tvst

DOI: 10.1167/tvst.6.4.13

Article

Evaluating Changes in Ocular Redness Using a Novel Automated Method

Francisco Amparo[1], Jia Yin[1], Antonio Di Zazzo[1], Tulio Abud[1], Ula V. Jurkunas[1], Pedram Hamrah[2], and Reza Dana[1]

[1] Massachusetts Eye and Ear Infirmary, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA
[2] Cornea Service, New England Eye Center, Department of Ophthalmology, Tufts Medical Center, Tufts University School of Medicine, Boston, MA, USA

Correspondence: Reza Dana, Massachusetts Eye and Ear Infirmary, Cornea Service, 243 Charles Street, Boston, MA 02114, USA. e-mail: reza_dana@meei.harvard.edu

Received: 22 February 2017
Accepted: 9 June 2017
Published: 20 July 2017

Keywords: ocular redness; conjunctival redness; hyperemia; continuous scale; objective measurement

Citation: Amparo F, Yin J, Di Zazzo A, Abud T, Jurkunas UV, Hamrah P, Dana R. Evaluating Changes in Ocular Redness Using a Novel Automated Method. Trans Vis Sci Tech. 2017;6(4):13, doi:10.1167/tvst.6.4.13

Purpose: To evaluate interobserver concordance in measured ocular redness among a group of raters using an objective computer-assisted method (ocular redness index [ORI]) and a group of clinicians using an ordinal comparative scale.

Methods: We conducted a prospective study to evaluate ocular redness in clinical photographs of 12 patients undergoing pterygium surgery. Photographs were acquired preoperatively, and at 1 week and 1 month postoperatively. One group of clinicians graded conjunctival redness in the photographs using an image-based comparative scale. A second group applied the ORI to measure redness in the same photographs. We evaluated redness change between time points, level of agreement among raters, and assessed redness score differences among observers within each group.

Results: Interobserver agreement using the image-based redness scale was 0.458 ($P < 0.001$). Interobserver agreement with the ORI was 0.997 ($P < 0.001$). We observed statistically significant differences among clinicians' measurements obtained with the image-based redness scale ($P < 0.001$). There were no significant differences among measurements obtained with the ORI ($P = 0.27$). We observed a significant change in redness between baseline and follow-up visits with all scoring methods. Detailed analysis of redness change was performed only in the ORI group due to availability of continuous scores.

Conclusion: Our findings suggest that the ORI scores provide higher consistency among raters than ordinal scales, and can discriminate redness changes that clinical observers often can miss.

Translational Relevance: The ORI may be a reliable alternative to measure ocular redness objectively in the clinic and in clinical trials.

Introduction

Ocular redness often is the earliest and most common clinical sign of ocular surface irritation and inflammation, and is of significant clinical relevance given that the degree of ocular symptoms and inflammation often correlates with the level of ocular redness experienced by patients. In some cases, ocular redness alone can be an indicator of the severity of the ocular surface condition, and is found as a criterion in algorithms that assess ocular surface disease, such as chemical burns, allergy, dry eye, or ocular graft-vs-host disease.[1,2] However, the value of ocular redness in the diagnosis, therapeutic decision-making, or assessent of response to therapy tends to be underused, in part due to the subjectivity of ocular redness assessment (e.g., "trace" vs. "3+") and the lack of a universally accepted and standardized measuring scale, which leads to the inability to compare among multiple or consecutive measurements.

A number of scales to measure ocular redness have been proposed, including ordinal, visual analog, illustrative-comparative, or complex automated systems.[3] However, to date most studies and clinical trials continue to use a combination of different categorical systems, with the most strict relying on comparative image-based scales that use standard photographs to depict different degrees of ocular redness.[3–7] A growing number of illustrative-comparative scales have been introduced in an attempt to reduce inter- and intraobserver variability, but variability among scales and inconsistent grading judgment among clinicians continue to be an obstacle.[3,5–9] We described a method to quantify ocular redness objectively that provides a continuous numerical centesimal score (0–100), the ocular redness index (ORI).[10] This semiautomated system demonstrated high performance measuring ocular redness and the scores obtained correlated strongly with those of validated image-based comparative scales, which are considered the least subjective and prone to bias ordinal redness scales used in the clinic.[10] In addition, we have shown that the ORI scores obtained by nonophthalmology-trained graders correlate strongly with the scores obtained by ophthalmologists who used image-based comparative scales.[10]

In the original report on the characteristics and performance of the ORI, we highlighted the limitations of our early studies, including the retrospective selection of clinical images used to evaluate ocular hyperemia, which restricted our capacity to evaluate ocular redness using the ORI in the same patient over time.[10] Additionally, the photographs evaluated in our initial study did not include a reference mark that allowed for color-balance standardization. The current study was designed to overcome the limitations described in our previous report, and to specifically validate the ability of the ORI to assess changes in ocular redness in a real clinical setting, where patients are evaluated prospectively. In this study we measured changes in ocular redness at different time points in patients undergoing a therapeutic intervention and evaluated the level of agreement among redness scores obtained by multiple observers using the ORI and another image-based scoring system.

Methods

Design

We designed a pilot study with the objective to enroll patients with an ocular surface pathology presenting with ocular redness that was expected to change due to a therapeutic intervention. Based on this premise, we enrolled patients with pterygia who had a surgical excision procedure planned, and recorded ocular redness with clinical photographs before and after surgical excision. We foresaw that ocular redness would follow a predictable pattern that we would be able to document through clinical imagery. Our rationale with this approach was that patients with pterygium present with an initial degree of redness that increases immediately after surgical excision due to inflammation and bleeding from the surgery, and later decreases with time. We recruited 12 consecutive pterygium patients with at least moderate redness who agreed to be photographed at presentation, within 1 week from surgery and at 1 month postoperatively. Conjunctival hyperemia was assessed in all patients at all visits by the attending clinicians using a 0+ to 3+ redness scale (0, none; 1, mild; 2, moderate; 3, severe).

All participants signed an informed consent. This study was conducted at the Cornea Service of the Massachusetts Eye and Ear Infirmary (Boston, MA), and was approved by the institutional review board. Research was conducted in accordance with the requirements of the Health Insurance Portability and Accountability Act, and the tenets of the Declaration of Helsinki.

Photographs

Photographs were obtained using a standard acquisition protocol for all patients at all visits. Photographs from the nasal conjunctiva were acquired using the SL-D7 Topcon photography system (Topcon Medical Systems, Inc., Oakland, NJ) at ×10 magnification and using the same light intensity and settings in all cases. Patients were instructed to gaze laterally while eyelids were held open to reveal the maximum amount of nasal conjunctiva while attempting to avoid specular reflection of the light source on the area of interest. The nasal conjunctiva was centered in the frame of the image, which included the corneal limbus, medial canthus, and both lid margins. To serve as a control for color standardization of all images before redness assessment, a matte white paper strip was included in the photograph at the plane of the lower lid.

Clinical Scoring of Conjunctival Photographs

Four ophthalmologists read and scored ocular redness in all photographs collected using the Validated Bulbar Redness grading scale (VBR).[7] The clinicians randomly evaluated the images in an entirely masked fashion and were unaware of the acquisition time point. The VBR scale consists of a set of five images illustrating five different degrees of ocular redness, ranging from normal (1) to very severe (5), where each image is assigned an entire value in an

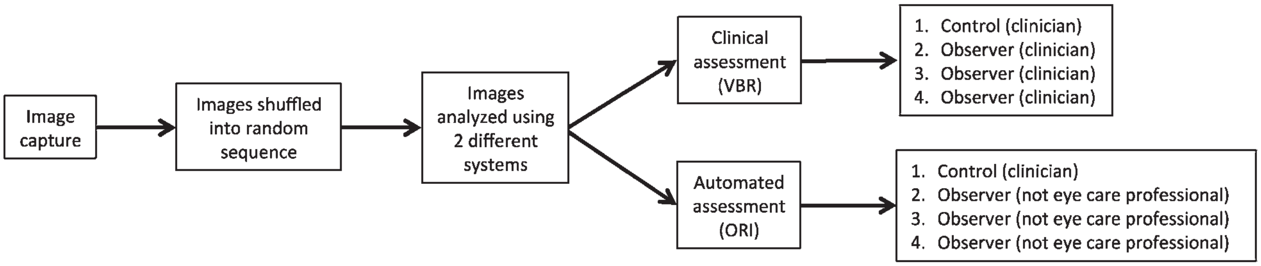
Image capture
Images shuffled into random sequence
Images analyzed using 2 different systems
Clinical assessment (VBR)
Automated assessment (ORI)
1. Control (clinician)
2. Observer (clinician)
3. Observer (clinician)
4. Observer (clinician)
1. Control (clinician)
2. Observer (not eye care professional)
3. Observer (not eye care professional)
4. Observer (not eye care professional)

Figure 1. Process for clinical and automated scoring of photographs.

order of ascending severity. All observers scored all images using the same computer monitor with the same brightness and color settings, the same room illumination, and with no time limit to complete their evaluations.

Automated Scoring of Conjunctival Photographs

To compare the performance of the automated system against the traditional method to score ocular redness, three graders without training in ophthalmology scored the same set of photographs using the ORI. In addition to these three graders, one of the clinicians who scored the photographs with the VBR scale also scored the images with the ORI; the observations from this observer were used as control observations against those obtained by nonclinicians. A total of four graders scored the photographs in each group (Fig. 1). The process to evaluate ocular hyperemia automatically using the ORI is as follows: the photograph file was opened in a computer and a white-balance function was presented to the operator to standardize the color in the image using the white reference mark included in the photograph during its acquisition. The observer defined the nasal conjunctival area to score with a seven-point region of interest selection tool so as to avoid lids, cornea, and other areas not intended for scoring. In this study the nasal conjunctiva exposed in the photographs was included in the evaluated area, with the exception of the portion of tissue invading the cornea (pterygium head) to maintain consistency among the images acquired before and after surgery. Finally, after selection of the area to evaluate, the observer obtained and recorded the redness score on a continuous centesimal (0–100) scale with the strike of a key.

The system used is based on a series of Java plug-ins for the image-processing platform ImageJ.[11] After color correction (white-balance) using the reference mark, the program identified the conjunctival area selected by the user and read the red-green-blue (RGB) values of each pixel, converting it to hue-saturation-value (HSV) space, and ultimately converted these values into a numeric centesimal value for redness in the selected area.[11]

Data and Statistical Analysis

We compared the mean redness scores at the different time points using the Wilcoxon matched-pairs signed rank test for categorical variables and the paired Student *t*-test for continuous variables. Then, we evaluated the correlation between the redness scores obtained with the two different systems by the control observer, as well as the correlation between redness scores obtained by the control observer with the VBR scale and those obtained by the attending ophthalmologists during the clinical visit (0+ to 3+ scale). Then, we evaluated the level of interobserver agreement with different methods to warrant a robust analysis: first, we calculated the coefficient of interobserver agreement for the image-based redness scale using the Cohen's kappa coefficient for ordinal variables (Light's solution for multiple raters), and for the automated digital redness system we calculated the coefficient of interobserver agreement using the intraclass coefficient of correlation (2-way random model). Finally, to confirm the results obtained with the coefficient of interobserver agreement, we looked for differences among the measurements from the different raters in each group. We applied the analysis of variance (ANOVA) for repeated measures in the automated continuous system group, and the Friedman test in the image-based ordinal scale group.

Finally, we reviewed the cases where clinicians using the image-based redness scale assigned the same score to images from two different visits, and quantified the redness difference recorded with the ORI in the same cases. Statistical significance was considered with a 2-tailed *P* value of less than 0.05 for all the analyses in this study.

Table 1. Clinical Image-Based Redness Scores

Observer	Pre Median (Mean ± SD; Range)	1 wk Median (Mean ± SD; Range)	4 wk Median (Mean ± SD; Range)	Change Baseline to Intervention	Change Intervention to 4 wk
1 (control)	3 (3.0 ± 0.6; 2–4)	4 (4.1 ± 1.0; 2–5)	2.5 (2.4 ± 1.4; 1–5)	$P = 0.03$	$P = 0.02$
2	3 (2.8 ± 0.6; 2–4)	4 (4.0 ± 0.7; 3–5)	2.5 (2.7 ± 1.1; 1–4)	$P = 0.01$	$P = 0.02$
3	3 (3.3 ± 0.8; 2–5)	5 (4.4 ± 0.8; 3–5)	3 (2.8 ± 0.8; 2–4)	$P = 0.03$	$P = 0.02$
4	3 (3.3 ± 0.6; 2–4)	5 (4.6 ± 0.7; 3–5)	3 (3.1 ± 1.0; 2–5)	$P = 0.01$	$P = 0.03$

Results

A total of 12 patients with pre- and post-intervention (pterygium excision) clinical photographs were enrolled in the study. A total of 33 clinical photographs were obtained and subjected to clinical and digital redness grading. The redness scores obtained by the attending ophthalmologists using the slit-lamp in the clinic (ordinal 0+ to 3+) showed an increase in redness of 0.6 units (28%; $P < 0.001$), from a median of 2+ (mean 2.2 ± 0.4) at baseline to a postintervention median score of 3+ (mean 2.8 ± 0.4), with a subsequent reduction of 1.2 units (54%; $P = 0.004$) from baseline to a 4-week post-intervention median score of 0.5+ (mean 1.0 ± 1.1).

The clinical and digital mean scores obtained by each observer in each group are presented in Tables 1 and 2. The median (ordinal variables) redness scores obtained with the image-based redness scoring system were 3 (moderate) for baseline, 4 and 5 (severe and very severe) post-intervention, and 2.5 and 3 (mild-moderate and moderate) 4 weeks after the intervention. There were statistically significant changes in all comparisons for all observers ($P < 0.05$; Table 1). The mean (continuous variables) redness scores obtained with the automated redness scoring system were 35.8 for baseline, 52.9 post-intervention, and 32.1 at 4 weeks after the intervention; there were statistically significant changes in all comparisons for all observers ($P < 0.001$; Table 2).

The availability of continuous scores obtained with the automated system allowed for precise calculation of change between visits, with a mean increase of redness after intervention of 17.1 points (48%; $P < 0.001$ vs. baseline) and a subsequent decrease of 20.8 points (58%) 4 weeks after intervention ($P < 0.001$ vs. post-intervention visit). There was a statistically significant correlation ($R = 0.66$; $P < 0.001$) between the ORI and VBR redness scores obtained by the control observer (clinician), as well as between the VBR scores obtained by the control observer and the scores obtained by the attending ophthalmologists (scale 0–3+) using the slit-lamp ($R = 0.60$; $P < 0.001$).

The interobserver agreement among clinical observers using the image-based redness scoring system and the Cohen's kappa coefficient was 0.458 ($P < 0.001$ for all comparisons), while the interobserver agreement among raters using the automated redness

Table 2. Automated Redness Scores

Observer	Preintervention Mean ± SD (Range)	1 wk Mean ± SD (Range)	4 wk Mean ± SD (Range)	Redness Increase Baseline to Intervention	Redness Decrease Intervention to 4 wk
1 (control)	35.6 ± 6.6 (25.9–48.2)	53.1 ± 10.3 (32.4–68.1)	31.8 ± 12.3 (4.0–46.9)	17.5 (+49%)*	21.3 (−60%)*
2	35.8 ± 6.3 (26.1–46.1)	52.8 ± 10.7 (32.5–69.5)	31.7 ± 12.1 (4.1–46.0)	17.0 (+47%)*	21.1 (−60%)*
3	35.8 ± 6.3 (25.7–46.2)	52.9 ± 10.3 (32.3–68.2)	32.2 ± 12.7 (3.8–48.8)	17.1 (+48%)*	20.7 (−58%)*
4	35.8 ± 6.0 (26.0–45.5)	52.9 ± 9.6 (33.9–67.8)	32.5 ± 12.3 (4.2–47.8)	17.1 (+48%)*	20.4 (−57%)*

* $P < 0.001$.

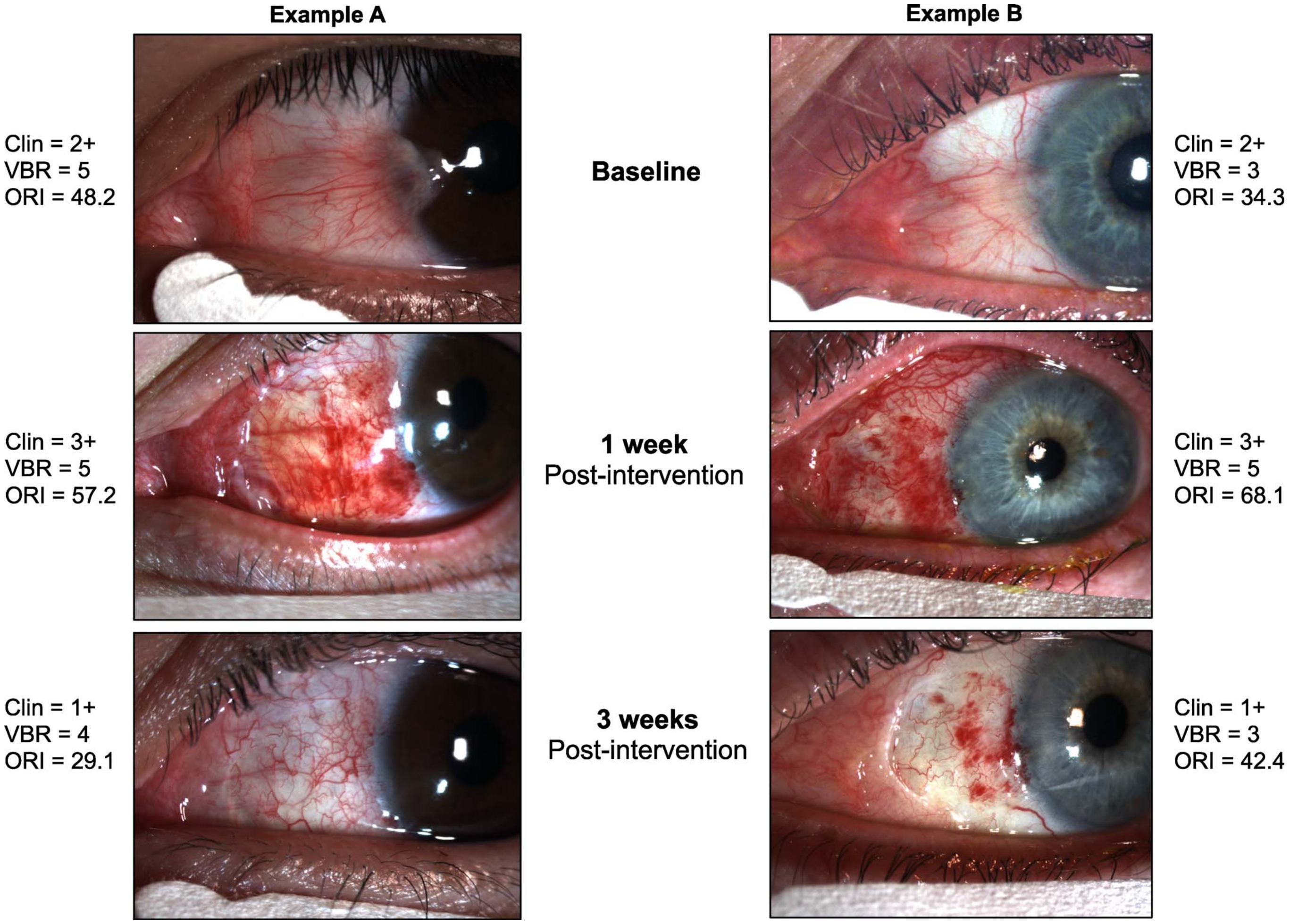

Figure 2. Examples of clinical noncomparative, clinical image-based, and automated ocular redness scores before and after pterygium surgery.

scoring system and the intraclass correlation coefficient was 0.997 (95% confidence interval [CI], 0.995–0.998; $P < 0.001$). The Friedman test revealed statistically significant differences among redness scores obtained by different observers using the VBR scale ($P < 0.001$). Meanwhile, the ANOVA for repeated measures revealed no statistically significant differences among the redness scores obtained by the different observers using the automated system ($P = 0.27$).

Finally, we assessed quantifiable differences recorded with the ORI among photographs where clinicians assigned the same redness score at two different time points with the VBR scale (17% of all comparisons). We found a mean difference of 11.2 points between pre- and post-intervention ORI scores, and that a statistically significant change of 26% went unrecognized by clinicians ($P < 0.001$).

Discussion

The ORI is a novel method designed to grade bulbar redness using a continuous centesimal (0–100) score. This method has the capacity to grade images obtained in a regular eye care setting objectively, circumventing the need for complex equipment or a trained human grader. In this study, we used a simple design to confirm the hypothesis that the ORI can be used as an alternative to the ordinal, partially subjective scales currently used to assess ocular redness in clinical trials. We enrolled patients undergoing a medical intervention that was expected to produce a significant change in the degree of ocular redness, and compared the redness scores obtained with the ORI and VBR scales (Fig. 2).

The redness scores obtained with the ORI strongly correlated with the scores obtained by trained clinicians using an image-based scale. The ORI was sensitive to changes in ocular redness in all patients and at all time points, which is something that trained eye care providers using the ordinal scale were unable to distinguish in 17% of the cases and for up to 26% of the changes (from baseline scores). Furthermore, the ORI showed a very high level of agreement among measurements from different observers, with minimal to no variation among the scores assigned by different observers to the same clinical image. In summary, both systems identified significant change in mean redness scores among the three time points; importantly, however, there was a significant level of disagreement among experienced clinicians' scores, which did not happen among graders using the automated system.

To strengthen the validity and generalizability of this study, we included several observers in each group, a white control in each photograph for prescoring color standardization, and a consecutive-visit prospective design to evaluate the same patient before and after an intervention expected to induce ocular redness change. The results are encouraging and suggest that this technique can help to obtain more precise and analyzable data in studies evaluating ocular redness, which has been relegated as a secondary outcome in studies and clinical trials due to the subjectivity and variability involved in its assessment.

The results of this study demonstrated the capacity of the ORI to discriminate minimal changes compared to image-based scales; this feature can be of even more value in conditions where ocular redness change or fluctuation is more discrete. It is important to note that it is critical to minimize variability in the process of image acquisition, including potential variations in technique, camera, and illumination settings. To address this, we adopted a simple protocol that enforced repeatability of conditions during image acquisition plus the inclusion of a white control for color standardization before ocular redness scoring. However, maintaining a consistent light source and variations in image acquisition technique are limitations that affect virtually all settings conducting ocular surface photography-based analysis.

Several automated methods for assessment of ocular redness through image analysis have been described,[3] some allocating significant resources to overcome the limiting factors present during image acquisition; however, the majority have not yet found broad applicability in the clinical trials scenario. The ORI originally was designed to be a functional alternative to the subjective, ordinal, common system used by clinicians to grade ocular redness more than as a flawless optical tool. Our aim was to develop an accessible tool that adapts to the average eye care setting and overcomes the limitations of ordinal scales that currently are used to assess ocular redness in clinical trials. A fundamental problem of calculation persists with the use of categorical and ordinal scales since measurements are approximate, intervals are not equal, and there is an inherent difficulty to assess agreement. With ocular redness, as with other clinical signs, categorization of the degree of severity leads to arbitrary discretization of the condition.

In the current study some clinicians noted the lack of enough categories in between the five degrees of redness offered by the image-based scale. Interestingly, in one report observers preferred grading scales with less categories, which apparently was related to practicality.[6] It has been proposed that smaller gaps between redness degrees in a reference scale reduce concordance, and, thus, studies that prioritize interobserver concordance among a large numbers of observers could benefit from scales with fewer categories.[12] However, the same studies suggest that trials that prioritize intraobserver consistency and detection of minimal changes would benefit from redness scales with more categories.[7,12] Many of these issues potentially are avoidable with the implementation of a continuous grading scale, such as the ORI.

Some investigators have proposed that appropriate ocular redness assessment must include a simultaneous weighed quantitative evaluation of color analysis and morphologic parameters of the eye's blood vessels to establish a universal score for the assessment of ocular redness.[9,13–15] While conceptually this is accurate, in practice it is very challenging. To date there is no "gold standard" for the assessment of bulbar redness, and most clinical trials continue using image-based or noncomparative, purely subjective scales. Interestingly, in the current study the correlation between scores obtained by clinicians using a 0 to 3+ ordinal scale and the VBR scale was only moderate, emphasizing the important limitation of observation-only redness assessment and confirming the poor quantitative accuracy of clinical grading already reported by other investigators.[9]

By comparing one of the most common techniques for ocular redness assessment with the ORI directly in the clinic and within the context of a prospective study, we not only compared its capacity to assess redness change against experienced clinical judgment, but also confirmed its direct clinical applicability. The ORI scores proved to be more consistent among different observers, making it a robust system when repeated observations from different sources are required. Additionally, the ORI circumvents other common sources of bias encountered when human observers grade ocular redness, such as the tendency of some observers to consistently over- or underestimate bulbar redness.

In summary, our results demonstrated that prospective assessment of ocular redness with the ORI is a reliable metric that is capable of discriminating redness changes that experienced clinical observers can miss, and demonstrated higher score consistency among observers than image-guided clinical scales.

Acknowledgements

The authors thank Joseph B. Ciolino, MD, for his assistance in writing the manuscript.

The Massachusetts Eye and Ear holds intellectual property pertaining to the use of the technology mentioned in this manuscript. Francisco Amparo and Reza Dana are listed as coinventors of this technology.

Disclosure: **F. Amparo**, (F); **J. Yin**, None; **A. Di Zazzo**; None; **T. Abud**, None; **U.V. Jurkunas**, None; **P. Hamrah**, None; **R. Dana**, (F)

References

1. The definition and classification of dry eye disease: report of the Definition and Classification Subcommittee of the International Dry Eye WorkShop (2007). *Ocul Surf*. 2007;5:75–92.
2. Ogawa Y, Kim SK, Dana R, et al. International Chronic Ocular Graft-vs-Host-Disease (GVHD) Consensus Group: proposed diagnostic criteria for chronic GVHD (Part I). *Sci Rep*. 2013;3:3419.
3. Baudouin C, Barton K, Cucherat M, Traverso C. The measurement of bulbar hyperemia: challenges and pitfalls. *Eur J Ophthalmol*. 2015;25:273–279.
4. McMonnies CW, Chapman-Davies A. Assessment of conjunctival hyperemia in contact lens wearers. Part I. *Am J Optom Physiol Opt*. 1987; 64:246–250.
5. IER. *IER Grading Scales*. Institute for Eye Research: Sydney, Australia; 2007.
6. Efron N, Morgan PB, Katsara SS. Validation of grading scales for contact lens complications. *Ophthalmic Physiol Opt*. 2001;21:17–29.
7. Schulze MM, Jones DA, Simpson TL. The development of validated bulbar redness grading scales. *Optom Vis Sci*. 2007;84:976–983.
8. Schulze MM, Hutchings N, Simpson TL. Grading bulbar redness using cross-calibrated clinical grading scales. *Invest Ophthalmol Vis Sci*. 2011; 52:5812–5817.
9. Fieguth P, Simpson T. Automated measurement of bulbar redness. *Invest Ophthalmol Vis Sci*. 2002;43:340–347.
10. Amparo F, Wang H, Emami-Naeini P, Karimian P, Dana R. The Ocular Redness Index: a novel automated method for measuring ocular injection. *Invest Ophthalmol Vis Sci*. 2013;54:4821–4826.
11. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9:671–675.
12. Bailey IL, Bullimore MA, Raasch TW, Taylor HR. Clinical grading and the effects of scaling. *Invest Ophthalmol Vis Sci*. 1991;32:422–432.
13. Schulze MM, Hutchings N, Simpson TL. The use of fractal analysis and photometry to estimate the accuracy of bulbar redness grading scales. *Invest Ophthalmol Vis Sci*. 2008;49:1398–406.
14. Sorbara L, Simpson T, Duench S, Schulze M, Fonn D. Comparison of an objective method of measuring bulbar redness to the use of traditional grading scales. *Cont Lens Anterior Eye*. 2007;30: 53–9.
15. Papas EB. Key factors in the subjective and objective assessment of conjunctival erythema. *Invest Ophthalmol Vis Sci*. 2000;41:687–91.