Training evaluation of a course in diabetic retinopathy screening

R Pauli, JW Huber, G Duncan, KP Shotliff*

Abstract

The success and effectiveness of diabetic screening programmes are dependent on the availability of appropriately trained image graders. This study was designed to evaluate graders enrolled on a locally devised, formal training course by means of a performance-based measure. The course consisted of four days of classroom-based tuition followed by three months of practice-based learning in the workplace. The aim was to establish whether trainees showed an improvement in their ability to grade images, and secondly whether test sets of images are useful in measuring training outcome. Thirteen trainees were required to grade a test set of 24 single images both before and after training. A significant improvement in sensitivity (from 35% before training to 45% after training) was observed as a result of training but at a cost of a decline in specificity. Trainees' confidence ratings measured on a five-point scale increased from an average of 2.4 to 4.1 ($p<0.01$). We concluded that the course needs to focus more on trainees' ability to discriminate between normal and abnormal images as well as improving grading accuracy in line with increased grading confidence. Test-based course evaluation can be seen to be a valuable instrument in establishing a quality standard for stated learning outcomes. In this research it has clearly indicated weaknesses of the training programme in its current form.

Eur Diabetes Nursing 2005; 2(2): 58–62.

Key words

diabetic retinopathy screening; training evaluation; quality assurance

Introduction

Diabetic retinopathy remains the most common cause of blindness in the working population of the Western world and fulfils the World Health Organization criteria for screening.[1] Using digital retinal photography, as recommended by the National Screening Committee in the UK and the American Diabetes Association, it is possible to detect sight threatening retinopathy at an early and potentially treatable stage. Treatment using laser therapy has been shown to reduce the progression of this complication of diabetes and so reduce blindness in a significant proportion of patients.

The National Service Framework for Diabetes in England and Wales (NSF)[2] and the subsequent National Screening Committee (NSC)[3] guidelines set clear targets for retinal screening and grading of the images obtained for all people with diabetes. An important factor in establishing and maintaining these targets is effective mechanisms for quality assurance. Whilst the need for evaluating actual screening programmes is well recognised,[4,5] training programmes for retinopathy screeners have been less scrutinised. Evaluation of screening performance at the level of the individual screener has been considered useful in other health screening programmes involving medical image interpretation. For breast cancer screening, a self-assessment scheme for radiologists and other film readers has been available in the UK for some time now.[6] This employs a test set of medical images which contains approximately half normal and abnormal cases and is used to evaluate both training outcome and performance of established screeners at regular intervals.

Training for retinopathy screening is currently organised at a local level, although there are efforts to standardise this through the development of a competence framework for education and training curricula within the UK.[7] From a quality assurance perspective, it is essential that training programmes are effective in enabling students to contribute confidently and accurately to everyday screening decisions. Whilst there are currently no standard criteria for levels of sensitivity and specificity which a screening programme should achieve, the National Institute for Clinical Excellence (NICE) has issued guidelines recommending that retinal screening should aim to achieve sensitivity above 80% and specificity above 95%. However, the evidence

Authors

R Pauli, PhD, Senior Lecturer
JW Huber, PhD, Principal Lecturer
School of Human and Life Sciences, Roehampton University, London, UK
G Duncan, PGC, Clinical Manager
Diabetic Retinopathy Screening Service, Mayday Healthcare NHS Trust, Croydon, UK
KP Shotliff, MD, FRCP, DCH, Consultant Physician, Chelsea & Westminster Hospital, London, UK

***Correspondence to:** R Pauli, School of Human and Life Sciences, Roehampton University, Holybourne Avenue, London SW15 4DJ, UK; e-mail: R.Pauli@roehampton.ac.uk

Received: 27 June 2005
Accepted in revised form: 11 August 2005

base reviewed shows that, in practice, actual levels of sensitivity and specificity are rather variable depending on both the screening methodology used and the grader background.[7]

In this paper we are presenting an evaluation of training outcome of a locally devised and delivered course in retinopathy screening. The course consists of a four-day course of lectures delivered in a hospital setting by two of the authors followed by a three-month distance learning period embedded in a retinopathy screening workplace. A rolling programme of courses is run, on average consisting of 12 students. The emphasis of the course is on actual screening of digital retinal images. It is aimed at health care professionals, from primary and secondary care, involved in or seeking to become involved in the delivery of retinal screening programmes. While this course includes material on the theoretical background to diabetic retinopathy, its primary focus is on teaching the image grading aspects of screening. For this reason, it concentrates on retinal photography with training in the grading of retinal disease. Theoretical learning outcomes are assessed through an essay. This includes knowledge of the results of the United Kingdom Prospective Diabetes Study (UKPDS) and other relevant studies. The practical skills in grading images are assessed through a grading exercise which requires students to grade images according to the NSC screening protocol. It is this practical element of the assessment that is under scrutiny in this paper. The use of a structured grading protocol for digital retinal photographs (Table 1) has been advocated for many years following publication of the Airlie House classification and has been reviewed again in the UK in the most recent NSC guidelines.[3] Retinal photographers and graders are now expected to use these criteria to grade digital images and to decide whether the person with diabetes should be screened annually or referred to an ophthalmologist, as well as the speed of that referral. The trainees on this course were instructed in the use of this grading scheme and trained to grade in accordance with it both during the four-day training and the three-month work placement.

Level R0 = No diabetic retinopathy
Level R1 = Background The following lesions >2 disc diameters (2DD) from fovea: • Microaneurysms • Retinal haemorrhages • ± hard exudates
Level R2 = Pre-proliferative The following lesions: • Venous beading • Venous loop or reduplication • Intraretinal microvascular abnormalities (IRMA) • Multiple deep, round or blot haemorrhages Cotton wool spots (CWS) careful search for above lesions of R2
Level R3 = Proliferative The following lesions: • New vessels on disc (NVD) • New vessels elsewhere (NVE) • Pre-retinal or vitreous haemorrhage(s) • Pre-retinal fibrosis ± tractional retinal detachment
Level M = Maculopathy The following lesions: • Hard exudates within 1DD of the fovea • Circinate or group exudates in the macula • Retinal thickening within 1DD of fovea • Any microaneurysms or haemorrhages within 1DD of the centre of the fovea only if associated with a best visual acuity ≤6/12
Level M (old system) Any of the above M grade lesions within 1DD and visual acuity better than 6/12
OTHER GRADES P = Evidence of previous laser treatment U = Ungradable or unobtainable images OL = Other non-diabetic eye disease

Table 1. Grading scheme for diabetic retinopathy

Aims

This study aimed to evaluate the effectiveness of the training programme in enabling trainees to classify retinal images correctly according to the standard grading scheme published by the NSC (see Table 1). Grading performance can be measured in terms of sensitivity, the probability of a true positive classification and specificity, and the probability of obtaining a true negative classification (usually these are expressed as percentage accura-

Grade	No. of images	Sensitivity pre-test (mean, SD)	Sensitivity post-test (mean, SD)	Sig. z-value*
R1	7	46 (27)	54 (13)	0.20
R2	6	19 (19)	44 (24)	**0.03**
R3	3	31 (25)	38 (23)	0.37
M	6	42 (34)	45 (33)	0.93
Overall sensitivity	22	35 (16)	45 (14)	**0.01**

*Based on Wilcoxon Signed Rank tests.

Table 2. Sensitivity in % for pre- and post-training test

cies). Traditionally, training evaluation has focused on trainees' perception of the delivery and value of the course. However, in the context of quality assurance of training courses it is more useful to measure actual improvement in performance as this reflects the long-term value of the course for the trainee as well as the stated learning outcomes more accurately. Ideally, training programmes should equip trainees to produce consistent levels of specificity and sensitivity independent of background and previous experience. Furthermore, the proposed performance evaluation can be used in a formative way to guide the development and improvement of the training course. A second aim was to assess the suitability of a test set of images for the purpose of training evaluation in the context of diabetic retinopathy screening. Whilst the validity of such test sets has been established in other screening programmes such as breast cancer screening,[8] use of a test set to evaluate retinopathy grading performance remains to be established.

Design and participants

The evaluation study employed a repeated measures design. Baseline grading performance was assessed at the start of the training course and measured again using the same images in different random order at the conclusion of training three months later. One training group consisting of 13 health professionals participated in the study. Participants varied considerably with respect to their previous experience of retinal screening, ranging from fairly experienced screeners to no previous experience of retinal screening at all.

Materials and procedure

Baseline and post-training performance were measured with a set of 24 single eye images from patients with known diagnosis based on the judgement of two independent experts. Twenty-two of the images presented with varying degrees of retinopathy, seven were classified as 'R1', six as 'R2', three as 'R3' and six as 'M' according to the grading scheme (Table 1). Two images were classified as normal ('R0'). Participants individually viewed each image on a 15-inch cathode ray tube screen. Grading was self-paced but an overall time limit of 75 minutes was imposed. Participants were required to record their observations on a standardised grading sheet on which they had to indicate the number of microaneurysms, presence of any other lesions and the diabetic retinopathy grade. In addition, they were required to provide a rating of their level of confidence with respect to each of these observations on a five-point scale ranging from 'very confident' to 'not at all confident'. Data collection for this study was integral to the course assessment and all data presented here were provided by participants in the context of their course assessment.

Results

Specificity and sensitivity were assessed for both baseline and post-training performance. Analysis was conducted using Wilcoxon Signed Rank tests for all pre- and post-training comparisons. Incomplete gradings were treated as missing data and omitted from the analysis. Table 2 shows sensitivity broken down into the four abnormal grading categories before and after training. Only one of the grading categories, R2, demonstrated a significant improvement in sensitivity as a result of training ($p<0.03$). Misinterpretations on intraretinal microvascular abnormalities (IRMAs) as new vessels and drusen as hard exudates were often noted at the initial assessment. All other grading categories yielded higher sensitivity in the post-training test than the pre-training test but these differences were not statistically significant. However, overall a significant improvement in sensitivity was observed in the post-training test ($p<0.01$). Specificity was calculated from the 'R0' category and was shown to decrease significantly from 69% (SD=33) before training to 39% (SD=22) after training ($p<0.01$). However, given the small number of normal images in this test set, these results must be interpreted with caution.

We further examined the accuracy of microaneurysm and other lesion identification. On the abnormal images, identification of other lesions increased from 69% (SD=12) to 83% (SD=11) ($p=0.004$). However, a decrease in accuracy was observed for micro-

aneurysm counts (37%, SD=14 before training; 31%, SD=9 after training; $p<0.05$).

Table 3 shows the mean confidence ratings for all image categories. All comparisons before and after training are highly significant at $p<0.01$.

We further examined the correlations between pre- and post-training confidence ratings which range from -0.17 for R3 to 0.39 for R0, suggesting individual differences in the way confidence ratings change in response to the course.

Grade	No. of images	Confidence rating pre-test*	Confidence rating post-test*
R0	2	2.9	4.4
R1	7	2.4	4.1
R2	6	2.2	4.0
R3	3	2.1	3.8
M	6	2.2	4.1

*Standard deviations are around 1.0 (min 0.6 to max 1.4). All differences are significant at $p<0.01$.

Table 3. Mean confidence ratings before and after training

Discussion

Overall performance at the end of the course falls some way short of the sensitivity and specificity recommendations put forward by NICE.[6] Although improvement is relatively weak overall in comparison to the baseline, there is a significant improvement both for the R2 category and for all abnormal categories combined. Significant improvement on these R2 images is especially important from a clinical perspective. The mean sensitivity reported for our trainees is in line with the range of sensitivity values reported for graders of similar background using single images in other studies.[8,9] Whilst it is clear that there is significant room for improvement in terms of training outcome, it needs to be acknowledged that the course was not devised to enable trainees to grade autonomously after completion, but that the training concept envisages at least another six months of supervised workplace grading before autonomous grading is recommended.

The observed decline in specificity is also of concern in this study, but may well result from a combination of factors, such as too small a sample of normal images (n=2) in the test set which resulted in false expectations of graders who may have assumed that the test set consists of abnormal images only, given the emphasis on recognising abnormality in the course. Experience with test set assessment in, for example, breast cancer screening has shown that observers have expectations about baselines under test conditions.[10] Review of the course delivery and content needs to consider whether more emphasis should be placed on discriminating normal and abnormal images as this appeared to be a real weakness in current trainees' performance.

Confidence ratings increased significantly in this study which is in contrast to actual performance when taking into account both specificity and sensitivity. Increases in confidence are frequently found as a result of training, and are also used as proxi-variables for actual performance. However, in the case of retinal screening, confidence ratings without performance data would give a biased picture of performance improvement. Course satisfaction surveys, another commonly used training outcome indicator, would be more reflective of inflated confidence ratings than actual image grading performance. These may, therefore, not provide accurate reflections of training quality when improvement in image grading performance is specified as the main learning outcome. In order to ensure that quality control of training courses is implemented and maintained appropriately (i.e. referenced to the stated learning outcomes) it is clearly very important to incorporate a performance-based evaluation.

The results of this study suggest that appropriate testing of specificity would necessitate development of a test set with an increased number of normal images to reflect the proportion of these in the screening population. An alternative is to build a test set with approximately half normal and half abnormal images which would also facilitate more sophisticated performance analysis based on ROC analysis and can be used to combine grading accuracy with confidence ratings into a single performance index.[11] We are currently devising a new test set of images which will serve to address these issues in our research.

A further issue that needs to be considered in future research is that heterogeneity of trainees may have skewed improvement measures negatively in this study. Trainees who have considerable previous grading experience may not show much improvement as a result of attending the course. Therefore lack of significance in mean performance before and after the course does not necessarily reflect on the quality of the course unless such individual differences are accounted for. This problem may be further exacerbated by the small sample available

for this study, which serves to increase the impact of possible individual differences.

In this study some students were not able to complete the grading task in the given time. The focus on currently running training programmes is on completing the grading of the entire image set within the given time frame. This is also important as the NSF requirements will lead to large workloads for future screeners.

Finally, it needs to be acknowledged that the relatively small sample of both trainees and images available for this study makes the statistical analysis reported somewhat vulnerable to error, particularly with respect to specificity. The research reported in this paper was conceived as a pilot study and our conclusions need to be confirmed by replication with further cohorts of trainees.

References

1. Diabetes care and research in Europe: The St Vincent Declaration. *Diabetic Med* 1990; **7:** 360.
2. National Service Framework Diabetes (NSF Diabetes). http://www.doh.gov.uk/NSF/diabetes
3. National Screening Committee (NCS). http://www.nscretinopathy.org.uk
4. Pandit RJ, Taylor R. Quality assurance in screening for sight-threatening diabetic retinopathy. *Diabetic Med* 2002; **19:** 285–291.
5. Garvican L, Scanlon PH. A pilot quality assurance scheme for diabetic retinopathy risk reduction programmes. *Diabetic Med* 2004; **21:** 1066–1074.
6. Gale AG, Savage CJ, Pawley EF, *et al.* Breast Cancer Screening: visual search and observer performance. In Kundel HL (ed). *Medical Imaging 1994: Image Perception.* Proceedings of The International Society for Optical Engineering (SPIE) 2166, 1994; 66–75.
7. National Institute for Clinical Excellence. *Management of Type 2 Diabetes. Retinopathy and Early Management.* London: NICE, 2002.
8. Pugh JA, Jacobson JM, van Heuven WAJ, *et al.* Screening for diabetic retinopathy. The wide-angle retinal camera. *Diabetes Care* 1993; **16:** 889–895.
9. Buxton MJ, Sculpher MJ, Ferguson BA, *et al.* Screening for treatable diabetic retinopathy: a comparison of different methods. *Diabetic Med* 1991; **8:** 371–377.
10. Pauli R, Hammond S, Cooke J, *et al.* Radiographers as film readers in screening mammography: an assessment of competence under test and screening conditions. *Br J Rad* 1996; **69:** 10–14.
11. Hammond SM, Davies IR. Indices of Performance. In Kundel HL (ed). *Medical Imaging 1994: Image Perception.* Proceedings of The International Society for Optical Engineering (SPIE) 2166, 1994; 170–178.