Published in final edited form as:

J Glaucoma. 2017 June ; 26(6): 505–510. doi:10.1097/IJG.0000000000000660.

Crowdsourcing to evaluate fundus photographs for the presence of glaucoma

Xueyang Wang, BS¹, Lucy I. Mudie, MBBS, MPH¹, Mani Baskaran, DO, DNB, PhD^{2,3}, Ching-Yu Cheng, MD, PhD^{2,3,5}, Wallace L. Alward, MD⁴, David S. Friedman, MD, MPH, PhD¹, and Christopher J. Brady, MD¹

¹Wilmer Eye Institute, Johns Hopkins School of Medicine, Baltimore, MD

²Singapore Eye Research Institute, Singapore National Eye Center, Singapore, Singapore

³Duke-NUS Medical School, Singapore, Singapore

⁴Department of Ophthalmology and Visual Sciences, University of Iowa Carver College of Medicine, Iowa City, IA

⁵Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Abstract

Purpose—To assess the accuracy of crowdsourcing for grading optic nerve images for glaucoma using Amazon Mechanical Turk (AMT) before and after training modules.

Methods—Images (n=60) from two large population studies were graded for glaucoma status and vertical cup-to-disc ratio (VCDR). In the baseline trial, users on AMT (Turkers) graded fundus photos for glaucoma and VCDR after reviewing annotated example images. In two additional trials, Turkers viewed a 26-slide PowerPoint training or a 10-minute video training and passed a quiz before being permitted to grade the same 60 images. Each image was graded by 10 unique Turkers in all trials. The mode of Turker grades for each image was compared to an adjudicated expert grade to determine accuracy as well as the sensitivity and specificity of Turker grading.

Results—In the baseline study, 50% of the images were graded correctly for glaucoma status and the area under the receiver operating characteristic (AUROC) was 0.75 (95% CI: 0.64–0.87). Post-PowerPoint training, 66.7% of the images were graded correctly with AUROC of 0.86 (95% CI: 0.78–0.95). Finally, Turker grading accuracy was 63.3% with AUROC 0.89 (95% CI: 0.83–0.96) after video training. Overall, Turker VCDR grades for each image correlated with expert VCDR grades (Bland-Altman plot mean difference = –0.02).

Conclusions—Turkers graded 60 fundus images quickly and at low cost, with grading accuracy, sensitivity and specificity all improving with brief training. With effective education, crowdsourcing may be an efficient tool to aid in the identification of glaucomatous changes in retinal images.

Keywords

teleglaucoma; crowdsourcing; image analysis

Introduction

Despite the large disease burden of glaucoma, widespread screening efforts in the community have been neither cost-effective nor sufficiently accurate.^{1–3} There is currently no single accepted method to detect glaucoma, and modalities such as tonometry, visual fields, and optic nerve head imaging each suffer from relatively low sensitivity and specificity.^{4–7} A particular challenge to efficient screening is the difficulty of interpreting images of optic nerves, and assessment of optic nerve images by highly trained personnel is expensive and logistically complex. While providing a training module has been shown to improve optic disc evaluation for glaucoma, interobserver agreement among ophthalmologists in the evaluation of glaucoma via disc photos is only fair to moderate ($\kappa = 0.20–0.68$).^{8–11}

Telemedicine is an expanding domain in ophthalmology that could mitigate the resource-intensive aspects of image analysis; non-mydratic fundus photography and remote interpretation have been used successfully in rural and remote settings for diabetic retinopathy (DR) screening.^{12,13} Moreover, computerized image analysis techniques have been developed for both DR and glaucoma, but none match or surpass the gold standard of expert interpretation.⁵

Crowdsourcing could be used to simplify grading and interpretation of optic nerve head images, as has been done to identify malaria parasites on images of thick blood smears and map out individual retinal ganglion cells via web-based games.^{14,15} One popular crowdsourcing platform is Amazon Mechanical Turk (AMT). AMT has been used to accurately grade fundus photographs for DR at very low cost.¹⁶ AMT has also been used to grade fundus photographs for glaucoma with minimal training provided to the Turkers, yielding a relatively high sensitivity but at the cost of an extremely low specificity.¹⁷

We present research employing new approaches to crowdsourcing via AMT to grade optic nerve images for the presence of glaucoma.

Materials and Methods

Fundus Photograph Datasets

A dataset of 60 non-mydratic fundus photographs centered between the optic nerve and macula was compiled from two large population-based studies: Singapore Epidemiology of Eye Diseases (SEED) and the Screening to Prevent Glaucoma Study (STOP). SEED is a large-scale population-based study in the Singapore Malay community conducted from 2004 to 2007 by researchers at the Singapore Eye Research Institute. Retinal fundus images were taken of both eyes of each subject in the study and were graded for image quality, vertical cup-to-disc ratio (VCDR), and glaucoma status in 4 categories: unlikely, possible, probable, and definite glaucoma. Glaucoma status was defined based on the International Society of

Geographical and Epidemiological Ophthalmology (ISGEO) classification.^{18–22} STOP is an ongoing population-based study in Baltimore in which retinal fundus photographs were graded for glaucoma status in 4 categories: unlikely, possible, probable, and definite. Each study supplied the images along with one or more ground-truth grading. The sample of images used for the present study was selected to include the highest quality and least ambiguous images (i.e., the images with the highest agreement among the multiple ground-truth grades), as well as to achieve balanced representation among the glaucoma categories and VCDR values.

All 60 images included in the present study were re-graded independently for image quality, glaucoma status in 3 categories (unlikely, possible, and probable or definite), VCDR, and presence of abnormal optic nerve features (hemorrhage, notch, and retinal nerve fiber layer (RNFL) defect) by two ophthalmologists (DSF and CJB) with consensus discussion used to resolve disagreements. The experts agreed on 88.3% of images, with 7 images requiring discussion to reach consensus. Agreement on 3-category glaucoma grading between experts was good ($\kappa W = 0.72$, $p < 0.001$, 95% CI: 0.64, 0.76). The images had a relatively uniform distribution of glaucoma status, with 24 unlikely, 19 possible, and 17 probable/definite glaucoma images (Figure 1). Most of the images were re-graded as fair quality (41.7%) or better (53.3%), with 3 (5.0%) images being graded by experts as poor quality. The distribution of expert VCDR grades had a relative uniform distribution from 0.1 to 0.9, with very few images with extremely low or high VCDR (Figure 2).

Crowdsourcing Interface

A customized interface was developed for Amazon Mechanical Turk (AMT <https://www.mturk.com>). AMT allows task administrators (Requesters) to upload small, discrete tasks called human intelligence tasks (HITs) for online workers (Turkers) to complete. Generally, each HIT takes seconds to minutes to complete and pays the Turker \$0.01–0.25 or more depending on its complexity, and HITs are often uploaded with multiple similar tasks as a “batch” for multiple repetitions by unique Turkers. In our interface, each image was graded by 10 unique Turkers with each receiving \$0.10 per image, with a maximum of 20 minutes of grading time allowed per image. There was no restriction or requirement on the number of images each Turker could grade. In general, Turkers with high approval ratings have access to the most highly paid HITs, and Requesters with a record of fair HIT approval are favored by Turkers. Throughout all trials in the present study, only Turkers who had a lifetime HIT approval rating of 99% and had previously completed at least 500 HITs could view and grade the images.

Baseline Trial with Minimal Training

The batch of 60 fundus photographs was uploaded to AMT and Turkers were asked to grade them for glaucoma status (unlikely glaucoma, possible glaucoma, and probable/definite glaucoma), image quality (poor, acceptable, excellent), nerve size (small, medium, large), VCDR (0.1–0.9), and features (hemorrhage, notch, myopic tilted disc, nerve fiber layer defect). Eight sample annotated images (2 unlikely, 2 possible, 2 probable, and 2 definite glaucoma) were available in the instructions as reference. No other teaching or explanations were provided.

Trial after Providing Training PowerPoint

We then developed a 26-slide PowerPoint training module (<http://tinyurl.com/jn3n84n>) that provided a basic introduction to glaucoma along with fundus photographs showing 3 categories of glaucoma severity (unlikely, possible, and probable/definite glaucoma). Features of the fundus, including small and large nerves, VCDR, tilted disc, notch, disc hemorrhages, and retinal nerve fiber layer defects, were explained with annotated images in animated Graphics Interchange Format images (GIFs).

In this second trial, Turkers were asked to view the PowerPoint before taking a quiz where they had to grade 6 images (2 unlikely, 1 possible, 1 probable, and 2 definite glaucoma) for glaucoma status. Turkers who were able to grade 4 or more images correctly received qualifications to complete the same task of grading the batch of 60 images as described above.

Trial after Providing Training Video

In a third trial, Turkers were asked to view the 9-minute video “What is Glaucoma & Why Does it Matter” by one of the investigators (WLMA) and grade 4 or more images correctly in the same quiz described above before they could receive qualifications to complete the same task of grading the batch of 60 images as described in the baseline trial.²³ The video was developed for resident education and covered similar information as the PowerPoint, including explanations and images of findings in fundus photographs of glaucomatous eyes. The disease burden and treatment of glaucoma were also briefly reviewed in the video. Turkers who already received qualifications after viewing the training PowerPoint in trial 2 could not participate in this trial.

Statistical Analysis

Throughout the three trials, the mode of the 10 Turker grades was calculated to serve as the Turker “consensus” score. Each image was compared to the expert grade to determine agreement as well as the sensitivity and specificity of Turker grading comparing unlikely glaucoma versus any abnormality. The area under the receiver operator characteristic (AUROC) was also calculated and compared across the three trials.²⁴ A Bland-Altman plot was also generated for the association between mean Turker VCDR grading and expert VCDR grading. Analysis of Variance (ANOVA) with Bonferroni correction was used to compare the differences in mean Turker VCDR grading between each group of images as graded by expert (No glaucoma, Possible glaucoma, Probable/definite glaucoma). A weighted kappa (k_w) statistic (with Cicchetti–Allison) weighting was used to compare agreement between Turker and expert glaucoma status grading. Stata Statistical Software (Release 14. College Station, TX: StataCorp LP) was used for all analyses.

This research was deemed IRB-exempt by the Johns Hopkins University School of Medicine Institutional Review Board.

Results

A total of 600 gradings were obtained for the batch of 60 fundus photographs in the baseline trial. In this trial without any Turker training, 51 unique Turkers took a total of 40 minutes to perform the task, with each Turker grading an average of 11.7 photographs (Table 1). In the post-PowerPoint training trial, 19 Turkers took 28 hours with each Turker grading on average 31.6 photographs. In the post-video training trial, 17 Turkers took 18 days with each Turker grading on average 35.2 photographs. The total cost of performing these classifications was \$84.00 for the baseline trial and \$238 for the post-PowerPoint trial and \$213 post-video training trials including the cost for the qualification quiz and the batch of 60 images.

In the baseline trial, 50% of the images were graded correctly, which improved to 66.7% and 63.3% with PowerPoint training and video training, respectively (Table 2). The agreement between Turker and expert 3-category grading for glaucoma status in the baseline trial was poor, with $\kappa W = 0.33$ (95% CI: 0.32 – 0.39), but improved in the post-Powerpoint ($\kappa W = 0.59$, 95% CI: 0.46 – 0.64) and post-video ($\kappa W = 0.58$, 95% CI: 0.45 – 0.62) training trials. The sensitivity of Turker grading for any abnormality (possible or probable/definite glaucoma) versus unlikely glaucoma also improved after both forms of training, from a baseline of 80.56% to 91.67% post-PowerPoint and 97.22% post-video training. Specificity improved after PowerPoint training from 66.67% to 70.83%, but decreased to 58.33% after video training. The 3-category area under the receiver operator characteristic (AUROC) increasing from 0.76 (95% CI: 0.64–0.87) at baseline to 0.87 (95% CI: 0.78–0.95) post-PowerPoint and 0.89 (95% CI: 0.83–0.96) post-video training. The improvement in AUROC was significantly better ($p=0.01$) post-video compared to baseline, but there was no significant difference in performance between post-PowerPoint and post-video training (Figure 3).

The proportion of images graded correctly by an individual Turker in the 60-image task was not correlated with their six-image quiz score (Spearman's rank correlation coefficient = 0.20, $p=0.4$, degrees of freedom = 17). Likewise, Turker experience with ophthalmic tasks, as measured by his or her lifetime number of completed HITs posted by the research group in previous studies,¹⁶ did not correlate with his or her proportion of images graded correctly in this task (Spearman's rank correlation coefficient = 0.11, $p=0.7$, degrees of freedom = 17). Furthermore, there was weak negative correlation between a Turker's experience and his or her quiz score (Spearman's rank coefficient = -0.50, $p=0.03$, degrees of freedom = 17).

Turker VCDR grades for each image were averaged to determine the consensus Turker VCDR grade for each image. Consensus Turker VCDR grades agreed with expert VCDR grades (mean difference = -0.02, limits of agreement = -0.26, 0.21) in a Bland-Altman plot (Figure 4). Furthermore, Turkers assigned lower VCDR values to images they graded as unlikely glaucoma than those that they graded as probable/definite glaucoma ($p<0.01$). The mean Turker VCDR grading on images experts called possible glaucoma was 0.66, while the mean Turker VCDR grading on images experts called probable/definite glaucoma was 0.88 (Table 3). Images that were graded consistently incorrectly at baseline and post-training

(Figure 5A–D) as well as images that were graded incorrectly at baseline but correctly post-training (Figure 5E–H) were distributed among all disease categories

Discussion

While the use of crowdsourcing for biomedical data analysis and interpretation is in its infancy, we found workers on AMT were able to identify glaucomatous features on fundus photographs with reasonable accuracy. The accuracy of grading for glaucoma status by Turkers was 50% without any additional training beyond providing a set of sample images. This improved to a grading accuracy of 67% after providing PowerPoint or video training. Turkers also seemed to be using VCDR in part to guide their grading for glaucoma status, with increasing VCDR values assigned to images with greater glaucoma severity. This is an encouraging trend that signals the training could have provided guidance in Turker grading.

While a grading accuracy of 67% is not sufficient for crowdsourcing grading of fundus photographs clinically, the improvement in Turker performance after training demonstrates that the crowd can become more accurate with education. Furthermore, since a 26-slide PowerPoint is just as effective as a 9-minute video in improving performance, it may be more efficient to provide future Turker trainings in PowerPoint format, which would allow more self-directed learning, and took substantially less time in our trials.

Although additional training improved Turker performance, our study is limited by the lack of information about the Turkers who participate in our tasks and the inherent variability in the crowd. Prior studies have shown that Turkers are younger and more highly educated than the internet using general-public.²⁵ One survey found that half of Turker respondents were located in the US, a third were in India, and the remainder from 64 other countries. The demographics of individual users varied by country; the majority of US Turkers were women using AMT as a source of supplemental income, whereas most Indian Turkers were men earning their primary income.²⁵ Another survey found that AMT in the United States tended to attract young Hispanic females and young Asian males and females as workers, and 90% of workers were from urban areas.²⁶ Since Turkers work anonymously and can freely choose which HITs they accept, the Requesters on AMT cannot control their exact cohort of Turkers at a certain period of time. Certain criteria can be set to restrict HITs to those who have more experience or higher approval ratings, and custom qualifications may be assigned to Turkers, but it would be difficult (and possibly contrary to the spirit of crowdsourcing) to control who exactly complete the HITs. Furthermore, while the demographics of Turkers may shift depending on the time of day of the release of HITs, ideally, the wisdom of the crowd can be harnessed any time with adequate training, but this has not been directly assessed.

Furthermore, our study is limited by the subjectivity of optic nerve head assessment using disc photos and the use of monoscopic disc photos. Interobserver agreement even among glaucoma specialist of evaluation of CDR in disc photos can vary widely, mainly due to the user-dependent interpretation of images. The use of stereoscopic rather than monoscopic photos has been shown to yield improved intra- and interobserver agreement in the evaluation of CDR,²⁷ but the need for a special viewer currently precludes its use in

crowdsourcing. Providing a standard definition of glaucoma such as the International Society for Geographical and Epidemiological Ophthalmology (ISGEO) system to the Turkers could reduce the subjectivity in glaucoma grading, but the introduction of such a grading scheme would require extra time in training and could discourage Turkers from investing their time into the task.

Our crowdsourcing interface could have introduced some error. It is possible that some of the Turkers who completed the baseline trial also participated in one of the post-training trials and took screenshots of their previous gradings for reference, but this is unlikely due to the time limitations and the economy of AMT incentivizing expedient work by the Turkers. Finally, because our focus was on identification of likely normal or abnormal discs, our grading accuracy may have been artificially inflated beyond what would have been seen without exclusion of ambiguous and low-quality images.

In the future, we hope to improve Turker grading accuracy via more rigorous training and convert crowdsourcing image analysis for glaucoma into a web-based game. Many of the images that were graded consistently incorrectly across all trials were of lower quality, which could have required greater effort from the Turkers and discouraged them from classifying them with higher accuracy. Since many of the images taken in the field when screening for glaucoma are of sub-optimal quality, we could provide more training to help Turkers identify pathology in poor quality image, or incentivize them further by giving bonuses for grading lower quality photographs accurately. Alternatively, we could ignore Turker grades of low quality images and require review by an expert for this subset of images. The training can also be modified to place greater emphasis on the more complex features of glaucoma, such as notching and RNFL defects. We also hope to follow the footsteps of MalariaSpot and EyeWire by building a web-based game platform so that anyone on the Internet can receive training for grading retinal images for glaucoma and then compete against each other in a game where high-scores are posted and achievements can be unlocked. We hope such “gameification” will not only increase engagement with our tasks, but also to promote awareness and increase the public’s understanding of glaucoma.

There are many existing large population based databases in which retinal photographs were taken (e.g., the National Health and Nutrition Examination Survey (NHANES), Atherosclerosis Risk in Communities Study (ARIC), and Cardiovascular Health Study) but were not analyzed for presence of glaucoma. If crowdsourcing can be further optimized, it could be a powerful tool to rapidly and accurately interrogate these large databases for the presence of glaucoma in participants. Furthermore, the crowd can be harnessed to grade retinal photographs in the field during large population screening studies, flagging concerning images to be forwarded to glaucoma experts in almost real-time at low cost. Alternatively, “the crowd” itself could be theoretically restricted to ophthalmologists or glaucoma specialists. Such an online platform might yield even higher sensitivity and specificity consensus grades, although the costs would likely need to be higher than employing general Turkers.

Our study demonstrated the use of AMT to crowdsource grading of 60 retinal photographs for the presence of glaucoma. While grading accuracy, sensitivity, and specificity did not

reach expert level, all three measures improved after providing training to the crowd. With enhancement, this method could provide clinical and research value while reducing costs.

Acknowledgments

This publication was made possible by the Johns Hopkins Institute for Clinical and Translational Research (ICTR) which is funded in part by Grant Number KL2TR001077 from the National Center for Advancing Translational Sciences (NCATS) a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the Johns Hopkins ICTR, NCATS or NIH.

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*. 2006; 90(3):262–267. [PubMed: 16488940]
2. Gupta P, Zhao D, Guallar E, et al. Prevalence of Glaucoma in the United States: The 2005–2008 National Health and Nutrition Examination Survey. *Glaucoma Prevalence in the United States. Investigative Ophthalmology & Visual Science*. 2016; 57(6):2905–2913.
3. Vajaranant TS, Wu S, Torres M, et al. The Changing Face of Primary Open-Angle Glaucoma in the United States: Demographic and Geographic Changes From 2011 to 2050. *American Journal of Ophthalmology*. 2012; 154(2):303–314.e303. [PubMed: 22541661]
4. Bettin P, Di Matteo F. Glaucoma: Present Challenges and Future Trends. *Ophthalmic Research*. 2013; 50(4):197–208. [PubMed: 24030362]
5. Healey PR, Lee AJ, Aung T, et al. Diagnostic Accuracy of the Heidelberg Retina Tomograph for Glaucoma: A Population-Based Assessment. *Ophthalmology*. 2010; 117(9):1667–1673. [PubMed: 20816247]
6. Maul EA, Jampel HD. Glaucoma Screening in the Real World. *Ophthalmology*. 2010; 117(9):1665–1666. [PubMed: 20816246]
7. Tielsch JM, Katz J, Singh K, et al. A Population-based Evaluation of Glaucoma Screening: The Baltimore Eye Survey. *American Journal of Epidemiology*. 1991; 134(10):1102–1110. [PubMed: 1746520]
8. Law SK, Tamboli Diana A, Ou Yvonne, et al. Development of a Resident Training Module for Systematic Optic Disc Evaluation. *Journal of Glaucoma*. 2012; 21(9):601–607. [PubMed: 21602706]
9. Jampel HD, Friedman D, Quigley H, et al. Agreement Among Glaucoma Specialists in Assessing Progressive Disc Changes From Photographs in Open-Angle Glaucoma Patients. *American Journal of Ophthalmology*. 2009; 147(1):39–44.e31. [PubMed: 18790472]
10. Azuara-Blanco A, Katz LJ, Spaeth GL, et al. Clinical agreement among glaucoma experts in the detection of glaucomatous changes of the optic disk using simultaneous stereoscopic photographs. *American Journal of Ophthalmology*. 2003; 136(5):949–950. [PubMed: 14597063]
11. Abrams LS, Scott IU, Spaeth GL, et al. Agreement among Optometrists, Ophthalmologists, and Residents in Evaluating the Optic Disc for Glaucoma. *Ophthalmology*. 1994; 101(10):1662–1667. [PubMed: 7936564]
12. Ng M, Nathoo N, Rudnisky CJ, et al. Improving Access to Eye Care: Teleophthalmology in Alberta, Canada. *Journal of Diabetes Science and Technology*. 2009; 3(2):289–296. [PubMed: 20144360]
13. Scanlon PH. The English national screening programme for sight-threatening diabetic retinopathy. *Journal of Medical Screening*. 2008; 15(1):1–4. [PubMed: 18416946]
14. Kim JS, Greene MJ, Zlateski A, et al. Space-time wiring specificity supports direction selectivity in the retina. *Nature*. 2014; 509(7500):331–336. [PubMed: 24805243]
15. Luengo-Oroz MA, Arranz A, Frean J. Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears. *J Med Internet Res*. 2012; 14(6):e167. [PubMed: 23196001]

16. Brady CJ, Villanti AC, Pearson JL, et al. Rapid Grading of Fundus Photographs for Diabetic Retinopathy Using Crowdsourcing. *J Med Internet Res*. 2014; 16(10):e233. [PubMed: 25356929]
17. Mitry D, Peto T, Hayat S, et al. Crowdsourcing as a Screening Tool to Detect Clinical Features of Glaucomatous Optic Neuropathy from Digital Photography. *PLoS ONE*. 2015; 10(2):e0117401. [PubMed: 25692287]
18. Shen SY, Wong TY, Foster PJ, et al. The Prevalence and Types of Glaucoma in Malay People: The Singapore Malay Eye Study. *Investigative Ophthalmology & Visual Science*. 2008; 49(9):3846–3851. [PubMed: 18441307]
19. Narayanaswamy A, Baskaran M, Zheng Y, et al. The Prevalence and Types of Glaucoma in an Urban Indian Population: The Singapore Indian Eye Study. *Investigative Ophthalmology & Visual Science*. 2013; 54(7):4621–4627. [PubMed: 23745009]
20. Chua J, Mani B, Liao J, et al. Determinants of Undetected Glaucoma in an Asian Community: The Singapore Epidemiology of Eye Disease (SEED) Study. *Investigative Ophthalmology & Visual Science*. 2014; 55(13):4274–4274.
21. Chua J, Baskaran M, Ong P, et al. Prevalence, risk factors, and visual features of undiagnosed glaucoma: The singapore epidemiology of eye diseases study. *JAMA Ophthalmology*. 2015; 133(8):938–946. [PubMed: 26043441]
22. Baskaran M, Foo RC, Cheng C, et al. The prevalence and types of glaucoma in an urban chinese population: The singapore chinese eye study. *JAMA Ophthalmology*. 2015; 133(8):874–880. [PubMed: 25974263]
23. Alward WLM. Welcome to the Iowa Glaucoma Curriculum. 2016. <http://curriculum.iowaglaucoma.org/>. Accessed February 15, 2016
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988; 44(3):837–845. [PubMed: 3203132]
25. Ipeirotis PG. Demographics of Mechanical Turk. *CeDER Working Papers*. 2010; 10(1)
26. Huff C, Tingley D. “Who are These People?”: Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents. *Research and Politics*. 2015 Jul-Sep;(1):1–12.
27. Lichter PR. Variability of expert observers in evaluating the optic disc. *Transactions of the American Ophthalmological Society*. 1976; 74:532–572. [PubMed: 867638]

Figure 1.
Histogram of expert glaucoma grades for fundus photograph dataset.

Figure 2.
Histogram of expert VCDR grades for fundus photograph dataset.

Figure 3.
Comparison of area under receiver operator characteristic for Baseline vs. Post-Training PowerPoint Trial vs Post-Training Video Trial.

Figure 4.
Bland-Altman plot of difference in Turker vs. Expert cup-to-disc ratio grading.

Figure 5.

A–D. Sample of images that were graded consistently incorrectly in all three trials. Expert grades: A-unlikely, B-unlikely, C-probable, D-definite glaucoma. **E–H.** Sample of images that were graded incorrectly at baseline without training but correctly post-PowerPoint or video training. Expert grades: E-unlikely, F-possible, G-probable, H-definite glaucoma.

Table 1

Cost and time of 600 total gradings of 60 fundus images

	Baseline Without Training	Post-PowerPoint Training Trial	Post-Video Training Trial
Number of unique Turkers	51	19	17
Time to completion of grading	40 minutes	28 hours	18 days
Total cost	\$84	\$238	\$213
Mean number of images graded per Turker	11.7	31.6	35.2

Table 2

Turker grading accuracy, agreement with expert, and AUC (3 category grading)

	Baseline Without Training	Post-PowerPoint Training Trial	Post-Video Training Trial
Percent of images graded correctly	50.0%	66.7%	63.3%
Weighted Kappa	0.33	0.59	0.58
AUROC (95% CI)	0.76 (0.64–0.87)	0.87 (0.78–0.95)	0.89 (0.83–0.96)
Sensitivity	80.56%	91.67%	97.22%
Specificity	66.67%	70.83%	58.33%

Table 3

Mean Turker CDR grading for images grouped by expert grading of glaucoma

Expert Grading for Glaucoma	Mean (SD) Turker CDR grading*
No glaucoma	0.40 (0.1)
Possible glaucoma	0.66 (0.1)
Probable/definite glaucoma	0.88 (0.1)

* P<0.001 calculated by ANOVA with Bonferroni correction