AMERICAN ACADEMY OF OPHTHALMOLOGY®

Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy

Rory Sayres, PhD,[1] Ankur Taly, PhD,[1] Ehsan Rahimy, MD,[2] Katy Blumer, BS,[1] David Coz, BS,[1] Naama Hammel, MD,[1] Jonathan Krause, PhD,[1] Arunachalam Narayanaswamy, PhD,[1] Zahra Rastegar, MD, PhD,[1] Derek Wu, BS,[1] Shawn Xu, BS,[3] Scott Barb, MD,[4] Anthony Joseph, MD,[5] Michael Shumski, MD, MSE,[6] Jesse Smith, MD,[7,8] Arjun B. Sood, MD,[9] Greg S. Corrado, PhD,[1] Lily Peng, MD, PhD,[1,*] Dale R. Webster, PhD[1,*]

Purpose: To understand the impact of deep learning diabetic retinopathy (DR) algorithms on physician readers in computer-assisted settings.

Design: Evaluation of diagnostic technology.

Participants: One thousand seven hundred ninety-six retinal fundus images from 1612 diabetic patients.

Methods: Ten ophthalmologists (5 general ophthalmologists, 4 retina specialists, 1 retina fellow) read images for DR severity based on the International Clinical Diabetic Retinopathy disease severity scale in each of 3 conditions: unassisted, grades only, or grades plus heatmap. Grades-only assistance comprised a histogram of DR predictions (grades) from a trained deep-learning model. For grades plus heatmap, we additionally showed explanatory heatmaps.

Main Outcome Measures: For each experiment arm, we computed sensitivity and specificity of each reader and the algorithm for different levels of DR severity against an adjudicated reference standard. We also measured accuracy (exact 5-class level agreement and Cohen's quadratically weighted κ), reader-reported confidence (5-point Likert scale), and grading time.

Results: Readers graded more accurately with model assistance than without for the grades-only condition ($P < 0.001$). Grades plus heatmaps improved accuracy for patients with DR ($P < 0.001$), but reduced accuracy for patients without DR ($P = 0.006$). Both forms of assistance increased readers' sensitivity moderate-or-worse DR: unassisted: mean, 79.4% [95% confidence interval (CI), 72.3%−86.5%]; grades only: mean, 87.5% [95% CI, 85.1%−89.9%]; grades plus heatmap: mean, 88.7% [95% CI, 84.9%−92.5%] without a corresponding drop in specificity (unassisted: mean, 96.6% [95% CI, 95.9%−97.4%]; grades only: mean, 96.1% [95% CI, 95.5%−96.7%]; grades plus heatmap: mean, 95.5% [95% CI, 94.8%−96.1%]). Algorithmic assistance increased the accuracy of retina specialists above that of the unassisted reader or model alone; and increased grading confidence and grading time across all readers. For most cases, grades plus heatmap was only as effective as grades only. Over the course of the experiment, grading time decreased across all conditions, although most sharply for grades plus heatmap.

Conclusions: Deep learning algorithms can improve the accuracy of, and confidence in, DR diagnosis in an assisted read setting. They also may increase grading time, although these effects may be ameliorated with experience. *Ophthalmology 2019;126:552-564 © 2018 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.aaojournal.org.

Diabetic retinopathy (DR), a common complication of diabetes, is a leading cause of preventable blindness worldwide.[1,2] Estimates indicate prevalence of DR ranging from 18% to 30% of type 2 diabetic patients and that of proliferative DR (PDR) ranging from 2.9% to 4.4% of such patients.[3] Most guidelines recommend at least annual DR screening for diabetic patients via retinal fundus photography or in-person dilated eye examinations.[4–6]

The literature indicates sizeable room for improvement in the accuracy of DR grading.[7–9] Sussman et al[8] reported measurements of accuracy across 23 physicians of different backgrounds using indirect ophthalmoscopy.

https://doi.org/10.1016/j.ophtha.2018.11.016
ISSN 0161-6420/18

They report an overall error rate of 61%, with a strong effect of physician background. Harding et al[9] measured sensitivity of 89% and specificity of 86% for experienced ophthalmologists on fundus photographs. By contrast, Lin et al[10] measured a sensitivity of 78% and specificity of 86% for digital fundus photography compared with color photography and found that mydriatic ophthalmoscopy had a lower sensitivity than either photography-based method. Together, these results indicate that existing screening methods can miss a substantial fraction of DR cases, leading to otherwise preventable vision loss.

Recent advances in machine learning promise to aid access to DR screening substantially and to improve diagnosis accuracy.[11–13] Systems developed using these methods (herein referred to as *algorithms*) have demonstrated specialist-level accuracy in diagnosis of DR severity. However, the impact of these models in clinical settings is not well understood. Previous attempts to use machine learning algorithms in a computer-assisted diagnosis setting have faced numerous challenges, including both over-reliance (repeating errors made by the model) and under-reliance (ignoring accurate algorithm predictions).[14–17] Some of these pitfalls may be avoided if the computer-assisted diagnosis system can explain its predictions.[16] A number of recently developed explanation methods focus on generating human-interpretable explanations of how predictions are made[18,19] and may be a powerful tool to aid physicians interpreting medical images, including retina images.[20,21]

In this study, we examined how a machine learning model for predicting DR severity improved doctor grading performance. We measured 3 primary outcomes: diagnosis accuracy compared with an adjudicated reference standard, subjective confidence in DR grading, and time spent grading. We examined these outcomes in both an unassisted context and with 2 levels of assistance: surfacing algorithm-predicted scores for each DR severity level (grades only) and showing both scores plus an explanation heatmap highlighting image regions that most contribute to the prediction (grades plus heatmap).

Methods

Grading Scale

We used the International Clinical Diabetic Retinopathy Disease Severity Scale in this study. The International Clinical Diabetic Retinopathy is a 5-point scale: no apparent retinopathy, mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, and PDR.[22] Referable DR is defined as moderate or worse DR[11,13,23] because management often changes from yearly screening to closer follow-up at moderate disease severity.[4]

Algorithm Development

The development of the algorithm used in this study is described in detail by Krause et al.[12] Briefly, the algorithm was trained using the Inception version 4 model architecture and a large dataset of more than 1.6 million retinal fundus images, then tuned (hyperparameters and checkpoint selection) on a set of 2000 images that had adjudicated labels agreed on by 3 retina specialists (one of whom was a coauthor, E.R.) as a reference standard. Training used the Tensorflow (https://www.tensorflow.org), framework. Training images were obtained from EyePACS clinics and 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya) among patients undergoing DR screening. The model used a 779×779 image size as input and was cotrained to make separate 5-class DR severity ratings, based on both partner-provided DR training labels and labels by an internal labeling process, as available. Hyperparameters used for the training process were tuned in an automated fashion to optimize the prediction performance on the tuning set.[24]

We also trained a model to predict gradeability for DR to assess the impact of this filtering and to exclude images that might be flagged as ungradable from analysis. This algorithm was trained using the same architecture as the DR model on a set of 140 000 images labeled by readers as being gradable or ungradable for DR.

Study Images and Reference Standard

For this study, we used an image set collected and graded using methods described in detail in Krause et al.[12] The image set consisted of 2000 total 45° fundus images from 1774 patients, randomly sampled from patients undergoing DR screening in EyePACS clinics between May and October 2015. Images were de-identified according to Health Insurance Portability and Accountability Act Safe Harbor before transfer to study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review IRB.

The reference standard was generated by 3 fellowship-trained retina specialists who first graded images independently and then participated in multiple rounds of adjudication until full consensus was reached. None of the adjudicators are included as readers in this study. Of the 2000 total images in our image set, adjudicated consensus was reached for 1804 images. The other 196 images were deemed to have insufficient image quality for adjudication and were excluded from the study. This criterion was fairly strict and was set by requiring at least 2 of 3 adjudicators determining that the image had sufficient quality to be gradable for both DR and diabetic macular edema. We further filtered 2 images that were found to be duplicates of one another, 1 image for which our automated prediction algorithm was unable to generate a prediction because of low image quality, and 6 images that were identified as ungradable by an algorithm trained to predict gradeability for DR. This left a remaining 1796 images in our evaluation set from 1612 unique patients. Demographic information about the patients included in this set is given in Table S1 (available at www.aaojournal.org).

The image set used was a tuning set, used in conjunction with other image sets to select hyperparameters during training. We chose this set because it had a significantly higher proportion of cases with DR than a held-out test set that also had been adjudicated (both described in Krause et al[12]). Because this was a tuning set, there was a possibility that the algorithm would be overfit to this set and would perform better than it might for a held-out set. To help control against this possibility, we analyzed algorithm performance between this set and the held-out test set and found that performance was comparable (Table S2, available at www.aaojournal.org).

Generating Highlights

We applied the trained deep learning algorithm on each of the 1804 fundus images being tested. This generated, for each fundus image, a set of 5 scores representing the relative strength of evidence for each DR class, constrained to sum to 1.

We then generated explanation heatmaps for a predicted severity level using the integrated gradients method, described in Sundararajan et al.[25] This method provides pixel-based maps that measure the contribution of each pixel in a fundus image to a predicted DR severity level. The contributions are measured relative to a baseline image, which is intended to provide no information into the model. For this study, we used a black image as the baseline. We verified that our trained algorithm predicted no DR for this baseline.

For each retina image, we generated a path of 50 steps, in which each step was interpolated between the blank baseline image and the target fundus image. For each DR level, we summed model gradients over each pixel and took the absolute value. We then surfaced the heatmap for the DR level with the highest score.

We visualized the highlights over the fundus image by converting the fundus image to grayscale and overlaying the explanation heatmap as a semitransparent green heat map (Fig 1). Readers could toggle on and off the explanation heatmaps to see the underlying retinal anatomic features more clearly.

In cases where the predicted DR level was none (i.e., no apparent DR), the resulting heatmap may not provide meaningful information because this is the same prediction as for the baseline image. We included heatmaps for these images in the grades plus heatmap condition to verify our hypothesis that these heatmaps would be substantially less useful than heatmaps for positive DR levels and to ensure that each experimental arm used the same full set of images.

Reader Study

A total of 10 ophthalmologist readers participated in this study, consisting of 4 United States board-certified general ophthalmologists (3 of whom are coauthors: A.B.S., N.H., and M.S.), 1 ophthalmologist trained outside the United States (United Kingdom and Germany board certified: Z.R.), 4 fellowship-trained retina specialists (2 coauthors: A.J. and J.S.), and 1 retina fellow in the middle of fellowship training (S.B.).

Each reader reviewed each image exactly once, in 1 of 3 conditions: unassisted, grades only, and grades plus heatmap. The assignment of image to experimental condition was counterbalanced across readers, so that each image had approximately the same number of readers for each condition. Because each image was read 10 times (once per reader) across 3 experiment arms, it was not possible for a single image's readings to be distributed evenly across arms. Each image was read by 3 readers for 2 conditions and by 4 readers for a third condition; the condition with the extra read varied randomly across images. As a result, the total number of reads for each condition varied slightly, between 5983 reads for unassisted and 5992 reads for grades plus heatmap. A screenshot of each condition is shown in Figure 1.

The algorithm grades were displayed for both grades only and grades plus heatmap conditions, in the form of histograms showing model confidence for each of the 5 DR severity levels. The grades plus heatmap arm included an overlay of explanatory highlight that could be toggled on and off in addition to the grade histogram.

Readers were briefed in the use of the tool before the trial began and given identical instructions. They were asked to grade DR as well as they could from the image. They were told that grade histograms provided a measure of evidence for each level of DR, as interpreted by the model, and that the heatmaps indicated regions providing evidence for the DR severity with the highest score. They were informed that accuracy and self-reported confidence were outcome measures for the study but were not told that time spent on task was an outcome measure.

Statistical Analysis

Statistical analyses reported herein were conducted using Python version 2.7.6 (http://www.python.org), with the numpy, scipy, statsmodels, and scikit-learn packages. The experiment described was performed exactly once, with no prior or later attempts at replication. All P values reported herein are uncorrected for multiple comparisons.

We conducted a logistic regression analysis on the data to verify the magnitude and statistical significance of our effects. We applied a mixed-effects logistic regression model using the statsmodels package in Python (https://pypi.org/project/statsmodels/). The outcome variable was the binary observation of a correct DR grade relative to the adjudicated reference standard. We modeled the experiment arm as a treatment variable using fixed effects for each of the grades-only and grades plus heatmap arms against a baseline estimate of accuracy for unassisted reads. We modeled the influence of reader identity and retina image as fixed effects.

Results

A comparison of the DR disease distribution determined by the retina specialist panel (reference standard) and the algorithm's grading are summarized in Table 1. Approximately 23% of the images showed mild or worse NPDR according to the reference standard. Overall, the model showed a 5-class accuracy of 88.4% (95% binomial confidence interval [CI], 87.9%−88.9%), with an accuracy of 96.9% (95% CI, 96.3%−97.3%) for images with no DR and 57.9% (95% CI, 55.3%−60.6%) for images with mild or worse NPDR.

Algorithmic Assistance Increases Sensitivity for Diabetic Retinopathy without Reducing Specificity

Our primary research question centered around whether clinicians, assisted by a deep-learning algorithm, can assess disease more accurately than either a clinician alone or the algorithm alone. We examined accuracy using a range of metrics. Our primary outcome measures were sensitivity and specificity for detecting different levels of DR and 5-class accuracy (the fraction of reads in which the 5-class International Clinical Diabetic Retinopathy grade exactly matched the reference standard).

We first examined sensitivity and specificity, which has been well characterized in the literature.[7–9] In our study sample, the algorithm showed a sensitivity of 91.5% and a specificity of 94.7% for moderate or worse NPDR. Compared with the algorithm, unassisted readers tended toward lower sensitivity, but higher specificity. Mean reader sensitivity was 79.4% (95% CI, 72.3%−86.5%) and mean reader specificity was 96.6% (95% CI, 95.9%−97.4%). These values are consistent with values reported in the literature.[7–9]

How does algorithmic assistance affect reader sensitivity and specificity? We found assistance increased sensitivity, without a corresponding decrease in specificity (Fig 2; Table 2). Figure 2 illustrates changes in sensitivity and specificity for referable DR classifications (moderate or worse NPDR), with and without each type of assistance. The sensitivity of assisted reads consistently exceeded that of unassisted reads for both types of assistance, whereas specificity remained largely unchanged (Fig 3B). We saw the same pattern of results when we examined other DR thresholds (Fig S1, available at www.aaojournal.org).

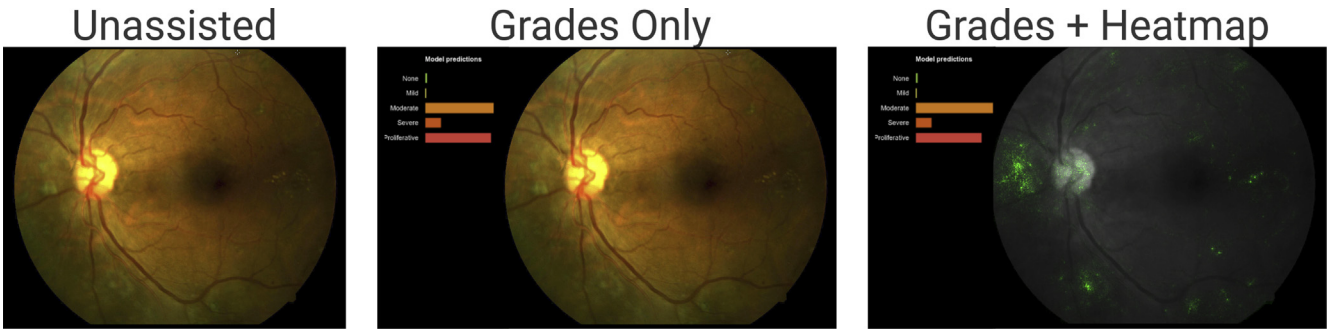

Figure 1. Illustration of experimental conditions: (**left**) unassisted, (**center**) grades only, (**right**) grades plus heatmap. Graders could toggle assistance on or off. For both grades-only and grades plus heatmap conditions, the view was the same as the unassisted view when assistance was toggled off.

Accuracy Improvements are Driven by Patients with Diabetic Retinopathy

We next examined 5-class DR accuracy. We assessed the overall impact of model assistance on this metric using a logistic regression analysis (Table 3). Across all reads, readers in the grades-only condition were significantly more accurate than unassisted readers ($P < 0.001$). Readers in the grades plus heatmap condition were not more accurate overall compared with unassisted readers ($P = 0.13$).

This was not unexpected, because we knew that heatmaps produced for cases predicted to have no DR would not necessarily provide meaningful information (see details in "Methods"). We hypothesized that these heatmaps may hurt reader performance by causing readers to second-guess themselves and overdiagnose patients with no DR.

To better assess the impact of assistance, we repeated our analyses separately for patients with no DR and patients with some level of DR (Fig 2; Table 2). We found that accuracy increases resulting from algorithmic assistance were driven by patients with DR. Accuracy for patients with no DR was high across conditions (range 92.5%–94.7% across readers and conditions) and did not increase significantly for grades-only assistance compared with unassisted ($P = 0.08$, 2-sided 1-sample t test on accuracy difference across images). Moreover, accuracy for patients with no DR decreased significantly for grades plus heatmap ($P = 0.007$), consistent with our hypothesis that heatmaps for patients with no DR can lead to overdiagnosis. Confusion matrices indicate that readers in the grades plus heatmap condition classified patients with no DR as having mild NPDR more often (Table S3, available at www.aaojournal.org), explaining why we did not see an impact on specificity for moderate or worse DR.

Benefit of Assistance Varies with Clinical Background

Although both types of model assistance seemed to improve reader performance, the increase may vary according to the clinical background of the reader (Fig 4). Without assistance, readers' 5-class accuracies for images with DR were 46.3% and 62.3% for general ophthalmologist and retina specialist, respectively. General ophthalmologists were significantly less accurate than the algorithm (57.9% accuracy for algorithm; $t = -5.53$; $P < 0.001$, 2-sided paired t test), whereas retina specialists were not significantly more accurate than the algorithm ($t = 1.85$; $P = 0.06$).

With assistance from the model, general ophthalmologists matched, but did not exceed, the model's accuracy (grades only: $t = -0.55$; $P = 0.58$; grades plus heatmap: $t = -0.84$; $P = 0.40$). With assistance, retina specialists significantly exceed the model's accuracy (grades only: $t = 4.34$; $P < 0.001$; grades plus heatmap: $t = 4.38$; $P < 0.001$).

Example Cases Affected by Grades Histograms Alone

To characterize further instances in which model assistance was the most helpful, we examined several representative cases in which revealing model grades alone to readers helped with grading (Fig 5). Assistance with model grades seemed to help readers in cases where there may be subtle lesions such as microaneurysms for mild cases, hemorrhages for moderate cases, or neovascularization for proliferative cases.

Figure 5A illustrates a case in which grades-only assistance prevented underdiagnosis of vision-threatening DR. This example was determined by adjudication to be proliferative because of

Table 1. Confusion Matrix Comparing the Adjudicated Reference Standard Grading and the Algorithm's Grading According to the International Clinical Diabetic Retinopathy Disease Severity Scale

Adjudicated Reference Standard Grading	No Apparent Diabetic Retinopathy	Mild Nonproliferative Diabetic Retinopathy	Moderate Nonproliferative Diabetic Retinopathy	Severe Nonproliferative Diabetic Retinopathy	Proliferative Diabetic Retinopathy	Total
No apparent DR	1355	23	14	1	3	1396
Mild NPDR	62	59	65	0	1	187
Moderate NPDR	12	5	130	2	1	150
Severe NPDR	1	0	13	16	1	31
PDR	0	0	4	1	27	32
Total	1430	87	226	20	33	1796

DR = diabetic retinopathy; NPDR = nonproliferative diabetic retinopathy; PDR = proliferative diabetic retinopathy.
The columns represent the reference standard grades. The rows represent the algorithm's grades.

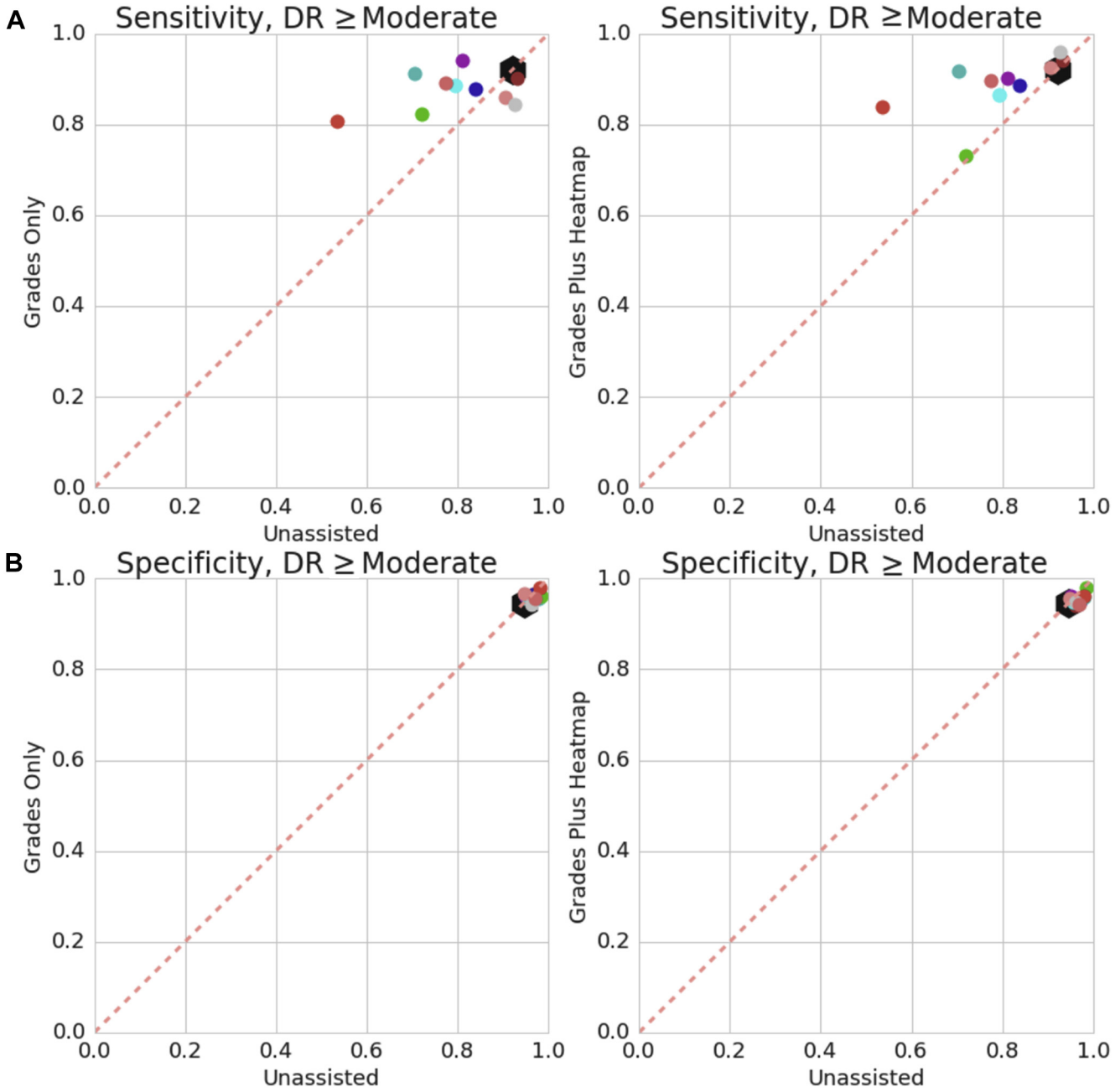

Figure 2. Scatterplots of reader (**A**) sensitivity and (**B**) specificity for each of the assisted conditions (y-axis) compared with the unassisted condition (x-axis) showing that model assistance increases reader sensitivity. Colored circles represent individual readers; large black hexagon represents the model. Each of the 10 data points (readers plus model) represents 1804 observations. DR = diabetic retinopathy.

neovascularization on the optic disc. Without assistance, 2 of 3 unassisted readers missed the neovascularization. (Note that the labeling tool supported zoom and contrast adjustments to ensure that the feature was not missed.) With assistance, all readers correctly assessed the case as PDR. In discussions after the study, each reader in the grades-only condition indicated noting the neovascularization. Overall, our data set contained 63 cases of vision-threatening (severe or proliferative) DR. Of those, we found 7 cases (11%) in which vision-threatening DR was missed by 1 or more unassisted readers but was identified by all assisted readers with grades only. This number increased to 15 cases (23%) when also considering assistance with grades plus heatmap.

We also saw situations in which assistance prevented underdiagnosis in NPDR cases. Figure 5B illustrates a case of severe NPDR in which unassisted readers consistently graded the image as moderate NPDR, whereas readers in the grades-only condition correctly identified it as severe NPDR. Circled regions on the image highlight subtle pathologic features that may be Intraretinal Microvascular Abnormalities, the distribution of prediction scores for both moderate and severe DR may result in more careful inspection of the image, increasing the chances of identifying pathologic features of this sort. We found 86 cases like this, in which grades-only readers were more accurate at detecting moderate NPDR or worse (40% of 213 total such cases in our data set).

Table 2. Sensitivity and Specificity for Moderate or Worse Diabetic Retinopathy for Each Reader in Each Experimental Condition

	Sensitivity (%)			Specificity (%)		
Labeler	*Unassisted*	*Grades Only*	*Grades Plus Heatmap*	*Unassisted*	*Grades Only*	*Grades Plus Heatmap*
GO1	79.41	88.73	86.49	95.64	95.25	94.72
GO2	90.54	85.92	92.65	94.72	96.58	95.64
GO3	53.52	80.88	83.78	98.10	98.10	96.04
GO4	81.08	94.12	90.14	95.09	96.20	96.20
GO5	70.42	91.18	91.89	97.91	95.64	95.85
RF	83.82	87.84	88.73	97.15	96.60	95.06
RS1	93.24	90.14	94.12	96.23	96.39	94.12
RS2	72.06	82.43	73.24	98.48	96.23	97.91
RS3	92.65	84.51	95.95	96.20	94.30	94.72
RS4	77.46	89.19	89.71	96.96	95.66	94.31
Model	92.08	91.55	91.02	94.60	94.82	94.66

Each row represents 1 labeler (physician reader or the model). G01 through G05 indicate the 5 general ophthalmologists, RF indicates the reader in the middle of a retina fellowship, and RS1 through RS4 indicate the 4 retina specialists. The algorithm had an overall sensitivity of 91.55% and specificity of 94.69%.

Cases in Which the Model Was Incorrect

We saw evidence in our data set of assisted doctors outperforming the model by identifying when the algorithm ultimately was incorrect but showing uncertainty in the call and using their own judgment to supersede the algorithm. Figure 5C, D shows examples of assisted doctors potentially using uncertainty conveyed by the model to overrule an inaccurate prediction. Figure 5C illustrates a case in which the reference standard grade was mild NPDR because of the presence of a microaneurysm inferior-temporal to the macula. The algorithmic prediction gave the highest score in this case to no DR (incorrect) with some score for mild NPDR. All unassisted readers

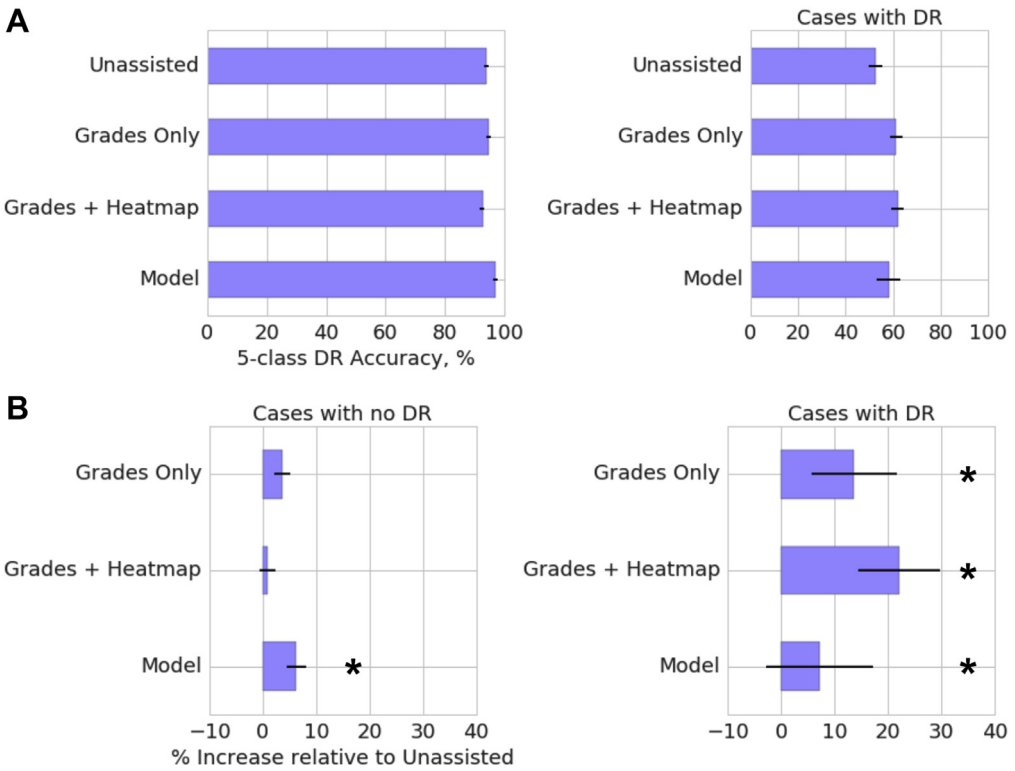

Figure 3. Graphs showing a lift in diabetic retinopathy (DR) grading accuracy with model assistance, driven by cases with DR. **A**, Overall accuracy for each experiment arm and raw model predictions. **B**, Improvement in accuracy versus unassisted reads. The y-axis represents the percent increase in DR grading accuracy relative to unassisted grading. Data are broken down by (left) cases with no DR and (right) cases with some level of DR. Error bars reflect 95% binomial confidence intervals across cases. Asterisks in (**B**) indicate significant accuracy increase relative to the unassisted condition (1-sample *t* test on difference in accuracy across cases: $P < 0.001$ for all significant bars; other bars are not significant, with $P > 0.01$). Each bar represents 16 236 observations across 9 readers.

Table 3. Summary of Logistic Regression Analyses of Diabetic Retinopathy Grade Accuracy Data

	Logit Coefficient Estimate	Odds Ratio	95% Confidence Interval	Standard Error	Z Score	P Value > |Z Score|
All cases						
Intercept	1.775	5.897	5.225–6.657	0.062	28.717	<0.001
Experiment arm	−0.376	0.686	0.392–1.202	0.286	−1.316	0.19
Grades only vs. unassisted	0.210	1.234	1.113–1.369	0.053	3.974	<0.001
Grades + heatmap vs. unassisted	0.078	1.081	0.977–1.196	0.052	1.512	0.13
Reader	0.001	1.001	0.986–1.015	0.007	0.078	0.94
Case	0.000	1.000	1.000–1.000	0.000	−1.623	0.10
Cases with no DR						
Intercept	3.145	23.215	18.796–28.674	0.108	29.187	<0.001
Experiment arm	−1.010	0.364	0.180–0.739	0.361	−2.800	0.005
Grades only vs. unassisted	0.146	1.157	0.969–1.380	0.090	1.616	0.11
Grades + heatmap vs. unassisted	−0.229	0.795	0.675–0.936	0.083	−2.748	0.006
Reader	−0.057	0.945	0.922–0.968	0.012	−4.605	<0.001
Case	0.000	1.000	1.000–1.000	0.000	−2.097	0.04
Cases with DR						
Intercept	−0.144	0.866	0.723–1.038	0.093	−1.552	0.12
Experiment arm	0.705	2.024	0.788–5.200	0.482	1.464	0.14
Grades only vs. unassisted	0.353	1.423	1.219–1.660	0.079	4.474	<0.001
Grades + heatmap vs. unassisted	0.388	1.474	1.263–1.720	0.079	4.918	<0.001
Reader	0.049	1.050	1.027–1.073	0.011	4.306	<0.001
Case	0.000	1.000	1.000–1.001	0.000	0.531	0.60

DR = diabetic retinopathy.
The intercept feature includes random effects of reader and retina image. The regression compares DR grade accuracy against overall accuracy for unassisted cases. Sample sizes for regression: 17 960 total observations; all data, 1796 images × 10 readers; cases with no DR, 13 960 observations; cases with DR, 4000 observations.

graded this image as no DR, whereas readers in the grades-only condition correctly graded mild NPDR. Figure 5D shows an example of moderate DR. There is a hemorrhage superior-temporal to the macula; without assistance, 2 of 3 readers missed this pathologic feature and graded the image as no DR; with assistance, all readers correctly graded the image. Again, the algorithm itself indicated uncertainty, with the strongest prediction for no DR and a secondary prediction for moderate DR. Figure 5C, D shows examples of doctors correctly evaluating cases in which the model was incorrect. These demonstrate readers avoiding overreliance. We next explored aggregate evidence of overreliance.

Overall, the readers agreed with the model 89.5% of the time (95% CI, 88.8%–90.3%) in the assisted arms. Without assistance, their answers agreed with the model 86.1% of the time (95% CI, 84.8%–87.3%). This indicates that assistance generally increases the similarity between readers' DR grades and model predictions.

If doctors overrely on the model, then we expect errors to increase when the model offers incorrect assistance. However, we did not observe this (Table 4). For the 88.3% of cases in which the model's predicted DR grade was correct, we saw a clear and significant improvement in accuracy in the assisted arms compared with the unassisted arm (2-sided proportions Z test: $P < 0.001$ for grades

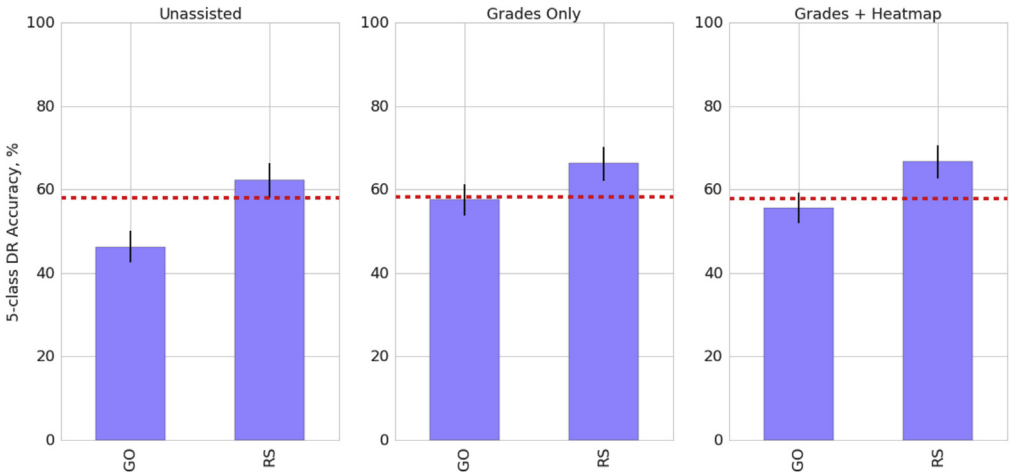

Figure 4. Bar graphs showing 5-class diabetic retinopathy (DR) grading accuracy for cases with any level of DR for (left) unassisted, (center) grades-only, and (right) grades plus heatmap conditions. Each plot breaks down performance by the reader's previous experience with retina images: general ophthalmologist (GO; n = 5) and retina specialists (RS; n = 4). Red dotted line indicates model performance. Error bars represent 95% binomial confidence intervals across all cases read by each set of readers.

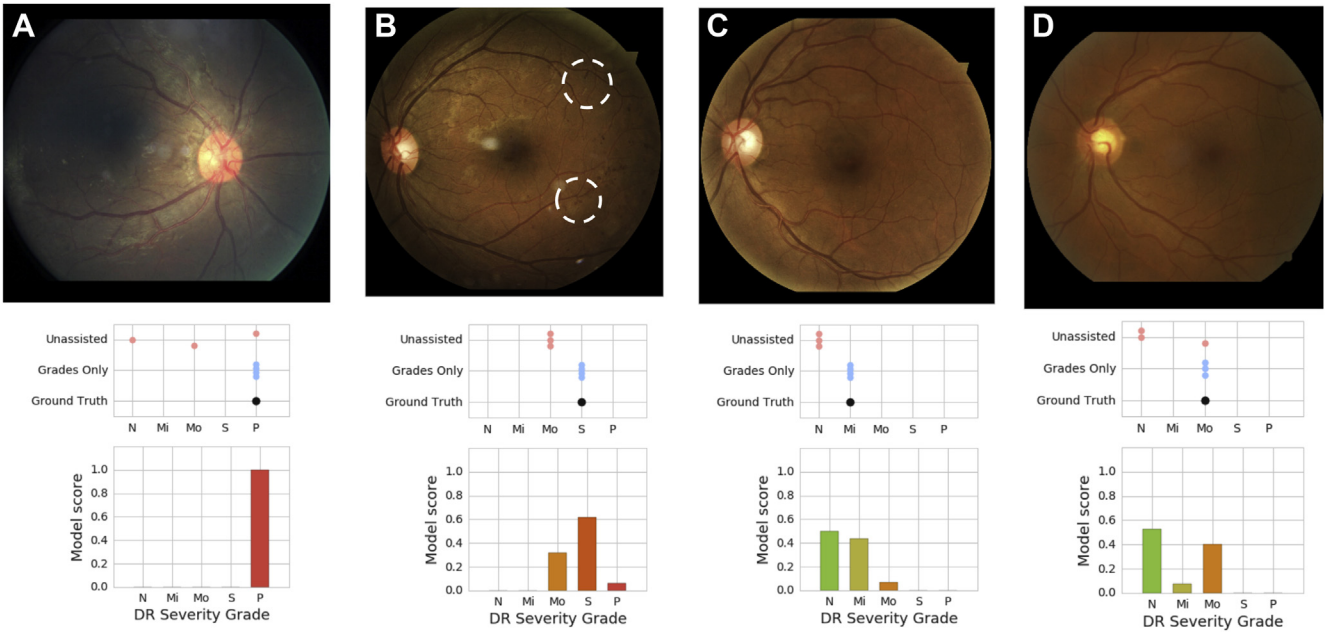

Figure 5. Example cases in which model grades improved diabetic retinopathy (DR) diagnosis over unassisted reads. **A**, Underdiagnosis of proliferative DR without assistance. **B**, Underdiagnosis of severe nonproliferative DR without assistance. **C**, **D**, Underdiagnosis without assistance, in which assisted readers consistently disagreed with an incorrect algorithmic prediction. For each panel, the top shows the fundus image; center shows readers' DR grades in the unassisted (red dots) and grades-only (blue dots) conditions (3 for each arm), along with the reference standard grade (black dot), for each image; and the bottom shows the distribution of model softmax scores for each DR class for each image. (Labels on the x-axis represent the 5 International Clinical Diabetic Retinopathy grades: none, mild, moderate, severe, and proliferative.) Circles in (**B**) highlight regions of pathologic features germane to the moderate versus severe distinction.

only, $P = 0.003$ for grades plus heatmap). However, for the 11.8% of cases in which the model's predicted DR grade was incorrect, reader accuracy did not change significantly with assistance compared with that of unassisted readers ($P = 0.07$, grades only; $P = 0.13$, grades plus heatmap).

Effects on Confidence and Time Spent on Task

Model assistance increased the self-reported confidence of readers (Fig 6), as well as the time taken for grading (Table 5). Model assistance increased the rate at which readers reported that they were very or extremely confident of their grades and decreased the rate of them being moderately or less confident (Fig 6). Confidence was significantly higher overall for both grades-only and grades plus heatmap conditions compared with the unassisted condition ($P < 0.001$, 2-sided Mann–Whitney U test), and was higher for the grades-only than grades plus heatmap condition ($P = 0.03$). For grades-only assistance, the shift mostly resulted in a sharp increase in extremely confident ratings, whereas for grades plus heatmap, the increase was distributed between very and extremely confident.

On average, grading time was consistently approximately 50 seconds for images both with and without DR ($P < 0.01$, Welch's t test; for all comparisons except grades only vs. unassisted: $t = 2.27$, $P = 0.02$). The addition of grades increased grading time on average by 8 to 15 seconds for cases without DR and over 20 seconds for images with DR. Grading times were significantly faster for grades plus heatmap compared with grades only ($t = 2.62$, $P = 0.008$). It should be noted that time spent on task was not considered a primary outcome measure in this study; readers were not instructed to grade quickly, which may have affected these effects.

Changes in Performance over the Course of the Experiment

The effects we observed, in particular the increased grading time with assistance, may vary over time, for instance, because of

Table 4. Reader Accuracy Stratified by Viewing Condition (Unassisted vs. Assisted) and Whether the Algorithm Was Correct

Case Type	Total Cases	Unassisted (95% CI)	Grades Only (95% CI)	P Value, Grades Only vs. Unassisted	Grades + Heatmap (95% CI)	P Value, Grades + Heatmap vs. Unassisted
Algorithm correct	15 870	91.1 (90.3–91.8)	94.4 (93.7–95.0)	$<10^{-10}$	92.6 (91.9–93.3)	0.003
Algorithm incorrect	2090	37.0 (33.5–40.6)	32.4 (29.0–36.0)	0.07	33.14 (29.7–36.7)	0.13

CI = confidence intervals.
When the algorithm was correct, reader accuracy improved. However, when the algorithm was incorrect, accuracy was not compromised, suggesting that readers were not overly reliant on the algorithm.

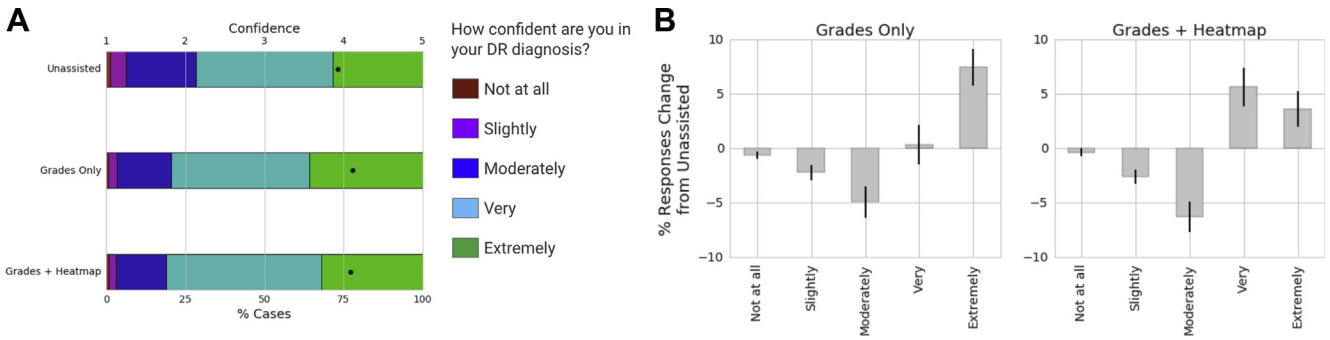

Figure 6. Graphs showing that model assistance increased diabetic retinopathy (DR) grading confidence. **A**, Distribution of reader confidence ratings across all cases for each experiment arm. The symbols represent the means and 95% confidence intervals on confidence scores, if we treat the responses as a 5-point interval scale. **B**, Changes in response rates (assisted vs. unassisted) for each confidence level for (left) grades-only and (right) grades plus heatmap conditions.

learning effects. Readers' initial performance grading with algorithmic assistance may differ with experience using the system. We therefore investigated whether DR grading performance changed over the course of our experiment (Fig 7; Table 6).

There was considerable case-by-case variability in terms of the time to grade an image. To assess overall trends, we divided the cases in each experiment block (605 images that readers graded in a given experiment condition) into bins of approximately 200 successive cases each. Although reader accuracy in the unassisted and grades-only conditions did not increase significantly over time (1-way analysis of variance: unassisted, $P = 0.52$; grades only, $P = 0.64$), accuracy improved in the grades plus heatmap condition ($P = 0.002$). Initial accuracy in this condition was comparable with unassisted readers at the beginning of a block but was significantly higher by the end of a block, comparable with the grades-only condition. This suggests that readers may become better at incorporating heatmap information to become more accurate over time.

This improvement also may reflect an increasing reliance on algorithm suggestions as readers come to trust them. We therefore examined changes in the rate of algorithm agreement (the fraction of cases where the readers' DR grade matched that of the algorithm) for each arm over the course of a block (Fig 7B). As expected, algorithm agreement did not increase over the course of a block for unassisted readers, because readers in this condition never saw the algorithm's predictions (negative control). Algorithm agreement also did not increase in the grades-only condition but increased significantly for grades plus heatmaps (Table 6). This indicates that the potential accuracy gain may be driven in part by reliance on the model for this condition.

Table 5. Time Spent by Readers for Each Type of Image Grading

	Mean Time Spent on Task, Seconds (95% CI)	Total No. of Reads
No DR		
Unassisted	50.92 (48.2–53.5)	4653
Grades only	58.1 (55.3–61.0)	4652
Grades + heatmap	64.7 (61.8–67.5)	4655
DR		
Unassisted	48.6 (44.1, 53.2)	1330
Grades only	71.4 (65.6–77.3)	1333
Grades + heatmap	70.3 (65.0–75.5)	1337

CI = confidence intervals; DR = diabetic retinopathy.

Finally, the overall increase in time spent on task with assistance occurred primarily at the beginning of each experiment block (Fig 7C). As readers became more experienced grading in a given condition, grading times tended to decrease (1-way analysis of variance: unassisted, $P = 0.006$; grades only and grades plus heatmap, $P < 0.001$). The decrease in grading time was stronger in assisted versus unassisted conditions (2-way analysis of variance, interaction between experiment arm and case bin: $F[2, 4] = 5.80$; $P < 0.001$).

Discussion

In this study, we investigated the effect of 2 types of visualization—model-predicted DR scores and explanation heatmaps—on the accuracy, speed, and confidence of readers grading fundus photographs for presence and severity of DR. We found a trend toward higher accuracy and confidence, but also higher grading times, with model assistance. As readers gained more familiarity with model assistance, there was a trend toward increased accuracy and decreased grading time (although the arms with model assistance still demonstrated a higher time on task). We also found that both types of assistance increased the sensitivity of readers for any level of DR without a significant impact on specificity.

Interobserver variability in human reading of medical imaging is well known.[26,27] Within the domain of DR screening, this variability can mean that many cases of vision-threatening DR may be missed. An early investigation in DR grading accuracy found a 49% error rate for internists, diabetologists, and medical residents in missing the diagnosis of PDR.[8]

Within our own data, we found that our readers tended to have high specificity (94–99% unassisted for moderate or worse DR; Table 2) but varied substantially in sensitivity (53%–93% unassisted). These values are in line with a number of published results, including large reviews of the DR diagnosis literature.[28–30] For instance, van der Heijden et al,[29] in evaluating an automated screening algorithm, reported sensitivities of 59% to 86% and specificity of 99% for 3 retina specialists evaluating the same image set for referable or worse DR.

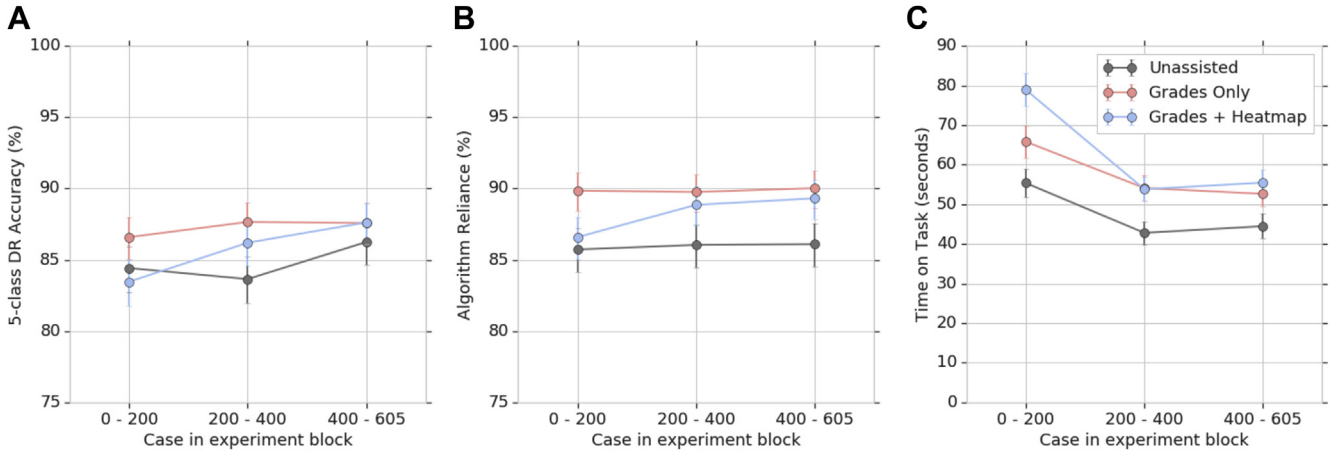

Figure 7. Graphs showing changes in reader performance over the course of the experiment. Each plot shows performance across cases, binned coarsely by when a reader graded a case within the experimental block: first 200 cases, second 200 cases, and remaining cases. (Some experimental blocks included 605 cases; the extra 5 cases are included in the third bin.) **A**, Five-class diabetic retinopathy (DR) accuracy. **B**, Algorithm reliance (rate at which graders gave the same DR grade as predicted by the algorithm). **C**, Time spent on task in seconds. Performance is broken down separately for each experiment arm: gray, unassisted; red, grades only; blue, grades plus heatmaps. Error bars reflect 95% confidence intervals.

The tendency of unassisted readers to prefer high specificity at the expense of sensitivity reflects a particular operating point, that is, a tradeoff between false-positive and false-negative results. Given ambiguous cases, unassisted readers are more likely to call lower levels of DR, missing some cases of true DR in order to minimize the number of non-DR cases incorrectly diagnosed as DR. The operating point of our model differs somewhat from the physician readers, with a higher sensitivity (91.2%), but marginally lower specificity (94.7%).

We observed that when readers saw model assistance, they also moved toward a higher-sensitivity operating point (Fig 2; Table 2). This raises the question of whether the model assistance "pulls" readers toward the model's operating point. If so, this would have substantial ramifications for the deployment of algorithmic assisted-read tools more broadly. It suggests that the choice of operating point for a model can influence tradeoffs in clinical outcomes. An alternative hypothesis is that assistance strictly increases sensitivity of assisted readers, without regard to the operating point of the model. We can not distinguish these hypotheses with the present data because both predict the pattern of results we observed. Future studies should investigate the extent to which operating point selection can influence tradeoffs between misses and false referrals in clinicians.

Although preliminary, this study has implications for how model assistance may impact the diagnosis of DR and other eye conditions in clinical settings. Intuitively, in clinical settings in which retinal fundus photographs are obtained and graded by a reader either remotely or on site, the use of an assisted tool such as this could increase the accuracy of the read at all levels of severity, including in treated populations.

The impact of algorithmic assistance is likely to vary with the context. In a screening context, assistance may help augment the abilities of nonspecialist clinicians with limited training on reading fundus images to make better referral decisions and to do so at scale. In this context, there may be a greater potential accuracy benefit to assistance, because nonspecialists likely will have a lower baseline accuracy than ophthalmologists. This benefit, and the related risks of overreliance or underreliance on the model, should be assessed with follow-up studies. Our observations around performance changes within the experiment are relevant here. Although time spent on task increased overall, we saw evidence that the increase in grading time diminished as readers graded images. With training, it is possible that assistance may be able to improve both the accuracy and speed of screening.

Within clinical eye care settings, algorithmic assistance like that examined in this study is more likely to enable consistency of care by helping to prevent attentional lapses. We demonstrated that assistance can help even well-trained specialists, increasing sensitivity without decreasing specificity. We observed that in our data, the cases in which assistance helped were cases in which pathologic features were visible but seemed to escape the attention of the grading ophthalmologist (Fig 5). As clinical environments become more data rich, the potential benefit of assistance may grow. For instance, increasing use of ultrawide-field imagery will require more sustained attention by experts. Algorithmic assistance can help to direct attention to concerning features consistently across large data sets.

Similarly, there is potential for algorithmic assistance for optometrists and other trained readers who are critical in caring for patients at risk of DR. Earlier work suggests that error rates may be much higher for nonspecialists, and therefore the potential benefit of assistance may be greater.[8] Although the present study focused on ophthalmologists and retina specialists, measuring these effects among trained but nonspecialist readers will inform the potential impact of assistance in a wider context.

In real-world circumstances, an assistive system also would determine whether images are of a sufficient quality

Table 6. Summary of Regression Analyses on Performance over the Course of an Experiment Block

	5-Class Accuracy		Algorithm Agreement		Time Spent on Task	
	Odds Ratio (95% CI)	*P Value*	*Odds Ratio (95% CI)*	*P Value*	*Coefficient*	*P Value*
Unassisted cases 201–400	0.94 (0.79–1.11)	0.49	1.02 (0.85–1.22)	0.80	−0.13	**0.06**
Unassisted cases 401–605	1.16 (0.96–1.37)	0.11	1.03 (0.86–1.22)	0.76	−0.12	**0.07**
Grades-only cases 0–200	**1.19 (1.00–1.42)**	**0.05**	**1.48 (1.22–1.79)**	**<0.001**	0.05	0.46
Grades-only cases 201–400	**1.31 (1.09–1.56)**	**0.003**	**1.46 (1.20–1.76)**	**<0.001**	−0.03	0.66
Grades-only cases 401–605	**1.30 (1.08–1.55)**	**0.004**	**1.50 (1.23–1.82)**	**<0.001**	−0.02	0.77
Grades + heatmap cases 0–200	0.93 (0.78–1.10)	0.41	1.07 (0.89–1.28)	0.44	0.24	**0.001**
Grades + heatmap cases 201–400	1.15 (0.96–1.37)	0.12	**1.32 (1.09–1.59)**	**0.004**	0.04	0.53
Grades + heatmap cases 401–605	**1.30 (1.08–1.55)**	**0.004**	**1.38 (1.14–1.67)**	**<0.001**	0.04	0.60

CI = confidence intervals.
Separate regression models were used to compare 5-class diabetic retinopathy grade accuracy (left 2 columns), algorithm reliance (middle 2 columns), and time on task (right 2 columns). Accuracy and algorithm agreement used logistic regressions; time spent on task used a linear regression against the log of time spent on task, truncated at the 98th percentile (448.6 seconds). Each regression compares performance for each factor against performance for the first 200 cases of the unassisted arm to enable examination of interactions between the bins with the assistance method. Boldface values are significant effects. Sample sizes for regression: 17 960 total observations, 1796 images × 10 readers.

for DR diagnosis. In this study, we were constrained to images that our adjudication panel determined were of adequate quality; for other images, we had no reference standard against which to evaluate reads. We performed a retrospective analysis with a model trained to predict image gradeability for DR and found that a substantial fraction (7%) of our image set would be flagged as gradable for DR by this model, but ungradable by our adjudication panel. The size of this set of images is likely the result of a conservative policy set by the adjudication panel. They excluded images in which any 1 of the 3 panel members considered the image of insufficient quality. This suggests that real-world assisted-read systems would benefit from allowing readers to decline to grade images because of low image quality. Future studies also will be required to assess the impact of gradability considerations on assistance.

Across all images, our results indicate that the grades-only condition provided a stronger benefit than grades plus heatmap. We showed that this effect was driven by a negative effect of heatmaps for cases with no DR: the heatmaps tended to cause readers to overcall these cases, in particular causing more false-positive grades of mild NPDR.

This negative effect was expected. The integrated gradients method is designed mainly to show evidence for positive predictions (pathologic features in the case of DR), but is not expected to be useful for negative predictions.[25] We included these cases in the grades plus heatmap condition to ensure that the same image set was graded in all experimental arms. The substantial negative effect we observed strongly demonstrates that using heatmaps when there is no predicted DR does not provide clinical benefit and may cause harm through overdiagnosis. However, we also observed that reader accuracy improved over the course of the experiment for grades plus heatmap. By the end of the experiment block, accuracy was comparable with the grades-only condition (Fig 7). This suggests that over time, clinicians learned to use the heatmaps for guiding diagnosis.

We also examined whether heatmaps may provide benefits in cases in which there are pathologic features. Our results indicate that, among cases with DR, grades plus heatmap show a benefit comparable with grades only, but not significantly greater (Figs 2–4; Tables 2–4). The relatively muted benefit of heatmaps in addition to algorithm-grade histograms may be the result of a number of factors. First, for the task of assessing DR severity in individual 45° fundus images, localization of features may well be less important than signals of what to look for. For instance, grades histograms may indicate difficult or uncertain cases by having prediction scores spread across several DR levels (e.g., Fig 5B–D). Seeing that a case is borderline or may have some evidence for multiple severity levels may be sufficient to trigger an exhaustive search of the image to identify any overlooked pathologic features. For evaluating larger images or several images, these tradeoffs may vary. Second, the method we used to generate heatmaps could be improved. Although qualitative feedback from readers on the heatmaps was positive, it may be that the methodology is either too aggressive or too cautious in highlighting image regions. Finally, cases with pathologic features, particularly vision-threatening DR, represented a relatively small fraction of this image set; identifying marginal increases in benefit from multiple sources of assistance may require testing with larger samples.

Deep learning has shown great promise in training algorithms that are highly accurate in terms of disease detection. This study attempts to extend this research by examining the interaction of physicians with different visualizations of deep learning model predictions. The results of this study are promising, suggesting that with increased transparency, model assistance can boost reader performance beyond what is achievable by the model or reader alone.

Acknowledgments

The authors thank Oscar Kuruvilla, Josh Carlson, and Jorge Cuadros, as well as Yun Liu for feedback on the manuscript and Scott Mayer McKinney for feedback on statistical analysis.

References

1. Zheng Y, He M, Congdon N. The worldwide epidemic of diabetic retinopathy. *Indian J Ophthalmol*. 2012;60:428–431.
2. Yau JWY, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*. 2012;35:556–564.
3. Thomas RL, Dunstan FD, Luzio SD, et al. Prevalence of diabetic retinopathy within a national diabetic retinopathy screening service. *Br J Ophthalmol*. 2015;99:64–68.
4. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. *Expert Rev Ophthalmol*. 2012;7:417–439.
5. Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American Diabetes Association. *Diabetes Care*. 2017;40:412–418.
6. American Academy of Ophthalmology. Diabetic retinopathy preferred practice pattern—updated 2017. https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017; 2017. Accessed 27.02.18.
7. Farley TF, Mandava N, Prall FR, Carsky C. Accuracy of primary care clinicians in screening for diabetic retinopathy using single-image retinal photography. *Ann Fam Med*. 2008;6:428–434.
8. Sussman EJ, Tsiaras WG, Soper KA. Diagnosis of diabetic eye disease. *JAMA*. 1982;247:3231–3234.
9. Harding SP, Broadbent DM, Neoh C, et al. Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool Diabetic Eye Study. *BMJ*. 1995;311:1131–1135.
10. Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *Am J Ophthalmol*. 2002;134:204–213.
11. 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;316:2402–2410.
12. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264–1272.
13. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.
14. Kohli A, Jha S. Why CAD failed in mammography. *J Am Coll Radiol*. 2018;15:535–537.
15. Taylor P, Potts HWW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer*. 2008;44:798–807.
16. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318:517.
17. Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 1997;39:230–253.
18. Xu K, Ba J, Kiros R, et al. *Show, attend and tell: neural image caption generation with visual attention*. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. 2015:2048–2057.
19. Fong RC, Vedaldi A. *Interpretable explanations of black boxes by meaningful perturbation. 2017 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA: IEEE Computer Society; 2017. https://doi.org/10.1109/iccv.2017.371. Accessed 13.4.2017.
20. Li L, Fredrikson M, Sen S, Datta A. Case study: explaining diabetic retinopathy detection deep CNNs via integrated gradients. http://arxiv.org/abs/1709, 09586; 2017 Accessed 27.02.18.
21. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*. 2018:1.
22. International Council of Ophthalmology. Resources: international clinical diabetic retinopathy disease severity scale, detailed table. Available at: http://www.icoph.org/resources/45/International-Clinical-Diabetic-Retinopathy-Disease-Severity-Scale-Detailed-Table-.html; 2010. Accessed 24.12.17.
23. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131:351–357.
24. Golovin D, Solnik B, Moitra S, et al. Google vizier: a service for black-box optimization. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '17*. New York: Association for Computing Machinery; 2017:1487–1495. Available at: https://doi.org/10.1145/3097983.3098043. Accessed June 30, 2017.
25. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*. 2017;70:3319–3328. http://arxiv.org/abs/1703.01365. Accessed 26.12.17.
26. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*. 2015;313:1122–1132.
27. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331:1493–1499.
28. Kawaguchi A, Sharafeldin N, Sundaram A, et al. Teleophthalmology for age-related macular degeneration and diabetic retinopathy screening: a systematic review and meta-analysis. *Telemed J E Health*. 2018;24:301–308.
29. van der Heijden AA, Abramoff MD, Verbraak F, et al. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96:63–68.
30. Ting DSW, Cheung GCM, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. 2016;44:260–277.

Footnotes and Financial Disclosures

Originally received: June 16, 2018.
Final revision: October 16, 2018.
Accepted: November 14, 2018.
Available online: December 13, 2018. Manuscript no. 2018-1367.

[1] Google Research, Google, LLC, Mountain View, California.

[2] Department of Ophthalmology, Palo Alto Medical Foundation, Palo Alto, California.

[3] Verily Life Sciences, South San Francisco, California.

[4] Department of Ophthalmology, Emory University, Atlanta, Georgia.

[5] Ophthalmic Consultants of Boston, Boston, Massachusetts.

[6] Magruder Laser Vision, Orlando, Florida.

[7] Denver Health Medical Center, Denver, Colorado.

[8] Department of Ophthalmology, University of Colorado School of Medicine, Aurora, Colorado.

[9] Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts.

*Both authors contributed equally as first authors.

Financial Disclosure(s):
The author(s) have made the following disclosure(s): R.S.: Employee — Google, LLC (Mountain View, CA).

A.T.: Employee — Google, LLC (Mountain View, CA); Stock options — Google LLC.

E.R.: Consultant — Google, LLC (Mountain View, CA).

K.B.: Employee — Google, LLC (Mountain View, CA).

D.C.: Employee — Google, LLC (Mountain View, CA).

N.H.: Consultant — Google, LLC (Mountain View, CA).

J.K.: Employee — Google, LLC (Mountain View, CA); Stock options — Google LLC.

A.N.: Employee — Google, LLC (Mountain View, CA).

Z.R.: Consultant — Google, LLC (Mountain View, CA).

D.W.: Employee — Google, LLC (Mountain View, CA); Stock options — Google LLC.

S.X.: Employee — Google, LLC (Mountain View, CA).

S.B.: Consultant — Google, LLC (Mountain View, CA).

A.J.: Consultant — Google, LLC (Mountain View, CA).

M.S.: Consultant — Google, LLC (Mountain View, CA).

J.S.: Consultant — Google, LLC (Mountain View, CA).

A.B.S.: Consultant — Google, LLC (Mountain View, CA).

G.S.C.: Employee — Google, LLC (Mountain View, CA).

L.P.: Employee — Google, LLC (Mountain View, CA).

D.R.W.: Employee — Google, LLC (Mountain View, CA).

Supported by Google, LLC Mountain View, California. The sponsor had a role in the study's approval for publication.

HUMAN SUBJECTS: No human subjects were included in this study. Images were de-identified according to Health Insurance Portability and Accountability Act Safe Harbor before transfer to study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review IRB.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Sayres, Taly, Narayanaswamy, Corrado, Peng, Webster

Analysis and interpretation: Sayres, Taly, Rahimy, Blumer, Coz, Hammel, Krause, Narayanaswamy, Rastegar, Wu, Xu, Barb, Joseph, Shumski, Smith, Sood, Corrado, Peng, Webster

Data collection: Sayres, Taly, Rahimy, Hammel, Narayanaswamy, Rastegar, Peng, Webster

Obtained funding: Sayres, Taly, Rahimy, Blumer, Coz, Hammel, Krause, Narayanaswamy, Rastegar, Wu, Xu, Barb, Joseph, Shumski, Smith, Sood, Corrado, Peng, Webster

Overall responsibility: Sayres, Taly, Rahimy, Blumer, Coz, Hammel, Krause, Narayanaswamy, Rastegar, Wu, Xu, Barb, Joseph, Shumski, Smith, Sood, Corrado, Peng, Webster

Abbreviations and Acronyms:
CI = confidence interval; **DR** = diabetic retinopathy; **NPDR** = nonproliferative diabetic retinopathy; **PDR** = proliferative diabetic retinopathy.

Correspondence:
Lily Peng, MD, PhD, Google Research, Google, LLC, 1600 Amphitheatre Way, Mountain View, CA 94043. E-mail: lhpeng@google.com.