044X_2005.xls). All cost parameters used in the cost-effectiveness model are presented in table 2, and are expressed in 2005/2006 sterling to aid comparison with our previous work.[7]

Cost-effectiveness was calculated by estimating the total grading costs and diagnosis costs for referred cases, and the number of appropriate outcomes and referable cases detected, for a population of 180 000—the approximate number of people screened annually in Scotland. Appropriate outcomes were defined as final decisions (recalls or referrals) appropriate to the actual grade of retinopathy present (figure 1). The outcomes reported reflect the overall sensitivity and specificity of the three-level grading system. The incremental cost per additional appropriate outcome and incremental cost per additional referable case detected were calculated for algorithm (b) versus algorithm (a).

### Sensitivity analysis for algorithm (b) versus algorithm (a)

Probabilistic sensitivity analysis was employed to estimate the distribution of cost and effect differences between the alternative automated algorithms.[16] Values were simultaneously selected for each parameter from an assigned distribution and the results recorded. The process was repeated 10 000 times to estimate the distribution of cost and effect differences between alternative strategies.

### Cost-effectiveness of manual versus automated grading

Since the test set for the above efficacy estimates was drawn from patients for whom level-one graders had noted lesions of observable/referable retinopathy, or which level-two graders had passed to level-three graders, it was not possible to assess the performance of level-one manual graders across the centres. Given this lack of efficacy data, we modelled several scenarios comparing manual grading with the superior automated algorithm (determined from the above analysis).

**Table 1** Key efficacy variables used in the model

| Variables derived from previous study[6 7] | Point estimate and 95% CI |
|---|---|
| Prevalence variables | |
| Prevalence (normal cases) | 0.681 (0.670 to 0.692) |
| Prevalence (mild retinopathy) | 0.266 (0.255 to 0.276) |
| Prevalence (observable retinopathy/maculopathy) | 0.013 (0.011 to 0.016) |
| Prevalence (referable retinopathy/maculopathy) | 0.040 (0.035 to 0.045) |
| Technical failure rate | 0.082 (0.076 to 0.089) |
| Efficacy of level-one manual grading | |
| Detection rate for technical failures | 0.969 (0.951 to 0.981) |
| Proportion of normal cases appropriately recalled | 0.920 (0.911 to 0.927) |
| Detection rate for mild retinopathy | 0.819 (0.800 to 0.837) |
| Detection rate for observable retinopathy/maculopathy | 1.000 (0.956 to 1.000) |
| Detection rate for referable retinopathy/maculopathy | 0.992 (0.971 to 0.998) |
| Variables derived from current study | |
| Efficacy of automated algorithm (a) (MA only) | |
| Detection rate for technical failures | 0.986 (0.974 to 0.993) |
| Proportion of normal cases appropriately recalled | 0.634 (0.619 to 0.649) |
| Detection rate for mild retinopathy | 0.796 (0.776 to 0.815) |
| Detection rate for observable retinopathy/maculopathy | 0.942 (0.884 to 0.971) |
| Detection rate for referable retinopathy/maculopathy | 0.950 (0.935 to 0.961) |
| Efficacy of automated algorithm (b) (MA, BH and exudates) | |
| Detection rate for technical failures | 0.988 (0.976 to 0.994) |
| Proportion of normal cases appropriately recalled | 0.632 (0.617 to 0.647) |
| Detection rate for mild retinopathy | 0.790 (0.769 to 0.809) |
| Detection rate for observable retinopathy/maculopathy | 0.933 (0.874 to 0.966) |
| Detection rate for referable retinopathy/maculopathy | 0.969 (0.957 to 0.978) |
| Efficacy of level-two manual grading | |
| Technical failures | |
| Proportion of technical failures correctly referred to slit-lamp* | 0.760 (0.722 to 0.795) |
| Proportion of technical failures incorrectly referred to level three | 0.099 (0.076 to 0.128) |
| Proportion of technical failures incorrectly recalled at 12 months | 0.133 (0.107 to 0.165) |
| Proportion of technical failures incorrectly recalled at 6 months | 0.008 (0.003 to 0.019) |
| No retinopathy | |
| Proportion of normal cases appropriately recalled* | 0.805 (0.757 to 0.846) |
| Proportion of normal cases incorrectly referred to slit-lamp | 0.044 (0.026 to 0.073) |
| Proportion of normal cases incorrectly referred to level three | 0.151 (0.115 to 0.196) |
| Proportion of normal cases incorrectly recalled at 6 months | 0.000 (0.000 to 0.013) |
| Mild retinopathy | |
| Proportion of mild cases appropriately recalled* | 0.915 (0.898 to 0.929) |
| Proportion of mild cases incorrectly referred to slit-lamp | 0.005 (0.002 to 0.011) |

Continued

**Table 1** Continued

| Variables derived from previous study[6 7] | Point estimate and 95% CI |
|---|---|
| Proportion of mild cases incorrectly referred to level three | 0.068 (0.055 to 0.083) |
| Proportion of mild cases incorrectly recalled at 6 months | 0.012 (0.007 to 0.020) |
| Observable retinopathy | |
| Proportion of observable cases appropriately recalled* | 0.448 (0.361 to 0.539) |
| Proportion of observable cases incorrectly referred to slit-lamp | 0.009 (0.002 to 0.047) |
| Proportion of observable cases incorrectly referred to level three | 0.440 (0.353 to 0.530) |
| Proportion of observable cases incorrectly recalled at 12 months | 0.103 (0.060 to 0.172) |
| Referable retinopathy | |
| Proportion of referable cases appropriately referred to level three* | 0.935 (0.917 to 0.949) |
| Proportion of referable cases incorrectly referred to slit-lamp | 0.012 (0.007 to 0.021) |
| Proportion of referable cases incorrectly recalled at 6 months | 0.043 (0.032 to 0.059) |
| Proportion of referable cases incorrectly recalled at 12 months | 0.010 (0.005 to 0.018) |
| Efficacy of level-three manual grading | |
| Technical failures | |
| Proportion of technical failures correctly referred to slit-lamp* | 0.596 (0.461 to 0.718) |
| Proportion of technical failures incorrectly referred to ophthalmology | 0.135 (0.067 to 0.253) |
| Proportion of technical failures incorrectly recalled at 12 months | 0.250 (0.152 to 0.382) |
| Proportion of technical failures incorrectly recalled at 6 months | 0.019 (0.003 to 0.101) |
| No retinopathy | |
| Proportion of normal cases appropriately recalled* | 0.889 (0.765 to 0.952) |
| Proportion of normal cases incorrectly referred to slit-lamp | 0.000 (0.000 to 0.079) |
| Proportion of normal cases incorrectly referred to ophthalmology | 0.067 (0.023 to 0.179) |
| Proportion of normal cases incorrectly recalled at 6 months | 0.044 (0.012 to 0.148) |
| Mild retinopathy | |
| Proportion of mild cases appropriately recalled* | 0.843 (0.750 to 0.906) |
| Proportion of mild cases incorrectly referred to slit-lamp | 0.000 (0.000 to 0.044) |
| Proportion of mild cases incorrectly referred to ophthalmology | 0.145 (0.085 to 0.236) |
| Proportion of mild cases incorrectly recalled at 6 months | 0.012 (0.002 to 0.065) |
| Observable retinopathy | |
| Proportion of observable cases appropriately recalled* | 0.451 (0.323 to 0.586) |
| Proportion of observable cases incorrectly referred to slit-lamp | 0.020 (0.003 to 0.103) |
| Proportion of observable cases incorrectly referred to ophthalmology | 0.314 (0.203 to 0.450) |
| Proportion of observable cases incorrectly recalled at 12 months | 0.216 (0.125 to 0.346) |
| Referable retinopathy | |
| Proportion of referable cases appropriately referred to ophthalmology* | 0.922 (0.902 to 0.938) |
| Proportion of referable cases incorrectly referred to slit-lamp | 0.005 (0.002 to 0.012) |
| Proportion of referable cases incorrectly recalled at 6 months | 0.027 (0.018 to 0.040) |
| Proportion of referable cases incorrectly recalled at 12 months | 0.047 (0.034 to 0.063) |

*Denotes appropriate level-two/ level-three grading outcomes for each category of retinopathy. Beta distributions were applied to all efficacy parameters in probabilistic sensitivity analysis. BH, blot haemorrhage; MA, microaneurysms.

In the first instance, we used efficacy and cost estimates previously obtained for manual level-one graders in Grampian[7] (tables 1 and 2) and factored in an additional cost of £33 110 per annum to account for the extra resources associated with quality assuring the fully manual system (see appendix 1 for full details). We then supplemented this by using a range of feasible sensitivity estimates for manual grading. In addition, we varied costs of level-one manual grading to reflect variation in staffing and grading rates reported across centres in Scotland, and assessed the impact of varying the sensitivity of the automated system within its confidence limits.

In addition, we assessed the impact of increasing the specificity of level-two graders following automated level-one grading; the specificity of the level-two graders might increase if their workload were to include a greater proportion of normal images, as would be the case if automated grading were implemented.

Given that implementation of automated grading may result in a small decrease in sensitivity, we developed a simple extrapolation model to assess the long term economic implications for any referable cases missed as a result. Also, since cost estimates for implementation of automated grading reflect the likely costs in Scotland, where there will be no licensing costs, we conducted sensitivity analysis to assess the impact on cost-effectiveness of increases in implementation costs. Details of these analyses are presented in appendix 2.

## RESULTS

### Cost-effectiveness of algorithm (b) versus algorithm (a)

Table 1 (variables derived from the current study) presents the efficacy findings for the alternative automated algorithms and for level-two and level-three manual graders. Compared with algorithm (a), algorithm (b) has a higher sensitivity for detecting cases of referable retinopathy.

Table 3 shows the results of the base case cost-effectiveness analysis for a cohort of 180 000 people with diabetes. Compared with automated algorithm (a), algorithm (b) leads to an increase in grading cost (+£7759) and an increase in the number of referable cases detected (+113). This equates to an incremental cost effectiveness ratio (ICER) of £68 pounds per additional referable case. A similar pattern of results is found when assessing cost-effectiveness in terms of the cost per appropriate screening outcome.

### Sensitivity analysis for algorithm (b) versus algorithm (a)

The results of 10 000 Monte Carlo simulations indicated a 4.1% chance of algorithm (b) being less costly and more effective (in terms of referable cases identified) than algorithm (a), a 95.1% chance of (b) being more effective and more costly, a 0.38% chance of (b) being more costly but less effective, and a 0.38%

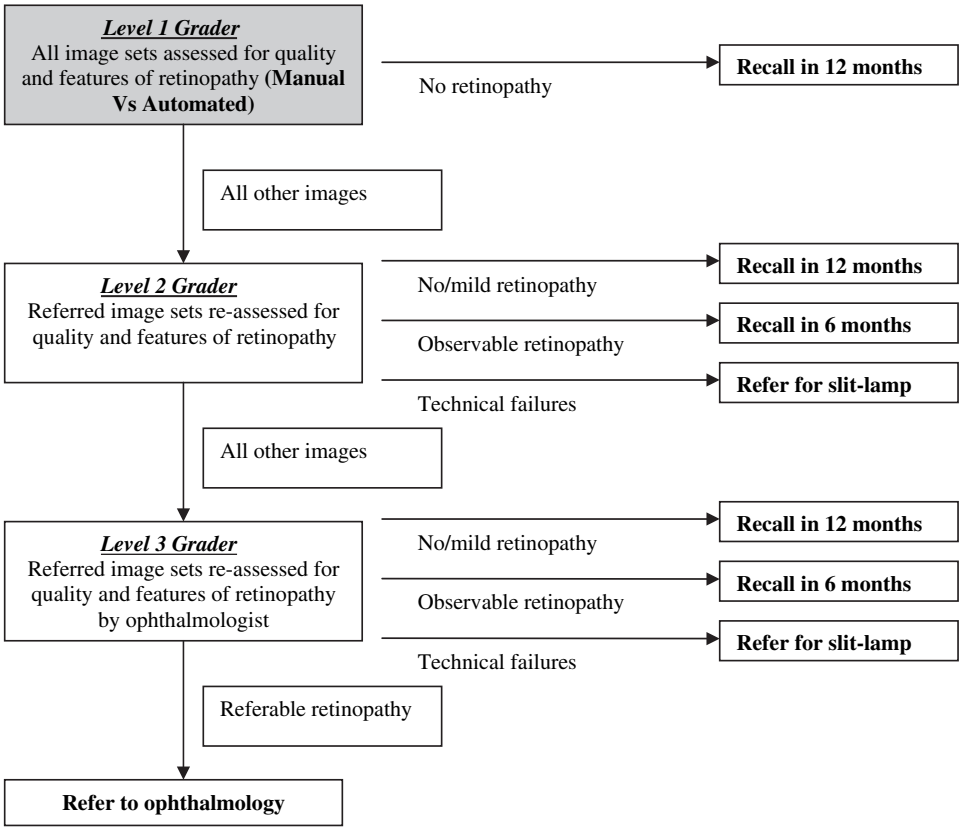


**Figure 1** Grading pathway recommended by the Health Technology Board for Scotland, 2002 (the shaded box indicates the choice between manual and automated grading). Level-one graders first of all assess image quality and identify whether or not images show any signs of retinopathy. Patients whose images show no signs of retinopathy are recalled for a further screening appointment a year later, whereas image sets showing signs of retinopathy, or poor quality image sets, are referred to level-two graders. The level-two graders review these image sets and pass on those with suspected referable retinopathy for final grading by a level-three grader (ophthalmologist). Level-two graders can also recall those with no retinopathy or mild retinopathy at 1 year, recall those with observable retinopathy at 6 months, and refer patients with poor quality image sets for a slit-lamp examination.

chance of (b) being less costly and less effective. Above a willingness to pay threshold of £68 per additional referable case detected, strategy (b) has the higher probability of being preferred compared with (a) on grounds of cost-effectiveness.

### Cost-effectiveness of automated grading (algorithm b) versus manual grading

Applying efficacy and cost estimates previously obtained for level-one graders in Grampian (tables 3 and 4, scenario 1), automated grading generates grading cost savings (£212 695) but identifies 123 fewer referable cases (ICER=£1727) and 734 fewer appropriate screening outcomes (for 180 000 people screened). The 734 additional inappropriate outcomes include the 123 cases of referable retinopathy, an additional 111 cases of observable retinopathy inappropriately recalled at 6 months, 203 additional false positive 6-month recalls, and 297 additional false ophthalmology referrals. However, the modelled false positive increases are based on the assumption that the specificity of level-two graders does not vary according to the case mix received from level-one grading. If we assume that level-two manual graders identify the additional normal cases passed upward by the automated system, with the specificity of manual level-one graders (0.920), then the difference in the false positive rates between the two systems is reduced (table 4; scenario 2). If the specificity of level-three graders were also to improve as result of there being more normal cases referred up through the system, then these differences in false positives would be reduced further.

Lowering the sensitivity of manual level-one graders (for referable retinopathy/maculopathy) from 99.2% to 98% (to be in keeping with the difference observed between manual and automated grading in our previous study) increases the ICER for manual grading versus algorithm (b) to £3834 per additional referable case (table 4; scenario 3). Lowering manual graders sensitivity to 97% (scenarios 5 and 6) results in algorithm (b) becoming the dominant strategy (ie, more effective and less costly). Holding the sensitivity of manual level-one grading constant at 99.2%, while varying the sensitivity of automated

**Table 2** Cost parameters and ranges used in the model

| Variable | Cost per patient and range for sensitivity analysis (£) | Distribution for probabilistic sensitivity analysis |
|---|---|---|
| Level-one manual grading | 1.48 (1.00–1.48) | |
| Additional QA costs (manual grading) | 0.18 | Gamma ($\alpha$ 14.5, $\beta$ 0.0123) |
| Level-one automated grading | 0.13 (0.065–0.26) | Uniform (min 0.065, max 0.26) |
| Level-two grading | 1.05 (1.05–1.41) | Gamma ($\alpha$ 3.15, $\beta$ 0.333) |
| Level-three grading (consultant ophthalmologist) | 2.41 | Gamma ($\alpha$ 19.36, $\beta$ 0.124) |
| Slit-lamp grading | 4.42 (3.16–5.69) | Gamma ($\alpha$ 130.24, $\beta$ 0.034) |
| Ophthalmology referral | 65.00* | |

*Average cost for an ophthalmology outpatient visit in Scotland http://www.isdscotland.org/isd/files/Costs_R044X_2005.xls.

**Table 3** Base case results for a cohort of 180 000. Cost per referable case detected and appropriate screening outcomes (including grading and referral costs)

| Strategy | Total grading/ diagnosis cost | Incremental cost | Referable cases detected | Additional referable cases detected | Appropriate screening outcomes | Additional appropriate screening outcomes | Incremental cost per additional referable case detected | Incremental cost per additional appropriate screening outcome |
|---|---|---|---|---|---|---|---|---|
| (a) Auto (MAs only) | £699202 | | 5998 | | 176149 | | - | - |
| (b) Auto (MAs, BHs, Exudates) | £706961 | £7759 | 6111 | 113 | 176253 | 104 | £68.37* | £74.70* |
| Manual grading | £919656 | £212695 | 6234 | 123 | 176987 | 734 | £1,727.22* | £289.45* |

Strategies are compared incrementally to row above.
*Subject to rounding error. BH, blot haemorrhage; MA, microaneurysms.

level one grading within its 95% confidence limit, the ICER for manual versus automated grading varied between £1460 and £2933 per additional referable case detected.

Lowering manual level-one grading costs to the estimated average across the five screening centres in Scotland (£1.00 per patient), reduces the ICER for manual grading to £2247 and £1028 per additional referable case detected—assuming 98% (scenarios 7 and 8) and 99.2% sensitivity for manual grading respectively.

The economic implications for any missed cases and the impact of automated implementation costs on long-term cost-effectiveness are assessed in appendix 2. Results from this exercise suggest that, over a 20-year time horizon, manual grading would cost between £25 676 and £267 115 per additional quality adjusted life year (QALY) gained.

## DISCUSSION

### Summary statement of results

This study has shown that the inclusion of automated algorithms to identify lesions of observable/referable diabetic retinopathy can improve the cost-effectiveness of automated grading. Compared with (a), algorithm (b) results in increased detection of cases of referable diabetic retinopathy (+113) without altering specificity. There is only a small annual cost increase (£7759) due to increased appropriate referrals to ophthalmology. The incremental cost per case detected for manual versus automated algorithm (b) is sensitive to small changes in the sensitivity of manual level-one graders (table 4).

### Strengths and weaknesses

This study adds to existing knowledge on the cost-effectiveness of automated systems by comparing the performance of alternative automated grading algorithms on a set of images from three screening centres in Scotland. A potential limitation is that the comparison between automated and fully manual grading relied on efficacy data for manual level-one graders obtained from a single centre. To address this limitation, we varied the assumptions surrounding sensitivity estimates for level-one manual graders. As indicated in table 4, if the average sensitivity of manual level-one graders (for referable retinopathy) across Scotland were 97%, then automated algorithm (b) would be the dominant option (more effective and less costly).

Further analysis was performed to consider the longer-term costs and consequences of any missed cases of retinopathy. Table 4 (scenario 1) suggests that automated grading might be expected to miss 123 additional cases of referable disease for a screening cohort of 180 000—assuming that manual level-one grading has sensitivity of 99.2%. This increase in the chance of false negatives is the result of automated grading missing 35 cases that were reference graded as referable disease in the clinical study upon which our cost-effectiveness models are based.[12] However, 26 (74%) of these cases were referable maculopathy (as defined by surrogate photographic markers) and Schofield *et al* have shown that that only 13.2% of such cases are subsequently found to have macular oedema (MO) requiring treatment.[17] Applying these proportions to the 123 modelled false negatives, and applying a relative risk of 0.017 for progression to blindness (in untreated versus treated MO),[4] 0.20 additional cases of visual loss would be expected from MO each year for every 180 000 individuals graded with the automated system ($123\times0.74\times0.13\times0.017$).

Of the nine cases of referable retinopathy missed in the clinical study, three were originally reference graded as referable background retinopathy (R3) and six were reference graded as proliferative retinopathy (R4). However, subsequent review by the lead clinician of the images graded as R4 found these to be showing collaterals or normal variations (see online supporting material: http://www.abdn.ac.uk/~mph605/bjo/). Grading is subjective and we do not believe these constitute clear errors, as defined by NHS Quality Improvement Scotland's (NHSQIS) Clinical Standards for Diabetic Retinopathy Screening.[18] Consensus between human graders on features of referable retinopathy is difficult to achieve as indicated in a recent study by Abramoff *et al*,[19] which showed that the sensitivity of the three retinal specialists varied between 62% and 85% on a random set of images.

In order to assess the potential economic implications of missing additional referable cases over a 20-year period, we developed a simple extrapolation model (see appendix 2 for details). Assuming no true cases of proliferative retinopathy are missed by the automated system, and that the non-proliferative cases of referable retinopathy are not at risk of progression within 1 year, we estimate that automated grading would result in a net cost saving of £2 940 256 over a period of 20 years, for a loss of only 11.01 QALYs. Furthermore, this analysis suggests that implementation costs for automated grading would have to be approximately 10-fold higher for the cost per QALY gained with manual grading to fall below £30 000 (the threshold applied by the National Institute for Health and Clinical Excellence to judge cost-effectiveness).

If we assume that 50% of missed referable retinopathy cases (R3 and R4) are in fact true proliferative cases (table AI) requiring immediate treatment, then net cost savings would be expected to drop to £1 877 082 and QALY losses would increase to 73.11 (equating to a cost per QALY gained of £25 676 for manual grading). Thus the cost-effectiveness of automated grading appears sensitive to the probability of missing true proliferative cases. However, it is reassuring to note that in a much larger recent validation study, where automated grading was compared with manual grading on 33 535 patients, automated grading was judged to have missed no additional cases of R3 or R4, as arbitrated by consensus between seven ophthalmologists. (http://www.ndrs.scot.nhs.uk/ExecGrp/Docs/2009%2006%2009%20Retinal%20autograding%20Waugh%20FINAL.pdf). We are therefore confident that adoption of automated

**Table 4** Deterministic sensitivity analysis of manual versus automated strategy (b) for a cohort of 180 000 people

| Scenario | Total cost | Referable cases | Appropriate recalls | FP 6-month recalls | FP referrals to ophthalmology | False negatives (observable) | False negatives (referable) | Incremental cost per additional referable case (manual vs automated (b1)) |
|---|---|---|---|---|---|---|---|---|
| Baseline (automated algorithm (b)) | £706961 | 6111 | 170142 | 864 | 1222 | 572 | 1089 | |
| 1. Manual grading: sensitivity for referable cases=99.2%[8] | £919656 | 6234 | 170753 | 661 | 925 | 461 | 966 | £1727* |
| 2. As (1) but with specificity of level 2 graders adjusted (see text) | £923784 | 6234 | 170696 | 684 | 959 | 461 | 966 | £1761* |
| 3. Manual L1 grading: sensitivity for referable cases=98% (assumes difference between manual and automated grading found in our previous study)[8] | £914912 | 6165 | 170753 | 661 | 925 | 461 | 1035 | £3834* |
| 4. As (3) but specificity of level 2 graders adjusted (see text) | £919039 | 6165 | 170696 | 684 | 959 | 461 | 1035 | £3910* |
| 5. Manual L1 grading: sensitivity for referable cases=97% | £901915 | 6107 | 170753 | 661 | 925 | 461 | 1093 | Manual system dominated |
| 6. As (5) but with specificity of level 2 graders adjusted (see text) | £915043 | 6107 | 170696 | 684 | 959 | 461 | 1093 | Manual system dominated |
| 7. Manual L1 grading: sensitivity for referable cases=98%; grading cost per patient=£1.00 | £828812 | 6165 | 170753 | 661 | 925 | 461 | 1035 | £2247* |
| 8. As (7) but with specificity of level 2 graders adjusted | £832939 | 6165 | 170696 | 684 | 959 | 461 | 1035 | £2322* |

Manual scenarios are compared incrementally to (b).
*Subject to rounding error. FN, false negative; FP, false positive.

level-one (disease/no disease) grading represents a cost-effective approach in the Scottish context.

It should be noted that the analysis reported here focuses on the cost-effectiveness of automating level-one disease/no disease grading within the three-tier grading system used in Scotland (figure 1). The findings may not be applicable to programmes relying on two-tier grading systems. This question requires further investigation.

An additional factor that has not been taken into consideration in the present analysis is the issue of increasing demand, and the limited capacity to meet this demand. As the prevalence of diabetes increases, full screening coverage may become untenable through reliance on manual grading alone (http://www.diabetes.org.uk/About_us/News_Landing_Page/People-with-diabetes-not-getting-retinal-screening/). Automated grading may enable screening programmes to meet the demand by performing the task of level-one grading thus releasing staff time for other tasks such as image capture and level-two grading.

### Relationship with previous study

Our group previously showed that automated grading compared favourably with manual level-one graders in Grampian.[7] Using the estimates for manual level-one grading from this previous study in the current analysis, the cost-effectiveness of automated grading declined compared with the previous estimate. This is because the sensitivity/specificity estimates obtained for automated grading were slightly lower than they were in the previous study. The slight decline in efficacy observed for automated grading may be due to random variation, variation in the type/quality of digital images obtained from different screening centres, differences in the population characteristics across the centres, or the inclusion of a different reference grader to assess efficacy.

### Policy implication

The question of whether automated grading is considered a cost-effective alternative to manual grading, depends on whether or not the anticipated cost savings associated with algorithm (b) are considered to outweigh the slightly higher probability of detecting referable cases with manual grading. This in turn depends on the probability of any missed cases progressing to more severe forms of disease within the screening interval (6 months–1 year), relative to the probability of progression with appropriate referral. Given the calculations outlined above, combined with the fact that a grading system using algorithm (b) would operate within safety standards set by NHSQIS,[18] automated grading is likely to be considered a cost-effective alternative to manual level-one grading in Scotland. NHSQIS recommend that for each grader, clear grading errors—failure to notice unequivocal signs of referable retinopathy or failure to notice that an image is of insufficient quality for grading—should not exceed 1 in every 200 patients screened.[18] Based on the estimated sensitivity for referable retinopathy and technical failures, and the estimated prevalence of these events, the expected clear error rate for algorithm (b) is 1 in 450, well within the recommended standard.

## CONCLUSION

The inclusion of automated exudate and haemorrhage detection can be considered more cost-effective than the previously developed algorithm based on MA/DH detection. With respect to the value of introducing automated detection systems into screening programmes, automated grading operates within the recommended national standard of error rate, is associated with

significant grading cost savings to the NHS, and is likely to be considered a cost-effective alternative to manual disease/no disease grading.

**Acknowledgements** We would like to thank Lorraine Urquhart, Julie Hughes (Grampian Retinal Screening Programme), Fiona Heggie (Greater Glasgow and Clyde Retinal Screening Service), Dr Meena Virdi (NHS Lanarkshire), and Dr Ken Swa (NHS Lothian) who all provided information for the cost survey.

**Funding** This study was funded by the Chief Scientist Office of the Scottish Governments Health Directorates (SGHD). The views expressed here are those of the authors and not necessarily those of the SGHD.

**Competing interests** Implementation in Scotland is being considered. If this occurs it is likely that there will be some remuneration for the University of Aberdeen, NHS Grampian and the Scottish Executive.

**Ethics approval** The North of Scotland Research Ethics Committee confirmed that a formal ethics application was not required for this study based on anonymised routine data.

**Contributors** JAO, PFS, ADF, KAG, GJP, GJW, SP, PM, WW and GL obtained funding for the study. ADF and KAG coordinated the study and obtained images for analysis. SP and GJW acted as the reference graders for the assessment of the efficacy of the alternative grading approaches. ADF developed the image processing computer algorithms underlying the automated algorithms. ADF and GJP conducted the analysis of the effectiveness data. GSS and PM designed and conducted the cost-effectiveness modelling, and took the lead on writing this paper. All authors commented, contributed to, and reviewed the drafts and final version of the paper.

**Provenance and peer review** Not commissioned; externally peer reviewed.

## REFERENCES

1. **Maberly D,** Walker H, Koushik A, *et al*. Screening for diabetic retinopathy in James Bay, Ontario: a cost-effectiveness analysis. *CMAJ* 2003;**168**:160–4.
2. **Davies R,** Roderick P, Canning C, *et al*. The evaluation of screening policies for diabetic retinopathy using simulation. *Diabet Med* 2002;**19**:762–70.
3. **James M,** Turner DA, Broadbent DM, *et al*. Cost-effectiveness analysis of screening for sight threatening diabetic eye disease. *BMJ* 2000;**320**:1627–31.
4. **Javitt J,** Aiello LP, Chiang Y, *et al*. Preventive eye care in people with diabetes is cost saving to the federal government. *Diabetes Care* 1994;**17**:909–17.
5. **Scottish Executive Health Department.** *Scottish diabetes survey 2006*. Edinburgh: SEHD, 2008.
6. **Philip S,** Fleming AD, Goatman KA, *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.
7. **Scotland GS,** McNamee P, Philip S, *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.
8. **Fleming AD,** Philip S, Goatman KA, *et al*. Automated assessment of retinal image quality based on clarity and field definition. *Invest Ophthalmol Vis Sci* 2006;**47**:1120–5.
9. **Fleming AD,** Philip S, Goatman KA, *et al*. Automated microaneurysm detection using local contrast normalization and local vessel detection. *IEEE Trans Med Imaging* 2006;**25**:1223–32.
10. **Fleming AD,** Goatman KA, Philip S, *et al*. Automatic detection of retinal anatomy to assist diabetic retinopathy screening. *Phys Med Biol* 2007;**52**:331–45.
11. **Diabetic Retinopathy Screening Implementation Group.** *Diabetic retinopathy screening services in Scotland: recommendations for implementation*. Edinburgh: Scottish Executive, 2003.
12. **Fleming AD,** Goatman KA, Philip S, *et al*. The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol* 2010; **94**:706–11.
13. **Fleming AD,** Goatman KA, Williams GJ, *et al*. Automated detection of blot haemorrhages as a sign of referable diabetic retinopathy, Medical Image Understanding and Analysis, Dundee, July 2008:235–9.
14. **Fleming AD,** Philip S, Goatman KA, *et al*. Automated detection of exudates for diabetic retinopathy screening. *Phys Med Biol* 2007;**52**:7385–96.
15. **Facey K,** Cummins E, Macpherson K, *et al*. *Organisation of services for diabetic retinopathy screening*. Glasgow: Health Technology Board for Scotland, 2002, Health Technology Assessment Report 1.
16. **Briggs AH.** Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 2000;**17**:479–500.
17. **Schofield CJ,** Ellis JD, Ellingford A, *et al*. Macular oedema is not predicted by perifoveal single blot haemorrhages. *Diabet Med* 2008;**25**:129–33.
18. **NHS quality improvement Scotland.** *Diabetic retinopathy screening—clinical standards*. Edinburgh: NHS Quality Improvement Scotland, 2004.
19. **Abramoff MD,** Niemeijer M, Suttorp-Schulten MSA, *et al*. 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.
20. **Meads C,** Hyde C. What is the cost of blindness. *Br J Ophthalmol* 2003;**87**:1201–4.
21. **van Hecke MV,** Dekker JM, Nijpels G, *et al*. Retinopathy is associated with cardiovascular and all-cause mortality in both diabetic and nondiabetic subjects. *Diabetes Care* 2003;**26**:2958.

## APPENDIX 1: ESTIMATE OF ADDITIONAL QUALITY ASSURANCE COSTS FOR THE MANUAL GRADING STRATEGY

1. Number of quality assurance (QA) reviews required annually per grader=500
2. Cost per QA review=£1.54
3. Estimated number of level-one graders in Scotland=53
   *Estimated QA costs for manual level-one graders
   =53×500×£1.54
   =£40 810
4. Number of QA reviews required for the automated system=2500 (assumption)
5. Number of patients to be screened per year in Scotland=180 000
6. Proportion of patients referred to level two with manual grading=0.38
7. Proportion of patients referred to level-two with automated grading=0.6
   *Estimated increase in the number of patients referred to level-two with the automated compared with manual grading
   =(180 000×0.6)−(160 000×0.38)
   =39 600
   *Estimated number of patients graded annually by one whole time equivalent (WTE) level-two grader
   =16 800 (based on grader working 42 weeks per year, 2 days a week, at rate of 80 patients per day)
   *Estimated number or level-two WTE graders required to grade additional referrals
   =39 600/16 800
   =2.36
   *Estimated actual numbers of graders required to grade additional referrals (assuming graders work half-time)
   =5
   *Estimated costs of quality assuring the automated grading strategy relative to the manual strategy
   =£1.54 (2500+(5×500))
   =£7700
   *Estimated additional QA cost of the manual strategy compared to the automated strategy
   =£40 810−£7700
   =£33 110
   =£0.18 per patient screened

## APPENDIX 2: COST/QALY ESTIMATES FOR AUTOMATED VERSUS MANUAL GRADING

In the clinical study upon which our cost-effectiveness estimates are based, automated grading missed 35 cases that were reference graded as referable disease. Of these cases, 26 (74%) were classed as referable maculopathy by the reference graders and nine (26%) were classed as referable retinopathy (three non-proliferative (R3) and six proliferative (R4)). These findings lead to our estimate of 123 additional referable cases being missed (91 referable maculopathy cases and 32 referable retinopathy cases) for a screening cohort of 180 000 individuals.

However, subsequent review by the lead clinician of the images graded as R4 found these to be showing collaterals or normal variations (see online supporting material http://www.abdn.ac.uk/~mph605/bjo/). Thus it is unlikely any of these patients would be at risk of vision loss. In addition, Schofield *et al* showed that only 13.2% of cases classified as referable maculopathy (as defined by surrogate photographic markers) turn out have macular oedema requiring treatment.[17]

**Table AI** Parameters used in the extrapolation model

| | | | |
|---|---|---|---|
| Cases missed | 123 | | |
| Expected grades | M2 | R3 | R4 |
| Proportions | 0.743 | 0.086 | 0.171 |
| Proportion with real pathology | 0.132 | 0.5 | 0.5 |
| Relative hazard for progression (untreated versus treated) | 0.017 | 0.0732 | 0.0732 |
| Cases of visual loss | 0.186 | 0.387 | 0.770 |
| Annual grading cost saving | £212695 | | |
| Annual social cost of blindness (1st year) | £7533 | | |
| Annual social costs of visual loss (subsequent years) | £7346 | | |
| Utility decrement associated with visual loss | 0.44 | | |
| Discount rate | 0.035 | | |

**Table AII** Extrapolation model

| Year | Incidence of additional cases of visual loss | Annual screening cost savings | Annual mortality adjusted increased cost of visual loss | Annual mortality adjusted QALY losses |
|---|---|---|---|---|
| 0 | 1.362 | £212695 | £10260 | 0.599 |
| 1 | 1.362 | £205502 | £19693 | 1.150 |
| 2 | 1.362 | £198553 | £28334 | 1.655 |
| 3 | 1.362 | £191839 | £36221 | 2.116 |
| 4 | 1.362 | £185351 | £43388 | 2.534 |
| 5 | 1.362 | £179083 | £49869 | 2.913 |
| 6 | 1.362 | £173028 | £55694 | 3.253 |
| 7 | 1.362 | £167176 | £60896 | 3.557 |
| 8 | 1.362 | £161523 | £65516 | 3.827 |
| 9 | 1.362 | £156061 | £69574 | 4.064 |
| 10 | 1.362 | £150783 | £73096 | 4.269 |
| 11 | 1.362 | £145685 | £76105 | 4.445 |
| 12 | 1.362 | £140758 | £78624 | 4.592 |
| 13 | 1.362 | £135998 | £80671 | 4.712 |
| 14 | 1.362 | £131399 | £82272 | 4.805 |
| 15 | 1.362 | £126956 | £83451 | 4.874 |
| 16 | 1.362 | £122662 | £84231 | 4.920 |
| 17 | 1.362 | £118514 | £84634 | 4.943 |
| 18 | 1.362 | £114507 | £84690 | 4.947 |
| 19 | 1.362 | £110635 | £84409 | 4.930 |
| | | £3128709 | £1251627 | 73.107 |
| | | | Net cost saving | £1877082 |
| | | | Net QALY loss | 73.107 |
| | | | ICER | £25676 |

ICER, incremental cost effectiveness ratio; QALY, quality adjusted life year.

In order to assess the potential economic implications any missed referable cases over a 20-year period, we developed an extrapolation model in Microsoft Excel (table AII).

Assuming a stable screening cohort of 180 000 patients, a constant prevalence of underlying referable disease (4%), and a constant false negative rate for automated grading (123 additional referable patients missed each year), we modelled costs and QALY losses associated with additional cases of blindness over a 20-year time horizon. The annual number of cases of vision loss were estimated by multiplying missed cases by annual risk ratios for progression in untreated versus treated patients.[4] Social costs associated with blindness were obtained from a Health Technology Assessment of treatment for age related macular degeneration, and applied to modelled incident cases.[20] Utility decrements associated with blindness were obtained from the literature.[15] Cumulative social costs associated with additional cases of blindness were then subtracted from the cumulative cost savings associated with automated grading, and the net cost was compared with cumulative QALY losses. Note the model does not account for possible cost and quality of life implications associated with less severe vision loss. Results were modelled for a cohort with age equal to the mean age the screening cohort in Scotland. Age and sex adjusted mortality rates from UK life tables were inflated to adjust for the increased risk of all causes of mortality in individuals with diabetic retinopathy,[21] and were applied to the cohort year on year. A discount rate of 3.5% was applied to future costs and QALYs.

## Findings

Assuming no true cases of proliferative retinopathy are missed by the automated system, and that the non-proliferative cases of referable retinopathy are not at risk of progression within 1 year, we estimate that automated grading would result in a net cost saving of £2 940 256 over a period of 20 years, for a loss of only 11.01 QALYs. Furthermore, this analysis suggests that implementation costs for automated grading would have to be approximately 10-fold higher for cost savings to be less than £30 000 per QALY lost (the threshold applied by the National Institute for Health and Clinical Excellence to judge cost-effectiveness).

However, if we assume that 50% of missed referable retinopathy cases (R3 and R4) are in fact true proliferative cases (table AI), then net cost savings would be expected to drop to £1 877 082 and QALY losses would increase to 73.11 (equating to a saving of £25 676 per QALY lost).

# Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy

G S Scotland, P McNamee, A D Fleming, et al.

*Br J Ophthalmol* 2010 94: 712-719 originally published online December 3, 2009
doi: 10.1136/bjo.2008.151126