DOI: 10.1111/j.1464-5491.2011.03504.x

Article: Care Delivery

External quality assurance for image grading in the Scottish Diabetic Retinopathy Screening Programme

K. A. Goatman[1], S. Philip[2], A. D. Fleming[1], R. D. Harvey[3], K. K. Swa[4], C. Styles[5], M. Black[3], G. Sell[3], N. Lee[3], P. F. Sharp[1] and J. A. Olson[2]

[1]School of Medicine and Dentistry, University of Aberdeen, [2]Diabetes & Endocrinology, Aberdeen Royal Infirmary, [3]NHS Highland, Inverness, [4]National Services Division, NHS Scotland, Edinburgh and [5]Queen Margaret Hospital, Dunfermline, UK

Accepted 16 October 2011

Abstract

Aims To develop and evaluate an image grading external quality assurance system for the Scottish Diabetic Retinopathy Screening Programme.

Method A web-based image grading system was developed which closely matches the current Scottish national screening software. Two rounds of external quality assurance were run in autumn 2008 and spring 2010, each time using the same 100 images. Graders were compared with a consensus standard derived from the top-level graders' results. After the first round, the centre lead clinicians and top-level graders reviewed the results and drew up guidance notes for the second round.

Results Grader sensitivities ranged from 60.0 to 100% (median 92.5%) in 2008, and from 62.5 to 100% (median 92.5%) in 2010. Specificities ranged from 34.0 to 98.0% (median 86%) in 2008, and 54.0 to 100% (median 88%) in 2010. There was no difference in sensitivity between grader levels, but first-level graders had a significantly lower specificity than level-two and level-three graders. In 2008, one centre had a lower sensitivity but higher specificity than the majority of centres. Following the feedback from the first round, overall agreement improved in 2010 and there were no longer any significant differences between centres.

Conclusions A useful educational tool has been developed for image grading external quality assurance.

Diabet. Med. 29, 776–783 (2012)

Keywords diabetic retinopathy, quality assurance, retinal screening, sensitivity, specificity

Abbreviation EQA, external quality assurance

Introduction

Compared with more established screening programmes, such as cervical and breast screening, diabetic retinopathy screening is still in its infancy. In 2002, the Health Technology Board for Scotland recommended a national screening programme for diabetic retinopathy, based on retinal photography [1]. In 2009/2010, the Scottish screening programme invited 210 015 people for screening, of whom 157 937 were successfully screened during that period (80.3% of the eligible population) [2].

Quality assurance is an essential component of a screening service. Regular internal quality assurance of image grading was recommended by National Health Service (NHS) Quality Improvement Scotland, which requires 500 images from every grader each year to be reviewed by a local top-level grader [3,4]. Several studies have investigated internal agreement elsewhere, for instance in Newcastle [5,6] and Bristol [7]. However, internal checks alone cannot measure agreement and consistency between centres or top-level graders: this requires some form of external quality assurance (EQA). Established national EQA programmes exist for cervical screening [8] and breast screening [9].

We describe a web-based system for external quality assurance of image grading developed for the Scottish Diabetic Retinal Screening Collaborative and we report results from the autumn 2008 and spring 2010 EQA rounds.

Methods

The external quality assurance process involved development of:

1. a web-based grading interface similar to the existing Siemens Soarian® software (Siemens AG, Healthcare Sector, Erlangen, Germany) used throughout Scotland;

Correspondence to: Dr Keith A. Goatman, Medical Physics Building, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK.
E-mail: k.a.goatman@abdn.ac.uk

2. a web-based remote access portal and secure user registration system;
3. an image viewing system allowing adjustment of zoom, brightness and contrast, histogram equalization and red-free mode display. An onscreen measuring tool was also available;
4. a database to store user grading;
5. software to record the use of controls, time and duration of operation by each user;
6. a set of images for users to grade;
7. software to analyse the grading.

Web-based software

Software was developed which closely matches the feature-based grading used in Scotland. Server-side programming used PHP and client-side scripting used Javascript. The system is compatible with all popular web browsers. Figure 1 shows the system in use, with the image display on the left, together with controls for contrast/brightness, zoom and red-free/colour display, and the feature grading panel on the right-hand side. The time taken for grading each image was recorded for the 2010 round. The sequence of any image adjustment controls used was also recorded. Each grader was presented with the images in a random order.

Following the first round, a number of additional features were added, including a 'sandbox' practice area (where graders can familiarize themselves with the system by grading as many example images as they wish), the ability to go back one image to correct accidental submissions, the recording of grading duration and online guidance notes.

The software implements the Scottish grading scheme, summarized in Table 1. The retinopathy grades are derived automatically from the selected features. There are eight possible grading outcomes: four of these outcomes require referral (M2, R3, R4 and R6), two indicate more frequent review with a 6-month interval (M1 and R2), while the remaining two categories (R0 and R1) result in re-screening in 12 months. Note that while the different UK national screening programmes use similar grading nomenclature, there are significant differences in usage between programmes (for example, in the English programme R3 is similar to the Scottish R4). Readers should refer to Table 1 to avoid any ambiguity.

Image grading test set

One hundred images from the NHS Grampian Diabetes Retinal Screening Programme were selected by a clinician with a special interest in diabetic retinopathy, but who was not a grader required to participate in the EQA.

Grading rounds

Two EQA rounds were run using the same image set: autumn 2008 (when participation was voluntary) and spring 2010. For both rounds, graders were given 3 weeks to grade the images. Following the first round, the top-level graders came together to survey the results and review any contentious images. Based on

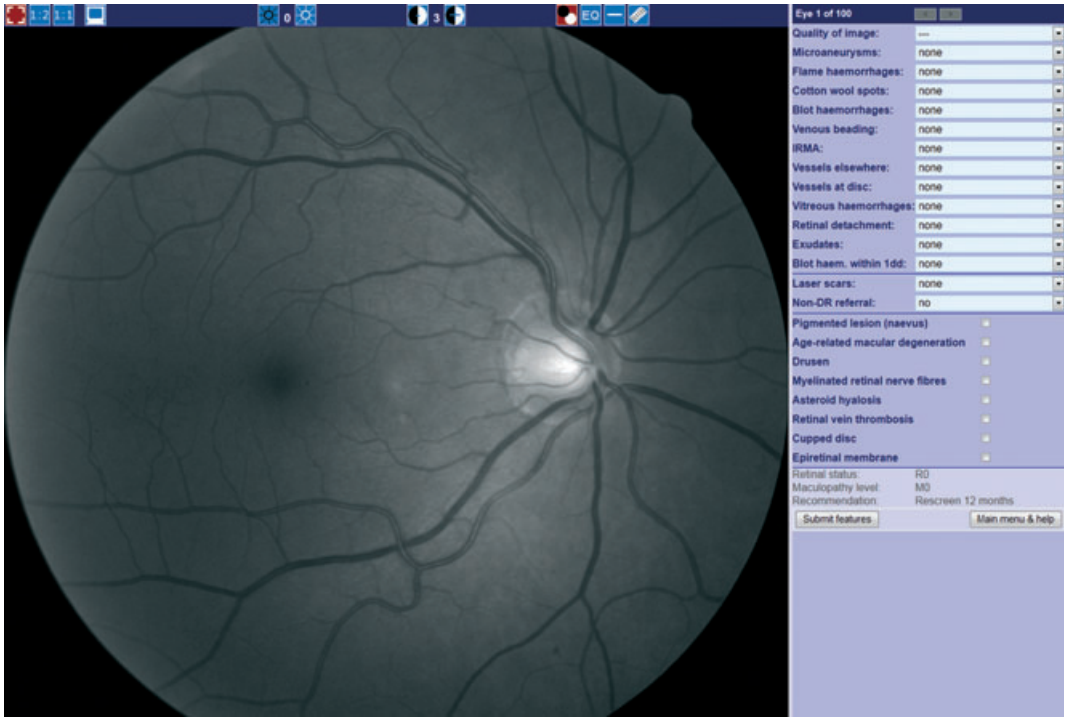

FIGURE 1 External quality assurance feature-based web browser grading screen ready for user to begin selecting features. Note the user has selected red-free mode.

Table 1 A summary of the grading scheme used in Scotland [15]

Grade	Description	Outcome
R0 (No visible retinopathy)	No visible diabetic retinopathy anywhere	Re-screen in 12 months
R1 (Mild retinopathy)	At least one dot, blot or flame haemorrhage, microaneurysm, exudate or cotton wool spot anywhere	Re-screen in 12 months
M1 (Observable maculopathy)	Exudate within a radius of > 1 but ≤ 2 disc diameters of the centre of the fovea	Re-screen in 6 months
R2 (Observable retinopathy)	Four or more blot haemorrhages (> Airlie House standard photograph 2a) in one hemi-field only (where the inferior and superior hemi-fields are delineated by a line passing through the centres of the fovea and optic disc)	Re-screen in 6 months
M2 (Referable maculopathy)	**Any blot haemorrhage or exudate within a radius of one disc diameter of the centre of the fovea**	**Refer to ophthalmology**
R3 (Referable retinopathy)	**Any of the following features:** **(1) four or more blot haemorrhages (> Airlie House standard photograph 2a) in the inferior and superior hemi-fields;** **(2) venous beading (> Airlie House standard photograph 6a);** **(3) intra-retinal microvascular anomalies (IRMA) (> Airlie House standard photograph 8a)**	**Refer to ophthalmology**
R4 (Proliferative retinopathy)	**Any of the following features:** **(1) active new vessels;** **(2) vitreous haemorrhage**	**Refer to ophthalmology**
R6 (Technical failure)	**Insufficient clarity or field of view for assessment**	**Slit-lamp examination**

Referable outcomes are shown in bold text.
Note that the grading categories do not correspond directly to the categories used in similar grading schemes (see, for example, those in England [16] and Wales [17]).

these discussions, guidance notes were issued before the second round with the aim of improving consistency. Except during the review process, graders did not have sight of the images again before the 2010 round.

The following guidance notes were given to graders before the second round.

1. Use the red-free, contrast and zoom tools.
2. Use the on-screen ruler to measure the disc diameter (vertical height) and distance from objects to the fovea.
3. Microaneurysms are spherical and have well-defined edges, whereas blot haemorrhages have less well-defined edges. A blot haemorrhage is defined as a haemorrhage with a diameter larger than the diameter of a vein at the disc margin.
4. Adequate clarity requires that third-generation blood vessels around the fovea must be visible. Adequate field of view requires the entire optic disc to be visible and the fovea to be at least two disc diameters from the image edge.

Consensus grading

A consensus grade was calculated for each image from the results of the 15 top-level graders. Consensus was deemed to have occurred when at least two-thirds of the top-level graders agreed whether the image was referable; otherwise the image was excluded from the analysis. Each grader was compared with the consensus standard.

Statistical analysis

The data were analysed to determine the following.

How well do the graders agree with the consensus?
Agreement was determined, firstly, for the eight individual grades in Table 1 (i.e. R0 graded as R0, R1 as R1, etc.) and, secondly, for the binary decision of whether an image is referable. The sensitivities and specificities for each top-level grader, together with 95% confidence intervals (CIs), were plotted. Sensitivities and specificities for all graders detecting referable images were shown on a receiver operator characteristic plot. Cohen's Kappa, a common measure of inter-observer agreement, was also calculated for each grader.

Are there differences in sensitivity and specificity between 2008 and 2010?
A Mann–Whitney *U*-test was used to find significant differences in the sensitivity, specificity or Kappa distributions between 2008 and 2010. A paired analysis was not attempted as a quarter of graders did not take part in both rounds and the 2008 results were anonymized.

Are there differences between centres?
The main goal of EQA is to detect significant grading differences between centres. A Kruskal–Wallis test was used to determine any significant differences in the sensitivity or specificity distributions between centres. Separate analyses were carried out for the 2008 and 2010 rounds. Where the Kruskal–Wallis test found significant differences, a post-hoc multiple-comparison

test, using Tukey's least significant difference method, was used to determine which centres differed.

Are there differences between grader levels?

A Kruskal–Wallis test was used to find any significant differences in the sensitivity or specificity distributions between the three grader levels. Where significant differences were found, Tukey's least significant difference method was used to find which levels differed.

The data were analysed using MATLAB 2010b and its Statistics Toolbox (The Mathworks Inc., Natick, MA, USA).

Confidentiality

Individual performance measures were only accessible to the individual grader, the national lead clinician, the relevant local programme manager and the system administrator. Similarly, published centre performance results were anonymized.

Results

Sixty-four graders (71%) completed the 2008 round and 84 graders (93%) completed the 2010 round. In the later round, the median time taken by graders was 3.5 h (interquartile range 2.5–4.7 h).

Consensus grading

Consensus was achieved for 90/100 images; the remaining 10 images were excluded from the analysis. The consensus standard contained 21 images graded R0, 24 graded R1, three graded M1, two graded R2, three graded R6, 22 graded M2, five graded R3 and 10 graded R4. Overall, 40/90 (44.4%) of the images had referable grades.

There was poorer consensus regarding which feature(s) made each image referable. The top-level graders were unanimous about the referable feature(s) in 37.5% (15/40) of referable images. In a further 52.5% (21/40) of images, there was a two-thirds majority. In the remaining 10% (4/40), images opinion was split regarding features such as a blot or exudate in the macula, or new vessels vs. intra-retinal microvascular anomalies.

How well do the graders agree with the consensus?

Table 2 shows the distribution of image grades in 2010 for the 7560 image-grading episodes (i.e. 84 graders each assessing 90 images). There was exact agreement between the standard and the graders across the eight grade categories in Table 1 in 5393/7560 [71.3% (95% CI 70.3–72.3%)] of images graded.

Regarding the binary decision of whether an image was referable or not, agreement was 6706/7560 [88.7% (95% CI 88.0–89.4%)]. Of the disagreements, 528/7560 [7.0% (95% CI 6.4–7.6%)] images were over-graded (i.e. grader referable but consensus non-referable) and 326/7560 [4.3% (95% CI 3.9–4.8%)] were under-graded. Of the referable grades, 208/256 [82.5% (95% CI 77.4–86.7%)] of the R6s were graded referable, 1667/1848 [90.2% (95% CI 88.8–91.5%)] of the M2s, 400/420 [95.2% (95% CI 92.8–96.9%)] of the R3s and 759/840 [90.4% (95% CI 88.2–92.2%)] of the R4s.

Individual grader Cohen Kappa values for referable/non-referable grading in 2008 ranged from 0.31 to 0.91 (median 0.75, interquartile range 0.67–0.78). In 2010, Kappa ranged from 0.51 to 0.93 (median 0.78, interquartile range 0.70–0.84).

Figure 2 shows the sensitivities and specificities, together with the 95% confidence intervals, for the fifteen top-level graders. Figure 3 shows a receiver operating characteristic plot for all the graders in 2010.

Are there differences between 2008 and 2010?

Grader sensitivities ranged from 60.0 to 100% (median 92.5%, interquartile range 11.3%) in 2008, and from 62.5 to 100% (median 92.5%, interquartile range 10%) in 2010. Specificities

Table 2 The distribution of image grading for the 2010 round

	Standard								
Graders	RO	R1	M1	R2	R6	M2	R3	R4	
RO	77.6%	5.5%	0.0%	0.0%	12.3%	0.3%	0.0%	0.6%	1764
R1	14.4%	76.2%	22.2%	31.5%	5.2%	6.8%	3.1%	8.5%	2016
M1	0.1%	3.5%	65.5%	2.4%	0.0%	2.1%	0.2%	0.0%	252
R2	0.0%	0.5%	1.2%	23.8%	0.0%	0.5%	1.4%	0.6%	168
R6	5.2%	2.1%	0.4%	0.6%	80.2%	2.9%	8.8%	3.6%	252
M2	1.2%	7.5%	4.8%	19.0%	1.2%	70.1%	17.6%	15.7%	1848
R3	0.8%	2.3%	4.4%	20.2%	0.4%	11.2%	64.3%	9.6%	420
R4	0.8%	2.5%	1.6%	2.4%	0.8%	6.1%	4.5%	61.4%	840
	1522	**2122**	**280**	**74**	**457**	**1720**	**664**	**721**	

Each column represents the standard grading and each row represents the grade allocated by the graders. Percentages are by column (of the total image gradings associated with each standard grade) so columns add up to 100%. The leading diagonal indicates where the standard and graders are in agreement. The 4 × 4 top-right segment represents under-grading (i.e. standard referable but grader non-referable) and the 4 × 4 bottom-left segment represents over-grading (i.e. grader referable but standard non-referable). The bold figure at the end of each row and column shows the number of images in that category.

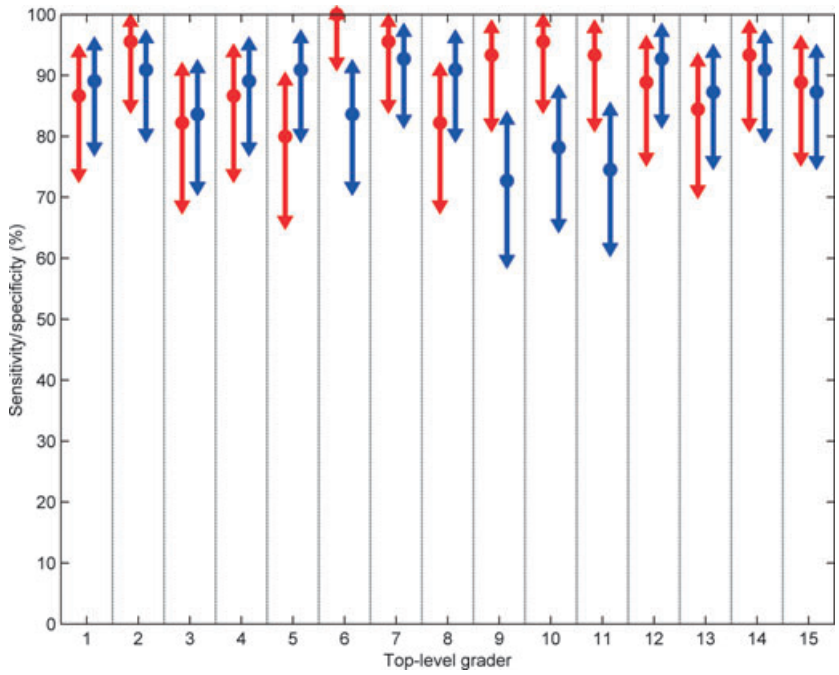

FIGURE 2 Sensitivities (left points) and specificity (right points) for the 15 top level graders who took part in the 2010 external quality assurance round. The arrows extend to the 95% confidence interval on the measurements.

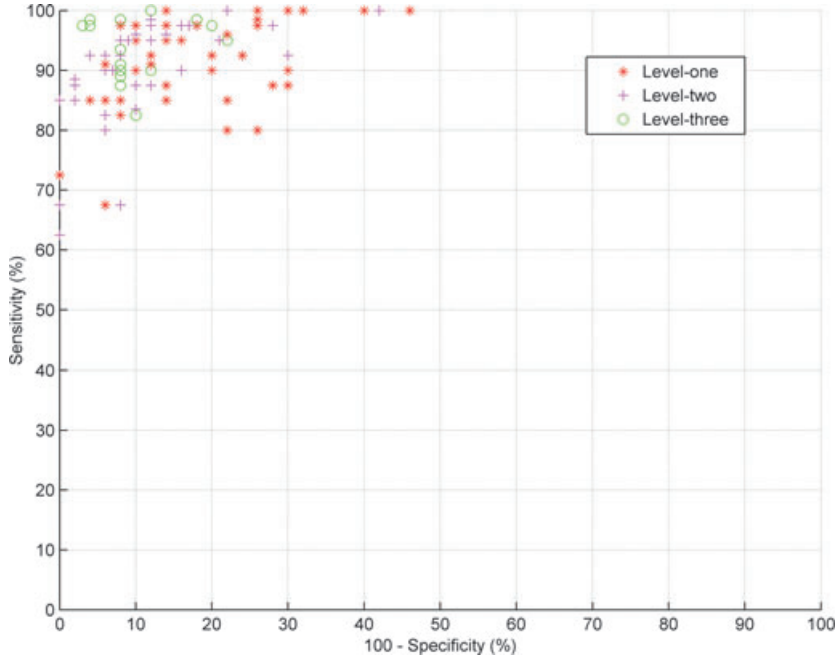

FIGURE 3 Receiver operating characteristic plot for detection of actionable images for each grader for the 2010 external quality assurance. Level-one graders are represented by asterisks (*), level-two graders by crosses (+) and level-three graders by circles (o). Note that jitter has been added to the data to reveal overlapping points.

ranged from 34.0 to 98.0% (median 85.0%, interquartile range 13.0%) in 2008, and 54.0 to 100% (median 88.0%, interquartile range 13.0%) in 2010. In 2010, 79/84 (94.0%) of graders had a sensitivity of at least 80%. 63/84 (75.0%) had a specificity of at least 80%, and 58/84 (69.1%) of graders had both a sensitivity and specificity of at least 80%. There was no significant

difference between the sensitivity distributions (P = 0.97) or specificity distributions (P = 0.093) between 2008 and 2010. In contrast, the inter-grader Kappa values were significantly higher in 2010 than 2008 (P = 0.015).

Are there differences between centres?

Table 3 lists the median centre sensitivity and specificities for 2008 and 2010. There were significant differences between the centre sensitivities in 2008 (P = 0.00016), but not between centre sensitivities in the 2010 round (P = 0.17). A multiple comparison test found the sensitivity of centre 4 in 2008 to be significantly lower (at the 5% level) than all the other centres, except centre 3. There were also differences between centre specificities in 2008 (P = 0.013), but not in 2010 (P = 0.20). A multiple comparison test found the specificity of centre 4 in 2008 was significantly higher (at the 5% level) than all centres except centres 1 and 8. There were no significant centre differences in Cohen's Kappa inter-grader agreement in either 2008 (P = 0.13) or 2010 (P = 0.055).

Are there differences between grader levels?

In 2010 there were 37 level-one graders, 32 level-two graders and 15 level-three graders. The median sensitivities and interquartile ranges for the level-one, -two and -three graders were 92.5% (12.5%), 92.5% (8.8%) and 95.0% (7.5%), respectively. There were no significant differences in sensitivities (P = 0.76). The specificities were 84.0% (16.0%), 90.0% (9.0%) and 92.0% (4.0%), respectively. There were significant differences in specificity between the three levels (P = 0.00017). A multiple comparison test found that the average specificity of the level-one graders was significantly lower than both the level-two and -three graders.

Discussion

Image grading EQA is one element of retinal screening quality assurance, complementing existing internal quality assurance. It can improve consistency between centres and allows graders to be compared using the same image set.

One of the challenges with this kind of exercise is creating a robust reference standard. Although the feature-based grading scheme appears an objective standard, a degree of subjectivity is involved mapping the continuous disease spectrum to discrete grades. For example, some of the criteria are based on the length of a 'disc diameter'. The optic disc is usually elliptical and the length varies depending on the angle and position of the measurement chord. The Scottish grading scheme mandates a vertical measurement for this reason. However, uncertainty still remains locating the precise edge of the disc. Often the uncertainty is compounded by also having to locate the fovea. Consequently, images having clear lesions approximately one disc diameter from the fovea are likely to generate disagreements, despite all the graders seeking to apply the criteria rigorously. Furthermore, the lesions themselves may be ambiguous. For example, it is often difficult to decide whether a red lesion within a disc diameter of the fovea is a dot/microaneurysm or a blot. If it is the latter, it is referable maculopathy, otherwise it is not. Exudates similarly give rise to uncertainty when they are small: is it instead a reflection artefact or small druse? Borderline, equivocal cases are not suitable for EQA; the test must distinguish inevitable uncertainty from clear grading errors.

One method for creating a reference grading is to convene a committee of experts to agree a grade for each image. Multiple observers should mean that little pathology is missed. However, there are problems with this approach. Firstly, selecting independent experts is not straightforward, as those with the most relevant experience are likely to be involved in screening and therefore should do the EQA themselves. Experts who are not involved in routine screening are likely, consciously or not, to apply slightly different criteria. Finally, the committee process is likely to disregard uncertainty, with the committee persuaded of 'correct' answers, even for genuinely uncertain images. In contrast, using the results from graders already taking part is easier and less expensive to set up than a separate committee. As

Table 3 Sensitivities and specificities for detecting referable images

	Autumn 2008		Spring 2010	
Centre	Sensitivity% (CI)	Specificity% (CI)	Sensitivity% (CI)	Specificity% (CI)
1	92.5 (86.4–96.0)	88.0 (81.8–92.3)	81.9 (75.2–87.1)	94.0 (89.8–96.5)
2	100.0 (96.9–100.0)	75.3 (67.9–81.5)	97.5 (92.9–99.1)	90.7 (84.9–94.4)
3	85.0 (79.9–89.0)	78.7 (73.7–82.9)	89.3 (86.1–91.9)	86.2 (83.0–88.8)
4	72.1 (66.1–77.4)	96.3 (93.6–97.9)	91.1 (87.2–93.9)	92.0 (88.7–94.4)
5	90.8 (88.5–92.6)	80.3 (77.7–82.6)	90.9 (89.2–92.3)	82.7 (80.8–84.5)
6	93.1 (88.1–96.1)	83.5 (77.7–88.0)	91.9 (86.6–95.2)	86.5 (81.1–90.6)
7	88.1 (82.2–92.3)	85.5 (80.0–89.7)	92.5 (88.5–95.2)	85.7 (81.2–89.2)
8	91.7 (87.5–94.5)	87.7 (83.5–90.9)	95.0 (91.5–97.1)	87.7 (83.5–90.9)
9	97.8 (95.6–98.9)	80.0 (75.8–83.6)	89.5 (86.3–92.1)	85.3 (82.1–88.0)
Mean (SD)	90.1 (8.1)	83.9 (6.3)	91.1 (4.3)	87.8 (3.6)

The values in parentheses indicate the 95% confidence intervals (CIs) for the measurements.
The number of graders at each centre has not been included to preserve anonymity. The smallest centre had three graders and the largest 32 graders.

participants grade independently of each other, it is possible to determine the genuinely equivocal images. This is the approach taken in this study, where the results from the top-level graders are used to determine the reference standard. Although in this study all the graders took part at the same time, the top-level graders could do the EQA earlier to establish the consensus standard before the other graders take part. This would allow more rapid feedback and prevent unnecessary inclusion of equivocal images.

It has been suggested that the reference standard in exercises like this should be seven-field stereo photography or final clinical outcomes [10]. However, the goal of the exercise is not to determine the sensitivity and specificity of the overall grading process (that has been carried out elsewhere [11–13]), but rather compare graders' ability to detect diabetic retinopathy in standard screening photographs.

There are several measures of agreement that could be used. For instance, given a standard grading for each image, one could calculate the number of images that agree exactly with the standard grade. However, exact agreement between the eight categories is clinically overly complicated, as more than one grade maps to a given care pathway; for example, both R0 and R1 result in a 12-month recall. An alternative is to score differences between graders based on weighted features [14]. However, while interesting to see which parts of the grading system are problematic, the primary purpose of screening is to identify patients requiring referral. The assessment here therefore uses the binary decision of whether or not graders correctly identify referable images.

Presenting the results in terms of sensitivity and specificity provides a clearer picture of how graders are performing than using a single metric, such as Cohen's Kappa. The sensitivity and specificity represent the trade-off between false negatives and false positives. While the Kappa score usefully indicated a significant improvement in centre agreement between 2008 and 2010, alone it is unable to differentiate centres with low sensitivity and high specificity from those with high sensitivity and low specificity.

Different levels of grader have different degrees of experience and different roles. The study investigated whether these differences affected their sensitivity and specificity. Although level-one graders were as sensitive as level-two and -three graders, they were less specific than the other levels. This is to be expected given their role as disease/no disease graders, and confirms that they are performing their role safely.

EQA should be an educational tool for graders. After completing the test, graders may login to compare their performance with other anonymous graders, and to step through their grading, comparing it with the consensus. As the grading is feature-based, they can see why an image has been given a particular grade. Any feature where the grader's selection does not match the consensus is highlighted, regardless of whether the image was correctly referred. By recording the use of image controls, such as red-free, it is possible to identify patterns of use in higher-performing graders, which can advise good practice. In 2008, centre 4 had a lower average sensitivity and a higher than average specificity. In each of the three categories of referable images (R3/R4, M2 and R6) this centre had one of the lowest detection rates, but in no single category was it significantly lower than all the other centres. At this centre the graders had set a high detection threshold, particularly in the case of technical failures: not a single false-positive technical failure was graded at this centre.

Clear and open feedback from the tests also increases confidence in the results. Both the committee- and consensus-derived standards can contain errors. The committee standard is more likely to include dogmatic grades for equivocal images, while the consensus standard is more likely to miss difficult features. This latter problem is easily dealt with by making the grading standard and results open to all graders. In this way, results may be discussed and, if necessary, corrected.

A relatively small test set is possible, compared with routine internal quality assurance, as the images were chosen with a tenfold higher proportion of referable disease than that in routine screening. Graders know they are under test conditions and, realizing the higher disease prevalence, may be more likely to mark disease features as present. The advantage of presenting the results on a receiver operating characteristic plot is that such shifts in grading will tend to produce a concomitant reduction in specificity. If the selected images are representative of referable and non-referable images, then this provides a good estimate of true grading performance. Eliminating equivocal examples should make the grading more representative of true performance.

In conclusion, a novel web-based EQA external quality assurance grading tool has been developed and tested using a consensus reference standard to compare grading centres in Scotland. There was closer agreement between the centres in the second round, suggesting that grader education and training is as important a role for the system as testing in order to ensure equivalent standards among grading centres.

Competing interests

Nothing to declare.

Acknowledgements

Funding for the project was provided by the Scottish Diabetic Retinopathy Screening Collaborative. The authors are grateful to Dr Gordon Prescott for statistical advice and to all the graders who provided helpful feedback after taking part in the EQA rounds.

References

1 Facey K, Cummins E, Macpherson K, Morris A, Reay L, Slattery J. *Health Technology Assessment Report 1: Organisation of Services for Diabetic Retinopathy Screening 2002*. Available at http://www.ndrs.scot.nhs.uk/Links/Docs/hta1.pdf Last accessed 1 November 2011.

2 Scottish Diabetic Retinal Screening Collaborative. *Annual Report 2009/2010*. Available at http://www.ndrs.scot.nhs.uk/Templates/DRS%20Collaborative%20Annual%20Report%202009-10.pdf Last accessed 1 November 2011.
3 NHS Quality Improvement Scotland. *Diabetic Retinopathy Screening March 2004*. Available at http://www.healthcare improvementscotland.org/default.aspx?page=12345 Last accessed 1 November 2011.
4 Slattery J. Sampling for quality assurance of grading decisions in diabetic retinopathy screening: designing the system to detect errors. *Int J Health Care Qual Assur* 2005; **18**: 113–122.
5 Pandit RJ, Taylor R. Quality assurance in screening for sight-threatening diabetic retinopathy. *Diabet Med* 2002; **19**: 285–291.
6 Arun CS, Young D, Batey D, Shotton M, Mitchie D, Stannard KP *et al*. Establishing ongoing quality assurance in a retinal screening programme. *Diabet Med* 2006; **23**: 629–634.
7 Patra S, Gomm EMW, Macipe M, Bailey C. Inter-observer agreement between primary graders and an expert grader in the Bristol and Weston diabetic retinopathy screening programme: a quality assurance audit. *Diabet Med* 2009; **26**: 820–823.
8 NHS Cervical Screening Programme. *External Quality Assessment Scheme for Gynaecological Cytopathology 2009*. Available at http://www.cancerscreening.nhs.uk/cervical/publications/nhscsp15.pdf Last accessed 1 November 2011
9 Scott HJ, Gale AG. Breast screening: PERFORMS identifies key mammographic training needs. *Br J Radiol* 2006; **79**: S127–S133.
10 Garvican L. Issues regarding quality assurance in the English National Screening Programme for Sight-Threatening Diabetic Retinopathy: response to paper by C. *Arun. Diabet Med* 2007; **24**: 688–690.
11 Harding S, Broadbent D, Neoh C, White M. Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight-threatening eye disease: the Liverpool Diabetic Eye Study. *Br Med J* 1995; **311**: 1131–1135.
12 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**: 528–534.
13 Scanlon PH, Malhotra R, Thomas G, Foy C, Kirkpatrick JN, Lewis-Barned N *et al*. The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy. *Diabet Med* 2003; **20**: 467–474.
14 Schneider S, Aldington SJ, Kohner EM, Luzio S, Owens DR, Schmidt V *et al*. Quality assurance for diabetic retinopathy telescreening. *Diabet Med* 2005; **22**: 794–802.
15 Scottish Diabetic Retinal Screening Collaborative. *Scottish Diabetic Retinopathy Grading Scheme 2007*. Available at http://www.ndrs.scot.nhs.uk/ClinGrp/Docs/Grading%20Scheme%202007%20v1.1.pdf Last accessed 1 November 2011.
16 UK National Screening Committee. *Essential Elements in Developing a Diabetic Retinal Screening Programme. Workbook version 4.3. 2009*. Available at http://www.retinalscreening.nhs.uk/userFiles/File/DiabeticRetinopathyScreeningWorkbookRelease43_2009-06-23.pdf Last accessed 1 November 2011.
17 Diabetic Retinopathy Screening Service for Wales (DRSSW). *DRSSW Grading Protocol*. Available at http://www.wales.nhs.uk/sites3/page.cfm?orgid=562&pid=25081 Last accessed 1 November 2011.