British Journal of Ophthalmology, 1987, **71**, 371–374

Trachoma grading: observer trials conducted in southern Malawi

JAMES M TIELSCH,[1] KEITH P WEST JR,[1] GORDON J JOHNSON,[2] TEFERRA TIZAZU,[3] LARRY SCHWAB,[4] MOSES C CHIRAMBO,[5] AND HUGH R TAYLOR[1]

From the [1]International Center for Epidemiologic and Preventive Ophthalmology, Wilmer Eye Institute, Johns Hopkins University School of Medicine, 600 N Wolfe Street, Baltimore, MD 21205, USA; [2]Department of Ophthalmology, Memorial University of Newfoundland, St John's, Newfoundland, Canada; [3]International Eye Foundation, Nairobi, Keyna; [4]International Eye Foundation, Bethesda, Maryland, USA; and [5]Kamuzu Central Hospital, Lilongwe, Malawi

SUMMARY A variety of grading schemes have been proposed for the clinical classification of inflammatory trachoma. During a population based study of ocular disease conducted in southern Malawi we tested a simplified version of the current WHO grading scheme. Intraobserver agreement statistics were less than satisfactory for three of four graders. Interobserver agreement when compared against either a well experienced standard ophthalmologist or a consensus grade improved over time for two of the three graders. However, initial agreement for all three graders was only fair to moderate. Previous studies of trachoma grading schemes support these unsatisfactory results. A new system of classification is needed that is both accurate and reliable in a field setting.

In regions of the world where trachoma is endemic it is a chronic, progressive disease with two overlapping phases. The inflammatory phase, indicative of active infection with ongoing transmission is found most often in children under the age of 10 years. The cicatricial phase, a consequence of chronic granulomatous inflammation, begins in childhood but does not usually become severe until later in adult life. The severity of scarring, and risk of subsequent blindness, is a function of the intensity and duration of the previous inflammatory phase. As such, the relative degree or severity of inflammation in a community is considered to be an excellent predictor of future risk of blindness in a population.

Many grading schemes have been devised over the last 50 years more accurately to classify persons or communities with varying severity of trachomatous inflammation. These have grown increasingly complex, using up to 22 signs for each eye with each sign scored from 0 to 3.[1] Even the current WHO recommended grading system uses eight signs scored from 0 to 3.[2] There are few published data on the reproducibility of these grading schemes between different observers or even between the same observer at different times.

During the course of a population-based survey of ocular disease conducted in the Lower Shire River Valley of Malawi we had the opportunity of examining intra- and interobserver reliability in the grading of trachomatous inflammation.

Material and methods

The population for the study came from a population-based prevalence survey of trachoma and other ocular diseases carried out in southern Malawi in the autumn of 1983. These studies took place in three villages adjacent to survey villages in the Lower Shire River Valley. A detailed description of the methods used in that investigation have been published elsewhere.[3] The three graders were fully qualified ophthalmologists with field experience in trachoma endemic areas. A special training programme on trachoma grading, including extensive photographic review and field examinations, was conducted by a specialist experienced in field studies of trachoma (HRT).

Correspondence to Dr James M Tielsch.

Three observer trials were conducted:

Study I. In study I 20 eyes of 10 children and 60 eyes of 30 adults were graded twice by each of the three graders and the specialist. Both children and adults were assigned study numbers and seen by each ophthalmologist in a different order. The order of the subjects was then changed for the second round of grading. The most experienced ophthalmologist in trachoma grading (HRT) served as a standard against which the other three graders were compared. The analysis for the study focused on both intra- and interobserver agreement.

Study II. Approximately two weeks after study I was concluded 50 eyes of 25 children were examined independently by each of the three graders. All three graders then examined the children jointly to come to a consensus grade for each child. Five children were unavailable for the consensus examination; however, all five were given the same grade by all three observers during the initial grading round. Only interobserver agreement was calculated for this study.

Study III. Three weeks after the conclusion of study II another 40 eyes of 20 children were examined by the three graders using the same methodology as in study II. Again only interobserver agreement was analysed.

METHODS OF EXAMINATION

Examinations were conducted in the open air in full sunlight with 2·5 × magnifying loupes. The upper lid of each eye was everted and the inflammatory response was graded and recorded appropriately. The grading scheme used for these trials was a simplification of the current WHO classification scheme. Table 1 compares this system with the existing WHO system. The new scheme did not require the separate grading of follicular and papillary response but maintained the WHO definitions for the categories that were retained. This system was considered to be easier to use in the field while disease was still classified into categories that had clinical and epidemiological relevance. The agreement analysis was conducted only for inflammatory trachoma. Cicatricial trachoma was not analysed because the field study focused primarily on young children, and the marginal distribution of scarring severity was overwhelmingly 'absent' or 'mild'.

Table 1 *Classification schemes for inflammatory trachoma*

Grade	*WHO scheme*	*Current scheme*
Absent	Five or fewer follicles in the entire tarsal conjunctiva	Same
Mild	More than five follicles in the entire tarsal conjunctiva, but less than five in zone 3	Same
Moderate	Five or more follicles in each of the three zones	Five or more follicles in zone 3
Severe	Conjunctiva thickened and opaque, normal vessels on the tarsus are hidden over more than half of the surface due to papillary hypertrophy and infiltration	Conjunctiva thickened and opaque, normal vessels on the tarsus are completely hidden due to papillary hypertrophy and infiltration

The kappa statistic was used in the analysis for both intra- and interobserver reliability.[4] Percentage agreement is an inappropriate measure of agreement in studies of this sort, because it does not correct for the amount of agreement expected from chance alone. When one category tends to be the most commonly occurring grade, percentage agreement is deceptively high. Kappa is a chance corrected measure of agreement that ranges between +1 (complete agreement) and −1 (complete disagreement). Values greater than 0 indicate agreement greater than expected by chance. Kappa statistics above 0·7−0·75 are considered by most authorities to indicate good agreement.[5] Where all types of disagreement are not the same, the standard kappa statistic can be somewhat conservative. For example, the disagreement between two observers when one observer grades inflammatory trachoma as mild and the other grades it as moderate is less than if the second observer graded it as severe. We therefore chose to use the weighted kappa statistic which permits an assignment of weights to differing levels of agreement. The weights we used are shown in Table 2. Complete disagreement was defined in the present situation when one observer reported no inflammation while another reported severe. Each less severe disagreement was then weighted more heavily in a linear fashion with complete agreement (none-none, mild-mild, etc.) given full weight. All analyses were conducted by person (not by eye) using the severity of the worst eye as the measure of inflammation.

Results

The prevalence rates of inflammatory trachoma are

Table 2 *Cell weights used for estimation of weighted kappa statistics*

Inflammatory grade	*Inflammatory grade*			
	Normal	*Mild*	*Moderate*	*Severe*
Normal	1	0·67	0·33	0
Mild	0·67	1	0·67	0·33
Moderate	0·33	0·67	1	0·67
Severe	0	0·33	0·67	1

Table 3 *Prevalence of inflammatory trachoma by person and eye*

Grade	*Study I**		*Study II†*		*Study III†*	
	Person	*Eye*	*Person*	*Eye*	*Person*	*Eye*
Normal	27 (67·5%)	56 (70·0%)	11 (44·0%)	23 (46·0%)	1 (5·0%)	2 (5·0%)
Mild	10 (25·0%)	18 (22·5%)	7 (28·0%)	14 (28·0%)	1 (5·0%)	2 (5·0%)
Moderate	2 (5·0%)	4 (5·0%)	7 (28·0%)	13 (26·0%)	10 (50·0%)	20 (50·0%)
Severe	1 (2·5%)	2 (2·5%)	0 (0·0%)	0 (0·0%)	8 (40·0%)	16 (40·0%)
Total	40 (100·0%)	80 (100·0%)	25 (100·0%)	50 (100·0%)	20 (100·0%)	40 (100·0%)

*Based on the grade of the standard ophthalmologist. †Based on the consensus of the graders.

presented in Table 3. There was very little difference in prevalence when examined by eye or by person within any one study. Study III had higher rates and more severe disease than studies I and II. This agrees with anecdotal observations made in these three locations during the conduct of the observer trials.

The analysis of intraobserver reliability examined in study I is presented in Table 4. The most consistent observer was the standard, with a kappa of 0·92; the graders ranged from 0·50 to 0·70. These kappa statistics are in the low to moderate range except for the standard, who showed excellent internal consistency.

Results for interobserver agreement are shown in Table 5. Graders 1 and 2 show a consistent increase in agreement over time; both were in the 'acceptable agreement' range by the time of study III. Grader 3 had the highest agreement in studies I and II, but fell off sharply in study III. This may reflect diagnostic drift or the small sample size. In an effort to describe the agreement based on all the data, a weighted average kappa was calculated for each examiner. These overall kappas are 0·65, 0·62, and 0·62 for graders 1, 2, and 3 respectively. While there is no basic difference among the graders, they still show agreement that is less than the acceptable range.

Discussion

All epidemiological studies are subject to measurement variability, whether they involve clinical observations, laboratory analyses, or interview based responses. Two main sources of variation in data such as these are: (*a*) sampling variability, which, if no bias is inherent in the data, can be estimated by standard statistical techniques; and (*b*) variability within and between observers. The magnitude of the latter effect is estimable only from studies undertaken specifically to evaluate it. Even then, drift in measurement may occur over time, leading to increasing unreliability in spite of earlier documentation that agreement was adequate. Standardisation of equipment, techniques, and lighting conditions can reduce this type of variability considerably. However, even when these conditions are met, the classification scheme must clearly define categories of severity.

Little work has been done to examine this issue in the various grading schemes for inflammatory trachoma. What studies have been done indicate that agreement is often less than satisfactory. Assaad and Maxwell-Lyons reported on interobserver trials for 574 persons examined in Taiwan[6] using a modification of the MacCallan classification. They found that, when two highly experienced ophthalmologists were compared, the percentage agreement was 68·3%. When we adjusted their findings for expected agreement, however, the weighted kappa was only 0·37 (95% confidence interval: 0·29, 0·46). Kupka and coworkers also used the same classification system for their work in Morocco.[7] They examined 100 adults and 100 school children using two graders. The

Table 4 *Intraobserver reliability for inflammatory trachoma by person*

Graders	*Weighted kappa*	*95% CI*
1	0·70	0·47–0·93
2	0·55	0·26–0·85
3	0·66	0·41–0·92
Standard	0·92	0·80–1·00

CI = confidence interval.

Table 5 *Interobserver agreement for inflammatory trachoma by person*

	Grader	*Weighted kappa*	*95% CI*
Study I*	1	0·58	0·31, 0·85
	2	0·51	0·30, 0·71
	3	0·64	0·38, 0·90
Study II†	1	0·64	0·41, 0·87
	2	0·61	0·41, 0·81
	3	0·71	0·51, 0·92
Study III†	1	0·80	0·56, 1·00
	2	0·86	0·67, 1·00
	3	0·47	0·11, 0·84

*Standard for comparison was the standard ophthalmologist.
†Standard for comparison was the consensus of the three graders.

school children were examined twice independently by each observer in order to estimate intraobserver variation. The interobserver percentage agreement was 53%, while the intraobserver percentage agreements were 67% and 64%. Weighted kappa statistics that we computed from their published data support these poor results. The interobserver kappa was 0·47 (95% CI: 0·36, 0·58) and the intraobserver kappas were 0·63 (95% CI: 0·49, 0·78) and 0·48 (95% CI: 0·33, 0·63). Dawson and coworkers using three observers examined 31 children in a treatment trial among American Indian children in Utah.[8] They did not report direct comparisons of observer grading, but found differential treatment effects when different observers were used. In another small treatment study Dawson and coworkers used two observers to compare the frequency of active versus inactive trachoma among 27 children.[9] The percentage agreement was 67%. With their data the kappa was 0·28 (95% CI: −0·03, 0·60). Observer trials were also conducted by the Royal Australian College of Ophthalmologists[10] during their National Trachoma and Eye Health Program. Three graders were compared for both intra- and interobserver agreement by means of photographs and a grading scheme indicating presence or absence of active follicular disease. Our calculations on their published data show that intraobserver agreement was good, with kappas ranging from 0·72 to 0·94. Interobserver agreement was not as good, however. The kappa was 0·54 (95% CI: 0·45, 0·63) using the extension of the method for more than two observers.[4] The most recent work in this area has been done by Brilliant and others in the Nepal Blindness Survey.[11] They conducted a rigorously designed assessment of observer variation for a variety of ocular conditions. Their trachoma grading system was a modification of the currently recommended WHO classification. Results revealed poor interobserver agreement for active inflammatory disease. Kappas for varying intensity levels ranged from 0·31 to 0·51 depending on whether inflammation was graded on a 4-point or 2-point scale. The kappa for overall intensity grade in the worst eye was 0·36.

Despite using a simplified version of the current WHO classification scheme and well trained and experienced ophthalmologists, our results are little better than those previously reported. While we noted that graders 1 and 2 improved to an acceptable level over time, grader 3 showed significant diagnostic drift, which emphasises the importance of restandardising clinical observers during the course of a protracted observation period.

The present study confirms that the current system for grading inflammation is inadequate. A new system of classification is needed if accurate and reproducible clinical observations are to be obtained for use in epidemiological studies of trachoma.

References

1 Tarizzo ML. *Field methods for the control of trachoma.* Geneva: World Health Organisation, 1973.
2 Dawson DR, Jones BR, Tarizzo ML. *Guide to trachoma control.* Geneva: World Health Organisation, 1981.
3 Chirambo MC, Tielsch JM, West KP Jr, *et al.* Blindness and visual impairment in southern Malawi. *Bull WHO* 1986; **64:** 567–72.
4 Fleiss JL. *Statistical methods for rates and proportions.* 2nd ed. New York: Wiley, 1981.
5 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33:** 159–74.
6 Assaad FA, Maxwell-Lyons F. Systematic observer variation in trachoma studies. *Bull WHO* 1967; **36:** 885–900.
7 Kupka K, Nizetic B, Reinhards J. Sampling studies on the epidemiology and control of trachoma in southern Morocco. *Bull WHO* 1968; **39:** 547–66.
8 Dawson DR, Elashoff RM, Hanna L, Hoshiwara I, Ostler HB. The evaluation of a controlled trachoma therapy trial with oral tetracycline. In: Nichols RL, ed. *Trachoma and related disorders.* Amsterdam: Excerpta Medica, 1971.
9 Dawson CR, Hanna L, Wood TR, Coleman V, Briones OC, Jawetz E. Controlled trials with trisulfrapyrimidines in the treatment of chronic trachoma. *J Infect Dis* 1969; **119:** 581–90.
10 Royal Australian College of Ophthalmologists. *The national trachoma and eye health program.* Sydney: Royal Australian College of Ophthalmologists, 1980.
11 Brilliant LB, Lepkowski JM, Musch DC. Reliability of ophthalmic diagnoses in an epidemiologic survey. *Am J Epidemiol* 1983; **118:** 265–79.

Accepted for publication 14 July 1986.