

IMPROVING CONSENSUS SCORING OF
CROWDSOURCED DATA USING THE RASCH MODEL:
DEVELOPMENT AND REFINEMENT OF A DIAGNOSTIC INSTRUMENT

by
Christopher J. Brady, MD

A thesis submitted to Johns Hopkins University in conformity with the requirements for
the degree of Master of Health Science in Clinical Epidemiology

Baltimore, Maryland

December, 2016

Abstract

Context:

Diabetic retinopathy (DR) is a leading cause of vision loss in working age individuals worldwide. While screening is effective and cost-effective, it remains underutilized, and novel methods are needed to increase detection of DR. This clinical validation study compared diagnostic impressions of retinal fundus photographs provided by volunteers on the Amazon Mechanical Turk (AMT) crowdsourcing marketplace with expert-provided gold-standard grading, and explored whether determination of the consensus of crowdsourced classifications could be improved beyond a simple majority vote (MV) using regression methods.

Methods:

One thousand two-hundred retinal images of individuals with diabetes mellitus from the Messidor public dataset were posted to AMT. Eligible crowdsourcing workers had at least 500 previously approved task with an approval rating of 99% across their prior submitted work. Ten workers were recruited to classify each image as normal or abnormal. If half or more workers judged the image to be abnormal, the MV “consensus” grade was recorded as abnormal. Logistic regression was used to determine if a more accurate “consensus” could be devised. Finally, Rasch analysis was used to calculate worker ability scores in a random 50% training set, which were then used as weights in a regression model in the remaining 50% test set. Outcomes of interest were the percent correctly classified images, sensitivity, specificity, and area

under the receiver-operator characteristic (AUROC) for the consensus grade as compared with the expert grading provided with the dataset.

Results: Using MV grading, the consensus was correct in 75.5% of images, with 75.5% sensitivity, 75.5% specificity, and an AUROC of 0.75 (95% Confidence Interval (CI) 0.73-0.78). A logistic regression model using Rasch-weighted individual scores generated an AUROC of 0.901 (95% CI 0.88-0.93) compared with 0.89 (95% CI 0.86-92) for a model using unweighted scores (Chi² p-value < 0.001). Setting a diagnostic cut-point to optimize sensitivity at 90%, % correct was 77.7%, sensitivity 90.3%, specificity 68.7%, and AUROC 0.80 (0.76-0.83).

Conclusions: Crowdsourced interpretations of retinal images provide rapid and accurate results as compared with a gold-standard grading. Creating a logistic regression model using Rasch analysis to weight crowdsourced classifications by worker ability improves accuracy of aggregated grades as compared with simple majority vote.

Acknowledgements:

Retinal images kindly provided by the Messidor program partners

(see <http://www.adcis.net/en/DownloadThirdParty/Messidor.html>). Research design was done in collaboration with Dr. Eliseo Guallar, Dr. David Friedman, and Dr. Lucy Mudie.

Table of Contents

Abstract.....	ii
Acknowledgements:	iv
List of Tables	vi
List of Figures	vii
Background	1
Crowdsourcing Background	3
Methods.....	4
Phase 1: Baseline Trial with Majority Vote Analysis.....	6
Phase 2: Logistic Regression Model.....	8
Phase 2 Outcomes	10
Phase 2 Statistical Analysis	10
Phase 3: Weighted Logistic Regression Model	11
Results.....	16
Phase 1 Results: Baseline Majority Vote	16
Phase 2 Results: Logistic Regression.....	17
Phase 3 Results: Weighted logistic regression model	22
Discussion.....	30
Works Cited.....	38
Curriculum Vitae	44

List of Tables

Table 1. Area under the receiver operator characteristic for each of the a priori and automated models.....	20
Table 2. Characteristics of different cut-off values using the final model compared with the naïve model and results from the Phase 1 baseline task.	21
Table 3. Characteristics of different cutpoint values using the weighted logistic model, as compared with the majority vote weighted cutpoint and the Phase 1 baseline task. (MV = majority vote).....	27
Table 4. Comparison of “easy” and “difficult” to grade images.....	29

List of Figures

Figure 1 Screenshot of the Amazon Mechanical Turk web interface for fundus photo grading.	5
Figure 2. Histogram of distribution of retinopathy grades in Messidor dataset.....	7
Figure 3. Flow diagram of images selected for Phase 2.	9
Figure 4. Screenshot of the Amazon Mechanical Turk web interface for Phase 2 fundus photo grading.....	9
Figure 5. Screenshot of the Volunteer Science hosted web interface for Phase 3 fundus photo grading.....	13
Figure 6. Simulated "red-free" retinal photograph created by deleting the red channel in Adobe Lightroom.	14
Figure 7. Receiver operator characteristic for the diagnosis of abnormal retinal photograph in the Phase 1 baseline analysis.....	17
Figure 8. Rate of image download by unique IP (internet protocol) address as a proxy for task viewing and completion demonstrates rapid access of the task.....	18
Figure 9. Comparison of area under the receiver operator characteristic between the naïve analysis and the final logistic model demonstrates an improved range of sensitivity and specificity of the model.....	21
Figure 10. Scatter plot of percentage of correctly graded images versus the number of tasks performed in the Phase 1 baseline task.	23
Figure 11. Histogram of Turker measure in log-odds units (logits) as determined by Rasch analysis used the random 50% (600 images) Training set.	24
Figure 12. Histogram of Turker weights expressed in odds of correctly classifying an average image correctly in the random 50% (600 images) Training set with the top and bottom centile (1%) truncated.	25
Figure 13. ROC generated from a logistic regression model using weighted consensus scores of the random 50% (600 images) Test set and a second using the non-weighted scores from the same data.	26
Figure 14. ROC from logistic regression model using weighted consensus scores using a dichotomization cutpoint designed to permit sensitivity of 90% shown alongside unweighted and Rasch-weighted majority vote cut-points.	27
Figure 15. Histogram displaying the distribution of image measures of the 1200 Messidor images. Images with more negative scores are “harder” to grade correctly, which images with more positive scores are “easier” to grade correctly.....	29
Figure 16. Representative retinal fundus images organized by progressive ease of grading correctly (A-E). A) The image reveals areas of chorioretinal atrophy (arrow), but is without lesions of diabetic retinopathy. B) This image reveals very subtle microaneurysms (arrows). C) This image reveals more obvious microaneurysms (arrowheads) and subtle hard exudates (arrow). D) This image reveals more apparent hard exudates (arrow). E) This image reveals obvious hard exudates (arrow) and more obvious hemorrhagic microaneurysms (arrowhead).	30

Background

Diabetes mellitus (DM) is a highly prevalent disease affecting over 415 million individuals worldwide, 80% of whom reside in low and middle income countries.¹ By 2040, the prevalence of DM is expected to reach 642 million, with the largest increases seen in countries with developing economies.¹ In the United States, 21.0 million people had known diabetes in the 2012, and another 8.1 million had undiagnosed diabetes.²

Diabetic retinopathy (DR) is an important complication of DM, currently affecting approximately 93 million people worldwide, with 28 million of these suffering from vision-threatening DR.³ It is estimated that the number of Americans with DR will reach 16 million by 2050, with 3.4 million of these individuals afflicted with vision-threatening DR.⁴

While DR is the leading cause of vision loss in working age individuals,⁴ screening for DR is an effective and cost-effective means of identifying the disease early, referring affected individuals for appropriate therapies, and preventing vision loss.⁵⁻⁸ Despite the increasing prevalence of DR, the annual increase in the number of practicing ophthalmologists is only 2%,⁹ largely in high-income countries.¹⁰ As a way of overcoming human resource shortfalls, and as a way to increase adherence with diabetic retinopathy screening recommendations more broadly, telehealth programs using non-mydratic fundus photography and remote interpretation are increasing.¹¹⁻¹³

In addition to improving screening uptake, telehealth may provide ways to reduce provider, payer, and societal costs.¹⁴⁻¹⁶ Among the costs of a telehealth program for diabetic retinopathy screening are the fundus camera, the telehealth software package, and the human resources needed for image acquisition and interpretation. Fundus photo interpretation costs in diabetic retinopathy screening may be high given labor-intensive interpretation protocols, and the need to interpret multiple images per patient. Computerized, semi-automated image analysis techniques have been developed which may be able to reduce physician workload and screening costs;¹⁷⁻¹⁹ however, these methods are not yet FDA-approved, nor in wide use clinically at this time. As telehealth expansion continues, novel, low-cost methods will be needed to interpret the large volume of fundus images expected with rising incidence of diabetes, especially in resource-poor settings and in large public health screenings.

The use of crowdsourcing in biomedical research is in its infancy, though some groups have used this method in public health research,²⁰ and to interpret biomedical images.²¹ Crowdsourcing has been used to categorize a number of fundus photos with a variety of diagnoses as normal or abnormal.²² In this trial conducted in the U.K. using untrained graders, the sensitivity was $\geq 96\%$ for normal versus severely abnormal and between 61-79% for normal versus mildly abnormal.²² In a proof-of-concept study, we have demonstrated that untrained crowdsourced workers can rapidly and accurately identify images with DR.²³ In this study we will seek to perform an external validation of our

method of crowdsourcing DR identification using a public dataset of 1200 retinal photographs, and explore methods of improving the determination of a consensus score from multiple individual crowdsourced grades including creating a logistic regression model that includes other data points collected at the time of the Turker grading, and a second model that weights the responses of Turkers based on ability in a training dataset using the Rasch model.

Crowdsourcing Background

Crowdsourcing, defined by Brabham, is “an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals.”²⁴ A subset of crowdsourcing, which Brabham terms “distributed-human-intelligence tasking,”²⁴ can involve subdividing larger tasks into small portions and then recruiting a group of individuals to each complete these small portions, and only collectively, the entire task.²⁴ Amazon Mechanical Turk (AMT) is an online distributed human intelligence market that allows access to thousands of people who can quickly accomplish small, discrete tasks for small amounts of money. Typical AMT tasks include tagging photos, translating words, or writing very short articles for websites. AMT has its own vocabulary used by workers (“Turkers”) and task administrators (“Requestors”). A “Human Intelligence Task” (HIT) is a small job which may be performed in a matter of seconds or minutes and, once the work is approved by the Requestor, may pay \$0.01-\$0.25 or more per task depending on the complexity of

the HIT. A group of HITs is called a “batch,” and is made up of similar HITs. Depending on the complexity of the task and the payment offered by the Requestor, a batch is often completed within minutes or hours of posting.²⁵


Several methods for aggregating multiple grades into a “consensus” score have been described. The simplest method, termed “majority vote” (MV), involves promoting the modal response to the crowdsourced determination.²⁶ In a binary classification scheme, whichever response is selected by half or more of respondents becomes the “consensus.” While this approach is computationally simple, the differential ability of workers is ignored as is differential difficulty of the unique tasks. Therefore other methods of aggregating scores have been explored that rely on patterns of individual worker responses over multiple tasks, and comparisons with expert annotations where available.²⁶⁻²⁹

Methods

An interface for fundus photo classification has been previously described for the AMT crowd-sourcing platform.²³ The United Kingdom national screening program grading scale³⁰ was chosen due its broad clinical telemedicine deployment. For the purposes of the study, terms from this scale were translated into plain language; “background” retinopathy was called “mild,” “preproliferative” was called “moderate,” and “proliferative” was called “severe.” “Maculopathy” is defined as abnormal on a training image with otherwise moderate disease, but is not coded separately. The AMT interface

was designed to provide training on grading of DR within each HIT. This training includes 7 images annotated with the salient features of each level of retinopathy in plain language. Turkers are presented with the following text: “This is a photo of the inside of the eye. We are looking to label eyes as healthy or unhealthy with respect to diabetes. Rate this eye.” Turkers can hover their mouse over the adjacent training images (2 normal, 1 mild, 1 moderate, 3 severe) while reviewing the active test image (Figure 1). This layout allows for all of the training and grading to occur in one browser window. In order to be eligible to view and complete the HITs Turkers needed to have successfully completed 500 prior HITs and have an overall HIT approval rate of 99%. Turkers receive US \$0.10 per image, with a 40% commission going to Amazon, for a total cost of US \$1.40 per image.

This is a photo of the inside of the eye. We are looking to label eyes as healthy or unhealthy with respect to diabetes. This task is difficult, but all good faith efforts will be accepted. Please answer both questions even if you are not sure. To close pop-out, move cursor left, or [tap here](#) on iPad.

	<p>Hover for examples:</p> <p>Normal</p> <p>Abnormal - Mild</p> <p>Abnormal - Moderate</p> <p>Abnormal - Severe Example 1</p> <p>Abnormal - Severe Example 2</p> <p>Abnormal - Severe Example 3</p>	<p>1) Rate this eye: NOTE: The color of the photo is not a marker of disease.</p> <p><input type="radio"/> Normal (Healthy) <input type="radio"/> Abnormal (mild, moderate, severe disease)</p> <p>2) Is this photo good enough quality to grade?</p> <p><input type="radio"/> Yes <input type="radio"/> No</p>
-------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Please leave any comments or feedback you have about this HIT or project:

Figure 1 Screenshot of the Amazon Mechanical Turk web interface for fundus photo grading.

Phase 1: Baseline Trial with Majority Vote Analysis

For the first phase of this project, 1200 images from the Messidor public dataset³¹ were posted for 10 unique binary annotations to provide external validation of the prior proof-of-concept study.* The Messidor dataset is composed of 800 mydriatic and 400 nonmydriatic retinal fundus photos of universally high quality and resolution. The images are supplied with ground truth grading on the following scale:

- 0: normal: no microaneurysms, no hemorrhages
- 1: 1-5 microaneurysms, but no hemorrhages
- 2: 6-14 microaneurysms OR 1-4 hemorrhages, but no neovascularization
- 3: 15 or more microaneurysms OR 5 or more hemorrhages OR presence of neovascularization (Figure 2)

* *Kindly provided by the Messidor program partners*
(see <http://www.adcis.net/en/DownloadThirdParty/Messidor.html>).

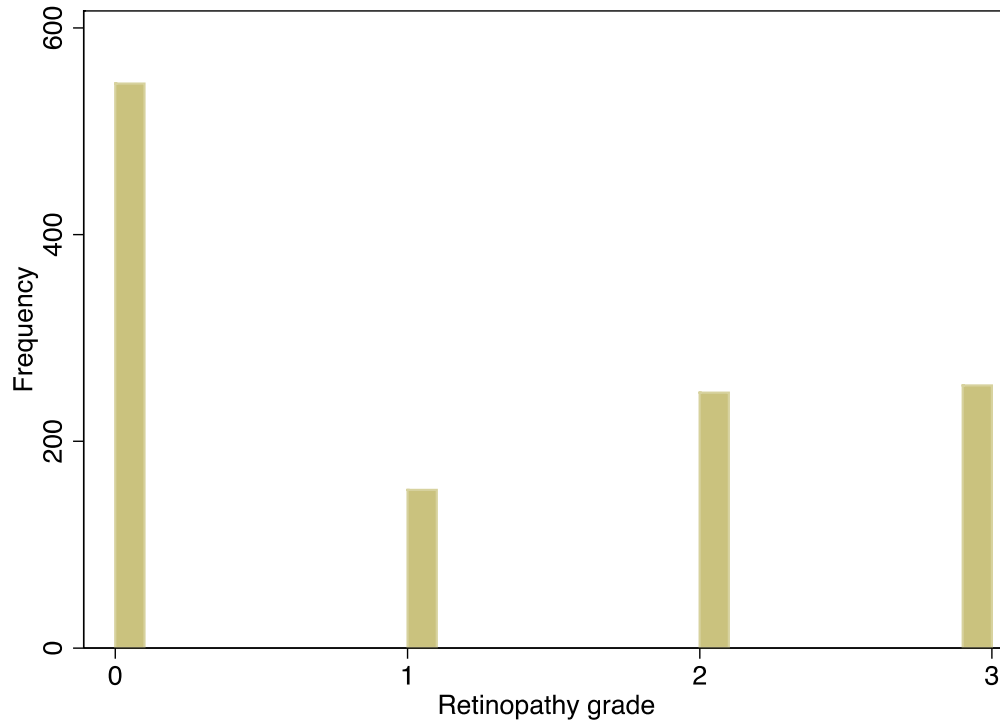


Figure 2. Histogram of distribution of retinopathy grades in Messidor dataset.

For the purposes of this study, the presence of 5 or fewer microaneurysms was felt to be clinically insignificant and thus we classified Messidor 0 and 1 images as “normal” and Messidor 2 and 3 images as “abnormal.”

To create the “majority vote” (MV) consensus, each image was assigned the grade of “abnormal” if half or more Turkers deemed an image “abnormal,” otherwise the image was classified as “normal.” Sensitivity, specificity and area under the receiver operator characteristic (AUROC) were calculated. This batch and grading scheme served as the “baseline” results for comparison with the regression models used in later phases of the research.

As there was no *a priori* rationale to suggest the mean Turker score (with rounding towards abnormal) would provide the most accurate approximation of the ground truth classification, 2 additional methods of generating consensus were explored.

Phase 2: Logistic Regression Model

For this phase, we sought to create a logistic prediction model to explore whether any of the additional data (in addition to the Turker classification of normal or abnormal) provided by AMT could be used to improve diagnostic accuracy. Additionally, the images were divided into 4 horizontal quadrants to a) force increased zoom, b) make the task more abstract, and c) permit 40 separate classifications per photograph.

One hundred fifty Messidor grade 3 images and fifty grade 0 images were randomly selected from the MESSIDOR dataset (Figure 3). Each image was split into 4 horizontal strips using Adobe Lightroom, each comprising 25% of the image, each of which was posted for 10 unique gradings (Figure 4). Along with the Turker's impression of retinopathy and a judgement of image quality, the time spent on the task, and number of prior retinal grading tasks were also recorded.

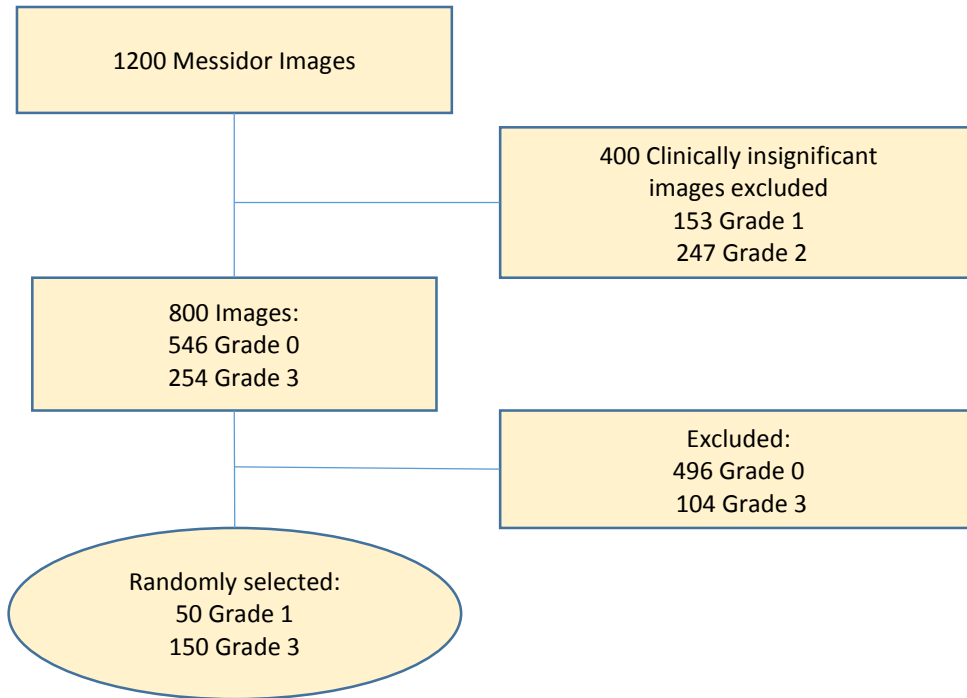



Figure 3. Flow diagram of images selected for Phase 2.

This is a portion of a photo of the inside of the eye. We are looking to label eyes as healthy or unhealthy with respect to diabetes. This task is difficult, but all good faith efforts will be accepted. Please answer both questions even if you are not sure. To close pop-out, move cursor left, or [tap here](#) on iPad.



<p>Hover for examples:</p> <p>Normal</p> <p>Normal Ex. 2</p> <p>Abnormal: Mild</p>	<p>Abnormal: Moderate</p> <p>Abnormal: Severe</p> <p>Abnormal: Severe Ex. 2</p> <p>Abnormal: Severe Ex. 3</p>	<p>1) Label this eye, based on the part of the photo that you can see: NOTE: The color/tint of the photo is not a marker of disease.</p> <p><input type="radio"/> Normal (Healthy)</p> <p><input type="radio"/> Abnormal (Mild, Moderate or Severe Disease)</p> <p>2) Is this photo good enough quality to grade?</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p>
------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Please leave any comments or feedback you have about this HIT or project:

You must ACCEPT the HIT before you can submit the results.

Figure 4. Screenshot of the Amazon Mechanical Turk web interface for Phase 2 fundus photo grading.

Phase 2 Outcomes

As in Phase 1, main outcomes were defined as percent agreement between the Turker consensus grade and the gold standard grade, as well as sensitivity and specificity for the detection of an abnormal retina. The area under the curve of the receiver operator characteristic (AUROC) was also calculated. For the naïve assessment for this phase, Turker consensus was defined as follows: if half or more of the Turkers marked the image quadrant as abnormal, the image strip was coded as abnormal. If one or more strips was coded abnormal in this task, the entire image was coded as abnormal. In the improved assessment, logistic regression was used to devise a predictive algorithm for image abnormality. The model prediction was also compared against the MV score from the baseline phase.

Phase 2 Statistical Analysis

Several *a priori* prediction models were explored. First, a logistic regression model was created including the raw Turker score (sum of 40 unique grades - 10 per image strip - where 0=normal; 1=abnormal), the total time spent on the image by all Turkers, the total prior experience by all Turkers and the total number of times the image was marked ungradeable after each continuous variable was examined for linearity using lowess non-parametric smoothed regressions (A priori 1). Variables with evidence of non-linearity were categorized or dichotomized. Additional models were created using the binary score for each strip quadrant as separate variables (A priori 2). Next, a model

using an ordinal categorical variable with the number of strips coded as abnormal was generated (A priori 3). Finally, a model using disjoint categorical variables for the number of abnormal quadrants was generated (A priori 4).

Next several automated model selection tools were utilized starting with all variables collected. First, the variance inflation factor (VIF) was checked on a model including all covariates, and items with $VIF > 8$ were excluded. VIF was rechecked until no items had $VIF > 8$. Next automated forward and backward predictor selection based on likelihood ratio tests, and forward and backward predictor selection based on Aikake's Information Criteria (AIC) were run. The best model from each automated method was then used for comparison.

All models were then compared using the receiver operator characteristic, and the model with the highest area under the curve (AUROC) was selected. Jackknife resampling was used to cross-validate the AUROC. Percent agreement, sensitivity, specificity and AUC were calculated and compared with the results from the naïve assessment.

Phase 3: Weighted Logistic Regression Model

For this phase, we recognized that among Turkers there is a range of ability, and among images there is a range of difficulty. In order to improve throughput, force Turkers to

grade images in multiples of 10 (rather than single images), and collect more data about the Turkers' interactions with the task for future phases, the project was migrated to a new online interface (Figure 5). In the new interface, the 1200 Messidor images were posted for binary grading first using the full color images, and then again with the images converted to grayscale with the red color channel removed in Adobe Lightroom (applied "B&W Preset" with "Green Filter," in "Black and White Mix" reduce Red to -75) (Figure 6). This was done to simulate "Red-free" images, which may allow for better detection of diabetic retinopathy.³² This allowed us to have a dataset with 30 grades per image, albeit captured under slightly different circumstances.


Figure 5. Screenshot of the Volunteer Science hosted web interface for Phase 3 fundus photo grading.

VOLUNTEER SCIENCE Feedback Play Games Take Surveys Learn More Sign In

You're playing: **Eyedentification**

Classification 10 of 10

Show Other Players Selections



Please Answer all the questions then click **Next** to continue.
Click the tabs to see examples of normal and abnormal eyes.
NOTE: *The color/tint of the photo is not a marker of disease.*

Questions Normal Normal Mild Moderate Severe Severe Severe

1) Label this eye:
 Normal (Healthy)
 Abnormal (mild, moderate, severe disease)

2) Is this photo good enough quality to grade?
 Yes
 No

+ Group Chat Previous Next

Zoom: In Full Out Clear My Selections

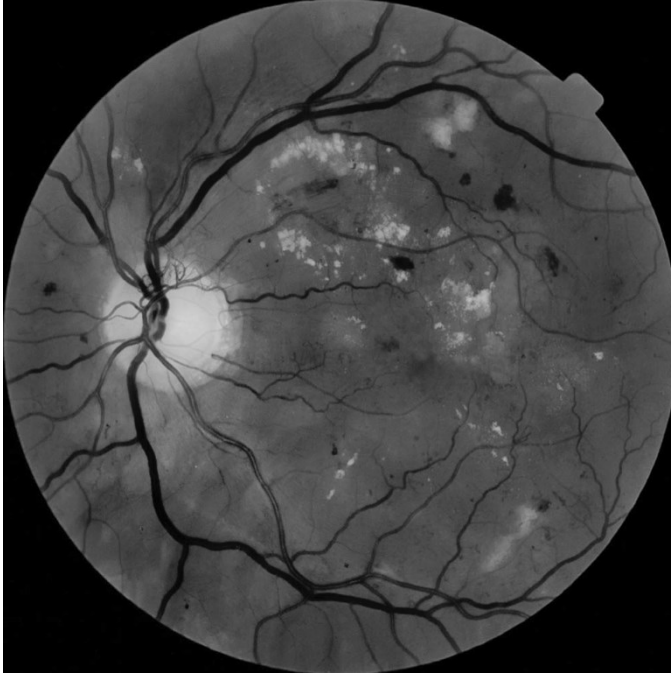


Figure 6. Simulated "red-free" retinal photograph created by deleting the red channel in Adobe Lightroom.

The dataset of 1200 images was randomly divided into 600 training and 600 test images.

Using the training images, a matrix of images and individual Turkers was created with each cell either being a missing datapoint (if that particular Turker did not grade that particular image), a "1" for a correct classification, or a "0" for an incorrect classification.

Rasch analysis was then performed to determine the "image measures" and "Turker measures."

The Rasch model, as described by Linacre,³³ specifies:

$$\log_e \left(\frac{P_{ni}}{1 - P_{ni}} \right) = B_n - D_i$$

Where, in the present study, P_{ni} is the probability of a given image n of difficulty B_n having a correct response provided by Turker i of skill level D_i . Therefore, the Turker's ability measure and the image's difficulty measure are expressed as log-odds units (logits), theoretically ranging from $-\infty$ to $+\infty$. The negative exponentiated Turker ability measure, then, is the odds that an image of average difficulty (i.e., $B_n=0$) would be categorized correctly by that particular Turker. This value was then multiplied by each of that Turker's categorizations from the test set (with abnormal = 1, normal = -1). The weighted scores were then summed for each image. In an initial analysis, the consensus image score was considered to be "abnormal" if greater than or equal to zero, and "normal" otherwise. Sensitivity, specificity and AUROC were calculated as above with comparison to the Baseline MV results. In a subsequent analysis, the consensus image score was included as a continuous variable in a logistic regression model to determine the ideal cut-off value for different values of percent correct, sensitivity and specificity.

Data were analyzed using Stata Statistical Software: Release 14 (StataCorp.

2015. College Station, TX: StataCorp LP) and Winsteps® Rasch measurement computer program. (Linacre, J. M. (2016). Beaverton, Oregon: Winsteps.com). The Johns Hopkins University Institutional Review Board (IRB) deemed this research IRB-exempt as non-human subjects research.

Results

Phase 1 Results: Baseline Majority Vote

A batch of 12,000 (1200 images x 10 repetitions) tasks was posted on AMT March 13, 2015, 11 am Eastern for a total cost of US \$1440 (\$12000 for Turker compensation, \$240 for Amazon commission). The grading was complete in 68 minutes, with 97% of images graded within 35 minutes. Tasks submitted without image grades were immediately re-posted so there were no missing data. The tasks were submitted by 281 unique Turkers, with each submitting a mean of 42.7 tasks (median 28, mode=1).

The MV consensus was correct in 75.5% of images. Sensitivity and specificity were 75.5%. The area under the receiver operator characteristic was 0.755 (

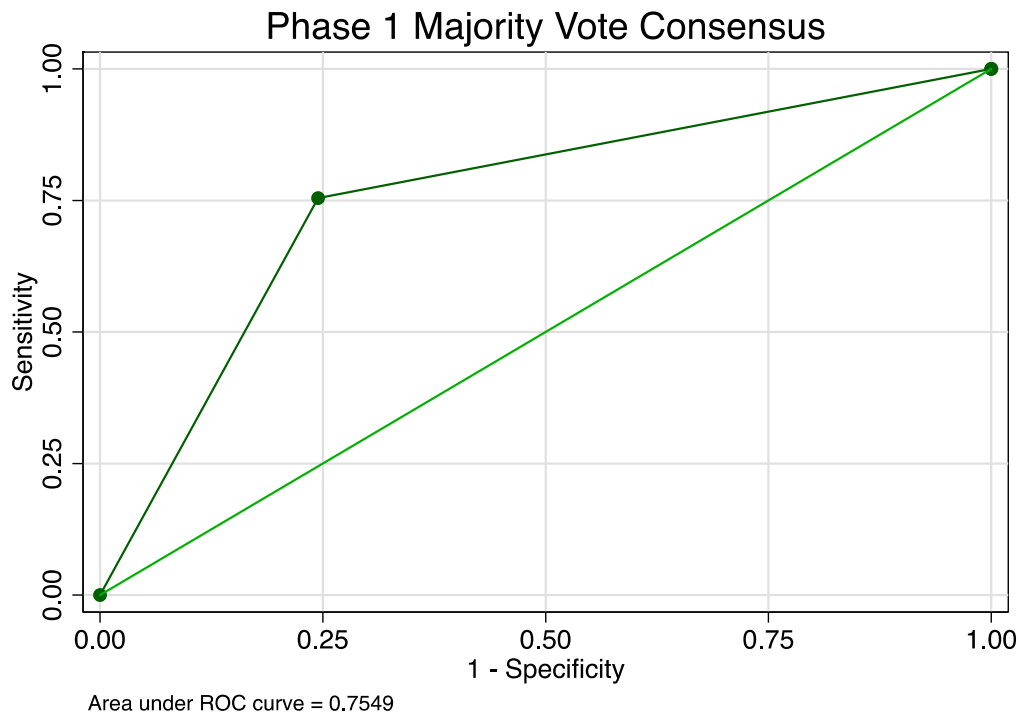


Figure 7).

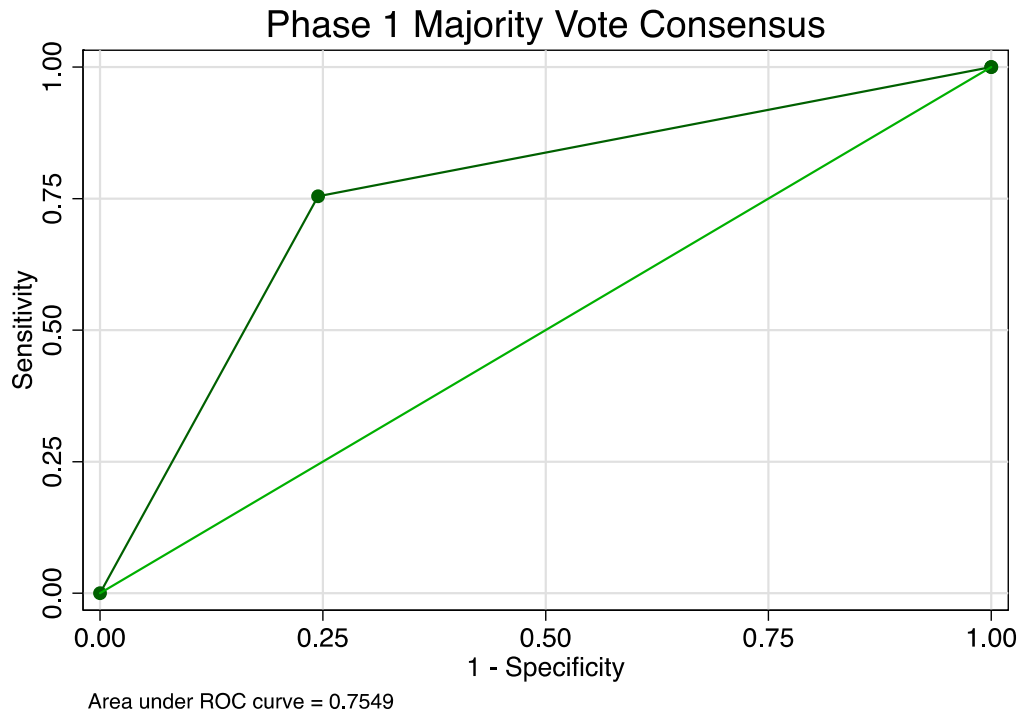


Figure 7. Receiver operator characteristic for the diagnosis of abnormal retinal photograph in the Phase 1 baseline analysis.

Phase 2 Results: Logistic Regression

A batch of 8000 tasks (4 strips x 400 images x 10 repetitions) was released on AMT March 18, 2016, 9 am Eastern for a total cost of US \$1120 (\$800 for Turker compensation, \$320 for Amazon commission). The gradings were completed in 53 minutes. The majority of images were graded within 15 minutes (Figure 8). Three of 8000 gradings were blank/missing, and due to the small amount of missing data (0.04%), this was ignored in the analysis.

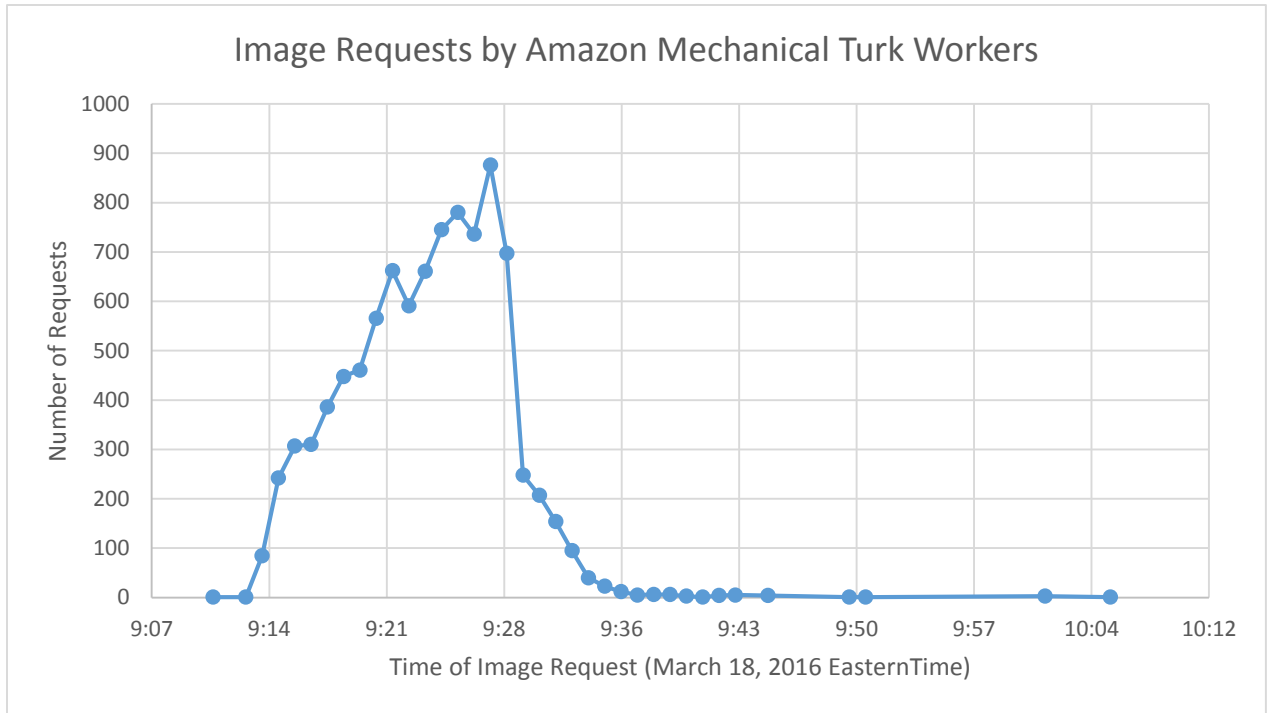


Figure 8. Rate of image download by unique IP (internet protocol) address as a proxy for task viewing and completion demonstrates rapid access of the task.

In our naïve phase 2 analysis, we assumed that if the consensus (MV) score of one quadrant of an image was determined to be abnormal, the overall image would be considered abnormal. In the naïve analysis, 82.5% of images were correctly classified, with 92.7% sensitivity and 52.0% specificity, and with an AUROC of 0.723 (Table 1, Table 2). The results for some images from the Phase 1 baseline analysis were correctly classified in 91.5% of images, with 92.3% sensitivity and 88.0% specificity, and with an AUROC of 0.903 (with similar results using the MV consensus method with the batches using the improved Volunteer Science interface with both color and simulated red-free images; data not shown).

When creating the logistic regression model, the relationship between abnormality and the raw score was deemed to be approximately linear (Supplemental figure 1). Total time spent was deemed to have evidence of non-linearity, and so this value was categorized into <3000 seconds, between 3000 and 5000 seconds and >5000 seconds (Supplemental figure 2). Likewise, total prior experience was deemed to have evidence of non-linearity, and so this value was dichotomized at 5000 prior tasks (Supplemental figure 3).

Results from the *a priori* and automated model selection tool are shown in Table 1/Supplemental figure 4. Prior to automated model selection, all variables were tested for collinearity. A variable for the sum of the consensus score for each strip (values 0-4) was found to be strongly collinear and was removed from this model. Next the raw score was found to be collinear with the individual strip binary scores, so these were not included together in the automated selection tool. Both forward and backward model selection using likelihood ratios returned the same model, raw score alone when excluding the individual strip scores, and the scores for strip 3 and strip 4 when excluding the raw total. The model including just the raw score was called “Auto 1” and then “Auto 2” was defined as using all four binary strip scores, as it did not make biological sense to exclude the strip 1 and 2 scores. Auto 2 was the same model as A priori 2. Using automated model selection based on AIC, the same models were found as with the likelihood ratio test.

	Area Under Receiver Operator Characteristic (95% CI)
Naïve baseline: 1 strip MV consensus abnormal = image abnormal	0.72
A priori 1: raw total score, time, experience, quality grade	0.87 (0.81-0.93)
A priori 2 / Auto 2: binary score for each quadrant	0.85 (0.79-0.92)
A priori 3: ordinal score for number of abnormal quadrants	0.84 (0.78-0.91)
A priori 4: disjoint categorical variable for number of abnormal quadrants	0.85 (0.79-0.91)
Final model (Auto 1): raw total score	0.87 (0.81-0.93)
Final model (Jackknife crossvalidation):	0.86 (0.79-0.92)

Table 1. Area under the receiver operator characteristic for each of the a priori and automated models.

Next all models were tested for goodness-of-fit. Due to low numbers of unique covariate patterns, Pearson’s Goodness of Fit test was used, and all models showed evidence of good fit with the exception of A priori model 4.

Based on the AUROC, and the parsimony of the model, Auto 1 was chosen as the final model. A chi-squared test comparing the AUROC between the naïve analysis and the final model was statistically significant ($p < 0.001$, Figure 9) A Jackknife crossvalidation was performed in order to get a validated AUROC (Table 1). In order to determine the ideal cut-off value for an abnormal test, values maximizing percent agreement, sensitivity, and specificity were explored (Table 2/Supplemental figure 5).

	% Correct	Sensitivity	Specificity	AUROC (95% CI)
Phase 1 baseline:	91.5%	92.3%	88.0%	0.90
Naïve method:	82.5%	92.7%	52.0%	0.72
Final model (Auto 1):				
(maximizing % correct)	83.0%	93.3%	52.0%	0.73 (0.65-0.80)
(sensitivity \geq 90%)	82.0%	91.3%	54.0%	0.73 (0.65-0.80)
(specificity \geq 90%)	75.0%	70.0%	90.0%	0.80 (0.74-0.86)

Table 2. Characteristics of different cut-off values using the final model compared with the naïve model and results from the Phase 1 baseline task.

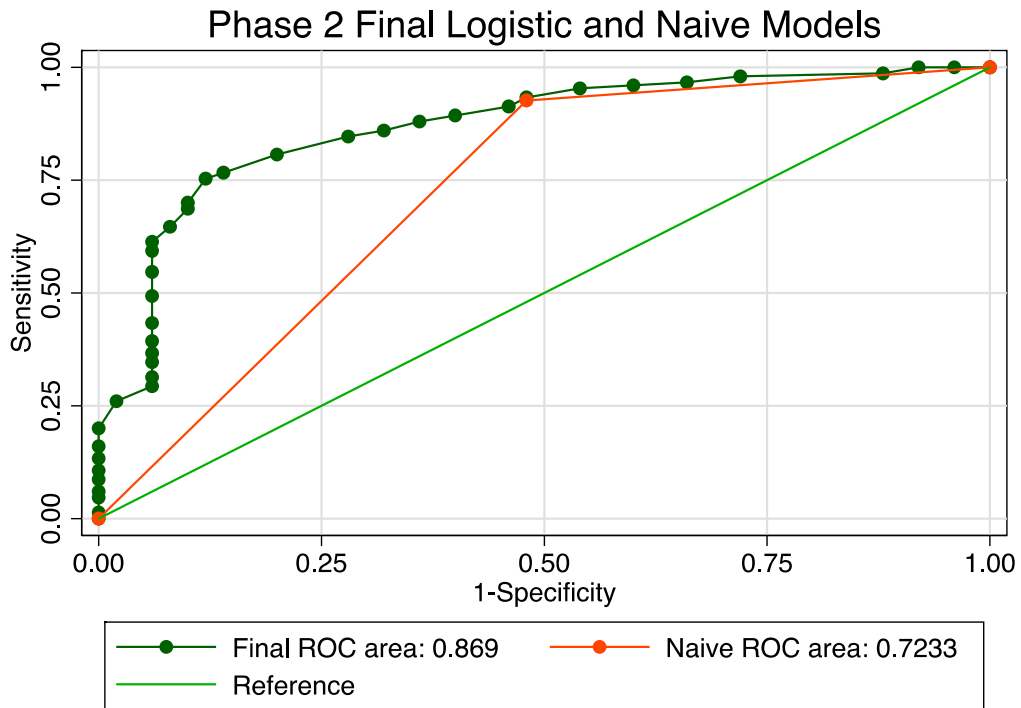


Figure 9. Comparison of area under the receiver operator characteristic between the naïve analysis and the final logistic model demonstrates an improved range of sensitivity and specificity of the model.

Phase 3 Results: Weighted logistic regression model

For this Phase, the focus shifted to the perspective of Turkers rather than on the images themselves. Exploration of Turker accuracy motivated an attempt to incorporate Turker ability into a predictive model. As demonstrated in Figure 10, and has been demonstrated in the literature³⁴ there is a distribution of Turker accuracy that is not necessarily related to the number of tasks performed. As such, any method that implicitly weights a consensus score based on number of tasks performed as does MV may reduce accuracy. In the Phase 1 baseline task, among the 281 unique Turkers median percentage of images graded correctly was 64.7% with an intraquartile range of 55.5% to 74.4%.

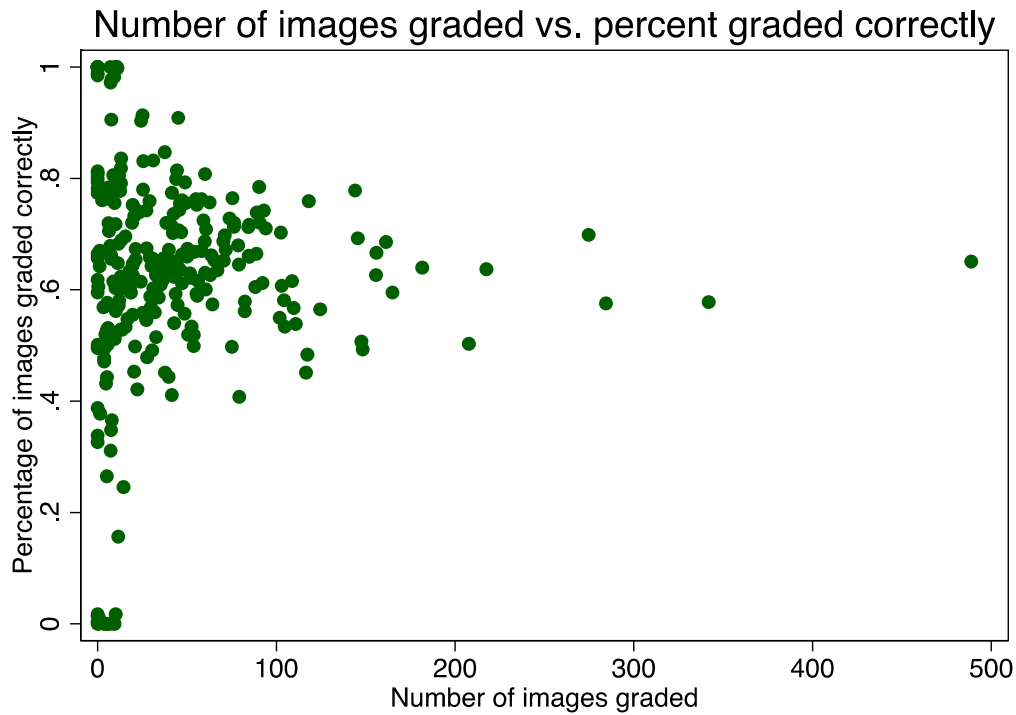


Figure 10. Scatter plot of percentage of correctly graded images versus the number of tasks performed in the Phase 1 baseline task.

Prior to performing Rasch analysis, the results of the improved Volunteer Science experiment (1200 color images + 1200 red-free images; cost \$2,558, completed over days) were merged with the Phase 1 baseline classifications to permit as many grades possible. In essence, we treated the Turkers as “test takers” taking a “test” involving grading multiple images. For stability, we excluded Turkers who had graded fewer 10 images within the training set of 600 images. Using Rasch analysis, we found Turker ability ranging from the most highly skilled at -3.75 logits, and the least skilled at 1.9 logits. The median ability is set in the model as zero, and the intraquartile range of

ability was -0.40 to 0.47 (

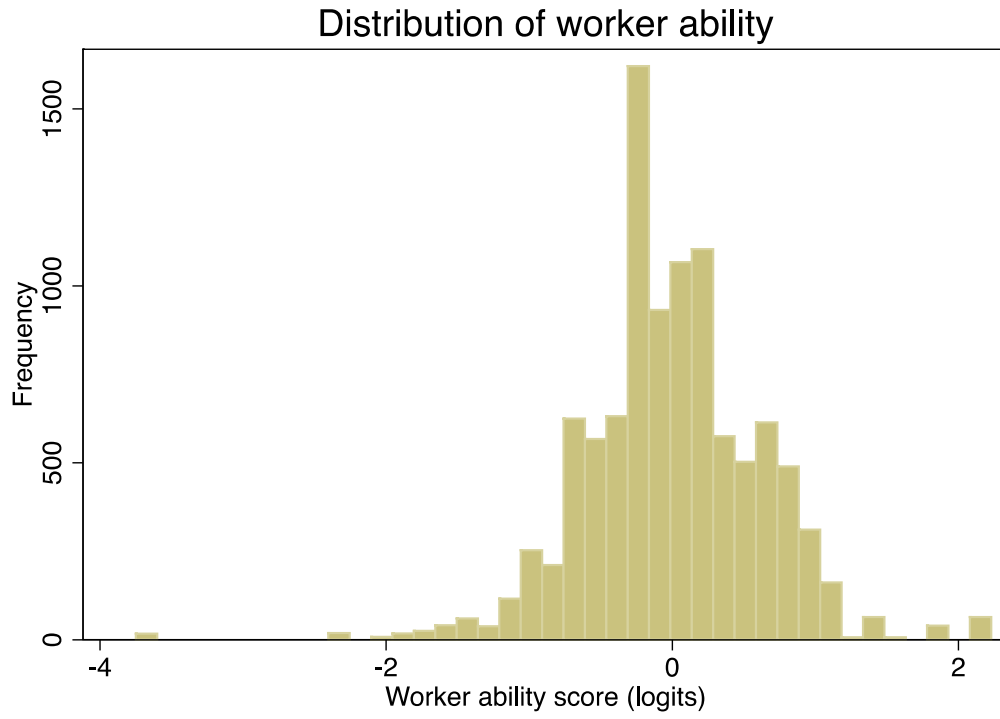


Figure 11).

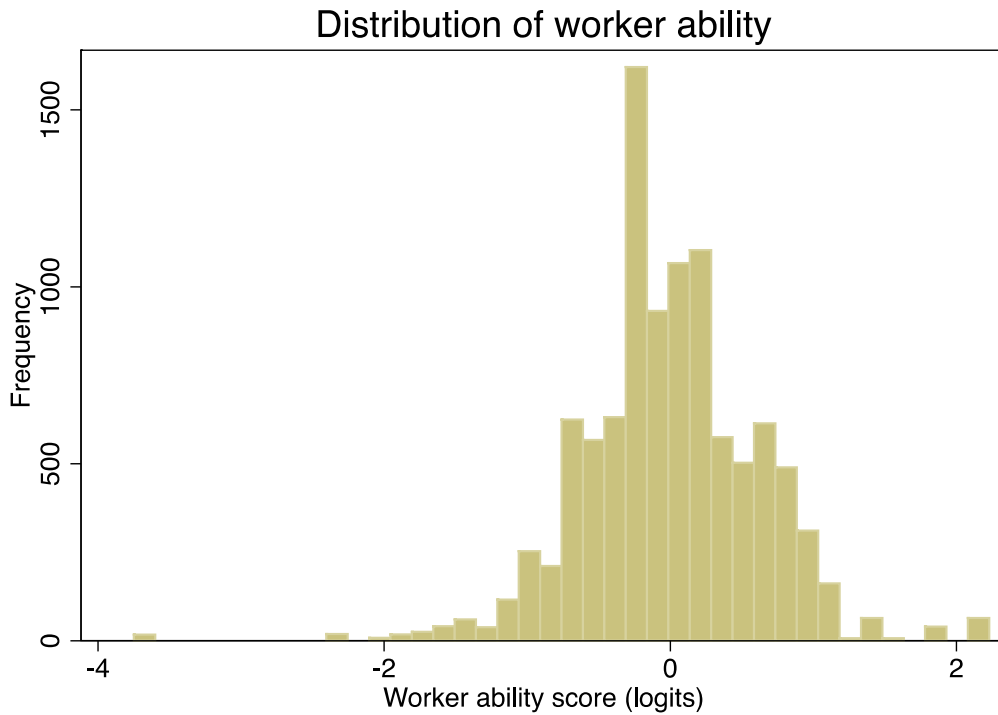


Figure 11. Histogram of Turker measure in log-odds units (logits) as determined by Rasch analysis used the random 50% (600 images) Training set.

After transformation, the Turker measure scores from log-odds to odds of correctly classifying an average difficulty image, weights outside the top and bottom centile were truncated to the level of the 1st and 99th centile to increase stability and

the effect of outliers (



Figure 12).



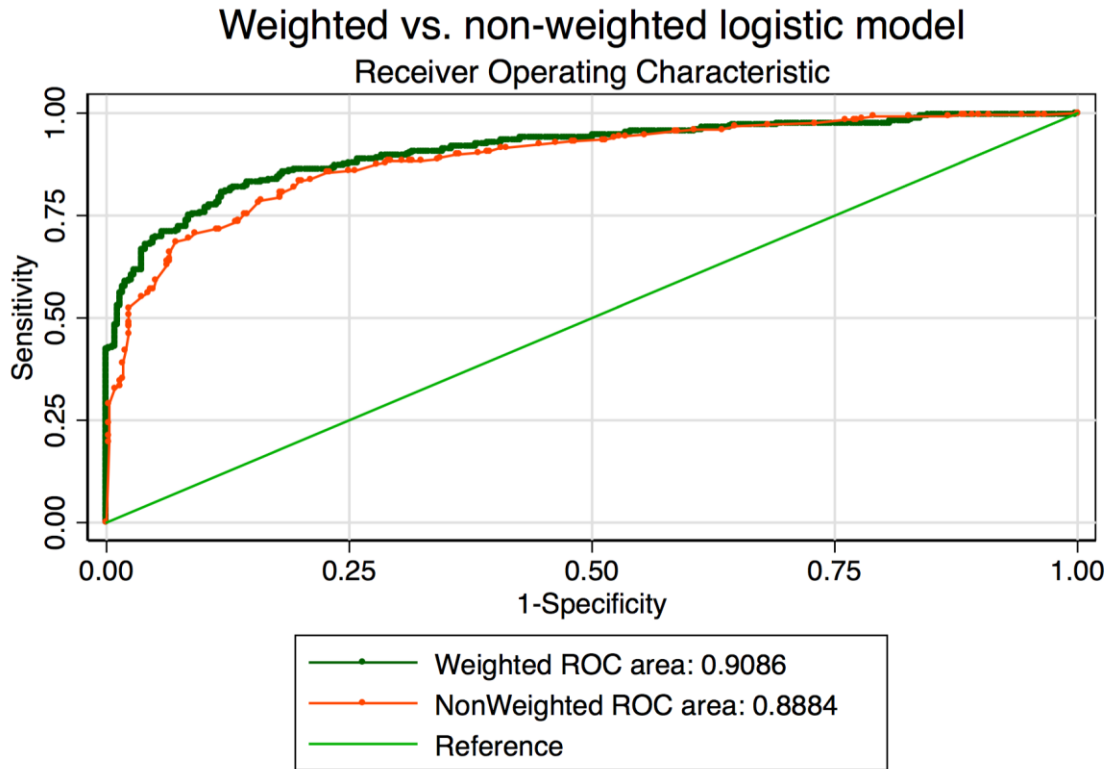
Figure 12. Histogram of Turker weights expressed in odds of correctly classifying an average image correctly in the random 50% (600 images) Training set with the top and bottom centile (1%) truncated.

When the Turker weights were applied to the classifications in the color images in the test set using the arbitrarily determined cut-off (0), the percent correctly classified improved to 80.7% with an AUROC of 0.817 from the 77.0% correct and AUROC of 77.1 determined in the phase 1 baseline task for the same images.

In order to determine the arbitrarily-determined cutoff could be improved, a logistic regression model using the consensus image score determined by the weighted Turker classifications was generated. Using this model, a much more granular ROC could be

generated. A similar ROC was generated from a separate regression model using the unweighted consensus classifications from the same batch (Figure 13). The AUROC's were 0.909 (95% Confidence Interval 0.883-0.934) and 0.888 (95% Confidence Interval 0.861-915) respectively (Chi² p-value < 0.001).

Figure 13. ROC generated from a logistic regression model using weighted consensus scores of the random 50% (600 images) Test set and a second using the non-weighted scores from the same data.



Examination of multiple dichotomization cutpoints revealed that choosing a cut-off would permit a minimum sensitivity of 90.3% allows for specificity of 68.7% and correctly classified at 77.7% with an AUROC of 0.80 (Table 3/

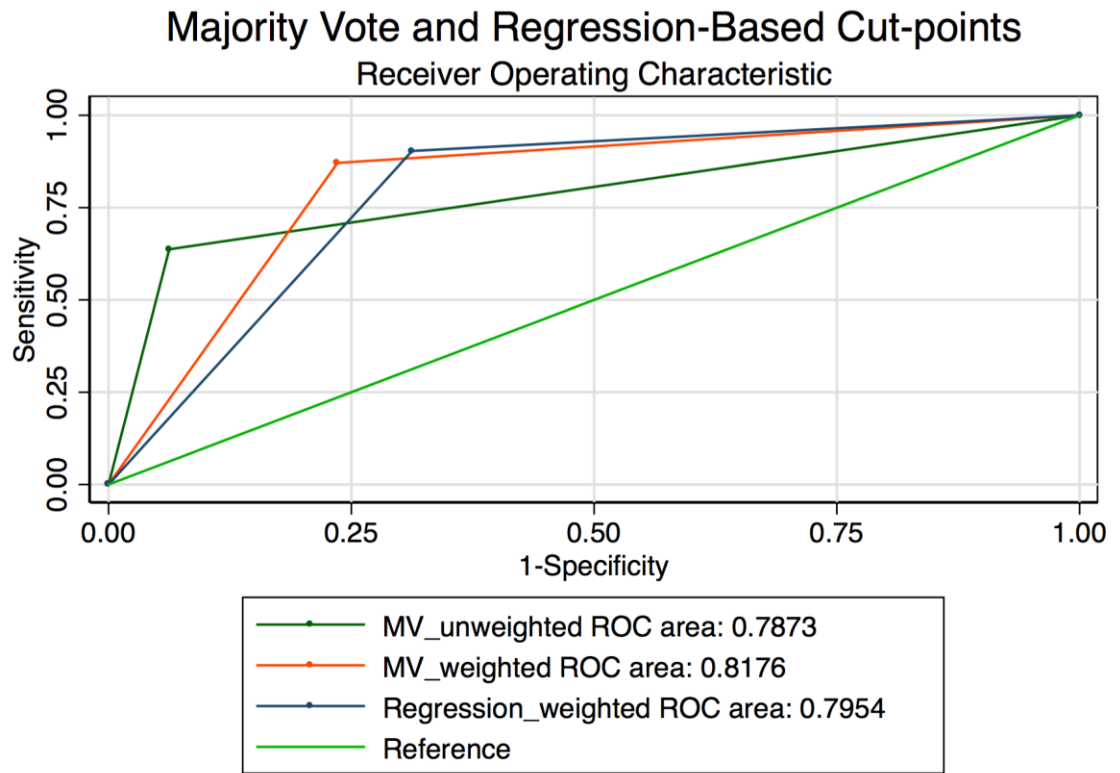


Figure 14).

	% Correct	Sensitivity	Specificity	AUROC (95% CI)
Phase 1 MV baseline:	75.5%	75.5%	75.5%	0.75
MV weighted cut-point:	80.8%	87.1%	76.4%	0.82
Weighted regression:				0.91 (0.88-0.93)
(maximizing % correct)	85.0%	81.1%	87.8%	0.84 (0.81-0.87)
(sensitivity \approx 90%)	77.7%	90.3%	68.7%	0.80 (0.76-0.83)
(specificity \approx 90%)	84.2%	75.8%	90.1%	0.83 (0.80-0.86)

Table 3. Characteristics of different cutpoint values using the weighted logistic model, as compared with the majority vote weighted cutpoint and the Phase 1 baseline task. (MV = majority vote)

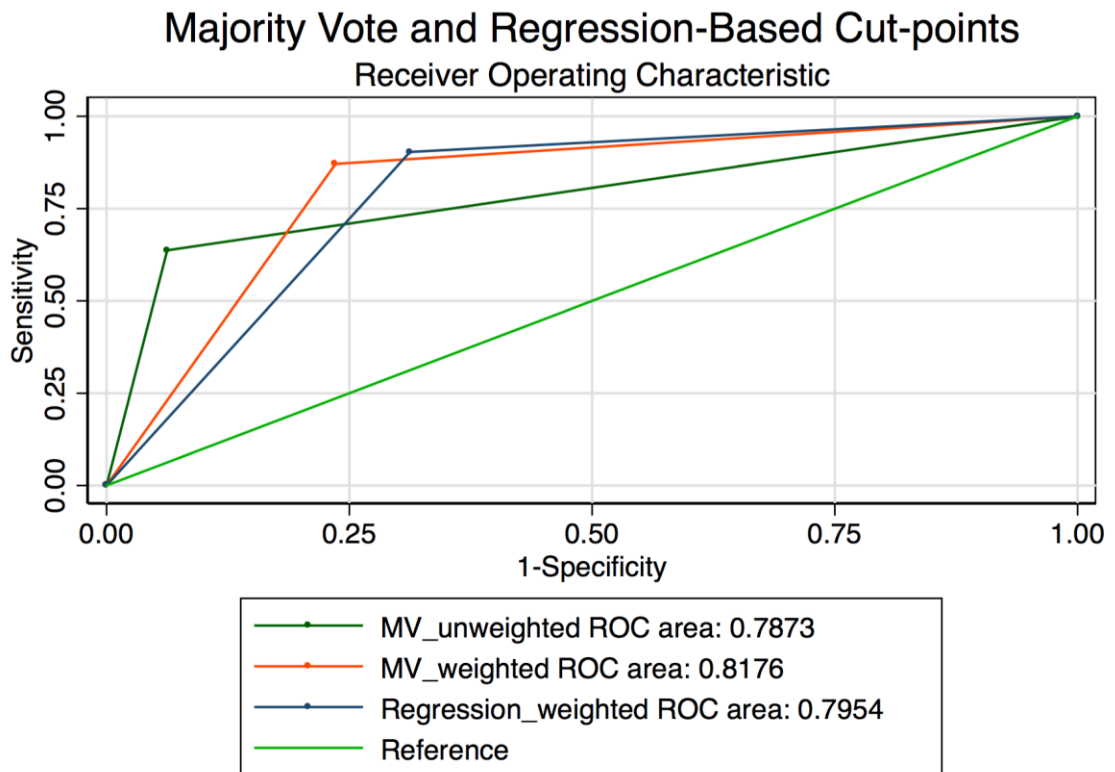


Figure 14. ROC from logistic regression model using weighted consensus scores using a dichotomization cutpoint designed to permit sensitivity of 90% shown alongside unweighted and Rasch-weighted majority vote cut-points.

Rasch analysis also allowed for a qualitative analysis of the retinal images. The images were sorted by image measure on the logit scale as generated by the Rasch analysis described earlier (

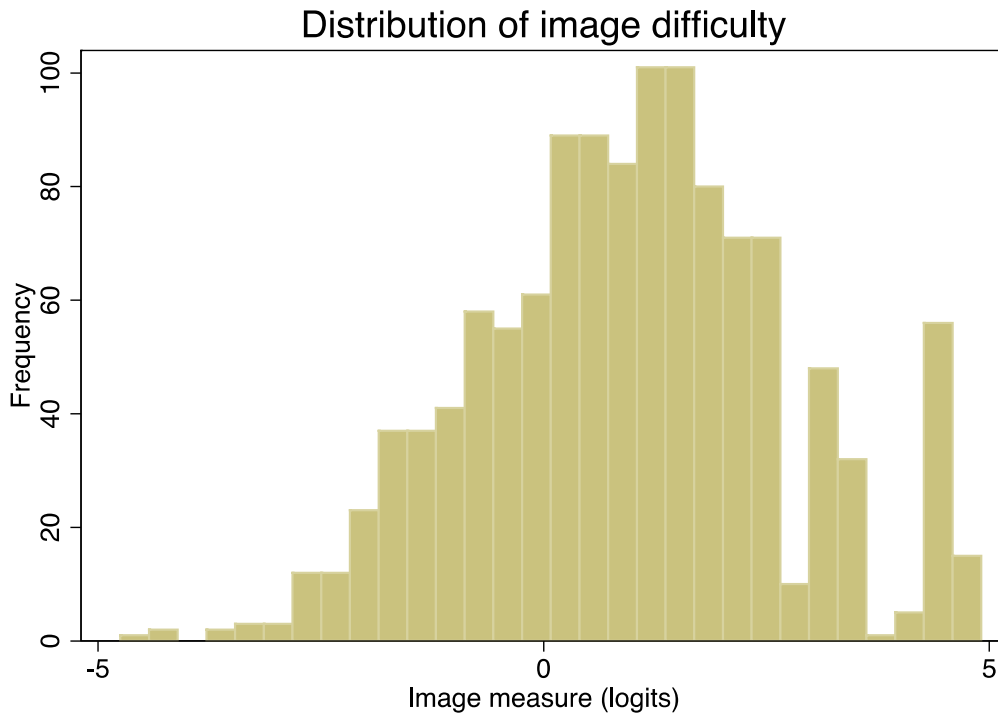


Figure 15). The 20 images with the lowest measures ranged from -4.75 to -2.56 logits, which corresponds to the log odds that a Turker of average ability (i.e., Turker measure = 0) would grade these images correctly. Among these only images 0, 1, or 2 Turkers (out of maximum of 30) graded correctly, and were designated as the most “difficult” to grade. The 20 images with the highest measures were selected as the “easiest” to grade. Twenty sequential images were then selected at the 3 quartiles as successively less difficult images to grade (Table 4/Figure 16). The “hardest” images were largely Messidor grade 0 and 1 images with some abnormal features, but without significant

diabetic retinopathy (e.g., chorioretinal atrophy, choroidal nevus, etc), that had been graded as abnormal by Turkers. Intermediate images were mostly Messidor grade 2 images with extrafoveal microaneurysms of subtle hard exudates, as well as Messidor grade 0 images without any non-diabetic pathology or distracting features. The “easiest” images were generally Messidor grade 3 with prominent hard exudates apparent.

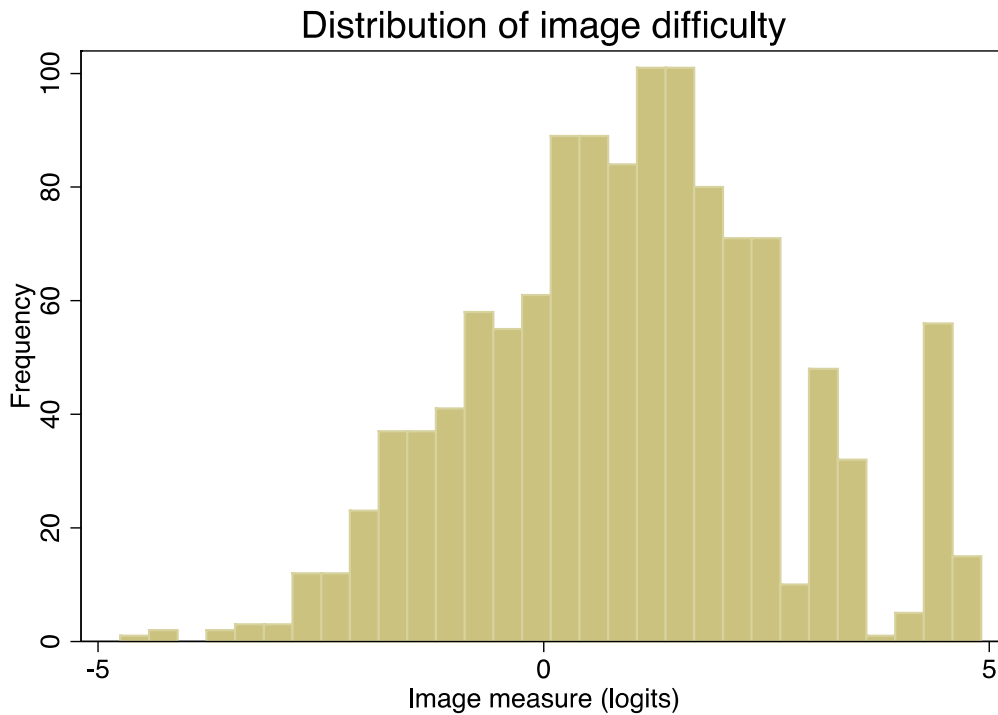


Figure 15. Histogram displaying the distribution of image measures of the 1200 Messidor images. Images with more negative scores are “harder” to grade correctly, which images with more positive scores are “easier” to grade correctly.

Difficulty	Measure score range (logits)	% images graded correct	Messidor grade (mode)
“Hardest”	-4.74 - -2.56	0-8.3%	1
Intermediate 1	-0.14 - -.04	43.4-53.9%	0

Intermediate 2	1.01 - 1.1	69.2-76.9%	0
Intermediate 3	2.04 - 2.09	85.2-88.9%	0
“Easiest”	4.5 - 4.91	100%	3

Table 4. Comparison of “easy” and “difficult” to grade images

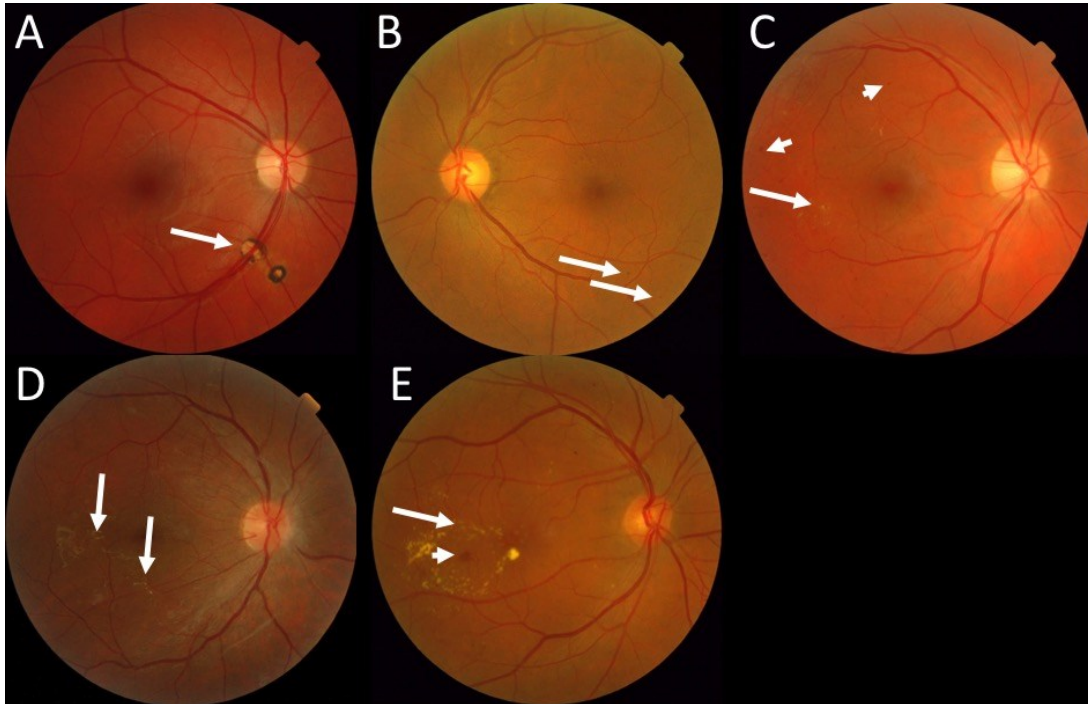


Figure 16. Representative retinal fundus images organized by progressive ease of grading correctly (A-E). A) The image reveals areas of chorioretinal atrophy (arrow), but is without lesions of diabetic retinopathy. B) This image reveals very subtle microaneurysms (arrows). C) This image reveals more obvious microaneurysms (arrowheads) and subtle hard exudates (arrow). D) This image reveals more apparent hard exudates (arrow). E) This image reveals obvious hard exudates (arrow) and more obvious hemorrhagic microaneurysms (arrowhead).

Discussion

In this study we have shown that crowdsourcing workers on a popular crowdsourcing platform, Amazon Mechanical Turk, are able to rapidly and accurately identify mild to moderate diabetic retinopathy in a large public dataset. While dividing the retinal

images into horizontal strips and using time spent grading images or prior experience does not seem to influence the accuracy of consensus grading, weighting Turker responses by their demonstrated ability does seem to improve accuracy.

There are many ways of defining a crowdsourcing consensus, or “divining the wisdom of the crowd.” For binary tasks, or categories that can be rationally dichotomized (as was done in this study by reducing 4 levels of disease to disease or no disease), one could take a simple majority vote (MV) approach such that the image receives the categorization rendered by half or more of respondents. To reach a consensus with categorical data, using the modal response may reduce the influence of outlier, or inattentive/malicious users. Both of these methods involve a post-hoc analysis of the data. Alternatively, one could allow consensus to be determined “on-the-fly,” such that if enough workers render the same or similar judgment of an image, the image is immediately coded with this classification, such that the full 10 responses need not be completed. In this study, we sought to compare the simple MV approach with several logistic prediction models.

Because data on the time spent completing the task and prior exposure to similar tasks is collected in addition to the grade for the current image task is collected when a crowdsourcing worker completes a task, in Phase 2 we sought to leverage this data to improve our determination of worker accuracy. Moreover, dividing each image into 4

segments allowed us to collect a “raw score” for each image on a semi-continuous scale of 0-40, as each worker renders a binary scale (0=normal, 1=abnormal), and each strip is graded by 10 unique workers. We hypothesized expanding our scale by taking additional image metrics into account would allow greater precision and flexibility to determine an ideal cutpoint for crowdsourcing as a diagnostic tool.

Interestingly, incorporating image metric such as time spent and prior worker experience did not substantially improve our ability to predict and abnormal retinal image. None of the automated model selection algorithms applied (either based on model likelihood ratio or AIC) included these variables in the prediction model. The logistic regression model did have better diagnostic characteristics over the range of different cut-off values for probability of image abnormality. However, in order to operationalize the model as a diagnostic (binary) tool, a single cut-off would need to be chosen, and as demonstrated in Table 2, no cut-off is objectively better than the naïve analysis. What is provided, however, is an opportunity to tailor the diagnostic characteristics to the particular clinical/research task at hand. In other words, an investigator could calibrate the model by choosing a cut-off to emphasize either sensitivity or specificity. We suspect that in future iterations of our crowdsourcing interface which allow for capturing more of the user interaction we may be able to improve upon our prediction model. Likewise, explorations of continuous outcomes

such as optic nerve cup-to-disc ratio may be more amenable to improved regression prediction versus a standard mean of crowdsourced values.

In phase 3 of this project, we sought to determine whether knowledge of an individual Turker's ability on a training set of images could be used to improve accuracy of the consensus grade in a separate test set of images. For this phase, we chose to use the Rasch model with "image difficulty" as the latent trait. In this way, we were able to determine the odds of each image being correctly classified by a Turker of average ability, and the odds of each Turker being able to correctly grade an image of average difficulty. This allowed us to weight a Turker's response to the images in the test set for use in a logistic regression model. This also allowed for a qualitative assessment of the retinal images from a unique perspective; ranked from difficulty to grade correctly rather than ranked by disease severity.

The use of weighting Turkers' responses in phase 3 showed a small, but significant improvement in the area under the receiver operator characteristic as compared with unweighted aggregation. This result was very encouraging and suggests several possible improvements that can be made to our crowdsourcing method. For example, if a returning Turker has previously had their "ability" calculated, this can be immediately applied to their new categorizations. If a new Turker begins a retinal grading task, they can be asked to perform a brief quiz to determine their "ability" prior to officially

grading images. This method may allow for a reduction in the number of annotations per image required to generate a stable estimate for each image.

A reasonable question is whether Rasch analysis was truly necessary. For an exploratory trial, the odds of correctly grading *all* images in the training set was compared to the Rasch derived measure, and these were quite similar (data not shown). The advantage of the Rasch measure is that 1) it takes into account the specific images graded and 2) it is amenable to being determined with a small number of images. Conceptually, we have thought of the 1200 image Messidor set as a large “item bank.” When a Turker grades several images, they are taking a test comprised of “questions” of known difficulty, which can be used to generate a weighted “score” that incorporates the item difficulty.

Rather than use whichever images are selected randomly for a given Turker’s “test,” we can apply computerized adaptive testing (CAT) methods to efficiently ascertain the Turker’s ability. Indeed, an immediate next phase of the present research is to validate image difficulty by creating a smaller item bank for use in CAT. Using the image difficulties calculated in the prior trial, images will be selected from the 100 images examined qualitatively in phase 3. We can require that a group of Turkers complete gradings on all the images in this bank to ensure no missing data. After ensuring adherence to the Rasch model, image difficulty measures will be re-calculated. Using a stepping algorithm, a short (Rasch-motivated item reduction can help determine

how short), CAT quiz can be created within our improved “Eyedentification” interface on AMT. Weighted consensus can then be compared using the phase 3 test set using naïve Turkers. Follow-up experiments will then seek to determine whether 10 unique categorizations are necessary for a stable prediction of disease status. Beyond the *a priori* desire for maximum efficiency, Amazon applies an additional 20% commission to tasks which use 10 or more unique graders per task, so the ability to use 9 workers or fewer would substantially reduce the cost of these experiments.

It is worth noting that the application of the Rasch model may have applications to the analysis of retinal imaging data beyond crowdsourcing. For example, in the analysis of images of premature babies at risk of retinopathy of prematurity, it has been demonstrated that there is very little agreement among experts on what constitutes so-called “Plus disease.”³⁵ A dichotomized (sometimes categorized with an intermediate “pre-plus”) grading of the level of vascular tortuosity of the posterior pole of the retina, Plus disease is the primary driver of whether a baby requires treatment for this blinding disease. Because of the lack of agreement among experts, it has been challenging to standardize grading and to calibrate automated grading systems. Treating “Plus-ness” as a latent trait could allow for the calculation of an interval scale measure of “Plus” as well as allow for calibration of experts’ gradings (correction of systematic bias).

Additionally, in a recent study exploring the use of deep learning artificial intelligence for retinal image interpretation published by researchers at Google³⁶ a stated limitation was their use of majority vote consensus grading of several ophthalmologists in both their 128,000+ image training set and 11,700+ image test sets. The authors acknowledged that much of the residual imprecision of the algorithm likely resides in “feeding” better gold standard data into the algorithm, creating an opening for similar methods as described here.

There are several limitations to crowdsourcing retinal image processing. Because users are anonymous, and cannot be directly selected by the researcher, there is no way to ensure high quality, highly conscientious workers each time work is posted. Indeed, the pool of workers can vary substantially over time and different trends in how workers engage with the site have become apparent to us over the course of the three years of this experiment. For example, we have recently noticed that many workers use automated “scripts” to accept/reserve large numbers of tasks at once, and then they can proceed at their own pace without concern for there being few tasks left for them. Indeed, this phenomenon may be the explanation for the “shark fin” shape with long tail in Figure 8. This “hoarding” has made metrics of time spent per image rather meaningless, but it is not clear that it has led to worse outcomes overall (data not shown). Regardless, researchers who wish to use crowdsourcing need to be aware of the “culture” of the crowdsourcing marketplace they choose.

Our current method used the supplied Messidor grade as the gold standard. While this is a high-quality, well-known data set, there were dramatic differences in how the images were graded compared to standard clinical and telemedicine grading schemes such as the one we used for training. Particularly, while we tried to eliminate clinically insignificant disease by defining the very mild disease category (Messidor 1) as “normal,” there was still the possibility of clinically very mild disease in the most severe Messidor category (e.g., 16 microaneurysms is Messidor 3, but, could be considered minimal retinopathy on most clinical grading scales).

There are several potential benefits to the use of crowdsourcing for the interpretation of visual data in ophthalmology. First, an inexpensive, rapid, and accurate system to reduce the number of images needing human grading in large public health screenings is needed. An approach which accurately identifies normal (or very mildly abnormal; allowing for some false negatives) fundi would be of great value and could reduce the skilled grader burden by up to 26-38% or more according to some investigators using artificial intelligence programs.¹⁹ A “first pass” to remove normal images is currently being done with an Artificial Intelligence (AI) solution in Scotland’s national screening program.³⁷ If appropriately validated, crowdsourcing could provide a similar service at lower cost, and with less infrastructure in resource-poor settings. Likewise, a means to

rapidly interrogate existing datasets with existing datasets could allow for nimble hypothesis generation for secondary data analyses.

Works Cited

1. IDF Diabetes Atlas, 7th edn. **International Diabetes Federation**, 2015. (Accessed 10/13/2014, at <http://www.diabetesatlas.org/>.)
2. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. US Department of Health and Human Services, 2014. (Accessed 10-13-2014, at <http://www.webcitation.org/6TIZswZP6> <http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf>.)
3. Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012;35:556-64.
4. Saaddine Jb HAANKZXKRBJP. Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United states, 2005-2050. *Archives of Ophthalmology* 2008;126:1740-7.
5. Beck RW. The burgeoning public health impact of diabetes: the role of the ophthalmologist. *Arch Ophthalmol* 2011;129:225-9.
6. Javitt JC, Canner JK, Frank RG, Steinwachs DM, Sommer A. Detecting and treating retinopathy in patients with type I diabetes mellitus. A health policy model. *Ophthalmology* 1990;97:483-94; discussion 94-5.
7. Javitt JC, Aiello LP. Cost-effectiveness of detecting and treating diabetic retinopathy. *Ann Intern Med* 1996;124:164-9.
8. Jones S, Edwards RT. Diabetic retinopathy screening: a systematic review of the economic evidence. *Diabet Med* 2010;27:249-56.
9. Silva PS, Cavallerano JD, Aiello LM, Aiello LP. Telemedicine and diabetic retinopathy: moving beyond retinal screening. *Arch Ophthalmol* 2011;129:236-42.
10. Bastawrous A, Hennig BD. The global inverse care law: a distorted map of blindness. *Br J Ophthalmol*. England2012;1357-8.
11. Mansberger SL, Shepler C, Barker G, et al. Long-term Comparative Effectiveness of Telemedicine in Providing Diabetic Retinopathy Screening Examinations: A Randomized Clinical Trial. *JAMA Ophthalmol* 2015;133:518-25.
12. Sim DA, Mitry D, Alexander P, et al. The Evolution of Teleophthalmology Programs in the United Kingdom: Beyond Diabetic Retinopathy Screening. *J Diabetes Sci Technol* 2016;10:308-17.
13. Kim J, Driver DD. Teleophthalmology for first nations clients at risk of diabetic retinopathy: a mixed methods evaluation. *JMIR Med Inform* 2015;3:e10.
14. Brady CJ, Villanti AC, Gupta OP, Graham MG, Sergott RC. Tele-ophthalmology Screening for Proliferative Diabetic Retinopathy in Urban Primary Care Offices: An Economic Analysis. *Ophthalmic surgery, lasers & imaging retina* 2014;45:556-61.
15. Au A, Gupta O. The economics of telemedicine for vitreoretinal diseases. *Curr Opin Ophthalmol* 2011;22:194-8.
16. Rein DB, Wittenborn JS, Zhang X, et al. The cost-effectiveness of three screening alternatives for people with diabetes with no or early diabetic retinopathy. *Health Serv Res* 2011;46:1534-61.
17. Abramoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol* 2013;131:351-7.

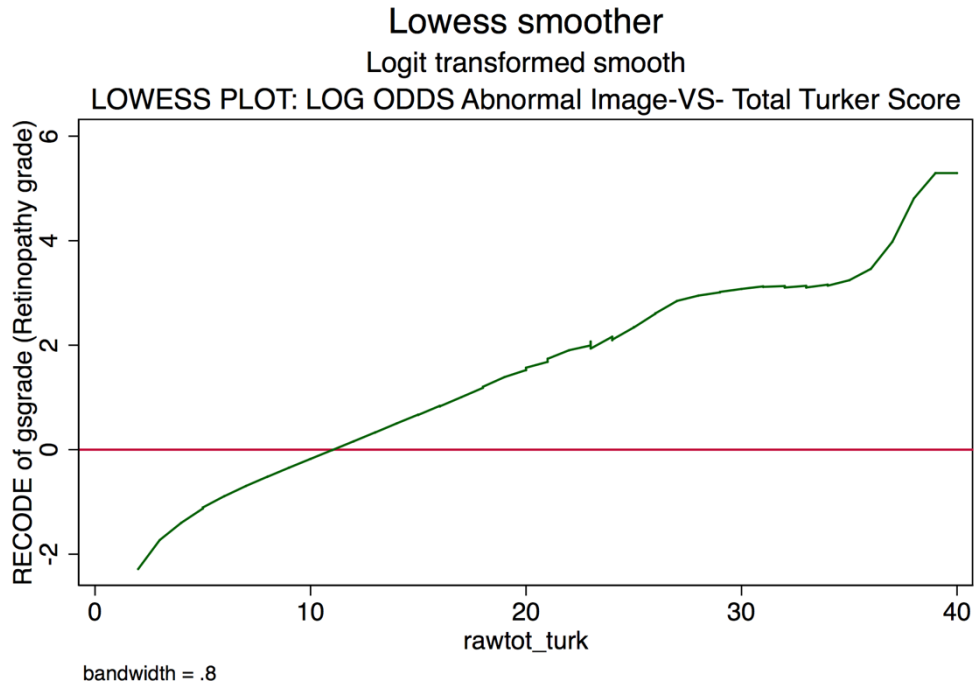
18. Trucco E, Ruggeri A, Karnowski T, et al. Validating retinal fundus image analysis algorithms: issues and a proposal. *Invest Ophthalmol Vis Sci* 2013;54:3546-59.
19. Goatman K, Charnley A, Webster L, Nussey S. Assessment of automated disease detection in diabetic retinopathy screening using two-field photography. *PloS one* 2011;6:e27524.
20. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. *Am J Prev Med* 2014;46:179-87.
21. Luengo-Oroz MA, Arranz A, Frea J. Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears. *Journal of medical Internet research* 2012;14:e167.
22. Mity D, Peto T, Hayat S, Morgan JE, Khaw KT, 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:e71154.
23. Brady CJ, Villanti AC, Pearson JL, Kirchner TR, Gupta OP, Shah CP. Rapid grading of fundus photographs for diabetic retinopathy using crowdsourcing. *Journal of medical Internet research* 2014;16:e233.
24. Brabham DC. *Crowdsourcing*. Cambridge, MA: MIT Press; 2013.
25. Amazon Mechanical Turk. at <http://www.mturk.com/>.
26. Whitehill J, Wu T-f, Bergsma J, Movellan JR, Ruvolo PL. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*; 2009. p. 2035-43.
27. Tang W, Lease M. Semi-supervised consensus labeling for crowdsourcing. *Special Interest Group on Information Retrieval 2011 Workshop on Crowdsourcing for Information Retrieval*, Beijing, China, July; 2011. p. 1-6.
28. Sheng VS, Provost F, Ipeirotis PG. Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*; 2008: ACM. p. 614-22.
29. Dalvi N, Dasgupta A, Kumar R, Rastogi V. Aggregating crowdsourced binary ratings. *Proceedings of the 22nd international conference on World Wide Web*; 2013: ACM. p. 285-94.
30. Shotliff K, Duncan G. Diabetic retinopathy: summary of grading and management criteria. *Pract Diab Int* 2006;23:418-20.
31. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Analysis and Stereology* 2014;33:231-4.
32. Venkatesh P, Sharma R, Vashist N, Vohra R, Garg S. Detection of retinal lesions in diabetic retinopathy: comparative evaluation of 7-field digital color photography versus red-free photography. *Int Ophthalmol* 2015;35:635-40.
33. Winsteps Tutorial 1. Winsteps.com, 2012. (Accessed 10/19/2016, at <http://www.winsteps.com/a/winsteps-tutorial-1.pdf>.)
34. Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the*

conference on empirical methods in natural language processing; 2008: Association for Computational Linguistics. p. 254-63.

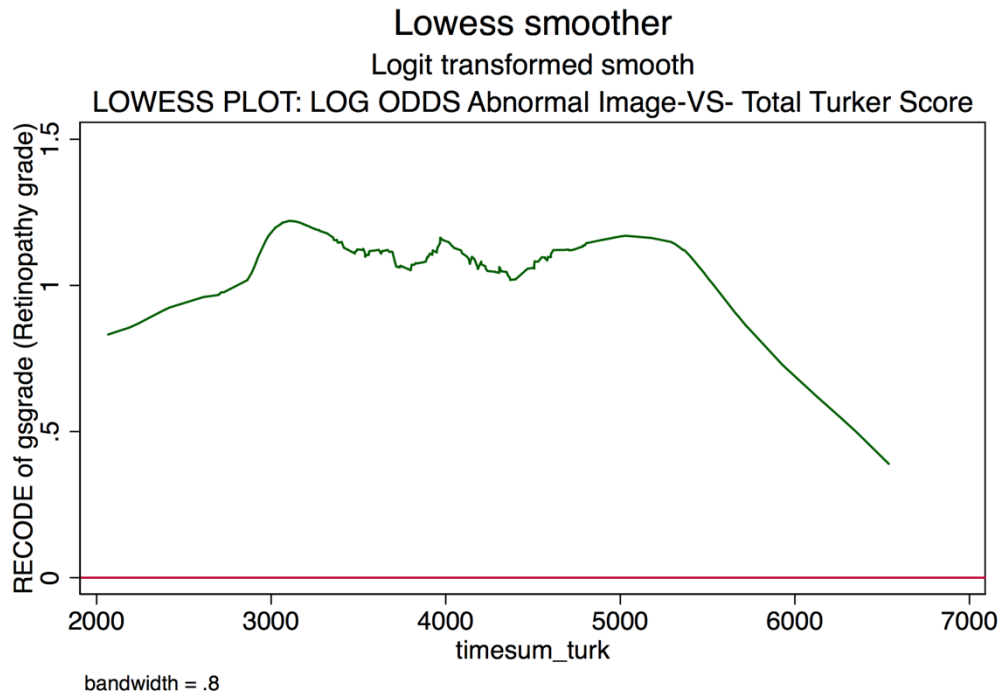
35. Hewing NJ, Kaufman DR, Chan RV, Chiang MF. Plus Disease in Retinopathy of Prematurity: Qualitative Analysis of Diagnostic Process by Experts. *JAMA Ophthalmol* 2013;1-7.

36. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama* 2016.

37. Medalytix Retinal Screening. (Accessed 12-3-2014, 2014, at [http://www.medalytix.com/.](http://www.medalytix.com/))

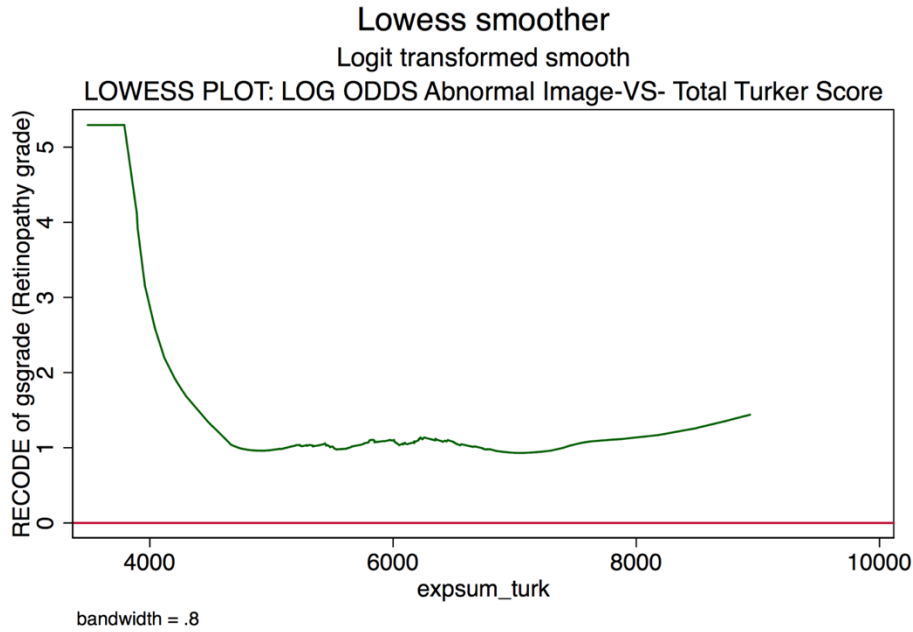


Supplemental figure 1. Grossly normal relationship between turker raw score and log odds of abnormality supports linear term for turker raw score in prediction



model.

Supplemental figure 2. Evidence of non-linearity in the relationship between total time spent on the task and logodds of abnormality supports categorizing this variable.



Supplemental figure 3. Evidence of non-linearity in the relationship between total experience score for all turkers grading the image and logodds of abnormality supports dichotomizing this variable.

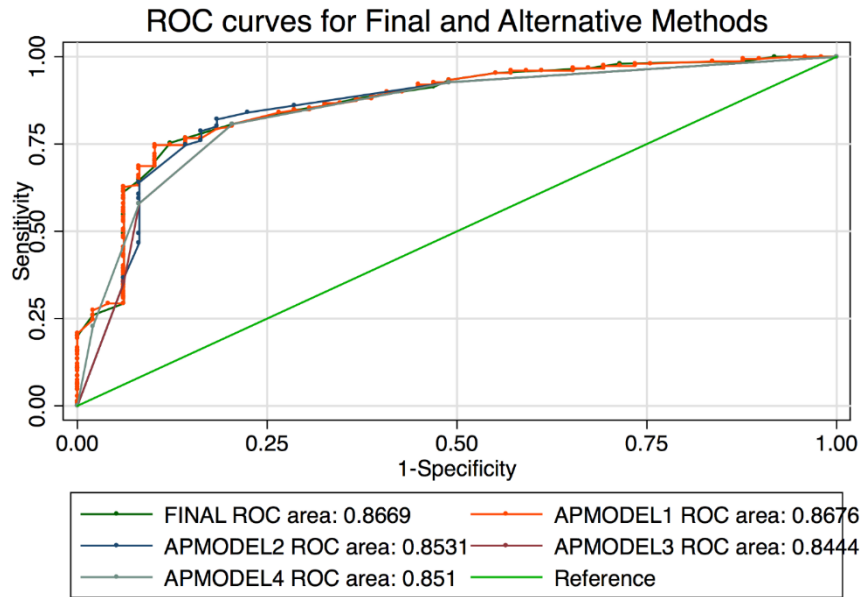
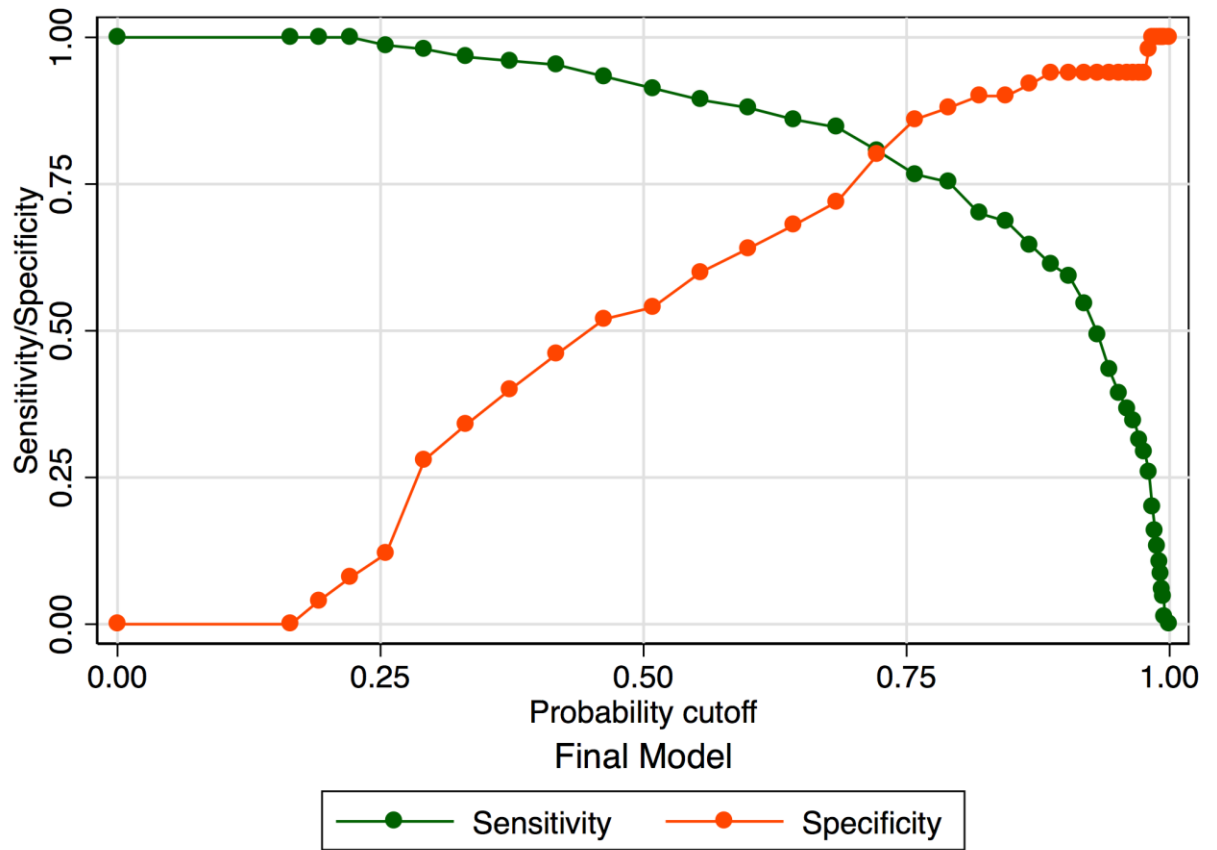


figure 4. Receiver operator characteristic for each model tested.

Supplemental



Supplemental figure 5. Sensitivity and specificity for various cut-off values for probability of abnormality using the final model.

Curriculum Vitae

Christopher J. Brady, was born February 9, 1979 in Worcester, Massachusetts. He is currently an assistant professor of ophthalmology at the Wilmer Eye Institute at Johns Hopkins University School of Medicine.

Dr. Brady received his bachelor degree from Columbia University, his medical degree from Johns Hopkins University School of Medicine and completed an internship in internal medicine at Johns Hopkins Bayview Medical Center. He completed his ophthalmology residency at Wills Eye Institute in Philadelphia, where he also completed a vitreoretinal fellowship. He specializes in the medical and surgical management of vitreoretinal diseases, including macular degeneration and diabetic retinopathy. His research focuses on the benefits of telemedicine for underserved populations and how new technologies can improve patient care.