Clinical Grading of the Upper Palpebral Conjunctiva of Non-Contact Lens Wearers

JANE MACKINVEN, BSc(Hons), CLAIRE L. MCGUINNESS, BSc(Hons), EVE PASCAL, PhD, MCOptom, and RUSSELL L. WOODS, PhD, MCOptom, FAAO

Private Practice, Glasgow, Scotland (JM), Private Practice, Northumberland, England (CLM), Department of Vision Sciences, Glasgow Caledonian University, Scotland (EP), and Schepens Eye Research Institute, Harvard Medical School, Boston, Massachusetts (RLW)

ABSTRACT: Purpose: To categorize the appearance of the normal upper palpebral conjunctiva using a grading scale and to investigate interobserver agreement with a grading scale modified to include decimal increments. Methods: Upper palpebral conjunctival appearance of 96 non-contact lens wearing subjects aged 18 to 75 years was assessed using a photographic grading scale that has a generic (zero to four) scale to score redness and roughness of the palpebral conjunctiva. Results: Median redness and roughness was about 1.25 units. About 5% of subjects had redness or roughness >2.0 units. Interobserver agreement improved during the study. By the end of the study, the standard deviation of the discrepancy scores was 0.12 to 0.19 units. Conclusions: Upper palpebral conjunctival redness or roughness >2.0 units are unusual. The grading scale can be used successfully with decimal rather than integer scale increments. For experienced clinicians, a change in grade of ≥0.5 units may be significant. (Optom Vis Sci 2001;78:13–18)

Key Words: clinical grading scales, palpebral conjunctiva, prevalence, interobserver agreement

Changes in the palpebral conjunctiva are a major complication of contact lens wear. Contact lens–induced papillary conjunctivitis is considered primarily to be a consequence of allergic stimuli and mechanical irritation. Changes in the palpebral conjunctival appearance can occur as a consequence of stimuli other than contact lenses, such as seasonal allergies and sutures. Better knowledge of the characteristics of the background appearance (i.e., in healthy eyes) and better methods of quantifying palpebral conjunctival appearance will improve the detection of contact lens–induced complications.

Clinical grading scales are useful methods for recording the results of an ocular health assessment and can be invaluable in evaluating change. Grading scales based on verbal description, although common in research, have not been widely adopted by clinicians. Most verbal description scales allow quantification of clinical signs using an ordinal scale. Typically, the absence of a sign is given a grade of zero, and numbers up to four are used to describe increasing levels of the sign. In the generic system proposed by Woods, grade 4 is described as “very severe changes that require intervention, usually medical,” whereas other grading systems provide a description for each condition. The equivalent description of grade 4 palpebral conjunctival appearance recommended by the U.S. FDA is “severe—localized or generalized papillae/follicles >1 mm in diameter with or without marked injection.” In recent years, photographic and pictorial grading scales have become available to practitioners. These grading scales are expected to be an improvement over verbal description scales because they should add a degree of objectivity. Both the CCLRU and Efron grading scales use the familiar zero- to four-unit scale. The CCLRU and Efron grading scales have photographs and drawings, respectively, that illustrate integer increments in each of the appearance scales. Expansion of the grading scale from five levels (e.g., by using decimals) should increase discriminability. Using an eight-level scale, Lofstrom et al. were able to make discriminations between groups that were not possible with a five-level scale. However, recently Efron has suggested that interobserver variability may be a limiting factor to the generalised use of grading scales in clinical practice.

One method of describing the background level of conjunctival appearance is to measure the prevalence of redness and roughness in a population known to have no obvious cause for abnormality, such as disease or contact lens wear. Unfortunately, the appearance of the normal, healthy, adult conjunctiva seems to have received scant attention because most research has assessed the changes associated with contact lens wear. In some studies, asym-
tomatic non-contact lens wearing individuals have acted as control subjects. Allansmith et al. found that among 68 non-contact lens wearers, 24% exhibited a satin-smooth conjunctival appearance, 69% had uniform papillae, and 7% had nonuniform papillae. In a much larger study, Korb et al. reported that 0.6% of their 500 control subjects had conjunctival papillae >0.3mm in diameter. Similarly, Saini et al. reported that all of the 20 non-contact lens wearing eyes in their study displayed a satin appearance of the upper palpebral conjunctiva.

Potvin et al. and Doughty et al. evaluated morphometry as an objective method of palpebral roughness assessment. They measured the size of fluorescein-highlighted features apparent on a photograph. Both subjective assessment and morphometry failed to distinguish between eight contact lens wearers and eight non-contact lens wearers. No study using a clinical grading scale has reported the prevalence of palpebral conjunctival appearance grades in a large sample population.

To evaluate background palpebral conjunctival appearance, we conducted a cross-sectional study, assessing redness and roughness with the CCLRU photographic grading scale. In addition, we measured interobserver agreement when this grading scale was interpolated to decimal units.

**METHODS**

**Subjects**

Ninety-six volunteers (51 females and 45 males) from the staff, student, and patient populations at a university in Scotland participated in this study. All were healthy non-contact lens wearers or had not worn contact lenses during the preceding 6 months. This was considered sufficient time for any lens-induced conjunctival changes to have resolved. Subjects were aged 18 to 75 years (median 22 years) and had no history of anterior segment disease or surgery. Of the 96 subjects, four gave a self-report of dry eye, 25 reported taking some form of medication, and 32 reported having experienced some form of allergy. No subject had clinically diagnosed dry eye. We found no association between these self-reported conditions and palpebral conjunctival appearance.

**Procedure: Prevalence Study**

Only the right eye of each subject was examined. Initially, the upper eyelid was everted and the palpebral conjunctiva was examined with a slitlamp biomicroscope under white light. The degree of redness observed in each of three zones (Fig. 1) was graded according to the CCLRU grading scale, with the two observers reassessed with the slitlamp under cobalt-blue illumination. The upper eyelid was everted again, and the palpebral conjunctiva was reassessed with the slitlamp under cobalt-blue illumination. The degree of roughness observed in each of the three zones was graded in a similar manner to that of redness. One of the two observers (JM and CM) examined each of 66 subjects. Because both observers examined the 30 subjects who were also included in the interobserver agreement study, the data collected by the first examining observer were used in the prevalence study. Before the study commenced, three of the authors discussed grading strategies and compared the grades assigned to the conjunctival appearance of human subjects, some of whom were contact lens wearers. The two observers (JM and CM) were trainee optometrists, and one of the other authors (RW) was an experienced user of clinical grading scales.

**Data Analysis**

In addition to analyses of the six appearance scores (two scales times three eyelid zones), we evaluated overall grading as might be recorded when only one score is assigned for each of redness and roughness. For both redness and roughness, we averaged the grades and reported the maximum grade. Averaging and maximum grade were strategies that were commonly used to assign corneal staining grades. Although the grade scores were approximately normally distributed (Fig. 2), because it is unknown whether the CCLRU grading scale represents an interval scale, nonparametric statistical tests were used. Because Barbeto and Simpson have argued that parametric statistical tests can be applied safely to such data, we analyzed the data using parametric statistics and found no differences from the nonparametric analyses presented here.

To analyze the interobserver agreement, frequency distributions of the discrepancy scores (i.e., the first observer’s grade minus the second observer’s grade) were examined for zone grades (i.e., in all three eyelid zones), averaged grade, and the maximum grade. This method of analysis makes certain assumptions about the data, in particular that the data are from an interval scale. These clinical grading scales are an ordinal system, but may not approximate an interval scale. In an interval scale, the difference between each

---

**FIGURE 1.**

Each of the three upper eyelid zones was graded separately for redness and roughness.
subsequent level is identical. For example, although the numerical difference between 0.5 and 1.0 units and between 3.0 and 3.5 units is the same in these clinical grading scales, we cannot be sure that the two differences have the same clinical meaning. If the grading scale does not approximate an interval scale, interobserver discrepancy scores are expected to vary with the level of measurement. Alternatively, if the grading scale approximates an interval scale, the discrepancy scores will be normally distributed, and the standard deviation of the discrepancy distribution will be independent of the range of the actual (raw) scores used in its determination. We found no systematic variation in our discrepancy scores (i.e., no correlation between the discrepancy score and the raw score). Therefore, we completed the analysis of the discrepancy distributions in the manner recommended by Bland and Altman and evaluated interobserver agreement from the standard deviations.

**RESULTS**

Similar, approximately normal, frequency distributions were found for palpebral redness (Fig. 2, top) and palpebral roughness (Fig. 2, bottom). As expected, the distributions based on the maximum grade were shifted to slightly higher scores than those based on average grade (average of the three zones). This difference was small because there were only minor differences between the distributions of grades assigned in each of the three zones (Table 1). Redness and roughness were significantly correlated (Spearman \( \rho \) ranged from 0.37 to 0.44), and there was a tendency for a greater difference between redness and roughness grades as the scores increased (Fig. 3). This could occur if the grading scale had unequal increments (e.g., logarithmic scaling) or if redness and roughness changed, at least partly, independently. There was no difference between the three zones in the distributions of redness grades (Kruskal-Wallis, \( p = 0.45 \)). More high-roughness scores were found in zone 1 than the other two zones (Kruskal-Wallis, \( p < 0.0001 \)). As shown in Table 2, between 3 and 9% of subjects had a grade \( > 2.0 \) units, a level suggested as the limit for normal palpebral conjunctival appearance.

With increasing age, there was a significant trend toward lower average roughness (Fig. 4, bottom; Spearman \( \rho = -0.36, p = 0.0005 \)) and a nonsignificant trend toward lower average redness (Fig. 4, top; Spearman \( \rho = -0.15, p = 0.14 \)). We are not aware of any studies that have reported a correlation between age and palpebral conjunctival appearance.

Observer 1 examined 51 of the 96 subjects. There was no difference between the distributions of roughness grades assigned by the two observers (Mann-Whitney, \( p = 0.79 \)). However, observer 1 recorded significantly more high-redness scores than observer 2 (Mann-Whitney, \( p = 0.007 \)). The difference was small; the median average-redness scores were 1.3 and 1.0 for observers 1 and 2, respectively. Because many of the subjects were seen by one observer only and there was no difference in the redness grades assigned by the two observers in the interobserver study (below), this difference may have been a real difference rather than a difference between the observers in their allocation of grades.

Early in the study, observer 2 recorded more high-roughness scores in zone 1 than observer 1 (Wilcoxon signed rank test, \( z = -2.3, p = 0.02 \)), and observer 2 therefore had more high scores for maximum roughness (Wilcoxon signed rank test, \( z = -2.5, p = 0.014 \)). The difference was small; the median roughness scores were 1.4 and 1.7 for observers 1 and 2, respectively. During the latter part of the study, there were no differences between the two observers in the distributions of roughness grades, suggesting that their grading skills improved during the study. There were no other significant differences in the distributions of grades assigned by the two observers (early or late in the study). There were no significant differences in the distributions of the grades assigned early and late in the study. For each of the three measures (grade in each zone, average grade, and maximum grade), we found significant differences in the interobserver agreement early and late in the study. The two observers improved the agreement of their grading during the study. Because there was no apparent difference in interobserver agreement between redness and roughness grades, these data were combined. The discrepancy distributions for the three measures were approximately normally distributed, both at the start and finish of the prevalence study (Kolmogorov-Smirnov, \( p > 0.21 \)). The standard deviations of the discrepancy distributions are shown in Table 3.

**DISCUSSION**

The recommendation that an upper palpebral conjunctival appearance of up to grade 2.0 units may be considered normal was
confirmed by our study because 91 to 97% of subjects were recorded as grade 2.0 units or less (Table 2). Because Allansmith et al.1 found that 7% of non-contact lens wearers had a nonuniform papillary appearance (some papillae of 0.4 to 0.8 mm in diameter) and Korb et al.6 found that 1% of non-contact lens wearers had elevated papillae larger than 0.3 mm, our interpretation is that those descriptions are similar to a CCLRU grading scale score of about 2.0 units.

The shapes of the frequency distributions for palpebral conjunctival appearance (Fig. 2) were quite different from those reported for corneal staining in a similar study. Dundas et al.15 reported a highly skewed distribution with <5% of subjects having a corneal staining score of 0.5 units. At least for these two clinical appearances, the grade that represents a normal appearance in the CCLRU grading scales differs between clinical appearance scales, unlike scales that use a single verbal description for all appearances.8 Consequently, care must be taken when comparing the grades attributed using different grading scales.

Because the majority of our subject sample were students and staff of a university in Scotland, it may not have been a good representation of the normal population found in a typical ophthalmic practice. Our sample had a median age of 22 years (range 18 to 75 years), an age typical of neophyte contact lens patients. Further research is required to determine whether the reduction in conjunctival redness with age has clinical significance.

We do not know whether clinicians applying a single grading score to the whole upper palpebral conjunctiva would average the score in the different zones, record the maximum grade, or use some other grading strategy. As with corneal staining,20 it is likely that different clinicians would use different grading strategies. As can be seen in Fig. 2, such an interobserver difference would constitute only a small systematic error when grading non-contact lens wearers. However, if contact lens–induced papillary changes occurred differentially in the different zones, interobserver agreement could be reduced by differences in grading strategy.

Although there are substantial benefits to the use of clinical

### TABLE 1.
The distribution of grades for palpebral redness and roughness were similar for all three zones (shown in Figure 1), except roughness of zone 1, which had significantly more higher roughness scores.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Redness</th>
<th>Roughness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone 1</td>
<td>Zone 2</td>
<td>Zone 3</td>
</tr>
<tr>
<td>0.0–0.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.2–0.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.4–0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.6–0.7</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>0.8–0.9</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1.0–1.1</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>1.2–1.3</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>1.4–1.5</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>1.6–1.7</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1.8–1.9</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2.0–2.1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2.2–2.3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2.4–2.5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2.6–2.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.8–2.9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### TABLE 2.
The percentage of subjects with a grade higher than 2.0 units.

<table>
<thead>
<tr>
<th></th>
<th>Redness</th>
<th>Roughness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone 1</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Zone 2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Zone 3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Average</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Maximum</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

**FIGURE 3.**
Average redness and average roughness were significantly correlated (Spearman ρ = 0.44). Similar correlations were found for maximum scores and for each score for each zone.
grading scales,5, 7, 8, 11, 12, 15, 24 use in private practice remains limited in many countries. Clinician reluctance may be due to a perceived difficulty of use. Our two observers, both inexperienced clinicians, were able to use the palpebral conjunctival appearance grading scale successfully. There were two small differences between the two observers. The difference in average-redness grade throughout the study (median 1.0 vs. 1.3) was probably a real difference between the subjects seen by the two observers. The difference in the roughness grades assigned for zone 1 (median 1.7 vs. 1.4) noted early in the study was not found late in the study. This is consistent with the reduction in the discrepancies between the two observers’ grades later in the study (Table 3). A similar study15 found no such change in the agreement between two similarly inexperienced observers. It appears that the moderate practice and interobserver discussion and comparison conducted before the commencement of our study was not adequate for good initial interobserver agreement. As shown in Table 3, by the end of the study, interobserver agreement was comparable to the interobserver standard deviations of 0.13 to 0.18 units reported previously for corneal staining15 and comparable to the intraobserver standard deviations of 0.19 to 0.21 units found when the roughness apparent in photographs of the palpebral conjunctiva was graded by experienced clinicians.12

Perhaps learning to make palpebral conjunctival appearance judgements is more difficult than learning to grade corneal staining. At the beginning of our study, the interobserver agreement was about the same as the standard deviation of 0.5 units reported by Efron14 for a large group of conference attendees who were asked to grade a projected photograph of the palpebral conjunctiva. The experience of that audience ranged from the novice to the expert, and the quality of the view available to each participant varied widely. By the end of our study, the interobserver agreement was similar to that found previously for corneal staining recorded by two trainee optometrists.15 Chong et al.12 found that experienced clinicians had similar discrepancy distributions for photographs of palpebral conjunctival roughness, 3- and 9-o’clock staining, and bulbar redness. Thus, after a little practice, experienced clinicians should not find making these judgements difficult when observing a real eye.

As predicted by Bailey et al.,13 decimalization of a cataract grading scale improved the interobserver and intraobserver agreement compared with the original integer unit scale.25 Similar improvement in interobserver agreement has been shown by decimalization of the CCLRU grading scale.15 Recently, Papas26 showed that a decimalized CCLRU grading scale for bulbar redness approximates an interval scale. This has the advantage that parametric statistics can be applied. This supports the results of our study and that of Dundas et al.15 because the discrepancy distributions were normally distributed. If interobserver and intraobserver discrepancy distributions of experienced clinicians have standard deviations of about 0.2 units, by the criteria of Bailey et al.,13 decimal grading scales are considered to have moderate to fine sensitivity. Also, a change in grade of 0.5 units or greater between observations (e.g., visits) made by experienced clinicians would represent a significant difference (because the 95% confidence limit is about twice the standard deviation). That discriminability is smaller than the maximum change in palpebral conjunctival redness and roughness of 1.0 units that Terry et al.9 suggested as defining successful contact lens wear.

ACKNOWLEDGMENTS
The authors have no proprietary interests in the products described. This study formed the basis of the honors project completed by JM and CM. We thank Dr. Trefford Simpson for his helpful comments on the manuscript.

Received June 12, 2000; revision received October 13, 2000.

REFERENCES

10. Cornea and Contact Lens Research Unit. CCLRU Grading Scales. Sydney, Australia: School of Optometry, University of New South Wales (available from Vistakon), 1996.

R. L. Woods
Schepens Eye Research Institute
20 Staniford Street
Boston, Massachusetts 02114
e-mail: rwoods@vision.eri.harvard.edu