Title

Inter-Rater Agreement between Trachoma Graders: Comparison of Grades Given in Field Conditions versus Grades from Photographic Review.

Journal

Ophthalmic epidemiology, 22(3)

ISSN

0928-6586

Authors

Gebresillasie, Sintayehu
Tadesse, Zerihun
Shiferaw, Ayalew
et al.

Publication Date

2015

DOI

10.3109/09286586.2015.1035792

Peer reviewed

HHS Public Access

Author manuscript

Ophthalmic Epidemiol. Author manuscript; available in PMC 2016 March 22.

Published in final edited form as:

Ophthalmic Epidemiol. 2015 ; 22(3): 162–169. doi:10.3109/09286586.2015.1035792.

Inter-rater agreement between trachoma graders: comparison of grades given in field conditions versus grades from photographic review

Sintayehu Gebresillasie[1], **Zerihun Tadesse**[1], **Ayalew Shiferaw**[1], **Sun N. Yu**[2], **Nicole E. Stoller**[2], **Zhaoxia Zhou**[2], **Paul M. Emerson**[3], **Bruce D. Gaynor**[2,4], **Thomas M. Lietman**[2,4,5,6], and **Jeremy D. Keenan**[2,4]

[1]The Carter Center Ethiopia, Addis Ababa, Ethiopia

[2]Francis I. Proctor Foundation, University of California, San Francisco, USA

[3]The Carter Center, Atlanta, GA, USA

[4]Department of Ophthalmology, University of California, San Francisco

[5]Department of Epidemiology & Biostatistics, University of California, San Francisco

[6]Institute for Global Health, University of California, San Francisco

Abstract

Purpose—Trachoma surveillance is most commonly performed by direct observation, usually by non-ophthalmologists using the World Health Organization (WHO) simplified grading system. However, conjunctival photographs may offer several benefits over direct clinical observation, including the potential for greater inter-rater agreement. This study assesses whether inter-rater agreement of trachoma grading differs when trained graders review conjunctival photographs versus when they perform conjunctival examinations in the field.

Methods—3 trained trachoma graders each performed an independent examination of the everted right tarsal conjunctiva of 269 children aged 0-9 years, and then reviewed photographs of these same conjunctivae in a random order. For each eye, the grader documented the presence or absence of follicular trachoma (TF) and intense trachomatous inflammation (TI) according to the WHO simplified grading system.

Results—Inter-rater agreement for grade of TF was significantly higher in the field (kappa coefficient, κ, 0.73, 95% confidence interval, CI 0.67-0.80) than by photographic review (κ=0.55, 95% CI 0.49-0.63; difference in κ between field grading and photo grading 0.18, 95% CI 0.09-0.26). When field and photographic grades were each assessed as the consensus grade from the 3 graders, agreement between in-field and photographic graders was high for TF (κ=0.75, 95% CI 0.68-0.84).

Corresponding Author: Jeremy Keenan, 513 Parnassus Avenue, Med Sci S334, Francis I. Proctor Foundation, University of California, San Francisco, Tel: 415-476-6323, Fax: 415-476-0527, jeremy.keenan@ucsf.edu.

Conflict of Interest: None of the authors report a conflict of interest.

Previous Publication Statement: This article has not been published previously and is not simultaneously being considered for any other publication.

Conclusions—In an area with hyperendemic trachoma, inter-rater agreement was lower for photographic assessment of trachoma than for in-field assessment. However, the trachoma grade reached by a consensus of photographic graders agreed well with the grade given by a consensus of in-field graders.

Keywords

inter-rater agreement; observer variation; trachoma; diagnosis; photography

Introduction

Trachoma grading is inherently subjective. Surveillance for clinically active trachoma is most commonly based on the World Health Organization (WHO) simplified grading system, which provides definitions for follicular trachoma (TF; ≥5 follicles measuring 0.5mm or greater on the central upper central tarsal conjunctiva) and intense trachomatous inflammation (TI; pronounced inflammatory thickening of the upper tarsal conjunctiva obscuring more than half the deep tarsal blood vessels).[1,2]

Although the WHO simplified system provides objective definitions for TF and TI, a subjective assessment is still required. For example, only the central upper tarsal conjunctiva should be examined for trachomatous signs, but the limits of this central area are imprecise. TF requires that only follicles 0.5mm or greater be counted, but in the field it can be difficult to determine follicle size and many are irregular in shape. Moreover, the definition for TF does not distinguish between superficial follicles and those buried under subconjunctival inflammation; some graders may count “buried” follicles, whereas others may not. In the case of TI, the WHO definition specifies that at least half the deep tarsal blood vessels must be obscured, but does not specify whether these are the primary, secondary, or tertiary branches of these vessels. Some observers may define an inflammatory reaction that obscures at least half the secondary and tertiary branches as TI, whereas others may define TI only when half the primary branches are obscured. Estimating the non-visible proportion and determining when ‘thickening’ is pronounced requires an element of individual judgment.

Because of these subjective components of trachoma grading, even experienced trachoma experts may have different opinions on what constitutes TF and TI. This variability may be compounded when grading trachoma in field conditions, where magnification and lighting conditions vary and where anxious children may limit the quality of the examination. A key feature of a diagnostic test is its reproducibility.[3] A more repeatable test for trachoma would allow for a more precise prevalence estimate, which is especially important at prevalence levels near treatment thresholds (eg, 10% TF triggering a mass azithromycin distribution). We hypothesized that inter-rater variability inherent in grading trachoma in field conditions might be reduced by grading conjunctival photographs instead.[4,5] To test this hypothesis, we assessed the inter-rater agreement between 3 trained trachoma graders who independently performed conjunctival examinations of a consecutive series of children aged 0-9 years in 2 settings: first, in field conditions, and then as a set of conjunctival photographs.

Materials and Methods

Ethics statement

We obtained ethics approval from the University of California, San Francisco Committee on Human Research and the Ethiopian Ministry of Science and Technology. The research adhered to the tenets of the Declaration of Helsinki.

General study design

In this study, 3 trained trachoma graders examined a series of study participants in the field, and then assessed conjunctival photographs from these same study participants several days later. We calculated agreement between the 3 graders in the 2 settings to determine whether photographic grading produced more reliable trachoma grades than in-field grading. We used proposed guidelines for the reporting of reliability studies when drafting this manuscript.[6]

Study setting

This is an ancillary study of an ongoing cluster-randomized clinical trial assessing the role of different treatment schedules of mass azithromycin for trachoma (clinicaltrials.gov identifier NCT01202331). We selected a convenience sample of 5 communities from the trial, each of which had received annual mass azithromycin treatments for 4 years followed by 2 years of annual azithromycin treatments targeted either to preschool children (3 communities) or to households containing a child with clinically active trachoma (2 communities). As per the trial protocol, a random sample of 60 children aged 0-9 years per community was selected from the preceding study census for monitoring. We performed monitoring in a different study community each day for 5 days, and included all consecutive children who presented during this time period in the current study.

Trachoma grading training

3 graders were included in the current study; a health officer (grader 1), clinical nurse (grader 2), and ophthalmologist (grader 3). Each had direct clinical experience with trachoma and each had performed trachoma grading at previous trial monitoring visits. At the beginning of the monitoring visit, the 3 graders each completed the same training workshop, conducted by a separate trial investigator. This training included a Microsoft PowerPoint presentation of the WHO simplified grading system, with emphasis on TF and TI.[1] Graders were subsequently tested on a series of 60 conjunctival images; each grader agreed sufficiently with an expert panel to allow participation as a trachoma grader for the study visit (Cohen's kappa coefficient, κ, for TF was 0.73, 0.77, and 0.90, for graders 1, 2, and 3, respectively; κ for TI was 0.65, 0.78, and 1.0 for graders 1, 2, and 3, respectively).

In-field grading

Each of the 3 trained graders examined the everted upper right tarsal conjunctiva for clinical signs of trachoma according to the WHO simplified grading system. We chose 3 graders because this number could feasibly examine a single study participant in the field, and allowed a consensus grade to be calculated. The presence or absence of TF and TI was

documented for each study participant. Each of the graders used 2.5x loupes and adequate light when grading, and each of the graders was aware of the age and sex of each participant. The conjunctiva was everted once and each of the 3 graders serially examined the conjunctiva in silence. We emphasized the importance of masking in this study; no discussion was allowed until each of the graders had confirmed that their grade had been documented. After all grades had been recorded, 1 of the graders took 3 photographs of the everted conjunctiva using a Nikon D90 digital SLR camera with a 105/2.8f macro lens (aperture priority, f-stop 40, ISO 400, native flash engaged, automatic white balance). The lid was not returned to its normal position until all 3 graders had examined the conjunctiva and the photographs were taken.

Photographic grading

A study investigator not participating in the grading for this study chose the best quality photograph for each study participant, relabeled the photograph names, and organized the photographs in a random order. Interspersed in a random order with the photographs for the current study was a random selection of 30 repeat photographs and an arbitrary selection of 40 photographs from a different study, which were included so that the graders would be masked to the prevalence of clinically active trachoma in the photograph set. All randomization procedures were accomplished with the RAND function in Microsoft Excel. The same 3 in-field graders independently assessed the digital photographs within 1 week of the original in-field grading. Each of the graders performed photographic grading on the same laptop computer; the laptop monitor illumination was set to the maximum level and no changes were made to any computer settings during this study. Grading was performed in a completely dark room.

Statistical analysis

We assessed agreement between the 3 graders using the free-marginal κ statistic described by Brennan and Prediger and its multi-rater counterpart described by Randolph.[7,8] The free-marginal κ statistic is recommended when raters are not instructed about the number of observations that should be assigned to each category (ie, when graders are free to assign observations to categories in any way they choose).[7] We calculated κ and bias-corrected bootstrapped 95% confidence intervals (CIs; 9,999 repetitions) separately for in-field grades and photographic grades.[9] We also performed similar κ statistics to determine agreement between in-field and photographic grades within the same grader, and compared the consensus in-field grade with the consensus photographic grade (with consensus defined as agreement by at least 2 of the 3 graders). We compared κ from in-field and photographic grades by calculating the difference in κ between the 2 settings, and assessed whether this difference was significantly different from 0 by constructing its bias-corrected bootstrapped 95% CI (9,999 repetitions). Sample size considerations were based on the CI of the inter-rater κ statistic; assuming 3 graders, an overall prevalence of TF of 45%, and an estimated κ of 0.6, then 259 participants would provide a 95% CI of ±10% (kapssi command in Stata). We used Stata software version 12.0 (StataCorp, College Station, TX, USA) for all statistical calculations.

Results

Inter-rater agreement

Each of the 3 graders successfully examined the conjunctiva of 269 children (median age 5 years, interquartile range 3-6 years; 51.7% female). In the field, at least 2 of the 3 graders agreed that TF was present in 120 children (44.6%) and that TI was present in 18 children (6.7%; Figure 1). Agreement between the 3 graders in the field was substantial, with κ=0.73 (95% CI 0.67-0.80) for TF and κ=0.91 (95% CI 0.87-0.95) for TI. By photographic review, at least 2 of the 3 graders agreed that TF was present in 119 children (44.1%) and that TI was present in 41 children (15.2%; Figure 1). Agreement between the 3 graders by photographic review was moderate, with κ=0.55 (95% CI 0.49-0.63) for TF and κ=0.76 (95% CI 0.69-0.82) for TI. Inter-rater agreement on the grades of TF and TI was comparable for each pair of graders in the field and by photographic review, although graders 1 and 2 were more likely to disagree with each other on the grade of TF and more likely to agree with each other on the grade of TI than either did with the third grader (Table 1). Inter-rater agreement was significantly better for in-field grades than photographic grades, both for TF (difference in κ 0.18, 95% CI 0.09-0.26) and for TI (difference in κ 0.15, 95% CI 0.08-0.23). Disagreement in photographic grading was mostly due to grader 1 undercalling TF, grader 2 overcalling TF, and grader 3 overcalling TI relative to the other graders (Figure 1).

Inter-rater agreement over time

To assess whether convergence of grades was achieved in the field, we divided the study population into deciles based on the order in which they were examined. As shown in Figure 2, there did not appear to be any improvement in agreement for the grading of TF in the field over the duration of the study, although κ for the in-field grades displayed less variability than that of the photographic grades.

Agreement between in-field and photographic grades

We compared in-field and photographic grades within each of the 3 graders (Figure 3). Agreement between in-field and photographic grades was generally substantial for TF (range κ=0.55-0.78) and TI (range κ=0.64-0.85). Disagreements for TF appeared to be grader-specific; relative to their respective in-field grades, grader 1 undercalled TF and graders 2 and 3 overcalled TF on photographic review. In contrast, all 3 graders overcalled TI on photographic review relative to their respective in-field grades. We also compared the in-field and photographic grade agreed on by a consensus of graders. The consensus grades had substantial agreement with each other for TF (κ=0.75, 95% CI 0.68-0.84) and TI (κ=0.80, 95% CI 0.73-0.87). Figure 4 shows a random selection of photographs that were discrepant between in-field and photographic consensus grades. The consensus of graders were much more likely to call TI on photographic review than on an in-field examination. Overcalling TI on photographs was more common in eyes that had been graded as TF in the field. For example, graders overcalled TI on photographs in 19 of 120 eyes (15.8%) that had a consensus in-field grade of TF, but only in 3 of 149 eyes (2.0%) that had a consensus grade of TF absent. In contrast, graders tended to call TF at a similar proportion in each setting (Figure 3).

Difficulty of classification

Graders unanimously agreed on the presence or absence of TF for 215 eyes in the field and 179 eyes on photographic review. Of 54 eyes with TF disagreements in the field, 34 (63.0%) also had disagreement on photographic review. In contrast, of the 215 eyes with no disagreement on TF in the field, only 56 (26.0%) had disagreement on photographic review (odds ratio, OR, 4.82, 95% CI 2.57-9.07, logistic regression). Results were similar for TI, though not statistically significant; 18 eyes with TI disagreement in the field, of which 6 (33.3%) also had disagreement on photographic review, compared with 251 eyes without disagreement on TI in the field, of which 42 (16.7%) had disagreement on photographic review (OR 2.49, 95% CI 0.88-7.00).

Intra-rater agreement for photographic grades

We assessed intra-rater agreement in a set of 30 repeat photographic images. Of these, graders 1, 2, and 3 diagnosed TF in 14, 16, and 15 of the original photographs, and TI in 7, 16, and 11 of the original photographs, respectively. Figure 5 shows contingency tables comparing the 2 sets of repeat images for each grader, stratified by the consensus in-field grade. Intra-rater agreement was perfect for each of the graders for both TF and TI, with the majority of grades also agreeing with those given in the field.

Discussion

In this study, we found that trachoma grading by 3 independent graders was more reproducible when done in field conditions compared with photographic review. Photographic grading generally overcalled TI relative to in-field grades, although this was most common among eyes that also had TF. The consensus photographic grade was highly concordant with the consensus in-field grade.

The original studies that tested the reproducibility of the WHO simplified grading system found substantial agreement between graders for the diagnosis of TF and moderate to substantial agreement for TI.[1,10] The current study is consistent with these earlier studies and confirms the high reproducibility of the WHO simple grading system, even by individuals who are not eye specialists.

Conjunctival photographs offer several potential advantages over in-field grades when assessing trachoma. Photographs capture an image that can be evaluated in ideal lighting conditions, without the time pressures of examining a child's everted eyelid in the field. Expertise is required to capture high quality images in the field, but field staff need not have expertise in trachoma grading; an advantage for trachoma programs that use non-ophthalmic health personnel for field work. Photographs can easily be assessed in a masked fashion, which is important for research studies and would be ideal for programmatic surveillance. Photographs also easily allow multiple independent grades and hence a consensus diagnosis, which could theoretically eliminate some of the inter-rater variability when grading for trachoma. This study was designed to test this last potential advantage. We had the same 3 graders grade for trachoma in field conditions, and then again from photographs of the same eyes presented in a random order. Contrary to our study hypothesis, we found that

photographic grades were significantly less reproducible between 3 graders than were in-field grades.

We can speculate why these 3 trachoma graders agreed less when grading photographs than when grading in the field. All field grades were performed in a masked fashion, but the graders were working together and could have communicated non-verbally. If this were the case, we would have expected convergence of grades over time, evident by an improvement in the in-field inter-rater κ as the graders performed more examinations. Instead, we found a relatively stable in-field κ. All 3 graders were experienced in the field, and all 3 had been tested on sets of conjunctival photographs during training sessions for the clinical trial. However, graders 1 and 2 had less photographic grading experience than grader 3. Graders 1 and 2 tended to have lower agreement between their in-field and photographic grades, suggesting that a lack of photographic grading experience may have contributed to the lower inter-rater agreement for photographic grading. Furthermore, all 3 had been trained on photographs and in the field by the same mentor, but they never met before commencing photographic grading to establish grading rules. Finally, in-field grading allows an examination in 3 dimensions; by moving in relation to the subject the grader is able to see whether follicles are raised, and gives more opportunity to see blood vessels. The ability for each examiner to dynamically examine the conjunctiva may have resulted in more agreement.

Graders generally overcalled TI on photographs compared to in the field, whereas they called a similar proportion of TF from photographs and while in the field. We found this relationship for each of the 3 individual graders, and also for the consensus grades. TI tended to be overcalled in eyes with TF, suggesting that this phenomenon would not result in an overestimate of the prevalence of clinically active trachoma if TI were included in the definition. Overcalling of trachomatous inflammation in photographs was reported in a previous study comparing in-field and photographic trachoma assessments, although that study did not find overcalling of TI, but rather overcalling of subtler grades of conjunctival inflammation.[5] It is unclear exactly why graders were more likely to call TI from photographs; however, this may be related to the difficulty of determining ‘thickening’ in a 2-dimensional photograph compared to a real-time, dynamic examination, to differences in lighting when viewing an everted conjunctiva versus when viewing a flash photograph, or due to the camera settings. Further research could be performed to determine the optimal camera settings at which TI is not overcalled.

We carefully designed this study to remove several potential types of variance. For example, we had the same 3 graders perform the in-field and photographic grades, and we used the same set of conjunctiva for both the in-field and photographic settings. All graders performed photographic grading with the same laptop computer monitor and in an equally dark environment. We also took care to minimize bias in the study. All in-field and photographic grading was performed masked to the results of the other graders. In-field and photographic grades were performed within 1 week of each other to minimize grading drift. In order to reduce the chances of the graders making a photographic diagnosis based on knowledge of the prevalence of trachoma encountered days earlier in the field, we included an additional 40 conjunctival photographs (15% of the sample size) from a different study.

This study supports the use of a consensus trachoma grade when using conjunctival photographs to monitor trachoma in areas with hyperendemic trachoma. Although the grades of individual graders were more likely to vary when judging photographs than when examining conjunctivae in the field, the consensus photographic grade achieved very high agreement with the consensus in-field grade. This suggests that when other benefits of photographic grading warrant its use (eg, as a masked outcome in research studies), the accuracy of photographic grading will be improved by having multiple graders independently grade each image. By definition, the consensus grade changed the results only for those cases that were more difficult to grade (ie, cases where 1 of the graders disagreed with the other 2 graders). This suggests that consensus grading would be most helpful in situations with a relatively high level of diagnostic uncertainty. For example, individuals treated with azithromycin may have persistent but smaller conjunctival follicles which could be graded differently by various graders; these cases could benefit from a consensus grade. Consensus grading may also be useful in programmatic settings where enhanced reliability is required. For example, the WHO recommends stopping trachoma programs when the prevalence of TF among children aged 1-9 years falls below 5%.[2] Communities with a prevalence approaching this 5% value would benefit from a more confident assessment of TF; something which consensus grading could provide. The advantages of consensus grading are not limited to photographic grading, as in-field grading would also be more reliable if done by consensus. However, consensus grading would likely be easier to implement and require less manpower if performed with photographs.

Despite the careful design, we acknowledge several limitations. The study was conducted in an area with hyperendemic trachoma, and its findings may not apply to areas with less prevalent trachoma. The use of the κ statistic depends in part on the proportion of the study population on which it is difficult to agree, with lower κ statistics generated when more of the observations are difficult cases.[11] We could not directly calculate the number of difficult cases in the current study because we only had 3 graders, and so all cases with disagreement had the same number of disagreements (ie, 1 grader disagreed with the other 2 graders). However, we did find that 34 of the 269 eyes (12.6%) caused disagreement for both in-field and photographic grading, suggesting that the proportion of difficult cases was not negligible. In addition, a limitation of the κ statistic is its reduction of an $r \times c$ contingency table into a single number, which results in a loss of information and inability to assess the degree of over- and under-calling by graders. It is important to point out that this study applies only to these 3 graders, who may not be generalizable to graders in general. The study is limited by the lack of information about ocular chlamydial infection, which is both the cause of clinical trachoma and also the state that trachoma grading attempts to capture. Finally, the study would have benefited from a larger sample size, as evidenced by the relatively wide CIs around our reliability estimates.

In summary, we did not find evidence to support our hypothesis that inter-rater agreement would be better for photographic trachoma grading than for in-field grading. Although unclear whether the study can be generalized outside the 3 experienced trachoma graders who participated in it, we did not find convincing evidence to support a change in practice based on inter-rater reliability alone. The study supports the use of photographic grading

when in-field grading is not feasible or when aspects of photographic grading are desirable (eg, masking), and suggests increased agreement with the in-field diagnosis when multiple trained graders assess each photograph.

Acknowledgements

We thank Melkam Andualem for her expert trachoma grading. We also thank Donald Everett (National Eye Institute, Bethesda, MD, USA), who was the program officer for the underlying clinical trial; the data safety and monitoring committee including William Barlow (University of Washington, Seattle, WA, USA; Chair), Donald Everett (National Eye Institute, Bethesda, MD, USA), Larry Schwab (International Eye Foundation, Kensington, MD, USA), Arthur Reingold (University of California, Berkeley, CA, USA), Serge Resnikoff (Brien Holden Vision Institute, Sydney, Australia, and International Health and Development, Geneva, Switzerland), and Patricia Buffler (University of California, Berkeley); the Goncha Siso Enese Woreda Health Office, including Abreham Tadesse and Tsegaye Tsehay; the head of Amhara Regional State Health Bureau, Ayeligne Mulualem; and the Ethiopian Federal Ministry of Health.

Financial support: This work was supported by the National Institutes of Health (NEI U10 EY016214 and NIH/NEI K23EY019071), the Bernard Osher Foundation, That Man May See, the Harper Inglis Trust, the Bodri Foundation, the South Asia Research Fund, Research to Prevent Blindness, and the International Trachoma Initiative. The trachoma control program in Amhara is supported by the Amhara Regional Health Bureau, the Lions-Carter Center SightFirst Initiative and many individual donors.

References

1. Thylefors B, Dawson CR, Jones BR, West SK, Taylor HR. A simple system for the assessment of trachoma and its complications. Bull World Health Organ. 1987; 65(4):477–483. [PubMed: 3500800]
2. Solomon, AW.; Zondervan, M.; Kuper, H.; Mabey, DC.; Foster, A. Trachoma control : a guide for programme managers. World Health Organization; Geneva: 2006.
3. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003; 138(1):W1–12. [PubMed: 12513067]
4. West SK, Taylor HR. Reliability of photographs for grading trachoma in field studies. Br J Ophthalmol. 1990; 74(1):12–13. [PubMed: 2306438]
5. Roper KG, Taylor HR. Comparison of clinical and photographic assessment of trachoma. Br J Ophthalmol. 2009; 93(6):811–814. [PubMed: 19304582]
6. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011; 64(1):96–106. [PubMed: 21130355]
7. Brennan RL, Prediger DJ. Coefficient Kappa - Some Uses, Misuses, and Alternatives. Educ Psychol Meas. 1981; 41(3):687–699.
8. Randolph, JJ. Joensuu Learning and Instruction Symposium. Joensuu; Finland: 2005. Free-Marginal Multirater Kappa (multirater κfree): An Alternative to Fleiss' Fixed- Marginal Multirater Kappa..
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar; 1977 33(1):159–174. [PubMed: 843571]
10. Taylor HR, West SK, Katala S, Foster A. Trachoma: evaluation of a new grading scheme in the United Republic of Tanzania. Bull World Health Organ. 1987; 65(4):485–488. [PubMed: 3500801]
11. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol. 2005; 58(7):655–661. [PubMed: 15939215]

Grader 1	Grader 2	Grader 3	In-field TF	In-field TI	Photographic TF	Photographic TI	
+	+	+	98	11	69	23	
-	+	+	4	0	36	11	← G1 undercalling
+	-	+	4	6	8	7	← G2 undercalling
+	+	-	14	1	6	0	← G3 undercalling
-	-	-	117	240	110	198	
+	-	-	13	4	3	0	← G1 overcalling
-	+	-	11	1	28	2	← G2 overcalling
-	-	+	8	6	9	28	← G3 overcalling

Figure 1. Comparison of trachoma grades from 3 examiners

3 experienced trachoma graders performed a series of 269 conjunctival examinations, then graded photographs of the same set of conjunctivae. Columns report all combinations of grades given by the 3 graders for follicular trachoma (TF) and intense trachomatous inflammation (TI), both for in-field and photographic grading. For non-unanimous examinations, 1 of the graders undercalled or overcalled disease relative to the other 2 graders, as pointed out to the right of the table (G1=grader 1, G2=grader 2, G3=grader 3).

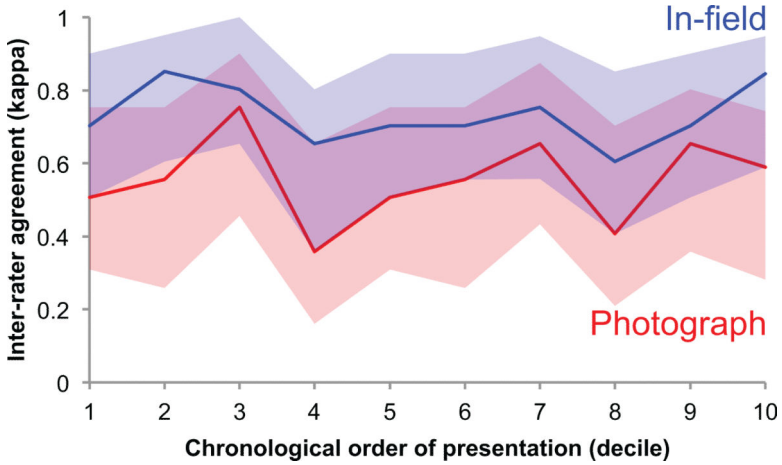

Figure 2. *Kappa* coefficients for 3 trachoma graders over the duration of the study

The *kappa* coefficient with 95% confidence intervals is shown for each decile of graded eyes, in the order that the eyes were examined in the field or by photographic review.

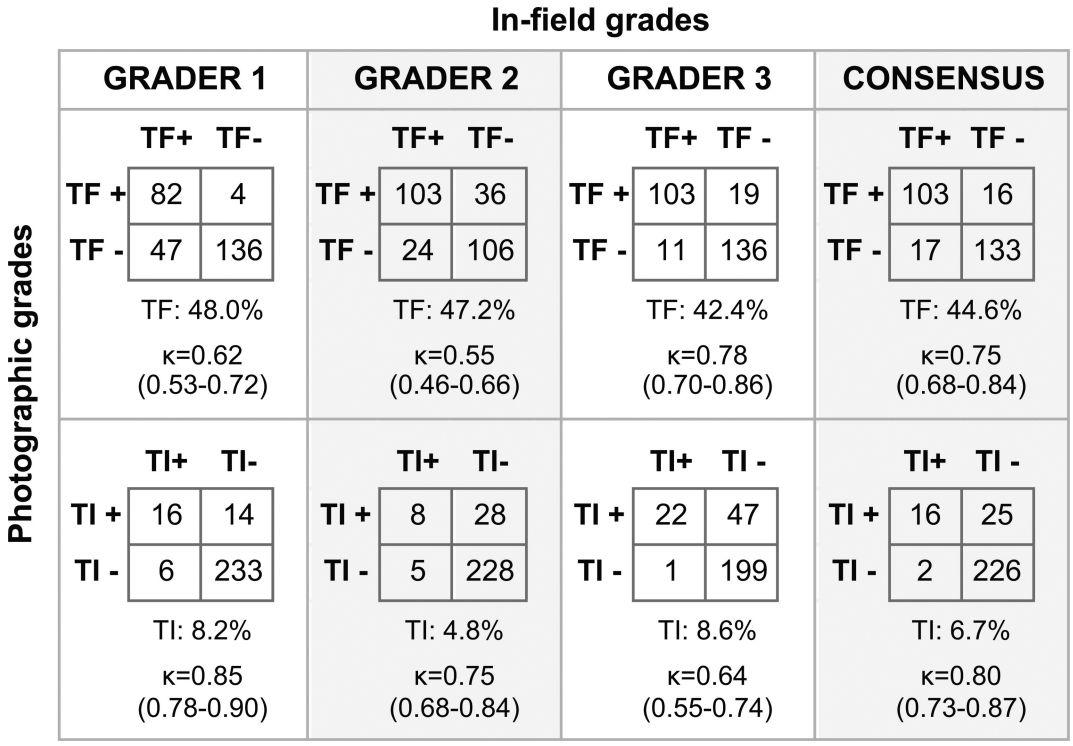

Figure 3. Comparison of in-field versus photographic trachoma grades

A 2×2 table is shown for each of 3 graders and for the consensus grade (the grade for which at least 2 graders agreed), for 269 conjunctival examinations. The prevalence of follicular trachoma (TF) or intense trachomatous inflammation (TI) as assessed by each grader from in-field grades is shown below each table. Also shown is the free-marginal kappa coefficient (κ) assessing agreement between in-field and photographic grades, with 95% confidence intervals in parentheses.

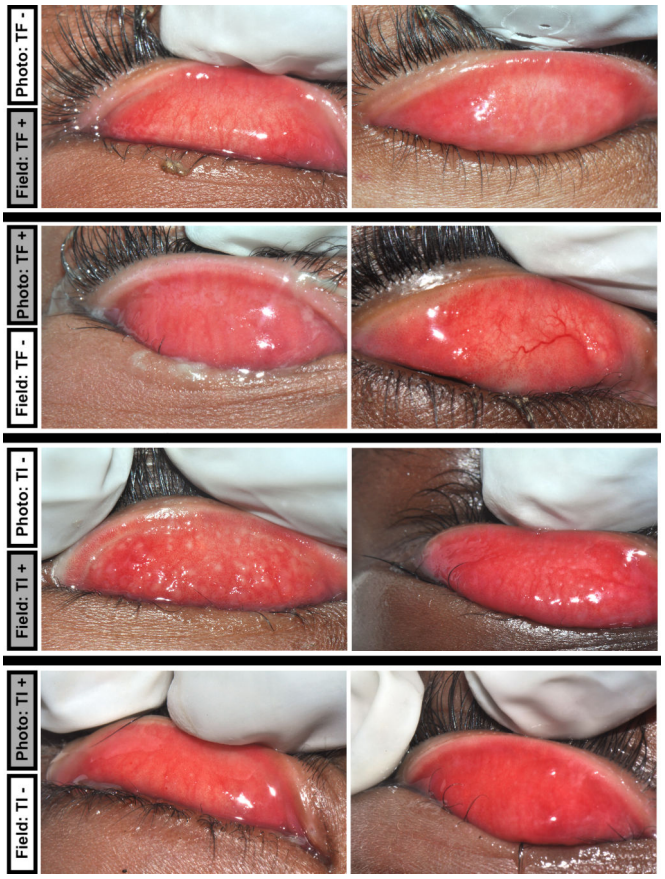

Figure 4. Disagreements between consensus in-field trachoma grade and consensus photographic trachoma grade among 3 graders

We randomly selected 2 cases from each category of discrepancy. The first row shows cases where the field consensus was follicular trachoma (TF) but photographic consensus was absence of TF; the second row shows cases where the field consensus was absence of TF but the photographic consensus was presence of TF; the third row shows cases where the field diagnosis was intense trachomatous inflammation (TI) but the photographic consensus was absence of TI (note, these are the only 2 examples where this was the case); the fourth row shows cases where the field diagnosis was absence of TI but the photographic diagnosis was presence of TI.

	IN-FIELD	GRADER 1			GRADER 2			GRADER 3		
TF	In-field consensus TF+ (N=17)		TF+	TF-		TF+	TF-		TF+	TF -
		TF +	14	0	TF +	15	0	TF +	14	0
		TF -	0	3	TF -	0	2	TF -	0	3
	In-field consensus TF- (N=13)		TF+	TF-		TF+	TF-		TF+	TF-
		TF +	0	0	TF +	1	0	TF +	1	0
		TF -	0	13	TF -	0	12	TF -	0	12
		κ=1.0			κ=1.0			κ=1.0		

	IN-FIELD	GRADER 1			GRADER 2			GRADER 3		
TI	In-field consensus TI+ (N=5)		TI+	TI-		TI+	TI-		TI+	TI-
		TI +	5	0	TI +	5	0	TI +	5	0
		TI -	0	0	TI -	0	0	TI -	0	0
	In-field consensus TI- (N=25)		TI+	TI-		TI+	TI-		TI+	TI -
		TI +	2	0	TI +	2	0	TI +	6	0
		TI -	0	23	TI -	0	23	TI -	0	19
		κ=1.0			κ=1.0			κ=1.0		

Figure 5. Intra-rater reliability of photographic trachoma grading, stratified by consensus trachoma grade given in the field

30 duplicate images were presented to each of 3 trachoma graders in a random order. Each grader demonstrated perfect agreement; most grades given by photographic review were consistent with the consensus grade given in the field. TF, follicular trachoma; TI, intense trachomatous inflammation.

Table 1

Pairwise kappa statistics for in-field and photographic trachoma grading of follicular trachoma (TF) and intense trachomatous inflammation (TI)

	Grader 1 vs grader 2, *k (95% CI)*	Grader 1 vs grader 3, *k (95% CI)*	Grader 2 vs grader 3, *k (95% CI)*
In-field			
TF	0.76 (0.67-0.83)	0.71 (0.63-0.80)	0.72 (0.65-0.80)
TI	0.92 (0.86-0.96)	0.92 (0.86-0.96)	0.90 (0.83-0.94)
Photographic			
TF	0.44 (0.34-0.55)	0.60 (0.51-0.70)	0.62 (0.53-0.72)
TI	0.85 (0.78-0.90)	0.71 (0.63-0.80)	0.72 (0.64-0.81)

CI, confidence interval