Published in final edited form as:
Ophthalmology. 2011 April ; 118(4): 768–771. doi:10.1016/j.ophtha.2010.08.027.

A Standardized Grading System For Scleritis

H. Nida Sen, MD, MHSc[1], **Amit A Sangave, BS**[1], **Debra A. Goldstein, MD**[2], **Eric B Suhler, MD, MPH**[3,4], **Denise Cunningham, FOPS, MS**[1], **Susan Vitale, PhD, MHS**[1], and **Robert B. Nussenblatt, MD, MPH**[1]

[1]National Eye Institute, National Institutes of Health, Bethesda, MD, USA

[2]University of Illinois at Chicago, Chicago, Illinois, USA

[3]Oregon Health Sciences University, Portland, Oregon, USA

[4]Portland VA Medical Center, Portland, Oregon, USA

Abstract

Objective—This study evaluated the performance of a standardized grading system for scleritis using standard digital photographs.

Design—Cross-sectional inter-observer agreement study

Participants—Photo archives from the National Eye Institute (NEI)

Methods—Three uveitis specialists from 3 different centers graded 79 randomly arranged images of the sclera with various degrees of inflammation. Grading was done using standard screen resolution (1024 × 768 pixels) on a 0 to 4+ scale, in two sessions: first, without using reference photographs; then with reference to a set of standard photographs (proposed grading system). The graders were masked to the order of images and the order of images was randomized. Inter-observer agreement in grading the severity of inflammation with and without the use of grading system was evaluated.

Main Outcome Measures—Inter-observer agreement

Results—The proposed grading system for assessing activity in scleritis demonstrated a good inter-observer agreement. Inter-observer agreement (pooled κ) was poor (0.289) without photographic guidance and improved substantially when the “grading system” utilizing standardized photographs was used (κ = 0.603).

Conclusion—Utilizing this system of standardized images for scleritis grading provides significantly more consistent grading of scleral inflammation in this study and has clear applications in clinical settings as well as in clinical research.

Assessment and standardization of the degree of activity in ocular inflammation is important both for patient care and clinical research. A quantitative, standardized grading of the severity of inflammation could serve as an inclusion criterion or as an outcome measure.

Corresponding author and reprint requests: H. Nida Sen, MD, MHSc, National Eye Institute, 10 Center Drive, Bldg: 10 Rm:10N112, Bethesda, MD 20892, senh@nei.nih.gov.

No authors have any financial/conflicting interests to disclose except Dr Eric Suhler who receives financial support from Celgene, Abbott, Genentech, Novartis, the Department of Veterans' Affairs, and institutional support from Research to Prevent Blindness.

The comparison of clinical research data requires reproducibility of such measures. It also requires that the methodology be practical for the clinical setting. Scleritis is a chronic, painful and destructive inflammatory disease of the sclera frequently associated with an infection or underlying systemic disease and is one of the most challenging conditions to manage in ophthalmology.[1] The Standardization of Uveitis Nomenclature (SUN) Working Group published standards for grading location and degree of activity of intraocular inflammation including endorsement of standardized photographs for grading vitreous haze[2]; however, no such system has been established for scleritis. In order to assess the performance of a new reference photograph-based scleritis grading system, we studied the level of agreement among uveitis specialists from different centers in grading severity of scleral inflammation with and without the aid of photographic guidance.

Methods

All photographs used were high-resolution (2544 × 1696 pixels) color images of sclera captured with the Canon EOS 20D digital camera mounted on a Haag-Streit BX900 slit lamp following a 10% phenylephrine instillation according to a standardized protocol at the National Eye Institute (NEI). These were then saved in a secure digital database (OIS WinStation 4000SL™, Sacramento, CA) and utilized in accordance with Declaration of Helsinki guidelines. Seventy-nine digital photographs of the sclera with various degrees of inflammation were selected from this digital photo-archive and downloaded in JPEG file format (1124×742 pixels). Three uveitis specialists from 3 different centers graded the photos on a 0 to 4+ scale, using a standard screen resolution (1024 × 768 pixels). The grading was done twice: first, without using photographic references (“session 1”); next, guided by standard photographic references (“session 2”). The same 79 photos were used for each session however the graders were masked to the order of images and the order of images was randomly assorted in both parts of the study.

Each grader was instructed on how to grade the photos in each session. For session 1, the graders were asked to grade each digital image on a 0 to 4+ (with a 0.5+ grade between 0 and 1+) scale where 0 represented no scleral inflammation and 4+ represented the most severe form of scleral inflammation, necrotizing scleritis. For session 2, the graders were asked to compare each digital photo to a similar one on the “grading system poster” provided to them at the end of the first session (Figure 1). This poster was developed by the investigators, who were not involved in grading, by choosing photos from NEI’s digital photo archive. Scleral inflammation in this poster was similarly graded with an ordinal scale of 0 (no scleral inflammation with complete blanching of vessels), 0.5+ (trace inflammation with minimally dilated deep episcleral vessels), 1+ (mild scleral inflammation with diffuse mild dilation of deep episcleral vessels), 2+ (moderate scleral inflammation with tortuous and engorged deep episcleral vessels), 3+ (severe scleral inflammation with diffuse significant redness of sclera ± obscuration of deep episcleral vessels with edema and erythema), and 4+ (necrotizing scleritis with or without uveal show) (figure 1).

Inter-observer agreement was assessed between each pair of graders (graders 1 and 2, graders 2 and 3, and graders 1 and 3), separately for session 1 and session 2. We also calculated the overall agreement among the 3 graders, separately for session 1 and session 2, by using pooled Kappa (κ). The kappa (κ) statistic is an index which compares the agreement against that which might be expected by chance. We computed kappa values using SAS 9.0 software (SAS Institute Inc., Cary, NC, USA). Possible values for Kappa can range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to −1 (complete disagreement). The convention is that values of kappa from 0.41 to 0.6 represent moderate agreement and values of 0.61 to 0.8 represent good or substantial

agreement[3]. In addition, to compare the 3-way agreement between session 1 and session 2, the number of agreeing grades (among the 3 graders) was evaluated.

Results

Results of the grading are shown in Table 1(available at http://aaojournal.org). The overall distributions of Grader 1's grades were similar between sessions 1 and 2. Grader 2 was less likely to assign grade 0 and more likely to assign grade 1 in session 2 than in session 1. Grader 3 did not use the 0.5 grade at all in session 1, but did use it in session 2.

Because grader 3 did not use the 0.5 category at all in session 1, and kappa requires a "square" table (i.e., the same number of categories must be used for the two entities whose agreement is being assessed), we combined the grade of 0.5 with the grade of 0 to compute the kappa statistics. The agreement of each pair of graders was assessed separately for sessions 1 and 2 (Table 2). Graders 1 and 2 had the closest agreement in session 1 (possibly due to grader 3's non-use of the 0.5 category). Pair-wise kappas for session 1 ranged from . 09 to 0.44. The pooled kappa for session 1 was 0.29. Agreement was higher in session 2 than in session 1 for all pairwise comparisons, with kappas ranging from 0.58 to 0.67. The pooled kappa for session 2 was 0.60 which is a substantial improvement.

Because the values of kappa for sessions 1 and 2 are not independent, we were not able to perform statistical testing to see whether agreement differed significantly between sessions. We therefore examined agreement in the following way: for each of the 79 photos, there are three session 1 grades (from graders 1, 2, and 3) and three session 2 grades (from graders 1, 2, and 3). For each photo in session 1, we recorded the number of graders that agreed exactly among the 3 graders (possible values were 0, 2, or 3). We did the same for each photo in session 2 and cross-tabulated the values (Figure 2a, available at http://aaojournal.org). Of the 11 photos for which there was no agreement in session 1, 7 had agreement of all 3 graders and 3 had agreement of 2 graders in session 2. Of the 50 photos for which two graders agreed in session 1, 29 had agreement of all 3 graders in session 2 and none had 0 agreeing graders. To perform statistical testing to see if the level of agreement tended to be higher in session 2 than in session 1, we collapsed the table in Figure 2a to create a 2×2 table and applied McNemar's test (which uses the discordant data to test the hypothesis that changes between sessions 1 and 2 occurred randomly, i.e., half the discordant cases should fall in the no agreement-agreement category and half should fall into the agreement-no agreement category). If categories 2 and 3 were grouped (i.e., any agreement versus no agreement) (Figure 2b, available at http://aaojournal.org), the probability that changes in the discordant cases occurred purely due to chance was 0.001, i.e., agreement was significantly more likely in session 2 than in session 1. If categories 0 and 2 were grouped (i.e., total agreement versus less-than-total agreement) (Figure 2c, available at http://aaojournal.org), the probability that changes in the discordant cases occurred purely due to chance was < 0.001, i.e., agreement was significantly more likely in session 2 than in session 1.

Discussion

Scleritis has been classified based on anatomical location (anterior, posterior) and nature of involvement (diffuse, nodular, necrotizing).[4] While this classification is helpful in the clinical setting and has implications for prognosis, within each category the severity varies. The SUN working group addressed standardization of grading of intraocular inflammation, including anterior chamber cells, flare and vitreous haze, with the latter defined by means of a photographic standard2. However, there is currently no standardized grading system for assessing severity or activity in scleritis. McCluskey and Wakefield proposed a system for

scoring the extent and severity of scleritis, based on common clinical signs5. Their scoring system takes into account clinical features such as the area of inflammation, pain, corneal involvement, and associated anterior chamber inflammation. While it grades pain and the area of inflammation, the system does not specifically address the degree or severity of inflammation. The detailed nature of this system makes it more reflective of the disease course rather than the inflammatory activity at one visit. In addition, the system requires the grading of eight components, each of which can be assigned subjective values from 0 to 2 or 0 to 4, which is likely to require more time from the grader.

In the absence of a standardized grading system, a clinician typically grades the overall scleritis activity on a none, mild, moderate, severe scale (or from 0 to 4), as is done with most other grading systems in ocular inflammatory disorders. Determination of the degree of inflammation is a critical step for the management of scleritis patients and allows for accurate assessment of response to therapy. Accurate, consistent assessment of the degree of inflammation is particularly important when patients are seen by different physicians at different visits, in comparing treatment effects between trials, and in interpreting published results from different groups. To the best of our knowledge, this is the first scleritis grading system utilizing photographic standards and focusing on the severity of scleral inflammation that has been formally evaluated for its utility. A large number of standardized photos were graded by 3 graders from 3 different centers with varying years of experience in the field of ocular inflammatory diseases. Our proposed grading system for scleritis is simple and easy to implement with 0 to 4+ grading scheme similar to that utilized for anterior chamber cell grading. A similar standard photographic grading system has been published by Nussenblatt and colleagues[6] for vitreous haze. Agreement using this system was tested in a pilot study with favorable results and since then it has been widely used and is now accepted by SUN working group as the standard for grading vitreous haze. In addition, agreement using this scale has since been re-evaluated and was found to be moderate to substantial.7

Most studies grade scleritis subjectively as “active” or “inactive” subjectively.[8] Although this may be satisfactory for patient care in some settings, a more consistent and less subjective method is desirable. The grading system proposed herein improved agreement between observers; interestingly, the improvement in agreement was more notable when there was a greater difference between years of experience of graders. A scleritis grading system similar to the current one has been previously used by our group in a pilot trial[9] and found to be helpful in adjudicating the degree of inflammation by different examiners.

As with every grading system, ours has limitations. The grading system was tested for agreement based on clinical photographs and not actual patients. It could be argued that comparing photos to photos may have inflated the level of agreement. On the other hand, this testing method may have avoided bias introduced by physicians’ knowledge of a particular patient’s disease course if the testing of the system were to be done in a clinical setting. In addition, this may serve as an advantage for clinical trials, allowing standard photographs to be graded independently by a reading center, eliminating individual investigator bias. These scleral photographs were taken 15–20 minutes following instillation of 10% phenylephrine, and each photograph reflected only one quadrant of the eye. In the future, further development of this grading system may include use of a composite score for each eye (based on, for example, grades 0 to 4 being assigned to each quadrant) and for each patient. In addition the current grading system needs further validation studies using a clinical outcome.

The graders in this study were practicing, uveitis-trained ophthalmologists, which may have overestimated the inter-observer agreement but these results are probably applicable to

uveitis specialists who participate in clinical trials, particularly given the graders were from different centers with different years of experience.

We believe that this grading system, while far from perfect, provides ophthalmologists with a reproducible, quantitative system and allows for more reliable and consistent assessment of scleral inflammation regardless of number of years of experience. It is simple and practical and can easily be implemented in clinical setting as well as clinical trials.

Acknowledgments

Support: National Eye Institute Intramural research program support (HNS, RBN, AS, SV). Dr. Suhler receives support from the Department of Veterans' Affairs.

References

1. Watson PG. Doyne Memorial Lecture, 1982. The nature and the treatment of scleral inflammation. Trans Ophthalmol Soc U K. 1982; 102:257–281. [PubMed: 6963521]
2. Standardization of Uveitis Nomenclature (SUN) Working Group. Standardization of uveitis nomenclature for reporting clinical data: results of the First International Workshop. Am J Ophthalmol. 2005; 140:509–516. [PubMed: 16196117]
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
4. McCluskey, PJ.; Wakefield, D. Scleritis and episcleritis. In: Pepose, JS.; Holland, GN.; Wilhelmus, KR., editors. Ocular Infection and Immunity. St. Louis, MO: Mosby; 1996. p. 642-662.
5. McCluskey PJ, Wakefield D. Prediction of response to treatment in patients with scleritis using a standardized scoring system. Aust N Z J Ophthalmology. 1991; 19:211–215.
6. Nussenblatt RB, Palestine AG, Chan C, Roberge F. Standardization of vitreal inflammatory activity in intermediate and posterior uveitis. Ophthalmology. 1985; 92:467–471. [PubMed: 4000641]
7. Kempen JH, Ganesh SK, Sangwan VS, Rathinam SR. Interobserver agreement in grading activity and site of inflammation in eyes of patients with uveitis. Am J Ophthalmol. 2008; 146:813–818. [PubMed: 18687418]
8. Taylor SR, Salama AD, Joshi L, et al. Rituximab is effective in the treatment of refractory ophthalmic Wegener's granulomatosis. Arthritis Rheum. 2009; 60:1540–1547. [PubMed: 19404964]
9. Sen HN, Sangave A, Hammel K, et al. Infliximab for the treatment of active scleritis [report online]. Can J Ophthalmol. 2009 44:e9–e12. http://article.pubs.nrccnrc.gc.ca/RPAS/rpv?hm=HInit&calyLang=eng&journal=cjo&volume=44&afpf=i09-061.pdf. [PubMed: 19506593]

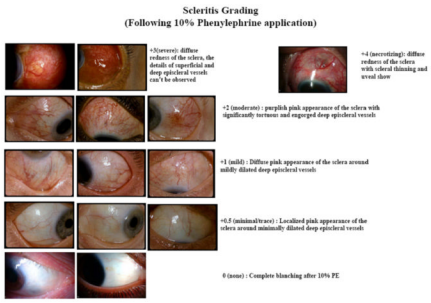

Figure 1. Scleritis grading system

Standardized digital photos of scleritis of varying severity is illustrated. Graders used this illustration as reference photos in session 2.

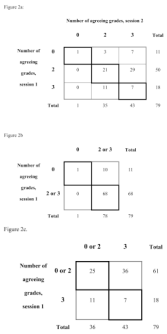

Figure 2. Comparing level of agreement for sessions 1 and 2

Zero represents no agreement among 3 graders, 2 represents 2 of the 3 graders agreeing and 3 represents full agreement by all 3 graders.

Ophthalmology. Author manuscript; available in PMC 2012 April 1.

Table 1

Distribution of grades assigned by each grader

Grade	Session 1: Without using standard photos: n (%)			Session 2: Using standard photos: n (%)		
	Grader 1	Grader 2	Grader 3	Grader 1	Grader 2	Grader 3
0	16 (20.2)	22 (27.8)	25 (31.6)	13 (16.5)	13 (16.5)	16 (20.2)
0.5	17 (21.5)	26 (32.9)	0 (0)	22 (27.8)	25 (31.6)	16 (20.2)
1	23 (29.1)	17 (21.5)	20 (25.3)	22 (27.8)	24 (30.4)	22 (27.8)
2	15 (19.0)	9 (11.4)	18 (22.8)	15 (19.0)	10 (12.7)	19 (24.0)
3	6 (7.6)	3 (3.8)	12 (15.2)	5 (6.3)	3 (3.8)	3 (3.8)
4	2 (2.5)	2 (2.5)	4 (5.1)	2 (2.5)	4 (5.1)	3 (3.8)
Mean (sd)	1.11 (0.99)	0.82 (0.91)	1.37 (1.22)	1.10 (0.95)	1.03 (1.00)	1.13 (0.99)

Distribution of grades in sessions 1 and 2 is shown with and without the use of standard reference photos (sd=standard deviation).

Table 2

Inter-observer Agreement

	Grader 1 vs 2	Grader 1 vs 3	Grader 2 vs 3	Pooled Kappa Graders 1, 2, and 3)
Unweighted (only exact agreement is considered)				
Session 1	0.442	0.348	0.086	0.2896
Session 2	0.587	0.672	0.576	0.6034
Weighted (agreement within ± 1 category is considered)				
Session 1	0.638	0.554	0.342	N.A.
Session 2	0.726	0.797	0.708	N.A.

Agreement among graders for sessions 1 and 2 when categories 0 and 0.5 are combined, (kappa statistic) is shown. Note the improvement in weighted kappa between session 1 and 2. N.A.=Not applicable (represents where a pooled kappa is not applicable)