Original Paper

A Virtual Reading Center Model Using Crowdsourcing to Grade Photographs for Trachoma: Validation Study

Christopher J Brady[1], MHS, MD; R Chase Cockrell[2], PhD; Lindsay R Aldrich[3], MPH; Meraf A Wolle[4], MPH, MD; Sheila K West[4], PhD

[1]Division of Ophthalmology, Department of Surgery, Larner College of Medicine at The University of Vermont, Burlington, VT, United States
[2]Division of Surgical Research, Department of Surgery, Larner College of Medicine at The University of Vermont, Burlington, VT, United States
[3]Larner College of Medicine at The University of Vermont, Burlington, VT, United States
[4]Dana Center for Preventive Ophthalmology, Wilmer Eye Institute, Baltimore, MD, United States

Corresponding Author:
Christopher J Brady, MHS, MD
Division of Ophthalmology
Department of Surgery
Larner College of Medicine at The University of Vermont
111 Colchester Avenue, Main Campus, West Pavilion, Level 5
Burlington, VT, 05401
United States
Phone: 1 802 847 0400
Email: christopher.brady@med.uvm.edu

Abstract

Background: As trachoma is eliminated, skilled field graders become less adept at correctly identifying active disease (trachomatous inflammation—follicular [TF]). Deciding if trachoma has been eliminated from a district or if treatment strategies need to be continued or reinstated is of critical public health importance. Telemedicine solutions require both connectivity, which can be poor in the resource-limited regions of the world in which trachoma occurs, and accurate grading of the images.

Objective: Our purpose was to develop and validate a cloud-based "virtual reading center" (VRC) model using crowdsourcing for image interpretation.

Methods: The Amazon Mechanical Turk (AMT) platform was used to recruit lay graders to interpret 2299 gradable images from a prior field trial of a smartphone-based camera system. Each image received 7 grades for US $0.05 per grade in this VRC. The resultant data set was divided into training and test sets to internally validate the VRC. In the training set, crowdsourcing scores were summed, and the optimal raw score cutoff was chosen to optimize kappa agreement and the resulting prevalence of TF. The best method was then applied to the test set, and the sensitivity, specificity, kappa, and TF prevalence were calculated.

Results: In this trial, over 16,000 grades were rendered in just over 60 minutes for US $1098 including AMT fees. After choosing an AMT raw score cut point to optimize kappa near the World Health Organization (WHO)–endorsed level of 0.7 (with a simulated 40% prevalence TF), crowdsourcing was 95% sensitive and 87% specific for TF in the training set with a kappa of 0.797. All 196 crowdsourced-positive images received a skilled overread to mimic a tiered reading center and specificity improved to 99%, while sensitivity remained above 78%. Kappa for the entire sample improved from 0.162 to 0.685 with overreads, and the skilled grader burden was reduced by over 80%. This tiered VRC model was then applied to the test set and produced a sensitivity of 99% and a specificity of 76% with a kappa of 0.775 in the entire set. The prevalence estimated by the VRC was 2.70% (95% CI 1.84%-3.80%) compared to the ground truth prevalence of 2.87% (95% CI 1.98%-4.01%).

Conclusions: A VRC model using crowdsourcing as a first pass with skilled grading of positive images was able to identify TF rapidly and accurately in a low prevalence setting. The findings from this study support further validation of a VRC and crowdsourcing for image grading and estimation of trachoma prevalence from field-acquired images, although further prospective field testing is required to determine if diagnostic characteristics are acceptable in real-world surveys with a low prevalence of the disease.

(J Med Internet Res 2023;25:e41233) doi: 10.2196/41233

KEYWORDS

trachoma; crowdsourcing; telemedicine; ophthalmic photography; Amazon Mechanical Turk; image analysis; diagnosis; detection; cloud-based; image interpretation; disease identification; diagnostics; image grading; disease grading; trachomatous inflammation—follicular; ophthalmology

Introduction

Trachoma

Despite intensive worldwide control efforts, trachoma remains the most important infectious cause of vision loss [1-3] and one of the overall leading causes of global blindness [4,5], with nearly 136 million people at risk of losing vision [6]. While the goal set by the World Health Organization (WHO) for the global elimination of trachoma as a public health problem by 2020 [7] was not met, this work contributed to dramatic reductions in the prevalence of the disease [8]. However, as with other diseases with the potential for elimination, the "last mile" may prove to be the most difficult [9,10]. With the decreased global prevalence of trachoma, there is growing recognition from the WHO and other major stakeholders that novel methods will be essential to achieve trachoma elimination [11-13].

Crowdsourcing

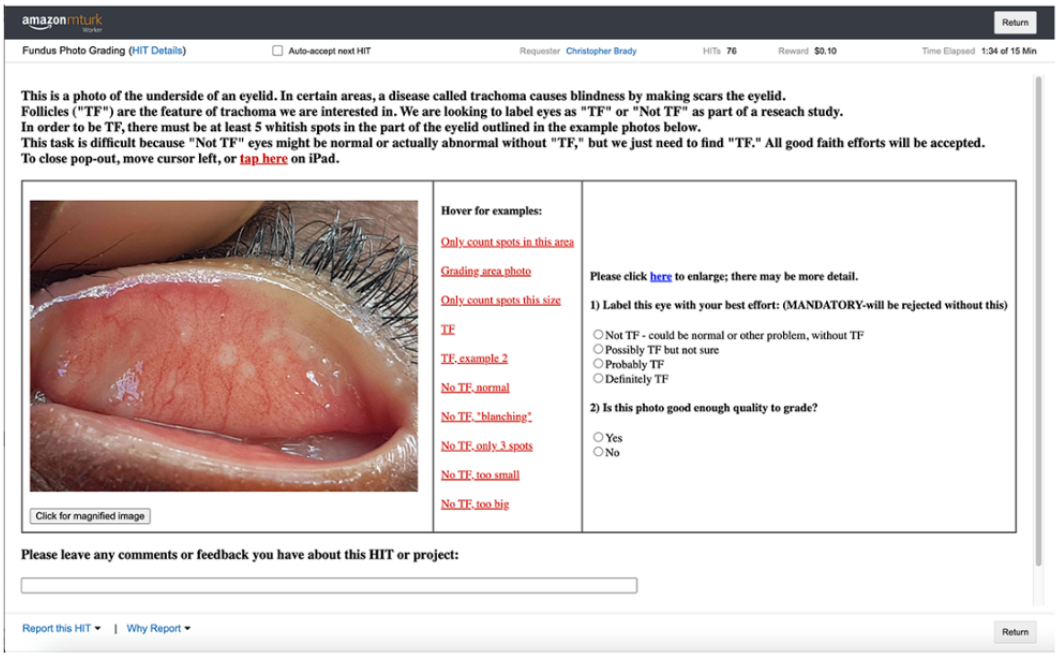
Ophthalmic photography and telemedicine are well-established clinical and research tools useful for many conditions including trachoma [14-17], but the use of imaging at scale in an elimination program that operates almost exclusively in rural areas of resource-poor countries remains unknown. In addition, expert-level grading of images in a conventional telemedicine reading-center model is labor-intensive and expensive, and the high volume of images generated by multiple concurrent elimination surveys has the potential to overwhelm existing grading capacity. For these reasons, crowdsourcing using untrained citizen scientists has been proposed for large image processing tasks and has been successfully used in our prior research on other ophthalmic conditions [18-20]. Previously, our group has demonstrated that a smartphone-based imaging system can acquire high-quality everted upper eyelid photographs during a district-level trachoma survey [21]. In this report, we describe an internal validation of the "virtual reading center" (VRC) paradigm in which crowdsourcing is used to provide a first-pass grading of images to eliminate those without TF (trachomatous inflammation—follicular), thereby substantially reducing the skilled grader burden for trachoma elimination telemedicine.

Methods

Crowdsourcing Platform

A previously designed Amazon Mechanical Turk (AMT) crowdsourcing interface for grading of retinal fundus photographs [18,20,22] was modified to allow for grading of external photographs of the everted upper eyelid (Figure 1). Annotated teaching images were included to allow AMT users to train themselves to identify the features of TF and provide their grade (0="definitely no TF"; 1="possibly TF"; 2="probably TF"; 3="definitely TF") within the same web frame. No additional training or qualification was required before AMT users could complete the tasks. Based on previous experiments on the smallest accurate crowd size [20], 7 individual grades were requested for each image, with US $0.05 compensation provided for each grade.

Figure 1. Amazon Mechanical Turk interface for active trachoma grading using crowdsourcing. HIT: human intelligence task; TF: trachomatous inflammation—follicular.



Ophthalmic Photographs

The data set of 2614 images used for this study was collected during a 2019 district-level survey in Chamwino, Tanzania, as part of the Image Capture and Processing System (ICAPS) development study [21]. During the ICAPS study, all images were assessed for gradability, operationally defined as sufficient resolution of the central upper tarsal plate to confirm or exclude the presence of 5 or more follicles by 2 experts. Gradable images then received a TF grade from 2 experts, with TF being defined as 5 or more follicles, at least 0.5 mm in diameter located in the central upper tarsal plate using the WHO simplified trachoma grading system [23]. Consensus was then decided by discussion of discordant grades. The consensus photo grade was then compared to the grade rendered by a certified field grader. All images used were completely deidentified photographs of a single everted upper eyelid (Figures S1 and S2 in Multimedia Appendix 1). The images were posted for grading on AMT in 2 batches (December 28, 2020, and March 9, 2021). For this validation study, we analyzed the set of 2299 gradable images with concordant expert and field ICAPS grades for TF (n=56 for TF-positive images). This concordant grade was considered the "ground truth" for analysis of VRC grading. Images without a concordant grade were excluded from the analysis.

Optimizing Crowdsourcing

The images were then randomly divided into equal training and test sets. The training set was used to explore and compare several ways of processing the crowdsourced output into a binary "No TF" or "TF" score. The 7 individual AMT grader scores (0-3) were summed to create a single raw score for each image (theoretical range: 0-21). A number of dichotomization "setpoints" along this range were then compared on the basis of their accuracy with the ground truth grade.

The best fitting grading method was defined using the WHO grading criteria from the Global Trachoma Mapping Project [24] as one that could identify TF with a kappa agreement of ≥0.7 with the ground truth grade in a simulated moderate prevalence sample and produce a prevalence within ±2% of the ground truth determined prevalence in the full sample [25]. Because a VRC model was anticipated using crowdsourcing as a first pass, with skilled grader overreads for positive images, minimizing the number of false-negative images while maximizing the reduction in skilled grader burden were secondary considerations during cutoff selection.

First, to mimic the current standard of care for certification of an in-person skilled grader, a simulated intergrader agreement test (IGA) [24] was conducted using 50 randomly selected images (TF n=20 based on ICAPS concordant grade). Cohen kappa agreement with the ground truth was calculated at each raw score in the simulated IGA. The score maximizing kappa was considered as the first possible setpoint to dichotomize the crowdsourced raw scores into "No TF" and "TF" for all the images in the training set.

For the next a priori dichotomization attempt, a "naïve" or "majority rules" setpoint of 50% of the raw score was applied in the training set, in which a raw score of 11 or greater was defined as TF. Finally, the receiver operating characteristic (ROC) was then analyzed and used to explore and sequentially optimize each of the diagnostic characteristics (sensitivity, specificity, percent correct, and kappa) of various cut-offs for the binary diagnosis of TF.

Individual grader performance was also compared to the ground truth to see whether several methods of excluding unreliable scores could improve score aggregation.

The best fitting grading paradigm was then applied to the test set to evaluate the final sensitivity, specificity, percent correct, kappa agreement, and percent TF positive compared with the ground truth values.

Ethical Considerations

This study was deemed nonhuman subjects research by the University of Vermont Institutional Review Board.

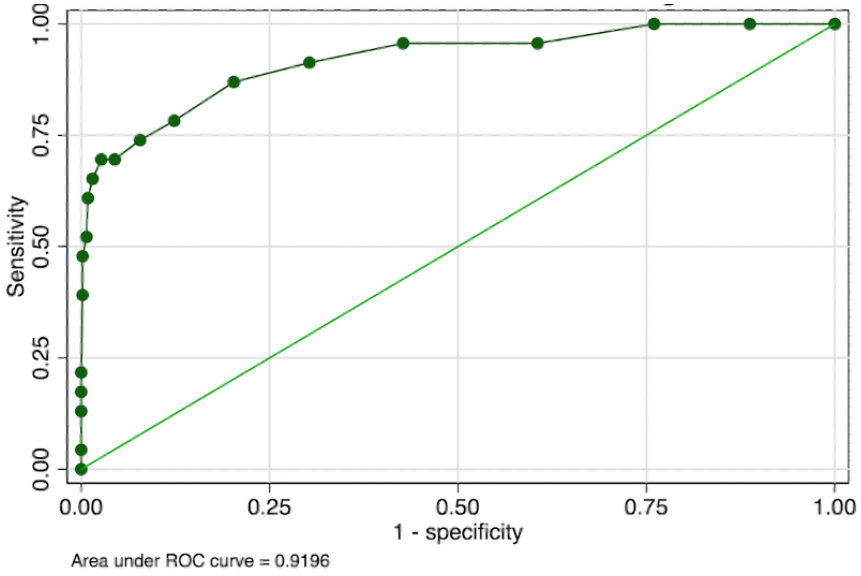
Results

Crowdsourcing Characteristics

The first batch of 1000 images was completed in 61 minutes, by 193 unique AMT users. The median time to grade an image was 18 (IQR 103) seconds, and the median number of images graded by each user was 15 (IQR 49) images with a mode of 1 image. The second batch of 1614 images was completed in 74 minutes, by 193 unique AMT users, with a median grading time of 16 (IQR 85) seconds. A total of 43 users worked on both batches. The median number of images graded per user was 22 (IQR 76) images, and the modal number was again only 1 image. The full distribution of images graded per user is shown in Figure S3 in Multimedia Appendix 1. The total cost for crowdsourced grading of the 2 batches, including AMT commission, was US $1098.

Summing the individual scores for the training images gave a range of 0-20 with a right-skewed distribution (Figure S4 in Multimedia Appendix 1). The area under the ROC (AUROC) for the raw score was 0.920 (95% CI 0.853-0.986) (Figure 2).

Figure 2. ROC of the crowdsourced raw score for the consensus ground truth grade. ROC: receiver operating characteristic.



Optimizing Crowdsourcing

The results of the various attempts to optimize the aggregation of raw scores into a binary TF score using the training set of 1149 images (TF n=23; prevalence 2.00%, 95% CI 1.27%-2.99%) are shown in Table 1. Each choice of a dichotomized setpoint allows for a different balance of diagnostic parameters along the ROC. In the IGA simulation, in which the prevalence of TF was enriched to 40% (95% CI 26.4%-54.8%) in a random subset of 50 images, the optimal kappa of 0.715 and an acceptable prevalence estimate of 46% (95% CI 31.8%-60.7%) were obtained with a cutoff of 6. When the cutoff of 6 was applied in the full sample with a TF prevalence of 2.00% (95% CI 1.27%-3.00%), kappa was reduced to 0.115 and the prevalence was overestimated by a factor of 10 (estimated prevalence: 21.6%, 95% CI 19.2%-24.1%), due to the high number of false-positive results.

The next model applied used a “truncated mean” approach, in which the highest and lowest AMT individual scores for each image were excluded from the image raw score to minimize the effect of outlier scores. The range of raw scores was now 0-15, with a likewise right-skewed distribution. Once again, the IGA simulation was performed, and the maximal kappa was obtained with a lower cutoff of 4.

Using the raw score with a truncated mean, the AUROC was 0.938 (Figure 3). Again, when the IGA-optimized cutoff was applied in the full sample, the diagnostic characteristics were reduced with a kappa of 0.162, despite a sensitivity of nearly 85% and specificity of over 90%. Because no cutoff met the prespecified criteria of kappa ≥0.7 in the IGA-simulated sample and matching prevalence of ±2% in the full sample, a VRC model was explored with skilled overreading of all positive images. Two cutoffs were explored: ≥4, which was selected from the IGA simulation with truncated means, and ≥5, which was selected as it allowed for a 90% reduction in skilled grader burden and still permitted a kappa ≥0.7 in the IGA simulation.

Using the more conservative cutoff of 4, the skilled grader was required to overread 16% of the entire data set (n=196 images), which took approximately 30 minutes using the same AMT interface. Kappa agreement with the ground truth score was 0.685, and the prevalence of TF in this sample was 2.52% (95% CI 1.7%-3.6%). Using ≥5 as a cutoff allowed for a bigger reduction in the skilled grader burden as only 10% of the images required overreading (n=115).

A separate set of raw score aggregation methods were attempted excluding the scores of individual AMT workers, who appeared to be less reliable. The pattern of crowdsourced responses (0-3 category usage) was reviewed at the level of each image (Figure 4A-C). Because the prevalence of TF in the data set was 2%, a reliable grader who completes a sufficient number of images should have had a median score of 0 (ie, most images should be graded as without TF). For stability, we attempted to exclude all grades from “variable” graders who had either graded fewer than 10 images (n=476 grades from 152 graders) or whose median score was >1 (n=616 grades from 14 graders; Figure 4D-F). The AUROC for this set of raw scores was 0.930 (95% CI 0.874-0.986). Neither of the cutoffs that permitted a kappa ≥0.7 in the IGA simulation generated a prevalence within ±2% of the true prevalence of the training set (data not shown).

Ultimately, the process chosen from the analysis of the training set to compute the dichotomized “No TF” versus “TF” score calculates a raw crowdsourced score after discarding the highest and lowest individual scores, designates images with truncated raw score ≥4 as preliminary positive, and then uses a skilled overread of all these preliminary positive images. We call this algorithm the “VRC method.”

For the final internal validation, the VRC method was then applied to the test set of 1150 images (TF n=33; prevalence 2.87%, 95% CI 1.98%-4.01%). In this set, 205 images required skilled overread constituting an 82% reduction in the skilled grader burden. After the overread, the kappa was 0.775, with a calculated TF prevalence of 2.70% (95% CI 1.84%-3.80%).

Table 1. Attempts to optimize the aggregation of raw scores using training set. In intergrader assessment (IGA) simulations, the true prevalence was 40%; otherwise, the training set gold standard prevalence was 2%.

Model and cutoff description (≥cutoff)	% correct	Sensitivity (%)	Specificity (%)	Kappa	Prevalence (%)	False positives, n	Skilled grader burden reduction (%)
IGA simulation							
Maximize kappa (6)	86	90	83	0.715	46.0	N/A[a]	N/A
Total raw score							
Maximized kappa (from IGA) (6)	80	87	80	0.115	21.6	3	78
Naïve/majority rules (11)	98	65	98	0.534	2.8	8	97
WHO[b] minimum accepted sensitivity (10)	97	70	97	0.450	4.0	7	96
Maximize kappa (full set) (14)	99	48	100	0.605	1.1	12	99
Mimic true prevalence (12)	98	61	99	0.587	2.1	9	98
Truncated mean approach (simulated IGA sample)							
Maximize kappa (4)	90	95	87	0.797	46.0	N/A	N/A
Truncated mean approach							
Maximized kappa (from IGA) (4)	85	91	84	0.162	17.1	2	84
Naïve/majority Rules (8)	98	61	98	0.497	2.8	9	97
WHO minimum accepted sensitivity (6)	94	74	95	0.326	6.5	6	93
Maximize kappa (full set) (11)	99	43	100	0.565	1.0	13	99
Mimic true prevalence (9)	98	52	99	0.524	1.9	11	98
Create 90% skilled grader burden reduction (5)	91	78	91	0.229	10.3	5	90
Virtual reading center model with overreads							
Truncated mean with maximized kappa from IGA; skilled overead of all positive images (n=196)	99	78	99	0.685 (0.786 in IGA sample)	2.5	5	84
Truncated mean with 90% skilled grader burden; with skilled overread of all positive images (n=115)	99	74	99	0.673 (0.741 in IGA sample)	2.4	6	90

[a]N/A: not applicable.

[b]WHO: World Health Organization.

Figure 3. ROC with the highest and lowest score truncated from the raw score showed a slightly better area under the curve compared with the raw score. ROC: receiver operating characteristic.

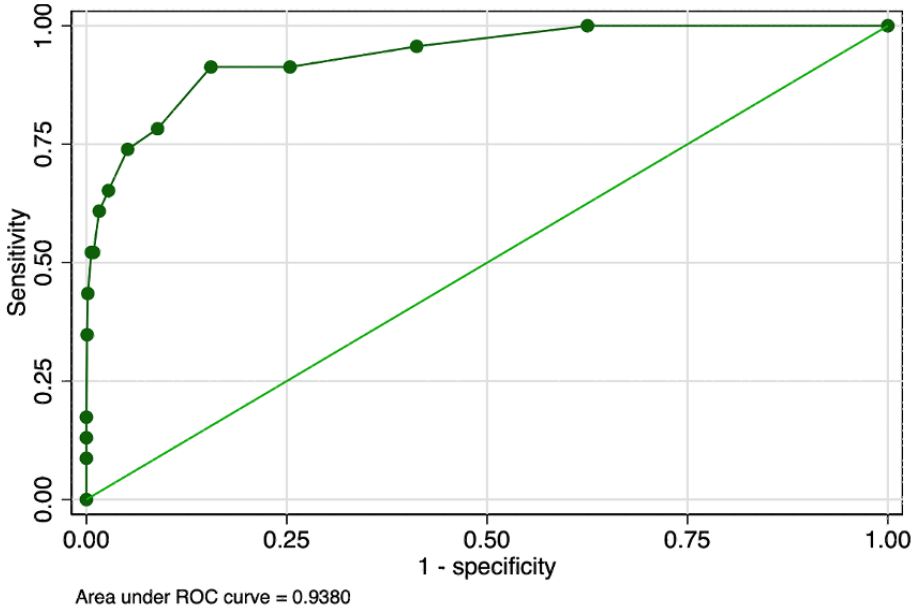
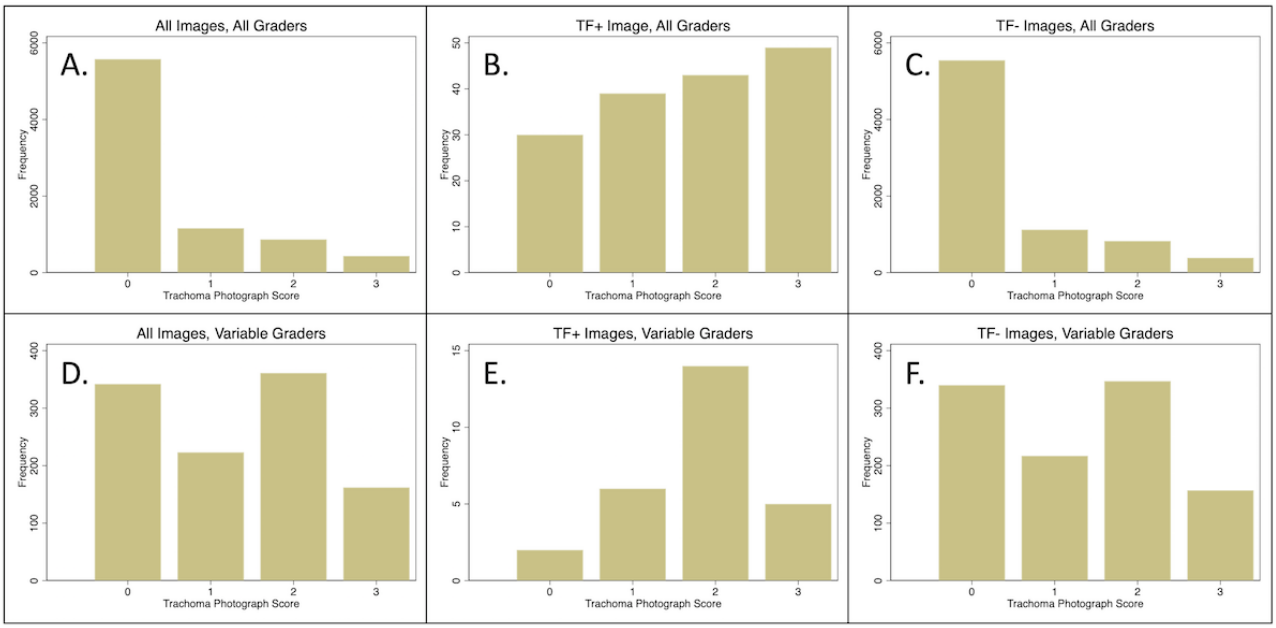


Figure 4. Image-level distribution of individual worker grades reveals the "fingerprint" of the entire data set (A and D) as well as images with (B and E) and without (C and F) active trachoma. In A-C, the entire set of grades is used demonstrating most images are without disease and those with disease had more scores indicating disease. In D-F, in which only less reliable grades are displayed, the difference between images with and without disease is less obvious. TF: trachomatous inflammation—follicular.



Discussion

Principal Findings

We developed and validated a decentralized, VRC using crowdsourcing to grade real-world smartphone-acquired images for the presence of trachoma. During the training of the VRC model, a "standalone" version of the VRC using only crowdsourcing was applied in a simulated skilled grader certification test (IGA) following WHO guidelines and was able to meet the minimum kappa agreement with an expert grader of >0.70. While we used the IGA guidelines originally developed for the Global Trachoma Mapping Project [24], which recommended a moderately high prevalence for the assessment (ie, 30-70%), more recent guidance has encouraged IGA with a minimum prevalence of 10% [26]. We see the greatest potential for applying telemedicine for trachoma grading in even lower prevalence environments and wanted to develop a tool that could produce prevalence estimates of ±2% with a TF prevalence of 5% or lower. Therefore, we tested the VRC using a data set with a low prevalence of 2%-3% for the training and test sets. In this environment, the most accurate VRC used crowdsourcing as a first pass, followed by skilled grading of positive images. This VRC model was able to maintain a kappa agreement of 0.775 with expert graders while generating a

prevalence estimate of 2.70% (compared with the true prevalence of 2.87%).

Trachoma Elimination

The current paradigm for clinical TF identification was developed in the milieu of high trachoma prevalence and prior to the era of global adoption of smartphone technology with wireless internet connectivity. Because the clinical sign of TF can be subtle, ongoing direct exposure to active disease is required for accurate field grading. In this sense, trachoma elimination programs can be conceived as the "victim of their own success" because as the disease is eliminated, it is harder to maintain a cadre of accurate field graders. While obviously a cause for celebration, the continued success of the enterprise is at risk without reevaluating and augmenting, or perhaps reinventing, acceptable district survey methods.

Ophthalmic Photography

Ophthalmic photography has long been used for standardized diagnosis and monitoring for clinical and research purposes in many conditions, including trachoma [14-17,27]. The theoretical advantages of applying photography in TF elimination programs include decentralized grading (and thus avoiding bias), access to expert consultation for difficult cases, auditability of findings, and the ability to re-examine images in light of future evolution in our understanding of the disease. There is sufficient interest in the potential role of ophthalmic photography in trachoma elimination that the WHO convened 2 informal workshops on the topic in 2020 and committed to support the development of the evidence base for innovations in trachoma photography and imaging interpretation. A recent systematic review found 18 articles in the English language, peer-reviewed literature which, in aggregate, demonstrate that agreement between clinical assessment and photo-rendered grade is at least as good as the agreement between intergrader agreement among clinical grades [17]. Likewise, several groups have demonstrated the feasibility of smartphone photography for TF diagnosis [14,21], including the ability to upload and transmit images to a cloud server from rural field locations [21]. These results support a potential role for applying a telemedicine paradigm for district surveys, though the use of imaging in an elimination program at scale remains unproven.

An important component of a successful telemedicine program is the management and interpretation of imaging data. In a conventional clinical telemedicine program, images are generally interpreted in a reading center by a skilled grader with supervision by a licensed clinician [28]. Such a model can be robust and cost-efficient in clinical medicine but may not be scalable in large public health programs. We have therefore validated a VRC paradigm that incorporates crowdsourcing and skilled human grading to produce grades that would meet the standards set by the WHO for certification of a skilled grader.

Use of Crowdsourcing

Crowdsourcing using unskilled internet users has been used to identify pathological features of diabetic retinopathy [19,20,29] and glaucoma [18,30] on fundus photographs as well as to classify surgical outcomes of ophthalmic surgery [31-33]. Aggregating the scores of a "crowd" of skilled graders has also been used in ophthalmology to generate robust consensus annotations of images for artificial intelligence (AI) algorithm training [34,35]. The advantages of crowdsourcing for image interpretation are that it capitalizes on human pattern recognition to identify features of the disease and ideally averages out individual biases (ie, over- or under-calling) to settle on an accurate classification. Using the AMT marketplace, large batches of >1000 images have been graded in parallel by hundreds of users in approximately 1 hour. We have shown that while this model is currently unable to provide a stand-alone classification for TF in a low prevalence environment; in a VRC model with skilled overreading of positive images, crowdsourcing accurately and efficiently identifies images free of TF. In the final validation model, only 205 of 1150 images identified as possible TF required a skilled grade.

Limitations

The VRC model using crowdsourcing has some disadvantages. Because there is limited gatekeeping on the pool of workers, there is the potential for fraudulent users to contaminate the grading. We tried to minimize this by running the data set in 2 batches separated by several months in time and saw very similar usage characteristics in terms of the time spent per image and the amount of time taken to complete the entire batch. Ongoing quality control checks would need to be built into the system to ensure stability as in any diagnostic paradigm. In this study, we also examined individual worker characteristics and found that some users were clear outliers in category usage since they classified many more images with pathology than their peers. While removing these individuals' 1160 grades improved our model, it was less stable and less efficient than simply truncating the highest and lowest grades. Specifically, by dropping 13.5% of the individual grades, only 478 of 1307 images had all 7 grades available, and 5 images had 3 or fewer grades (Figure S5 in Multimedia Appendix 1), in contrast with 1294 images with all grades available in the full training set. Methods to validate individual workers in real-time could be applied in future VRC models.

Perhaps, the biggest challenge to implementing telemedicine for TF is ensuring the data coming into the VRC system are of the highest possible quality. The data set used for this validation had 9% ungradable images [21], and while these were posted to AMT for crowdsourced assessment, gradability was not a required element of the task and there was limited instruction on what constituted an ungradable image. As such, workers were generally unable to identify expert-identified ungradable images as such (data not shown). Future refinements in photographic technique will likely improve the value of VRC grading.

Conclusions and Future Directions

There is growing enthusiasm for automated classifiers built using AI methods to identify ophthalmic disease [36-39], culminating in the 2018 US FDA authorization of the first autonomous AI system in any field of medicine [40]. While we feel strongly that AI will likely contribute to or replace human grading for many visual tasks, we suggest for current use a VRC using crowdsourcing that can be flexible enough to incorporate AI as it matures. Furthermore, we recognize the realities of

global priorities and funding timelines, with funding now set to expire in 2030, which is a short time to validate an autonomous AI system. We believe that a VRC system that meets or exceeds the standards set by the WHO and Tropical Data, and has value to add over current systems, could be deployed without delay to meet the needs of global elimination programs that are currently struggling to train and standardize skilled graders.

Acknowledgments

CJB was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Numbers P20GM103644 and P20GM125498.

Data Availability

All data generated or analyzed during this study are included in this published paper and in Multimedia Appendices 2 and 3.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary figures.
[DOCX File , 3222 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Open data.
[XLSX File (Microsoft Excel File), 10 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data codebook.
[XLSX File (Microsoft Excel File), 512 KB-Multimedia Appendix 3]

References

1. Burton MJ, Faal HB, Ramke J, Ravilla T, Holland P, Wang N, et al. Announcing The Lancet Global Health Commission on global eye health. Lancet Global Health 2019 Dec;7(12):e1612-e1613. [doi: 10.1016/s2214-109x(19)30450-4]
2. Ferede AT, Dadi AF, Tariku A, Adane AA. Prevalence and determinants of active trachoma among preschool-aged children in Dembia District, Northwest Ethiopia. Infect Dis Poverty 2017 Oct 09;6(1):128 [FREE Full text] [doi: 10.1186/s40249-017-0345-8] [Medline: 28988539]
3. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, Vision Loss Expert Group of the Global Burden of Disease Study. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. Lancet Glob Health 2017 Dec;5(12):e1221-e1234 [FREE Full text] [doi: 10.1016/S2214-109X(17)30393-5] [Medline: 29032195]
4. Lietman TM, Oldenburg CE, Keenan JD. Trachoma: time to talk rradication. Ophthalmology 2020 Jan;127(1):11-13. [doi: 10.1016/j.ophtha.2019.11.001] [Medline: 31864470]
5. Williams LB, Prakalapakorn SG, Ansari Z, Goldhardt R. Impact and trends in global ophthalmology. Curr Ophthalmol Rep 2020;8(3):136-143 [FREE Full text] [doi: 10.1007/s40135-020-00245-x] [Medline: 32837802]
6. World Health Organization. WHO Alliance for the Global Elimination of Trachoma by 2020: progress report on elimination of trachoma, 2020. Wkly Epidemiol Rec 2021;96(31):353-364 [FREE Full text]
7. WHO Alliance for the Global Elimination of Trachoma. Report of the third meeting of the WHO Alliance for the Global Elimination of Trachoma. World Health Organization. 1998. URL: https://apps.who.int/iris/handle/10665/65933 [accessed 2022-06-21]
8. West SK. Milestones in the fight to eliminate trachoma. Ophthalmic Physiol Opt 2020 Mar;40(2):66-74. [doi: 10.1111/opo.12666] [Medline: 32017172]
9. Reaching the Last Mile Forum: Keynote Address. World Health Organization. 2019. URL: https://www.who.int/director-general/speeches/detail/reaching-the-last-mile-forum [accessed 2022-06-21]
10. Whitty CJM. Political, social and technical risks in the last stages of disease eradication campaigns. Int Health 2015 Sep;7(5):302-303. [doi: 10.1093/inthealth/ihv049] [Medline: 26311754]
11. Network of WHO collaborating centres for trachoma: second meeting report. World Health Organization. 2017. URL: https://apps.who.int/iris/handle/10665/258687 [accessed 2022-06-21]

12. Borlase A, Blumberg S, Callahan EK, Deiner MS, Nash SD, Porco TC, et al. Modelling trachoma post-2020: opportunities for mitigating the impact of COVID-19 and accelerating progress towards elimination. Trans R Soc Trop Med Hyg 2021 Mar 06;115(3):213-221 [FREE Full text] [doi: 10.1093/trstmh/traa171] [Medline: 33596317]
13. Report of the 23rd meeting of the WHO alliance for the global elimination of trachoma by 2020. World Health Organization. 2021. URL: https://apps.who.int/iris/handle/10665/341049 [accessed 2022-06-21]
14. Nesemann JM, Seider MI, Snyder BM, Maamari RN, Fletcher DA, Haile BA, et al. Comparison of smartphone photography, single-lens reflex photography, and field-grading for trachoma. Am J Trop Med Hyg 2020 Dec;103(6):2488-2491 [FREE Full text] [doi: 10.4269/ajtmh.20-0386] [Medline: 33021196]
15. Snyder BM, Sié A, Tapsoba C, Dah C, Ouermi L, Zakane SA, et al. Smartphone photography as a possible method of post-validation trachoma surveillance in resource-limited settings. Int Health 2019 Nov 13;11(6):613-615 [FREE Full text] [doi: 10.1093/inthealth/ihz035] [Medline: 31329890]
16. West SK, Taylor HR. Reliability of photographs for grading trachoma in field studies. Br J Ophthalmol 1990 Jan;74(1):12-13 [FREE Full text] [doi: 10.1136/bjo.74.1.12] [Medline: 2306438]
17. Naufal F, West SK, Brady CJ. Utility of photography for trachoma surveys: a systematic review. Surv Ophthalmol 2022;67(3):842-857. [doi: 10.1016/j.survophthal.2021.08.005] [Medline: 34425127]
18. Wang X, Mudie LI, Baskaran M, Cheng CY, Alward WL, Friedman DS, et al. Crowdsourcing to evaluate fundus photographs for the presence of glaucoma. J Glaucoma 2017 Jun;26(6):505-510. [doi: 10.1097/IJG.0000000000000660] [Medline: 28319525]
19. Brady CJ, Mudie LI, Wang X, Guallar E, Friedman DS. Improving consensus scoring of crowdsourced data using the Rasch model: development and refinement of a diagnostic instrument. J Med Internet Res 2017 Jun 20;19(6):e222 [FREE Full text] [doi: 10.2196/jmir.7984] [Medline: 28634154]
20. Brady CJ, Villanti AC, Pearson JL, Kirchner TR, Gupta OP, Shah CP. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. J Med Internet Res 2014;16(10):e233 [FREE Full text] [doi: 10.2196/jmir.3807] [Medline: 25356929]
21. Naufal F, Brady CJ, Wolle MA, Saheb Kashaf M, Mkocha H, Bradley C, et al. Evaluation of photography using head-mounted display technology (ICAPS) for district trachoma surveys. PLoS Negl Trop Dis 2021 Nov;15(11):e0009928 [FREE Full text] [doi: 10.1371/journal.pntd.0009928] [Medline: 34748543]
22. Brady C, Bradley C, Wolle M, Mkocha H, Massof R, West S. Virtual reading center can remotely identify follicular trachoma. Investig Ophthalmol Visual Sci 2020;61(7):492.
23. Solomon AW, Kello AB, Bangert M, West SK, Taylor HR, Tekeraoi R, et al. The simplified trachoma grading system, amended. Bull World Health Organ 2020 Oct 01;98(10):698-705 [FREE Full text] [doi: 10.2471/BLT.19.248708] [Medline: 33177759]
24. Solomon AW, Pavluck AL, Courtright P, Aboe A, Adamu L, Alemayehu W, et al. The global trachoma mapping project: methodology of a 34-country population-based study. Ophthalmic Epidemiol 2015;22(3):214-225 [FREE Full text] [doi: 10.3109/09286586.2015.1037401] [Medline: 26158580]
25. Report of the third global scientific meeting on trachoma. World Health Organization. URL: https://apps.who.int/iris/handle/10665/329074 [accessed 2022-06-21]
26. Courtright P, MacArthur C, Macleod C, Dejene M, Gass K, Harding-Esch E. Tropical Data: Training System for Trachoma Prevalence Surveys. London: International Coalition for Trachoma Control; 2019.
27. Solomon AW, Bowman RJC, Yorston D, Massae PA, Safari S, Savage B, et al. Operational evaluation of the use of photographs for grading active trachoma. Am J Trop Med Hyg 2006 Mar;74(3):505-508 [FREE Full text] [Medline: 16525114]
28. Horton MB, Brady CJ, Cavallerano J, Abramoff M, Barker G, Chiang MF, et al. Practice guidelines for ocular telehealth-diabetic retinopathy, third edition. Telemed J E Health 2020 Apr;26(4):495-543 [FREE Full text] [doi: 10.1089/tmj.2020.0006] [Medline: 32209018]
29. Mitry D, Peto T, Hayat S, Morgan JE, Khaw K, Foster PJ. Crowdsourcing as a novel technique for retinal fundus photography classification: analysis of images in the EPIC Norfolk cohort on behalf of the UK Biobank Eye and Vision Consortium. PLoS One 2013;8(8):e71154 [FREE Full text] [doi: 10.1371/journal.pone.0071154] [Medline: 23990935]
30. Mitry D, Peto T, Hayat S, Blows P, Morgan J, Khaw K, et al. Crowdsourcing as a screening tool to detect clinical features of glaucomatous optic neuropathy from digital photography. PLoS One 2015;10(2):e0117401 [FREE Full text] [doi: 10.1371/journal.pone.0117401] [Medline: 25692287]
31. Rootman DB, Bokman CL, Katsev B, Rafaelof M, Ip M, Manoukian N, et al. Crowdsourcing morphology assessments in oculoplastic surgery: reliability and validity of lay people relative to professional image analysts and experts. Ophthalmic Plast Reconstr Surg 2020;36(2):178-181. [doi: 10.1097/IOP.0000000000001515] [Medline: 31789786]
32. Paley GL, Grove R, Sekhar TC, Pruett J, Stock MV, Pira TN, et al. Crowdsourced assessment of surgical skill proficiency in cataract surgery. J Surg Educ 2021;78(4):1077-1088 [FREE Full text] [doi: 10.1016/j.jsurg.2021.02.004] [Medline: 33640326]

33. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 2019 Jun;14(6):1097-1105. [doi: 10.1007/s11548-019-01956-8] [Medline: 30977091]
34. Schaekermann M, Hammel N, Terry M, Ali TK, Liu Y, Basham B, et al. Remote tool-based adjudication for grading diabetic retinopathy. Transl Vis Sci Technol 2019 Nov;8(6):40 [FREE Full text] [doi: 10.1167/tvst.8.6.40] [Medline: 31867141]
35. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
36. Kim MC, Okada K, Ryner AM, Amza A, Tadesse Z, Cotter SY, et al. Sensitivity and specificity of computer vision classification of eyelid photographs for programmatic trachoma assessment. PLoS One 2019;14(2):e0210463 [FREE Full text] [doi: 10.1371/journal.pone.0210463] [Medline: 30742639]
37. Campbell JP, Mathenge C, Cherwek H, Balaskas K, Pasquale LR, Keane PA, American Academy of Ophthalmology Task Force on Artificial Intelligence. Artificial intelligence to reduce ocular health disparities: moving from concept to implementation. Transl Vis Sci Technol 2021 Mar 01;10(3):19 [FREE Full text] [doi: 10.1167/tvst.10.3.19] [Medline: 34003953]
38. Abràmoff MD, Leng T, Ting DSW, Rhee K, Horton MB, Brady CJ, et al. Automated and computer-assisted detection, classification, and diagnosis of diabetic retinopathy. Telemed J E Health 2020 Apr;26(4):544-550 [FREE Full text] [doi: 10.1089/tmj.2020.0008] [Medline: 32209008]
39. Socia D, Brady CJ, West SK, Cockrell RC. Detection of trachoma using machine learning approaches. PLoS Negl Trop Dis 2022 Dec;16(12):e0010943 [FREE Full text] [doi: 10.1371/journal.pntd.0010943] [Medline: 36477293]
40. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 2018;1:39 [FREE Full text] [doi: 10.1038/s41746-018-0040-6] [Medline: 31304320]

Abbreviations

AI: artificial intelligence
AMT: Amazon Mechanical Turk
AUROC: area under the receiver operating characteristic
CPT: current procedural terminology
FDA: Food and Drug Administration
ICAPS: Image Capture and Processing System
IGA: intergrader agreement test
ROC: receiver operating characteristic
TF: trachomatous inflammation—follicular
VRC: virtual reading center
WHO: World Health Organization

Edited by A Mavragani; submitted 26.07.22; peer-reviewed by J Keenan, G Lim; comments to author 11.01.23; revised version received 30.01.23; accepted 19.02.23; published 06.04.23

Please cite as:
Brady CJ, Cockrell RC, Aldrich LR, Wolle MA, West SK
A Virtual Reading Center Model Using Crowdsourcing to Grade Photographs for Trachoma: Validation Study
J Med Internet Res 2023;25:e41233
URL: https://www.jmir.org/2023/1/e41233
doi: 10.2196/41233
PMID: