Title

Importance of Including Borderline Cases in Trachoma Grader Certification

Journal

American Journal of Tropical Medicine and Hygiene, 91(3)

ISSN

0002-9637

Authors

Gaynor, Bruce D
Amza, Abdou
Gebresailassie, Sintayehu
et al.

Publication Date

2014-09-01

DOI

10.4269/ajtmh.13-0658

Peer reviewed

Am. J. Trop. Med. Hyg., 91(3), 2014, pp. 577–579
doi:10.4269/ajtmh.13-0658

Importance of Including Borderline Cases in Trachoma Grader Certification

Bruce D. Gaynor,* Abdou Amza, Sintayehu Gebresailassie, Boubacar Kadri, Baido Nassirou, Nicole E. Stoller, Sun N. Yu, Puja A. Cuddapah, Jeremy D. Keenan, and Thomas M. Lietman

F. I. Proctor Foundation, Department of Ophthalmology, Department of Epidemiology and Biostatistics, Institute for Global Health, University of California, San Francisco, California; Programme National de Lutte Contre la Cecité Niamey, Niger; The Carter Center, Addis Ababa, Ethiopia

Abstract. We assessed trachoma grading agreement among field graders using photographs that included the complete spectrum of disease and compared it with cases where there was consensus among experienced graders. Trained photographers took photographs of children's conjunctiva during a clinical trial in Ethiopia. We calculated κ-agreement statistics using a complete set of 60 cases and then recalculated the κ using a consensus set where cases were limited to those cases with agreement among experienced graders. When the complete set of 60 cases was used, agreement was moderate (κ = 0.61, 95% confidence interval [95% CI] = 0.56–0.67). When the consensus set was used, agreement improved significantly (κ = 0.75, 95% CI = 0.68–0.80). The κ of the consensus set was higher than the complete set by 0.14 (95% CI = 0.12–0.16) ($P < 0.001$). If testing sets remove difficult-to-grade cases, agreement in trachoma grading may be higher than actually seen in population-based trachoma surveys.

INTRODUCTION

The World Health Organization (WHO) aims to eliminate trachoma as a public health problem by 2020 in large part by using mass antibiotic distributions. Guidelines for starting and stopping antibiotic treatments are entirely based on eye examination using the WHO simplified grading scale.[1] Trachoma programs prioritize treatment areas by clinical grades, and the WHO's Ultimate Intervention Goals (UIGs) assess grades when determining trachoma elimination as a public health problem.[2,3] Individual- and community-randomized clinical trials have typically had clinical exam as primary or secondary outcomes.[4–9]

Unfortunately, grading of clinical trachoma can have considerable variability. Even with rigorous training, different examiners grade cases differently. Experienced graders may not agree on the clinical grade of borderline cases.[10] Photography allows for construction of training sets of clinical trachoma with whatever characteristics are desired. For example, training and testing sets can include the complete spectrum of cases or be limited to those cases where there is consensus among experienced graders.

Agreement with experienced trachoma graders using a κ-statistic is the most common method currently used for certifying competence of field graders.[11–13] The prevalence of trachoma follicular (TF) and trachoma intense (TI) are the signs used for assessing the need for a control program and monitoring its results.[14] Here, we assessed trachoma grading agreement using photographs from a trachoma-endemic area in Ethiopia that included the complete spectrum of disease, and then, we reassessed agreement when cases were limited to those cases in which there was consensus among three experienced graders.

METHODS

We performed a cluster-randomized clinical trial evaluating different mass treatment strategies for trachoma in Amhara, Ethiopia from 2006 to 2013.[15,16] In fall of 2011, after 5 years of mass antibiotic treatment per study protocol, a trained photographer took a minimum of two photographs of the right averted superior tarsal conjunctiva of each child participant 0–5 years of age using a Nikon D-series camera with a Micro Nikkor 105 mm f/2.8 lens (Nikon Corporation, Chiyoda, Tokyo, Japan). Photographs were reviewed and selected for inclusion in this study if they were well-focused, centrally positioned, horizontally oriented, and without a tear lake. After this initial photo screening, a testing set of 60 cases was selected from the screening set that was evenly divided between clinically active and clinically inactive trachoma (TF and/or TI by the WHO system)[1] using preliminary grades from one experienced grader. Photos were chosen for the testing set, because they were clear representatives of TF and/or TI. Borderline images were not excluded from the testing set if they met inclusion criteria.

We trained 27 potential graders on the WHO simplified grading system for trachoma studies in Ethiopia (14 graders), Niger (9 graders), and Tanzania (4 graders). All graders had at least one previous training session before the training associated with this study. The 60 cases were graded by three experienced graders (B.D.G., J.D.K., and T.M.L.) masked to each other's assessment. A consensus grade was defined as the grade on which two or three of three graders agreed. Cases were classified into consensus and complete (all 60 cases were included) groups, and the 27 potential graders were tested on these two groups.

ANALYSIS

κ-Statistics for 27 graders for the complete set of 60 slides were calculated. κ-statistics were then recalculated for consensus cases, in which the experienced graders were in agreement. Bootstrap 95% percentile confidence intervals were determined by resampling graders ($N = 999$). A P value for the difference between paired groups (a grader's κ with the consensus test set versus the κ with the complete test set) was determined by bootstrap resampling of the graders. Density plots of the distribution of κ-values for the graders were made using a Gaussian kernel and Scott's rule of thumb for kernel width. All calculations were performed in Mathematica 9.0 (Wolfram Research, Champaign, IL).

*Address correspondence to Bruce D. Gaynor, F. I. Proctor Foundation, Department of Ophthalmology, University of California at San Francisco, 513 Parnassus, Med Sci 334C, San Francisco, CA 94143. E-mail: bruce.gaynor@ucsf.edu

RESULTS

According to the census in November of 2011, 824 children ages 0–5 years were enrolled in the study from 24 villages. All children were photographed, and 799 (97.0%) individuals had at least one photograph that satisfied inclusion criteria for quality. Sixty (7.3%) photographs were then selected for the complete set in this study. There was consensus among the three experienced graders in 45 (75%) of the TF cases and 44 (73%) of the TI cases within the complete set. The grading agreement was spread evenly among the three experienced graders; two graders agreed in 52 (87%) of the TF cases, whereas two different graders agreed in 50 (83%) of the TI cases. When the complete set of 60 cases was used in testing 27 trachoma graders, the mean κ-score for TF was 0.611 (95% confidence interval [95% CI] = 0.561–0.672) (Figure 1A). When cases were limited to those cases in which there was consensus between the experienced graders, the mean κ-score for TF increased to 0.748 (95% CI = 0.683–0.804). Consensus cases for TF yielded a κ-score of 0.137 (95% CI = 0.121–0.155) higher than complete cases ($P < 0.001$).

When the complete set of 60 cases was used, the mean κ-score of 27 graders for TI was 0.605 (95% CI = 0.519–0.684) (Figure 1B). When cases were limited to those cases in which there was consensus among experienced graders, the mean κ-score for TI increased to 0.789 (95% CI = 0.688–0.885). Consensus cases for TI yielded a κ-score of 0.185 (95% CI = 0.140–0.224) higher than complete cases ($P < 0.001$).

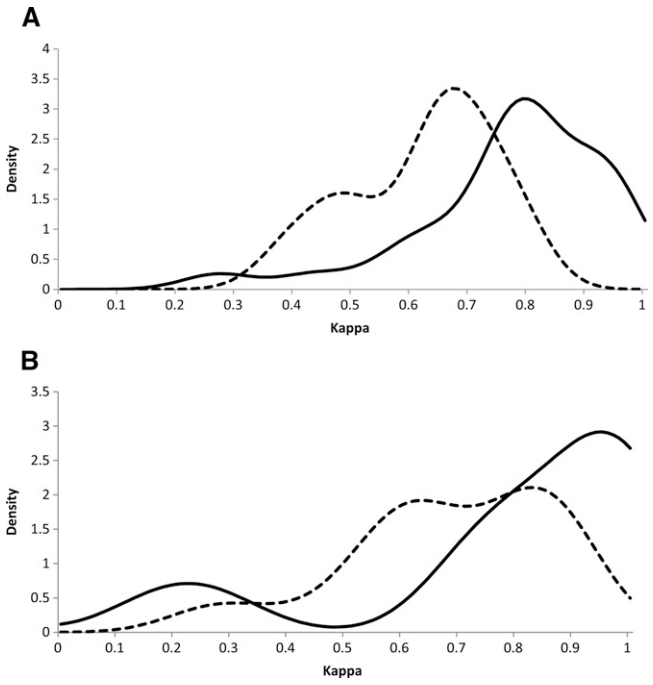

FIGURE 1. (**A**) Density of the κ-statistic of 27 trachoma graders for TF test photographs with the complete spectrum of disease (dashed curve) versus test photographs limited to those photographs for which experienced graders were in consensus (black curve). Note that far more graders would be certified with a threshold of 0.6 if the consensus set were used. (**B**) A similar density of κ for 27 trachoma graders for TI.

DISCUSSION

When training prospective field workers for trachoma grading, we found higher agreement when cases were limited to those cases in which there was consensus among experienced graders than when a complete set of cases was used where there was no consensus among experienced graders. We found a 0.14 difference between the consensus and complete cases in κ-statistics. Thus, agreement with a consensus test set is higher than would be expected in a population-based survey that includes the complete spectrum of disease.

To ensure that grader training is applicable to our clinical setting, the cases in the training should be representative of those cases likely to be seen. The clinical findings in test sets should be as easy or difficult to discern as in our own clinical population. Neither obvious nor very subtle findings should be overrepresented. Studies in which participants with ambiguous or borderline results are undersampled or not included at all are not generalizable to populations in which most trachoma programs operate.

Agreement with experts or κ-statistics have been reported in other trachoma grading studies, and these statistics show a high level of agreement.[11,12,17,18] However, it is not clear if borderline cases were included or if cases were limited to consensus cases. Intergrader reliability compares the results of measurements made by different graders. It could be argued that inclusion of a test photograph where expert graders are split 50% to 50% does not provide information; whether a trainee guesses normal or active, they will agree with experts 50% of the time. Exclusion of these cases, however, can artificially inflate apparent agreement, at least when using the κ-statistic. There are other measures that can capture how graders perform on marginal cases. The reliability error obtained as a decomposition of the Brier Score assesses whether people grade 50% of cases active approximately 50% of the time, 20–80% of cases active approximately 20% of the time, etc.[19] Reliability is a measure of how close predicted probabilities are to true probabilities. For binary events, uncertainty is maximum when the event occurs 50% of the time, and the uncertainty is zero if the event always occurs.

This study has some limitations that might affect its generalizability. We used three experienced graders to determine unanimous and discrepant cases, although other experienced graders may agree and disagree on different cases. We only used cases from one country (Ethiopia) with hyperendemic trachoma, and the spectrum of disease is different in meso- or hypoendemic areas.

Photography has been used for trachoma grading, and technology has advanced with digital imaging.[4,6,13,20–22] Photo grading has advantages over field grading. (1) Exam conditions can be standardized. (2) Grading sets can be created with sufficient case variability to increase the chance of obtaining highly qualified graders, regardless of existing disease prevalence. (3) Fewer graders need to be trained and regularly recertified. (4) Photo grades can be audited at any later time. Digital photography has advantages over film photography, including (1) photos can be reviewed in real time and accepted or rejected, increasing acquisition of high-quality images, (2) camera settings can be tested and adjusted quickly as needed, and (3) film purchase and laboratory processing is unnecessary.

In the future, programs may be guided by laboratory testing for infection (e.g., polymerase chain reaction), but programs will continue to use the clinical exam until then in the form of either field examination or photographic examination. Testing of graders using cases that reflect sufficient disease variability is critical to ensure that only high-quality graders obtain certification.

Trachoma grading test sets that include the complete spectrum of clinical disease result in lower intergrader agreement than sets limited to cases where there is consensus among experienced graders. Testing limited to easier cases will exaggerate agreement compared with testing over the complete spectrum of disease, which may result in lower certification thresholds; therefore, unqualified graders are more likely to be used. Overestimating the quality of graders could result in overtreatment, undertreatment, or incorrectly certifying elimination of trachoma as a public health problem. Including borderline cases ensures that only qualified trachoma graders are certified.

Received November 12, 2013. Accepted for publication January 31, 2014.

Published online July 7, 2014.

Acknowledgments: The authors thank the data and safety monitoring committee, including William Barlow (University of Washington, Washington, DC; Chair), Donald Everett (National Eye Institute, Bethesda, MD), Larry Schwab (International Eye Foundation, Kensington, MD), Arthur Reingold (University of California, Berkeley, CA), and Serge Resnikoff (Brien Holden Vision Institute, University of New South Wales, Sydney, Australia), who were generous with their time and advice and met before, during, and after this study. Trial registration: ClinicalTrials.gov NCT00792922.

Financial support: This work was supported by National Institutes of Health Grants NIH/NEI U10 EY016214 and K23 EYO19881-01 and National Institutes of Health/University of California San Francisco-Clinical & Translational Science Institute (CTSI) Grant Number KL2 TR000143.

Disclaimer: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The corresponding author had full access to all of data in the study and had final responsibility for the decision to submit for publication.

Authors' addresses: Bruce D. Gaynor, Nicole E. Stoller, Sun N. Yu, Puja A. Cuddapah, Jeremy D. Keenan, and Thomas M. Lietman, F. I. Proctor Foundation, University of California, San Francisco, CA, E-mails: Bruce.Gaynor@ucsf.edu, Nicole.Stoller@ucsf.edu, Sun.Yu@ucsf.edu, Puja.Cuddapah@ucsf.edu, Jeremy.Keenan@ucsf.edu, and Tom.Lietman@ucsf.edu. Abdou Amza, Boubacar Kadri, and Baido Nassirou, Programme Faculté des Sciences de la Santé/Université Abdou Moumouni de Niamey, Programme Nationale des Soins Oculaire, Niger, E-mails: dr.amzaabdou@gmail.com, boubacarkadri@gmail.com, and nasbeido@yahoo.fr. Sintayehu Gebresailassie, The Carter Center, Addis Ababa, Ethiopia, E-mail: sint_selam@yahoo.com.

REFERENCES

1. Thylefors B, Dawson CR, Jones BR, West SK, Taylor HR, 1987. A simple system for the assessment of trachoma and its complications. *Bull World Health Organ 65:* 477–483.
2. Mariottii SP, 2003. Proceedings of the 2nd Global Scientific Meeting on Trachoma; Geneva, Switzerland.
3. WHO, 2012. *Accelerating Work to Overcome the Global Impact of Neglected Tropical Diseases: A Roadmap for Implementation*. Geneva: World Health Organization.
4. Emerson PM, Lindsay SW, Alexander N, Bah M, Dibba SM, Faal HB, Lowe KO, McAdam KP, Ratcliffe AA, Walraven GE, Bailey RL, 2004. Role of flies and provision of latrines in trachoma control: cluster-randomised controlled trial. *Lancet 363:* 1093–1098.
5. Jha H, Chaudary JS, Bhatta R, Miao Y, Osaki-Holm S, Gaynor B, Zegans M, Bird M, Yi E, Holbrook K, Whitcher JP, Lietman T, 2002. Disappearance of trachoma from Western Nepal. *Clin Infect Dis 35:* 765–768.
6. Michel CE, Roper KG, Divena MA, Lee HH, Taylor HR, 2011. Correlation of clinical trachoma and infection in Aboriginal communities. *PLoS Negl Trop Dis 5:* e986.
7. Solomon AW, Harding-Esch E, Alexander ND, Aguirre A, Holland MJ, Bailey RL, Foster A, Mabey DC, Massae PA, Courtright P, Shao JF, 2008. Two doses of azithromycin to eliminate trachoma in a Tanzanian community. *N Engl J Med 358:* 1870–1871.
8. West S, Muñoz B, Lynch M, Kayongoya A, Chilangwa Z, Mmbaga BB, Taylor HR, 1995. Impact of face-washing on trachoma in Kongwa, Tanzania. *Lancet 345:* 155–158.
9. Schachter J, West SK, Mabey D, Dawson CR, Bobo L, Bailey R, Vitale S, Quinn TC, Sheta A, Sallam S, Mkocha H, Faal H, 1999. Azithromycin in control of trachoma. *Lancet 354:* 630–635.
10. See CW, Alemayehu W, Melese M, Zhou Z, Porco TC, Shiboski S, Gaynor BD, Eng J, Keenan JD, Lietman TM, 2011. How reliable are tests for trachoma?–a latent class approach. *Invest Ophthalmol Vis Sci 52:* 6133–6137.
11. Amza A, Kadri B, Nassirou B, Stoller NE, Yu SN, Zhou Z, Chin S, West SK, Bailey RL, Mabey DCW, Keenan JD, Porco TC, Lietman TM, Gaynor BD; Partnership PRET, 2012. Community risk factors for ocular chlamydia infection in Niger: pretreatment results from a cluster-randomized trachoma trial. *PLoS Negl Trop Dis 6:* e1586.
12. Hassan A, Ngondi JM, King JD, Elshafie BE, Al Ginaid G, Elsanousi M, Abdalla Z, Aziz N, Sankara D, Simms V, Cromwell EA, Emerson PM, Binnawi KH, 2011. The prevalence of blinding trachoma in northern states of Sudan. *PLoS Negl Trop Dis 5:* e1027.
13. Solomon AW, Bowman RJ, Yorston D, Massae PA, Safari S, Savage B, Alexander ND, Foster A, Mabey DC, 2006. Operational evaluation of the use of photographs for grading active trachoma. *Am J Trop Med Hyg 74:* 505–508.
14. Solomon AW, Zondervan M, Kuper H, Buchan JC, Mabey DCW, Foster A, 2006. *Trachoma Control: A Guide for Programme Managers*. Geneva: World Health Organization.
15. Gebre T, Ayele B, Zerihun M, Genet A, Stoller NE, Zhou Z, House JI, Yu SN, Ray KJ, Emerson PM, Keenan JD, Porco TC, Lietman TM, Gaynor BD, 2012. Comparison of annual versus twice-yearly mass azithromycin treatment for hyperendemic trachoma in Ethiopia: a cluster-randomised trial. *Lancet 379:* 143–151.
16. Ayele B, Gebre T, Moncada J, House JI, Stoller NE, Zhou Z, Porco TC, Gaynor BD, Emerson PM, Schachter J, Keenan JD, 2011. Risk factors for ocular *Chlamydia* after three mass azithromycin distributions. *PLoS Negl Trop Dis 5:* e1441.
17. Harding-Esch EM, Edwards T, Mkocha H, Munoz B, Holland MJ, Burr SE, Sillah A, Gaydos CA, Stare D, Mabey DC, Bailey RL, West SK, 2010. Trachoma prevalence and associated risk factors in the Gambia and Tanzania: baseline results of a cluster randomised controlled trial. *PLoS Negl Trop Dis 4:* e861.
18. Stare D, Harding-Esch E, Munoz B, Bailey R, Mabey D, Holland M, Gaydos C, West S, 2011. Design and baseline data of a randomized trial to evaluate coverage and frequency of mass treatment with azithromycin: the Partnership for Rapid Elimination of Trachoma (PRET) in Tanzania and The Gambia. *Ophthalmic Epidemiol 18:* 20–29.
19. Murphy AH, 1996. General decomposition of MSE-based skill scores: measures of some basic aspects of forecast quality. *Am Meteorological Soc 124:* 2353–2369.
20. Roper KG, Taylor HR, 2009. Comparison of clinical and photographic assessment of trachoma. *Br J Ophthalmol 93:* 811–814.
21. West SK, Taylor HR, 1990. Reliability of photographs for grading trachoma in field studies. *Br J Ophthalmol 74:* 12–13.
22. Bhosai SJ, Amza A, Beido N, Bailey RL, Keenan JD, Gaynor BD, Lietman TM, 2012. Application of smartphone cameras for detecting clinically active trachoma. *Br J Ophthalmol 96:* 1350–1351.