EXTENDED REPORT

Incremental nature of anterior eye grading scales determined by objective image analysis

J S Wolffsohn

Br J Ophthalmol 2004;**88**:1434–1438. doi: 10.1136/bjo.2004.045534

Correspondence to:
Dr J S Wolffsohn,
Neurosciences Research Institute, Aston University, Aston Triangle, Birmingham B4 7ET, UK;
j.s.w.Wolffsohn@aston.ac.uk

Accepted 30 March 2004

Aim: To use previously validated image analysis techniques to determine the incremental nature of printed subjective anterior eye grading scales.
Methods: A purpose designed computer program was written to detect edges using a 3×3 kernal and to extract colour planes in the selected area of an image. Annunziato and Efron pictorial, and CCLRU and Vistakon-Synoptik photographic grades of bulbar hyperaemia, palpebral hyperaemia roughness, and corneal staining were analysed.
Results: The increments of the grading scales were best described by a quadratic rather than a linear function. Edge detection and colour extraction image analysis for bulbar hyperaemia ($r^2 = 0.35-0.99$), palpebral hyperaemia ($r^2 = 0.71-0.99$), palpebral roughness ($r^2 = 0.30-0.94$), and corneal staining ($r^2 = 0.57-0.99$) correlated well with scale grades, although the increments varied in magnitude and direction between different scales. Repeated image analysis measures had a 95% confidence interval of between 0.02 (colour extraction) and 0.10 (edge detection) scale units (on a 0–4 scale).
Conclusion: The printed grading scales were more sensitive for grading features of low severity, but grades were not comparable between grading scales. Palpebral hyperaemia and staining grading is complicated by the variable presentations possible. Image analysis techniques are 6–35 times more repeatable than subjective grading, with a sensitivity of 1.2–2.8% of the scale.

Grading scales are well established as aids to assist in the monitoring of anterior eye characteristics. They require a given ocular feature to be gauged relative to predetermined images chosen to represent different degrees of the condition of interest on an ordinal scale. Such scales vary in the number of images and conditions of interest and can be descriptive,[1 2] artistically rendered,[3] photographic,[4–6] or computer generated.[7] However, even with the use of a grading scale, there is a wide discrepancy between observers grading the same image and on repeat grading by the same observer.[8–10] Interpolating between grading images (such as to one tenth of a unit) increases discrimination,[11–13] but relies on a linear incremental increase in severity between grades.

More recently computerised image analysis techniques have been used for grading anterior eye characteristics. Different studies have used a combination of thresholding,[8 14–18] edge detection,[14 19 20] smoothing,[8 14 19 21] colour extraction,[8 15 18 21] and morphometry and densiometry[22] to grade bulbar hyperaemia. These image analysis techniques have been used to try to determine how clinicians grade bulbar hyperaemia, with one study suggesting that the number of vessels and the proportion of the image occupied by vessels are more important than relative colouration[15] whereas another indicated both these factors were integral to grading.[8] However, the correlation between the computer image analysis techniques used and clinician grading was not linear, and was more discrepant for higher grades of bulbar hyperaemia.[8] Less research has been conducted on the objective grading of palpebral hyperaemia and corneal staining, although it has been noted that there are significant differences between observers in subjective grading of these features.[23 24] Doughty and colleagues have examined palpebral roughness by measuring the size of fluorescein highlighted features[25 26] and Miyata and colleagues assessed staining severity using anterior fluorophotometry.[27] Wolffsohn and Purslow[28] examined the range of different image analysis methods used previously and showed that colour extraction and edge detection using a 3×3 kernal were the most repeatable and robust to changes in image luminance for bulbar and palpebral hyperaemia and fluorescein staining.

The objective computer image analysis grading techniques used in these studies have not yet been generally used in clinical practice. Although computers and slit lamp biomicroscope cameras are becoming more common in hospitals and eye care practices, printed grading scales have the advantage of being inexpensive and portable. Therefore this study aimed to determine, by objective image analysis, whether commonly used clinical pictorial and photographic subjective grading scales are incremental in nature.[9]

METHODS

A purpose designed computer program was written (Labview and Vision Software, National Instruments, Austin, TX, USA) to objectively quantify changes in ocular physiology from stored image files of the anterior eye, using the techniques previously found to be most repeatable and robust to changes in luminance (fig 1):[28]

- Colour extraction: the relative intensity of the red, green, and blue colour planes was extracted and the ratio of red (for hyperaemia) or green (for staining) to overall intensity calculated.
- Edge detection: each pixel was compared to its neighbours and spatial filters used to alter pixel values with respect to variations in light intensity of their neighbourhood. Non-linear Sobel and Robert 3×3 filter kernels were used as previously identified.[28] The number of pixels highlighted by a $^{15}/_{256}$ greyscale threshold was divided by the total area to give the percentage of area detected as containing edges.

To examine the incremental nature of the Annunziato,[29] Efron (Millennium Edition),[9] and Vistakon-Synoptik[4] grading scales, the printed images of bulbar hyperaemia,

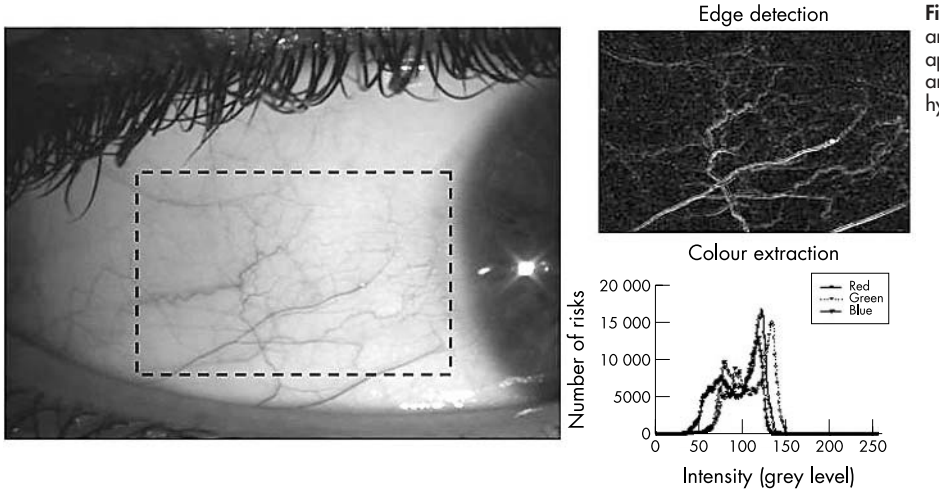

Figure 1 Example of edge detection and colour extraction techniques applied to a selected (dashed rectangle) area of an image of bulbar hyperaemia.

palpebral hyperaemia (also referred to as papillary conjunctivitis), and corneal staining extent were scanned at 600 dpi stored in tagged image format (TIFF) and analysed. Vistakon-Synoptik palpebral conjunctivitis images were analysed selecting the palpebral hyperaemia and the area with reflections separately to distinguish between hyperaemia and roughness. Original 700×525 pixel JPEG images of bulbar hyperaemia, lid redness, and roughness (white light and sodium fluorescein) and corneal staining extent from the CCLRU grading scale were analysed. Compression of a non-glossy TIFF image into the high quality JPEG format of the CCLRU grading scale was found to not significantly affect the image analysis techniques used. An area of approximately 90 000 pixels (about 6.0 mm^2) covering the area of interest was outlined manually three times for each scale grade.

Statistical analysis

Analysis of variance was used to examine overall effects and Tukey's pairwise multiple comparison test to assess individual differences between scale grades. The results were fitted using linear ($y = mx+c$) and quadratic ($y = ax^2+bx+c$) functions, with the variance assessed by Pearson's Product Moment Correlation. Image analysis discrimination was described by the standard deviation of repeated measures divided by the scale range.

RESULTS

Bulbar hyperaemia, palpebral hyperaemia roughness, and corneal staining grade images were best described by a quadratic rather than linear or other curve fitting functions (table 1). Edge detection and red colouration significantly differed with increasing bulbar hyperaemia scale grades ($p<0.001$), although for the Annunziato and Vistakon-Synoptik scales, the edge detection increments were smaller between higher grades (fig 2).

Red colouration increased with increasing palpebral hyperaemia scale grades ($p<0.001$), although the increments between grades were less regular with photographic scale (CCLRU and Vistakon-Synoptik) grades. However, for palpebral hyperaemia edges detected increased with the Efron scale ($F = 131.0$, $p<0.001$), decreased with the CCLRU scale ($F = 1.6\times10^4$, $p<0.001$) and despite varying between grades, did not progress incrementally in the Annunziato ($F = 306.5$, $p<0.001$) and Vistakon-Synoptik ($F = 49.4$, $p<0.001$) scales (fig 3).

Palpebral roughness in photographic scales depicted by reflections (CCLRU and Vistakon-Synoptik) showed a general increase in edges detected and red colouration with

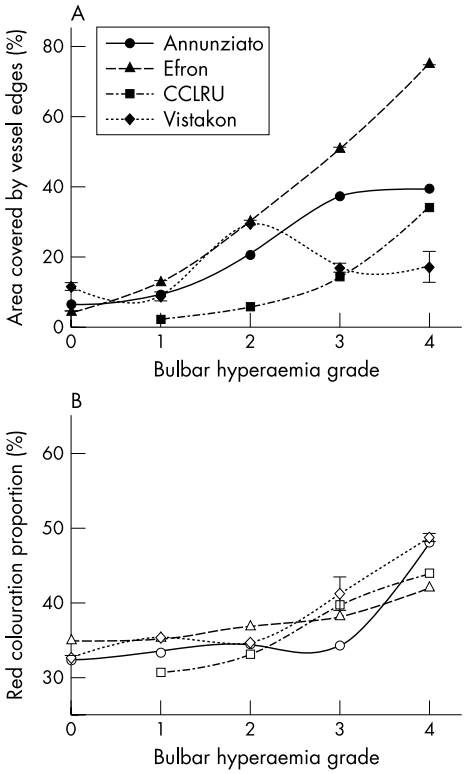

Figure 2 Edge detection (filled symbols) and relative red colouration (open symbols) with grading scale images of bulbar hyperaemia. Error bars = 1 SD.

increasing scale grades, although the increments between grades were non-uniform ($p<0.001$). Palpebral roughness depicted by fluorescein staining viewed with cobalt blue illumination through a Wratten filter (CCLRU), resulted in an increase in edges detected ($F = 264.2$, $p<0.001$) and a decrease in green colouration ($F = 778.9$, $p<0.001$) with increasing scale grade, although the highest grade shows an apparent decrease in severity (fig 4).

Green colouration increased with increasing corneal staining scale grade for the Annunziato ($F = 5763.8$, $p<0.001$) and Efron ($F = 1.3\times10^4$, $p<0.001$) scales, but decreased with the CCLRU scale ($F = 306.5$, $p<0.001$) and did not progress incrementally with the Vistakon-Synoptik scale ($F = 665.9$,

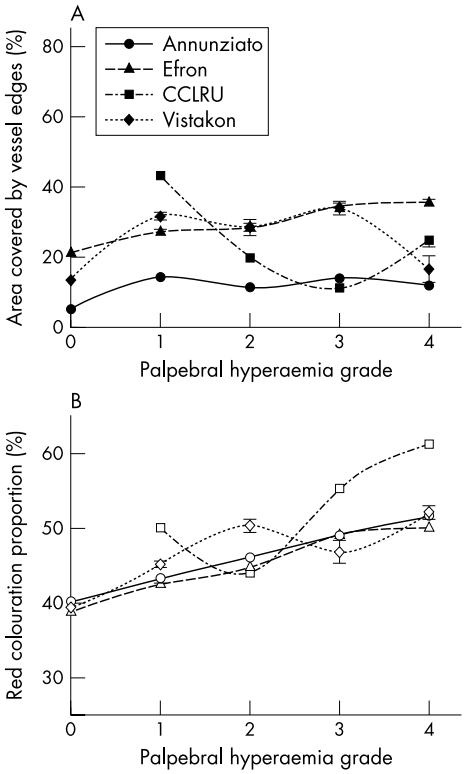

Figure 3 Edge detection (filled symbols) and relative red colouration (open symbols) with grading scale images of palpebral hyperaemia. Error bars = 1 SD.

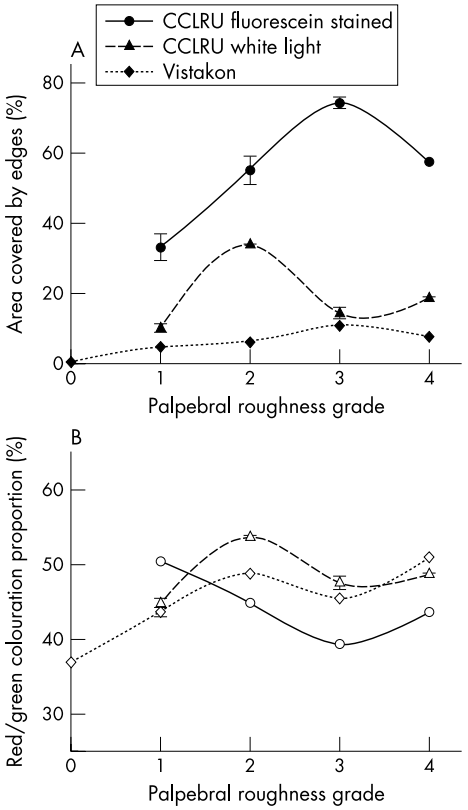

Figure 4 Edge detection (filled symbols) and relative red/green colouration (open symbols) with grading scale images of palpebral roughness. Error bars = 1 SD.

$p<0.001$). Edges detection showed a general increase with increasing corneal staining grade ($p<0.001$), although there was not a systematic incremental change (fig 5).

The variability between repeat highlighting of the bulbar conjunctiva, palpebral conjunctiva and corneal area was generally small (table 2). There was no significant difference in discrimination between edge detection and colour extraction for each of the grading scales (2.8 (SD 3.8)% *v* 1.2 (SD 2.5)%, $p = 0.15$).

DISCUSSION

Validated image analysis techniques of edge detection and colour extraction showed that bulbar hyperaemia, palpebral hyperaemia roughness, and corneal staining grade images were quadratic rather than linear in nature. This results in the lower end of the scale being more sensitive (a smaller change between grades) than the upper end of the scale. As most eyes only have minimal hyperaemia and corneal staining[14 23 24] this approach to grading scale design could be considered appropriate, but could lead to errors if clinicians interpolate between scale grade images to improve sensitivity.[11–13] For example, if a clinician decides an eye had bulbar hyperaemia halfway between grade 0 and grade 1 on the Efron Millennium edition grading scale, they would note a grade of 0.50, whereas the grade identified by image analysis is 0.48 (using the quadratic fit of scale grade [x] against edge detection $[y] = 2.5x^2+8.0x+3.9$, $r^2 = 0.99$). Obviously the difference is only slight and could not be considered of clinical significance. However if the linear nature of grades 0–4 was followed by the clinician ($y = 20.6x-9.4$, $r^2 = 0.99$) the interpolated subjective grade would be 0.84. Although the difference is again within the variability of clinical grading[7] and relative change will govern clinical decision making, individual grading strategies will increase the variance between individuals, and hence

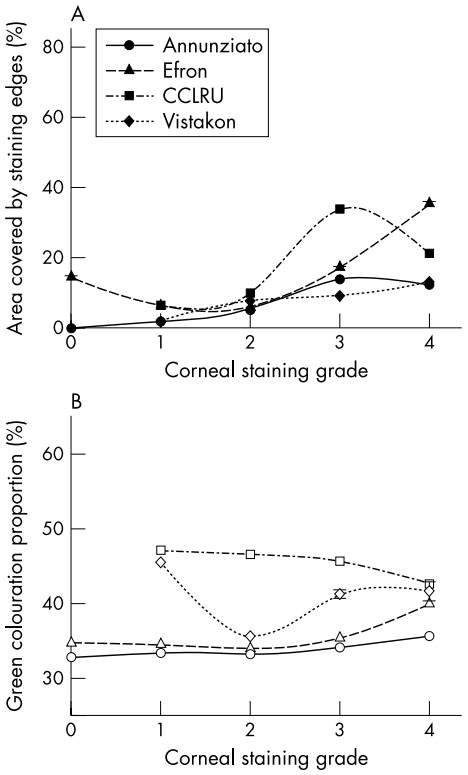

Figure 5 Edge detection (filled symbols) and relative green colouration (closed symbols) with grading scale images of corneal staining. Error bars = 1 SD.

Table 1 Correlation (r^2) between image grade and edge detection or relative colouration image analysis for the Annunziato, Efron, CCLRU, and Vistakon-Synoptik grading scales

		Linear correlation		Quadratic correlation	
		Edge detection	Colouration	Edges detection	Colouration
Bulbar hyperaemia	Annunziato	0.87	0.69	0.89	0.88
	Efron	0.97	0.89	0.99	0.99
	CCLRU	0.83	0.97	0.89	0.98
	Vistakon	0.14	0.85	0.35	0.97
Palpebral hyperaemia	Annunziato	0.32	0.99	0.71	0.99
	Efron	0.95	0.97	0.96	0.98
	CCLRU	0.39	0.62	0.84	0.99
	Vistakon	0.02	0.75	0.84	0.81
Palpebral roughness	CCLRU white light	0.32	0.53	0.94	0.92
	CCLRU Nafl	0.01	0.04	0.30	0.39
	Vistakon	0.73	0.77	0.84	0.89
Corneal staining	Annunziato	0.88	0.83	0.88	0.95
	Efron	0.49	0.55	0.99	0.95
	CCLRU	0.51	0.86	0.65	0.99
	Vistakon	0.95	0.04	0.96	0.57

decrease the statistical power of clinical research studies or the ability of clinicians to monitor small changes over time.

Edge detection techniques are local rather than global in nature and examine the surrounding pixels to determine the presence of edges (vessels or areas of staining). Colour extraction has face validity[28] and examines global relative colouration (red for hyperaemia and green for staining). Both techniques were strongly correlated with increasing bulbar hyperaemia scale grades, although for higher grades the Annunziato and Vistakon-Synoptik scales rely on an increase in red colouration in isolation, rather than in combination with an increased number of blood vessels. As with the other scales analysed, grades are not comparable between grading scales as has previously been shown objectively.[9] Hence clinicians should note the grading scale used when grading on clinical records.

Palpebral hyperaemia scale images were well described by colour extraction techniques. However, although all the scales are in agreement that red colouration increases with scale grade, with pictorial scales (Efron and Annunziato) blood vessels become more prominent with initial scale grades and are then replaced by increasing severity of papillae (both identified by edge detection). In comparison, blood vessels vary in prominence between photographic scale grades (CCLRU and Vistakon-Synoptik). Palpebral roughness in photographic scales was depicted by reflections (CCLRU and Vistakon-Synoptik) or by fluorescein staining viewed with cobalt blue illumination through a Wratten filter (CCLRU). The intensity, incidence angle, and type of illumination will affect the reflections as well as the apparent size and shape of the papillae, and therefore the non-uniform change with increasing photographic scale intensity may be expected. Highlighting the papillae with fluorescein would appear a more appropriate method for determining palpebral roughness as previously described,[25 26] causing an increase in edges detected and a decrease in the fluorescein coverage, despite an apparent decrease in severity with the highest CCLRU scale grade. Further investigation of the palpebral response to stimuli such as antigens, toxic chemicals, and mechanical effects is required to determine whether the

Table 2 Image analysis standard deviation (SD) of repeated measures and discrimination (SD of repeated measures divided by the scale range) of edge detection (% coverage) and relative colouration (%) for the Annunziato, Efron, CCLRU, and Vistakon-Synoptik grading scales

		SD repeated measures		Discrimination (%)	
		Edge detection	Colouration	Edges detection	Colouration
Bulbar hyperaemia	Annunziato	0.15	0.07	0.4	0.4
	Efron	0.32	0.01	0.5	0.2
	CCLRU	0.07	0.14	0.4	0.5
	Vistakon	1.77	0.01	8.6	4.3
Palpebral hyperaemia	Annunziato	0.24	0.07	3.5	0.6
	Efron	0.84	0.12	5.6	1.1
	CCLRU	0.09	0.07	0.5	0.5
	Vistakon	1.93	0.01	9.5	6.1
Palpebral roughness	CCLRU white light	2.36	0.36	5.7	1.3
	CCLRU Nafl	0.90	0.48	3.8	6.7
	Vistakon	0.36	0.27	0.1	1.9
Corneal staining	Annunziato	0.08	0.01	0.6	0.4
	Efron	0.34	0.02	1.6	0.4
	CCLRU	0.66	0.18	4.4	4.1
	Vistakon	0.24	0.01	1.7	3.0

response is similar and how it should be best depicted or photographed for grading purposes.

Staining can differ in intensity (dependent on factors such as the amount of fluorescein instilled, tear film production and drainage, depth of the wound), area, shape, and segmentation. Therefore an epithelial scratch, punctate staining, and confluent ulceration could all have similar intensity of green colouration and edge detected area. All the staining (extent) scales analysed, except the Vistakon-Synoptik scale, depicted more than one type of staining and therefore assessing the ability of image analysis measures to determine the severity of staining is complicated. A general increase in edges detected with increasing scale grade was seen in the all the scales analysed, although the change in green colouration was more variable. However, image analysis would have merit in monitoring changes in individual patients with time and the computer could also count the number of segments identified (to identify between punctate and confluent type staining), provide a ratio of the longest to the shortest axis (to give an indication of shape) in addition to the measures of edges detected (a stable indicator of area) and green colouration (stain intensity).

The image analysis techniques were highly repeatable for both pictorial and photographic scale grades, having a 95% confidence interval of between 0.02 (colour extraction) and 0.10 (edge detection) scale units (on a 0–4 scale). Compared with reported values of clinician subjective grading variability using these grading scales,[7 9 12 15] image analysis techniques are approximately 6–35 times more repeatable, with a sensitivity of 2.8 to 1.2% of the scale (respectively). This study again highlights the high repeatability of image analysis techniques and their ability to assess a range of indicators of anterior ocular physiology.[28]

The occasional apparent reversal in severity in several of the scales could arise from deficiencies in the scale images, such as the lack of an appropriate photographic image taken with similar perspective and illumination or from scale designers considering a range of feature characteristics to assess the grade of an image. Image analysis of a particular feature may require the assessment of a number of characteristics to provide a more simplistic condition grade, comparable with (although having better repeatability and sensitivity than) subjective techniques. There has been much discussion and debate in the literature concerning the merits and relative difficulties of constructing photographic versus pictorial grading scales, with the suggestion that painted grading scales, although not as realistic as photographs, allow more control in depicting incremental increases in severity that are clear and unambiguous to the clinician.[3 7 11] This study generally supports the notion that pictorial grading scales have more incremental increases between grades than photographic scales, although image analysis cannot assess the realism of an image. In the future, image analysis techniques could allow grading of real time or stored images and comparison with population norms without incurring the limitation of photographic or pictorial subjective grading.

In conclusion, the printed grading scales analysed were quadratic in nature, having a higher sensitivity for grading features of low severity. Grading features such as palpebral hyperaemia, palpebral roughness, and corneal staining is complex and there is a compromise between the simplicity of a single scale and the ability to fully describe and monitor changes in the feature. Edge detection and colour extraction image analysis techniques are highly repeatable and offer the potential for more repeatable and sensitive grading than using printed subjective grading scales.

REFERENCES

1 **Mandell RB**. Slit lamp classification system. *J Am Optom Assoc* 1987;**58**:198–201.
2 **Woods R**. Quantitative slit lamp observations in contact lens practice. *J Br Contact Lens Assoc* 1989;Scientific Meeting: 42–5.
3 **Efron N**. Grading scales for contact lens complications. *Ophthal Physiol Opt* 1998;**18**:182–6.
4 **Andersen JS**, Davies IP, Kruse A, *et al. Handbook of Contact Lens Management*. Jacksonville, Florida, USA: Vistakon, 1996.
5 **McMonnies CW**, Chapman-Davies A. Assessment of conjunctival hyperaemia in contact lens wearers. Part 1. *Am J Optom Physiol Opt* 1987;**64**:246–50.
6 **Lofstrom T**, Anderson J, Kruse A. Tarsal abnormalities: a new grading system. *CLAO J* 1998;**24**:210–15.
7 **Chong T**, Simpson TL, Fonn D. The repeatability of discrete and continuous anterior segment grading scales. *Optom Vis Sci* 2000;**77**:244–51.
8 **Fieguth P**, Simpson TL. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci* 2002;**43**:340–7.
9 **Efron N**, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthal Physiol Opt* 2001;**21**:17–29.
10 **Terry R**, Sweeney DF, Wong R, *et al*. Variability of clinical researchers in contact lens research. *Optom Vis Sci* 1995;**12**:16.
11 **Bailey IL**, Bullimore MA, Raasch TW, *et al.* Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci* 1991;**32**:422–32.
12 **Twelker JD**, Bailey IL. Grading conjunctival hyperaemia using a photography-based method. *Invest Ophthalmol Vis Sci* 2000;**41**:s927.
13 **Lloyd M**. Lies, statistics, and clinical significance. *J Br Contact Lens Assoc* 1992;**15**:67–70.
14 **Owen CG**, Fitzke FW, Woodward EG. A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae. *Ophthal Physiol Opt* 1996;**16**:430–7.
15 **Papas EB**. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci* 2000;**41**:687–91.
16 **Chen PCY**, Kovalcheck SW, Zweifach BW. Analysis of microvascular network in bulbar conjunctiva by image processing. *Int J Microcirculation Clin Exp* 1987;**6**:245–55.
17 **Guillon M**, Shah D. Objective measurement of contact-lens induced conjunctival redness. *Optom Vis Sci* 1996;**73**:596–605.
18 **Simpson TL**, Chan A, Fonn D. Measuring ocular redness: first order (luminance and chromaticity) measurements provide more information than second order (spatial structure) measurements. *Optom Vis Sci* 1998;**75**:279.
19 **Villumsen J**, Ringquist J, Alm A. Image analysis of conjunctival hyperaemia: a personal computer based system. *Acta Ophthalmol* 1991;**69**:536–9.
20 **Maldonado M**, Arnau V, Martinez-Costa R, *et al.* Reproducibility of digital image analysis for measuring corneal haze after myopic photorefractive keratectomy. *Am J Ophthalmol* 1997;**123**:31–41.
21 **Willingham FF**, Cohen KL, Coggins JM, *et al.* Automatic quantitative measurement of ocular hyperaemia. *Curr Eye Res* 1995;**14**:1101–8.
22 **Horak F**, Berger U, Menapace R, *et al.* Quantification of conjunctival vascular reaction by digital imaging. *J Allergy Clin Immunology* 1996;**98**:495–500.
23 **Begley CG**, Barr JT, Edrington TB, *et al.* Characteristics of corneal staining in hydrogel contact lens wearers. *Optom Vis Sci* 1996;**73**:193–200.
24 **Mackinven J**, McGuinness CL, Pascal E, *et al.* Clinical grading of the upper palpebral conjunctiva of non-contact lens wearers. *Optom Vis Sci* 2001;**78**:13–18.
25 **Potvin R**, Doughty MJ, Fonn D. Tarsal conjunctival morphometry of Asymptomatic soft contact lens wearers and non-wearers. *Int Contact Lens Clinics* 1994;**21**:225–30.
26 **Doughty MJ**, Potvin R, Pritchard N, *et al.* Evaluation of the range of areas of the fluorescein staining patterns of the tarsal conjunctiva in man. *Doc Ophthalmol* 1995;**89**:355–71.
27 **Miyata K**, Amano S, Sawa M, *et al.* A novel grading method for superficial punctate keratopathy magnitude and its correlation with corneal epithelial permeability. *Arch Ophthalmol* 2003;**121**:1537–9.
28 **Wolffsohn JS**, Purslow C. Clinical monitoring of ocular physiology using digital image analysis. *Contact Lens Ant Eye* 2003;**26**:27–35.
29 **Annunziato T**, Davidson RG, Christensen MT, *et al. Atlas of Slit Lamp Findings and Contact-Lens Related Anomalies*. Fort Worth, TX, USA: Southwest Independent Institutional Review Board, 1992.