Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy

Jonathan Krause, PhD,[1] *Varun Gulshan, PhD,*[1] *Ehsan Rahimy, MD,*[2] *Peter Karth, MD,*[3] *Kasumi Widner, MS,*[1]
Greg S. Corrado, PhD,[1] *Lily Peng, MD, PhD,*[1,*] *Dale R. Webster, PhD*[1,*]

Purpose: Use adjudication to quantify errors in diabetic retinopathy (DR) grading based on individual graders and majority decision, and to train an improved automated algorithm for DR grading.

Design: Retrospective analysis.

Participants: Retinal fundus images from DR screening programs.

Methods: Images were each graded by the algorithm, U.S. board-certified ophthalmologists, and retinal specialists. The adjudicated consensus of the retinal specialists served as the reference standard.

Main Outcome Measures: For agreement between different graders as well as between the graders and the algorithm, we measured the (quadratic-weighted) kappa score. To compare the performance of different forms of manual grading and the algorithm for various DR severity cutoffs (e.g., mild or worse DR, moderate or worse DR), we measured area under the curve (AUC), sensitivity, and specificity.

Results: Of the 193 discrepancies between adjudication by retinal specialists and majority decision of ophthalmologists, the most common were missing microaneurysm (MAs) (36%), artifacts (20%), and misclassified hemorrhages (16%). Relative to the reference standard, the kappa for individual retinal specialists, ophthalmologists, and algorithm ranged from 0.82 to 0.91, 0.80 to 0.84, and 0.84, respectively. For moderate or worse DR, the majority decision of ophthalmologists had a sensitivity of 0.838 and specificity of 0.981. The algorithm had a sensitivity of 0.971, specificity of 0.923, and AUC of 0.986. For mild or worse DR, the algorithm had a sensitivity of 0.970, specificity of 0.917, and AUC of 0.986. By using a small number of adjudicated consensus grades as a tuning dataset and higher-resolution images as input, the algorithm improved in AUC from 0.934 to 0.986 for moderate or worse DR.

Conclusions: Adjudication reduces the errors in DR grading. A small set of adjudicated DR grades allows substantial improvements in algorithm performance. The resulting algorithm's performance was on par with that of individual U.S. Board-Certified ophthalmologists and retinal specialists. *Ophthalmology 2018;125:1264-1272 © 2018 by the American Academy of Ophthalmology*

Supplemental material available at www.aaojournal.org.

Diabetic retinopathy (DR) and diabetic macular edema (DME) are among the leading causes of vision loss worldwide. Several methods for assessing the severity of diabetic eye disease have been established, including the Early Treatment Diabetic Retinopathy Study Grading System,[1] the Scottish Diabetic Retinopathy Grading System,[2] and the International Clinical Diabetic Retinopathy (ICDR) disease severity scale.[3] The ICDR scale is one of the more commonly used clinical scales and consists of a 5-point grade for DR: no, mild, moderate, severe, and proliferative.

The grading of DR is a fairly complex process that requires the identification and quantification of fine features such as microaneurysms (MAs), intraretinal hemorrhages, intraretinal microvascular abnormalities, and neovascularization. As a result, there can be a fair amount of grader variability, with intergrader kappa scores ranging from 0.40 to 0.65.[1,4–7] This is not surprising because grader variability is a well-known issue with human interpretation of imaging in other medical fields such as radiology[8] or pathology.[9]

Several methods have been proposed for resolving disagreements between graders and obtaining a reference standard. One approach consists of taking the majority decision from a group of 3 or more independent graders. Another consists of having a group of 2 or more graders work independently and then having a third generally more senior grader arbitrate disagreements, with that individual's decision serving as the reference standard. Last, a group of 3 or more graders may first grade independently and then collectively discuss disagreements until there is full consensus on the final grade. The difference between various methods of resolving disagreements has not been examined extensively.

Deep learning[10] is a family of machine learning techniques that allows computers to learn the most predictive features directly from images, given a large dataset of labeled examples, without specifying rules or features explicitly. It has also been applied recently in medical imaging, producing highly accurate algorithms for a variety of classification tasks, including melanoma,[11] breast cancer lymph node metastasis,[12,13] and DR.[14–16] Because the network is trained to predict labels that have been paired with the images, it is imperative that the labels accurately represent the state of disease found in the image, especially for the evaluation sets.

In this study, we examine the variability in different methods of grading, definitions of reference standards, and their effects on building deep learning models for the detection of diabetic eye disease.

Methods

Development Dataset

In this work, we built upon the datasets used by Gulshan et al[14] for algorithm development and clinical validation. A summary of the various data sets and grading protocols used for this study is shown in Figure S1 (available at www.aaojournal.org). The development dataset used consists of images obtained from patients who presented for DR screening at EyePACS-affiliated clinics, 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya), and the publicly available Messidor-2 dataset.[17,18] Images from the EyePACS clinics consisted of 45° retinal fundus images that were primary (or posterior pole-centered), nasal (disc-centered), and temporal field of view. The rest of the data sources only had primary field of view images. The images were captured using the Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, and Topcon NW using 45° fields of view. All 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 Institutional Review Board.

There were 3 types of grades used in the development dataset: grading by EyePACS graders, grading by ophthalmologists, and adjudicated consensus grading by retinal specialists. Images were provided at full resolution for manual grading. The images provided by EyePACS clinics were graded according to the EyePACS protocol.[19] In the EyePACS protocol, all 3 images of an eye (primary, nasal, and temporal) were graded together and assigned a single grade. We assigned the grade provided for the entire eye as a grade for each image. Although the eye-level grade may not correspond exactly with the image-level grade, this method allowed for training on different fields of view, which helps the algorithm tolerate some deviation from the primary field of view. The second source of grades was the development dataset used by Gulshan et al,[14] where each image was independently graded by ophthalmologists or trainees in their last year of residency. The grading interface used is shown in Fig S2 and Fig S3 (available at www.aaojournal.org). The third type of grades was obtained using our subspecialist grading and adjudication protocol on a small subset of images. Our subspecialist adjudication protocol consisted of 3 fellowship-trained retina specialists undertaking independent grading of the dataset. Next, each of the 3 retina specialists reviewed the images with any level of disagreement in a combination of asynchronous and live adjudication sessions. For algorithm development, the development dataset was divided into groups: a "train" set and a "tune" set. The "train" set consisted of all the images that were not adjudicated and was used to train the model parameters. The "tune" set consisted of images with adjudicated consensus grades and was used to tune the algorithm hyperparameters (e.g., input image resolution, learning rate) and make other modeling choices (e.g., network architectures). All of the images in the "tune" set were of the primary field.

Clinical Validation Dataset

The clinical validation dataset consisted of primary field of view images that were acquired from EyePACS clinics (imaged between May and October 2015). These did not overlap with any images or patients used in the development dataset.

The clinical validation set was graded using the same adjudication protocol as the tuning set (face-to-face adjudication by 3 retina specialists). In addition, the clinical validation set was graded by 3 ophthalmologists. These ophthalmologists were distinct from the retina specialists who adjudicated the set. After grading, all disagreements between the adjudicated consensus of retinal specialists and majority decision of the ophthalmologists were reviewed manually by a retinal specialist, who also assigned a likely reason for the discrepancy.

Algorithm Training

Our deep learning algorithm for predicting DR and DME was built upon the architecture used by Gulshan et al.[14] We used a convolutional neural network[20] that predicted a 5-point DR grade, referable DME, and gradability of image. The input to the neural network was an image of a fundus, and through the use of many stages of computation, parameterized by millions of numbers, the network output a real-valued number between 0 and 1 for each prediction, indicating its confidence.

The parameters of a neural network were determined by training it on a dataset of fundus images. Repeatedly, the model was given an image with a known severity rating for DR, and the model predicted its confidence in each severity level of DR, slowly adjusting its parameters over the course of the training process to become more accurate. The model was trained via distributed stochastic gradient descent and evaluated on a tuning dataset throughout the training process. The tuning dataset was used to determine model hyperparameters (parameters of the model architecture that cannot be trained with gradient descent). Finally, we created an "ensemble" of models, training 10 individual models and combining their predictions, which improves performance and robustness.

Algorithmic Improvements

We made a number of improvements to the core neural network, compared with the study by Gulshan et al.[14] First, we trained our model on a larger set of images obtained from EyePACS.[21] Because it would be prohibitively expensive to relabel all of these images with U.S.-licensed ophthalmologists (as done by Gulshan et al[14]), we also use labels determined by the EyePACS grading centers to train our DR classifier. To use both our ophthalmologist grading and the EyePACS grading effectively, we treated the labels from these 2 independent sources as separate prediction targets, training our model to predict both an EyePACS grade for DR and a grade determined by our own labeling procedure, if present.

By using this larger training set, we performed a more extensive search for well-performing hyperparameters using a Gaussian process bandit algorithm.[22] One significant change in model hyperparameters that resulted was an increase in the resolution of images the model used as input: The model used in this work

Table 1. Baseline Characteristics

	Development		Clinical Validation
	Train	*Tune*	*EyePACS-2*
Images (No.)	1 665 151	3737	1958
Patient Demographics			
Unique individuals (No.)	238 610	2643	999
Age (average ± SD)	53.5±11.6	54.3±11.1	54.9±10.9
Female/total patients where gender was known	140 183/230 556 (60.8%)	1101/1767 (62.3%)	603/997 (60.5%)
Image Quality Distribution			
Fully gradable/total images where image quality was assessed	1 343 726/1529 771 (87.8%)	3547/3737 (94.9%)	1813/1958 (92.6%)

	Development				Validation	
	Train		Tune		EyePACS-2	
Disease Severity Distribution	*No.*	*%*	*No.*	*%*	*No.*	*%*
Total images where DR was assessed	1 662 646	100.0%	3547	100.0%	1813	100.0%
No DR	1 164 368	70.0%	2417	68.1%	1478	81.5%
Mild	152 938	9.2%	458	12.9%	125	6.9%
Moderate	249 138	15.0%	497	14.0%	144	7.9%
Severe	50 640	3.0%	106	3.0%	50	2.8%
Proliferative	45 562	2.7%	69	1.9%	16	0.9%
Total images where DME was assessed	252 544	100.0%	3547	100%	1813	100%
Referable DME	28 807	11.4%	281	7.9%	79	4.4%

DME = diabetic macular edema; DR = diabetic retinopathy; SD = standard deviation.
Summary of image characteristics and available demographic information in the development and clinical validation datasets. The adjudicated reference standard was used for computing the DR and DME distributions on the "Tune" and "EyePACS-2" datasets, and the majority reference standard was used for the "Train" dataset. For DR, the majority was taken over both the EyePACS partner grade, and the grading done by our ophthalmologists (using whichever grades are available). For DME, the majority is taken over grades from our ophthalmologists. For Image quality distributions, the EyePACS partner grade was used for the "Train" dataset and an adjudicated image quality standard was used for the "Tune" and "EyePACS-2" datasets. The "Tune" dataset and "Validation" dataset differ in their definition of DME used during adjudication: the "Tune" dataset used a standard of hard exudates within 2 disc diameter (DD) of the fovea (to approximate the macular OCT box), whereas the "Validation" dataset used a standard of hard exudates within 1 DD of the fovea.

has an input resolution of 779×779 pixels, a large increase over the 299×299 pixels used by Gulshan et al.[14] This has the effect of increasing the effective resolution of lesions. For example, an MA or small hard exudate may occupy only 1 pixel or less within a 299×299 pixel image,[23,24] but at this higher resolution, it would occupy approximately 6.8 pixels. The effective resolution of other lesions is similarly increased when using higher input resolution.

Other significant changes included changing model architecture from "Inception-v3"[25] to "Inception-v4"[26] and predicting a 5-class rating for DR (Gulshan et al[14] only predicted binary referables).

Evaluation

One advantage of using machine learning-based systems for predicting DR is that they explicitly output a confidence in the range (0–1) for each image. By applying a threshold to this confidence to determine referability, an algorithm's sensitivity and specificity can be tuned for a particular use case; as the threshold is lowered, the algorithm becomes more sensitive, but less specific. Performance across many different thresholds is depicted in a receiver operating curve and is summarized with the area under the curve (AUC). We used as our primary model evaluation metric for DR the AUC at different DR severity thresholds, for example, the AUC for predicting moderate and above DR. In addition, for some experiments on 5-class prediction we evaluated using the quadratic-weighted Cohen's kappa.[27]

The algorithm's sensitivity and specificity when predicting a 5-class output cannot be tuned with a single threshold, though.

Therefore, we used a cascade of thresholds, building on the model's predicted confidence in each of the severity levels for DR. First, a threshold was picked for proliferative DR: If the model's predicted probability of an image corresponding to proliferative DR was above a specified threshold, chosen to reach a target sensitivity on the tuning set, the image was classified as proliferative DR. Next, another threshold was picked for the model's predicted probability of severe DR and applied to all images not classified as proliferative DR. A similar procedure was applied for both moderate and mild DR. Finally, all remaining images whose probabilities did not pass any of the thresholds were classified as no DR.

Table 2. Comparison of Retinal Specialist Grades before and after Adjudication on the Validation Dataset

	Majority Decision of Retinal Specialist Grading before Adjudication				
	No	*Mild*	*Moderate*	*Severe*	*Proliferative*
Adjudicated Consensus					
No	1469	4	5	0	0
Mild	58	62	5	0	0
Moderate	22	3	118	1	0
Severe	0	0	13	36	1
Proliferative	0	0	0	1	15

Confusion matrix for diabetic retinopathy between the grade determined by majority decision and adjudicated consensus.

Table 3. Agreement between Each Retina Specialist and the Adjudicated Reference Standard on the Validation Dataset

	DR		
	Sensitivity	*Specificity*	*Quadratic-Weighted Kappa*
Retinal Specialist A	74.6%	99.1%	0.82
Retinal Specialist B	74.4%	99.3%	0.89
Retinal Specialist C	82.1%	99.3%	0.91
Majority Decision (Retinal Specialists)	88.1%	99.4%	0.91

DR = diabetic retinopathy.
Retina specialists correspond to those who contributed to the final adjudicated reference standard. Sensitivity and specificity metrics reported are for moderate or worse DR. Agreement between the preadjudication 5-point DR grade and the final adjudicated grade is also measured by the quadratic-weighted kappa.

Results

Grading and Adjudication

The baseline characteristics of the development and clinical validation datasets are described in Table 1. The training portion of the development set consisted of more than 1.6 M fundus images from 238 610 unique individuals. The tune portion of the development set consisted of adjudicated consensus grades for 3737 images from 2643 unique individuals. The clinical validation set consisted of 1958 photos from 998 unique individuals. Compared with the clinical validation set, the development set had a slightly higher proportion of abnormal images.

A comparison of the grades generated by the majority decision from independent grading by the retinal specialists in the panel before the adjudication process versus the adjudicated reference standard is outlined in Table 2. Most grades remained within 1 step of each level of severity. For DR, there were 27 instances (1.5% of images) where the difference was 2 steps, and none that were 3 or

Table 4. Comparison of Ophthalmologist Grades versus Adjudicated Grades from Retina Specialists on the Validation Dataset

	Majority Decision of Ophthalmologist Grading				
	No	*Mild*	*Moderate*	*Severe*	*Proliferative*
Adjudicated Consensus					
No	1432	32	10	0	1
Mild	67	38	19	1	0
Moderate	20	14	94	14	2
Severe	0	0	9	41	0
Proliferative	0	0	1	2	13

	Majority of Ophthalmologist Grading	
	Not Referable	*Referable DME*
Adjudicated Consensus		
Not Referable	1710	17
Referable DME	13	65

DME = diabetic macular edema.
Confusion matrix for diabetic retinopathy and DME between the grade determined by majority decision of the ophthalmologists and the adjudicated consensus of retinal specialists.

more. Adjudication more often yielded new grades with higher levels of severity. The weighted-kappa score for the grade determined by the majority decision of retinal specialists and by adjudication was 0.91 (Table 3).

To better characterize the difference in each retinal specialist's preadjudication grade and the adjudicated consensus for detecting moderate and above DR, we measured the sensitivity and specificity of each specialist (Table 3). Although all 3 specialists were very specific (~99%), sensitivity ranged from 74.4% to 82.1%, corresponding to weighted-kappa scores of 0.82 to 0.91. A calculation of the nonadjudicated majority decision of the panel of 3 retina specialists showed higher sensitivity (88.1%) than grades from individual specialists (Table 3), leaving 11.9% of further cases with moderate or above DR that required the adjudication process to detect. During independent grading, some retinal specialists used a different variation for grading referable DME in which they considered any hard exudate within the macular OCT box (~2 disc diameters) as referable. This discrepancy was resolved during adjudication, and all images where referable DME was potentially present were reassessed to match the standard protocol (where hard exudates within 1 disc diameter was considered referable DME). For this reason, the independent grading of the retinal specialists for DME was not included in subsequent analyses.

In addition to having retinal specialists grade and adjudicate the clinical validation datasets, we also had 3 U.S. board-certified ophthalmologists grade the same set (Table 4). Quadratic-weighted kappa values were generally good (0.80−0.84), but somewhat lower for ophthalmologists than for the retina specialists (Table 5). The majority decision of the ophthalmologists yielded a higher agreement (weighted kappa: 0.87) than individual ophthalmologists alone. A comparison of all graders (3 retinal specialists and 3 ophthalmologists) showed that disagreements were more common in cases in which the adjudicated consensus yielded referable disease (Fig 1). An analysis combining the 3 retinal specialists and 3 ophthalmologists is given in Table S1 (www.aaojournal.org).

A summary of the causes of errors in the nonadjudicated grades generated from the majority decision of the ophthalmologists is presented in Table 6. The most common causes were missed MAs (36%), artifacts (20%), and disagreement regarding whether a lesion was an MA or a hemorrhage (16%). Thirty-six of 193 disagreements differed by at least 2 grades (e.g., no DR vs. moderate or moderate vs. proliferative). Of these, overgrading, cases in which the majority decision of the ophthalmologists was more severe than the final adjudicated consensus grade, accounted for 14 (7.3%) of the disagreements, and undergrading accounted for 22 (11.4%) of the cases.

Model Results

We compare the results of our model with the model used by Gulshan et al[14] in Table 7, using the majority decision or adjudicated consensus of retinal specialist grades. For moderate or worse DR, we observed an improvement in AUC with increasing image resolution as well as with the other algorithmic improvements outlined in the "Methods" section for both methods of computing reference standard. However, this improvement in AUC was larger when using the adjudicated consensus as the reference standard (from 0.934 to 0.986 AUC). When predicting mild or worse DR (Fig 2), increasing image resolution showed only a slight gain (<1%) when using majority decision as the reference standard, while when evaluating with adjudicated consensus grades we observed that increasing resolution improved the model's AUC from 0.954 to 0.984. However, increasing resolution beyond approximately

Table 5. Agreement between Ophthalmologists Graders with the Adjudicated Reference Standard on the Validation Dataset

	DR			DME	
	Sensitivity	*Specificity*	*Quadratic-Weighted Kappa*	*Sensitivity*	*Specificity*
Ophthalmologist A	75.2%	99.1%	0.84	81.5%	98.7%
Ophthalmologist B	74.9%	97.9%	0.80	62.7%	98.6%
Ophthalmologist C	76.4%	97.5%	0.82	86.4%	99.1%
Majority Decision (Ophthalmologists)	83.8%	98.1%	0.87	83.3%	99.0%

DME = diabetic macular edema; DR = diabetic retinopathy.
Sensitivity and specificity metrics are for moderate or worse DR and referable DME for each grader. Agreement between the adjudicated grade and the 5-point scale is also measured by the quadratic-weighted kappa.

450×450 pixel input seemed to have marginal performance gains. Another benefit of evaluating with adjudicated consensus grades was in reduced metric variability, where we observed reduced metric variance (confidence intervals) across input resolutions. This was particularly important when making changes to the algorithm because it made it easier to determine with confidence if a change in the algorithm was actually an improvement.

We illustrate the performance of our final model using the adjudicated consensus as the reference standard in Table 8. On 5-class DR prediction, our model achieved a quadratic-weighted kappa of 0.84, which is on par with the performance of individual ophthalmologists and individual retinal specialists. There were 257 disagreements between the adjudicated consensus and the model. Of the 56 (21.8%) disagreements that were 2 or more steps, 53 (20.6%) were the result of overgrading, that is, the model predicted a higher grade than the adjudicated consensus, and 3 (1.2%) resulted from undergrading.

In addition to 5-point grading analysis for DR, we also demonstrate the model's performance in binary classification tasks.

Figure 3 summarizes the model's performance as well as that of the ophthalmologists and retinal specialists in detecting mild or worse DR, moderate or worse DR, severe or worse DR, and proliferative DR. For all of these classification tasks, the algorithm's performance was approximately on par with that of individual ophthalmologists and retinal specialists. For mild or worse DR, the algorithm had a sensitivity of 0.970, specificity of 0.917, and AUC of 0.986.

Discussion

Deep learning has garnered attention recently because of its ability to create highly accurate algorithms from large datasets of labeled data without feature engineering. However, great care must be taken to evaluate these algorithms against as high-quality "ground truth" labels as possible. Previous work[14] used majority decision to generate ground truth, without an adjudication process. The present work
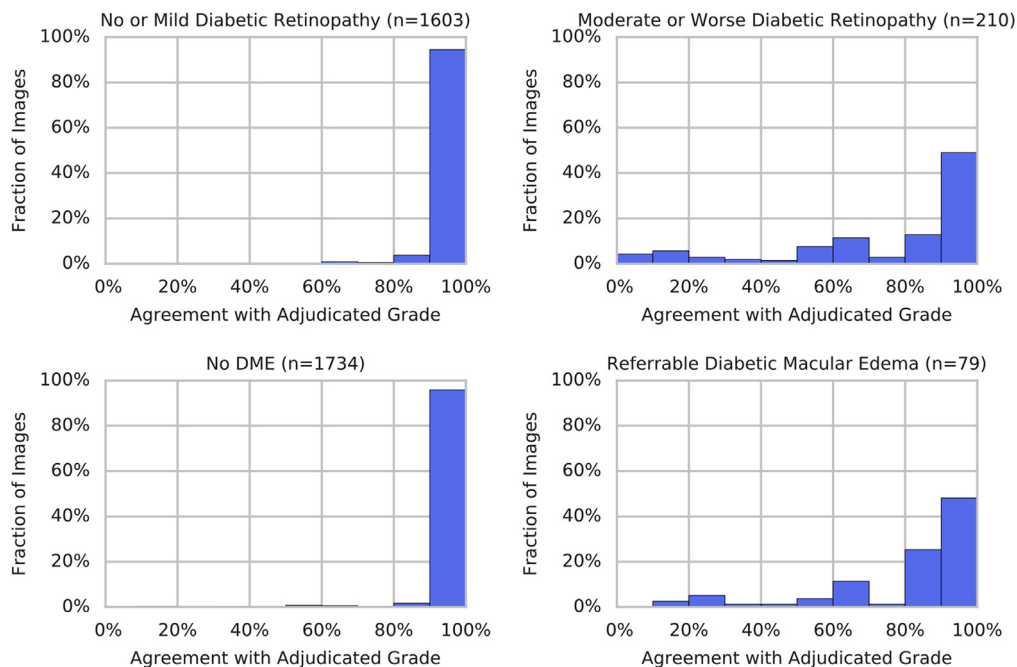

Figure 1. Grader agreement based on the adjudicated consensus grade for referable diabetic retinopathy (DR) and diabetic macular edema (DME). Independent grading of all 3 retinal specialists and all 3 ophthalmologists are included in this analysis.

Table 6. Reasons for Difference between Adjudication of Retinal Specialist and Majority Decision from Ophthalmologist Graders

	Adjudication Consensus Grade (Retina Specialists) - Majority Decision Grade (Ophthalmologists)					
	−4	−2	−1	1	2	Total
Artifact vs. not		5	28	1	5	39
Extent of lesions		1	16	9		26
Hemorrhage vs. MA		1	13	13	3	30
Hemorrhage vs. not		4		4	11	19
IRMA vs. not		1				1
Missed hemorrhage			2			2
Missed MA			6	63		69
Missed NVD/NVE				2	1	3
PRP vs. not	1	1			1	3
Other					1	1
Total	1	13	65	92	22	193

IRMA = intraretinal microvascular abnormality; MA = microaneurysm; NVD/NVE = neovascularization of the disc or neovascularization of elsewhere; PRP = panretinal photocoagulation.
Disagreements between the adjudicated consensus and majority decision were examined and characterized by a retinal specialist. Positive numbers denote that the adjudication grade was more than the majority decision of ophthalmologist grade, and vice-versa for negative numbers.

suggests that an adjudication process yielding a consensus grade from multiple retina specialists provides a more rigorous reference standard for algorithm development, especially for the identification of artifacts and missed MAs, ultimately leading to superior algorithms and improved detection of disease.

The adjudication sessions of the 3 retina specialists yielded considerable insights into the grading process. First, image artifacts, especially those resembling typical pathology such as MAs, were a common source of disagreement. These disagreements were consistent across many images, because artifacts are often caused by consistent lens artifacts or dust spots. The adjudication process was critical in correcting these errors. Likewise, when true pathology was present, a frequent cause of grader disagreement was in differentiating an MA from an intraretinal hemorrhage, which in the absence of other findings, represented the difference between labeling an image as mild or moderate DR. In addition, there was significant subjective variance in the precise definition of the boundaries between grades at the outset of the

adjudication process. Although traditional grading definitions are fixed, it became very clear that there are gray areas where these definitions are not sufficient to make an unambiguous call on DR severity, for example, on whether a macula-centered image with numerous hemorrhages was more likely to represent moderate or severe DR. Once all 3 retina specialists converged on a more precise set of definitions, staying within but also refining the traditional definitions, many of these initial disagreements were resolved. On conclusion of the adjudication process, each of the retina specialists indicated that the precision used in the adjudication process was above that typically used in everyday clinical practice.

A comparison of the performance of algorithm and manual grading also yielded interesting insights. First, although manual grading by ophthalmologists led to an approximately even ratio of overgrading (7.3%) and undergrading (11.4%), the grading by the algorithm had a higher ratio of overgrading (20.36%) than undergrading (1.2%). This can be attributed to the decision to operate the algorithm at a very high-sensitivity set point, which would be desirable for a screening algorithm to maximize the detection of referable DR. However, a combination of the algorithm for initial screening coupled with manual review of the resulting positive images would likely yield a system that is both highly sensitive and specific, which may help limit the number of diabetic patients being referred who do not actually have referable disease.

It is also worth noting that although full-resolution images were used for grading and adjudication, training models using an input resolution greater than 450×450 pixels did not yield a significant boost in model performance. This suggests neural networks may not require as high of a resolution input image to detect DR lesions as humans do. However, this is an observation made in the setting of a specific grading scale (ICDR) using only single nonstereoscopic 45° color fundus photographs. This observation may not generalize if other finer grading scales are used, such as Early Treatment Diabetic Retinopathy Study, if the input images are different (stereoscopic, ultrawide-field), or if the ground truth incorporated signals from other clinical data, such as other types of retinal imaging like OCT or fluorescein angiography.

There are several limitations to this study that may represent further avenues for research. One key limitation is that we used hard exudates as a proxy for DME, which

Table 7. Differences in the Final Area under the Curve Observed for Algorithms Trained in the Image Resolution Algorithm Selection Experiments for Moderate or Worse Diabetic Retinopathy

	Original Model	Original Model, High Resolution	Our Full Model
Majority decision of retinal specialists ground truth	0.954	0.976	0.986
Adjudicated consensus ground truth	0.934	0.962	0.986

Majority decision was based on the grades of the 3 retinal specialists. "Original Model" refers to the model and training data of Gulshan et al,[14] "Original Model, High Resolution" uses higher resolution input images (779×779 pixels), and "Our Full Model" incorporates all changes described in the "Algorithmic Improvements" section in the main text.
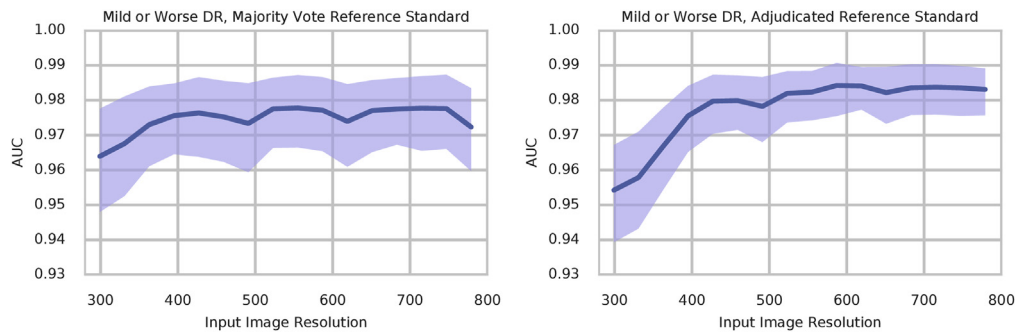
Figure 2. Image resolution input to model versus area under the curve (AUC) for mild and above DR. **Left:** Using majority decision of retinal specialists as the reference standard. **Right:** Using the adjudicated consensus grade of retinal specialists as a reference standard. Shaded areas represent a 95% confidence interval as measured via bootstrapping. DR = diabetic retinopathy.

generally requires stereoscopic images or OCT to properly diagnose and characterize. Including OCT imaging in the process of determining ground truth for DME will be valuable in future studies.

In addition, we used only 3 retinal specialists to establish ground truth and 3 U.S. Board-Certified ophthalmologists as a comparison group. However, the kappa scores, sensitivity, and specificity of the graders in this study are in line with, if not slightly better than, what has been reported in other studies[6,7,28,29] but we found a wider range for sensitivity, especially when it comes to detecting mild disease. For example, using adjudication as the ground truth, the 3 retinal specialists in Abramoff et al[29] had preadjudication

sensitivities of 71% to 91% and specificities of 95% to 100% for calling referable DR (defined as moderate or worse DR or referable DME). Likewise, our retinal specialists had preadjudication sensitivities of 74% to 82% and specificities of more than 99% for moderate or worse DR. Gangaputra et al[6] reported sensitivities of 53% and 84% and specificities of 97% and 96% for clinical grading versus reading center grading for detecting mild or worse DR. The retinal specialists and ophthalmologists in our study had sensitivities between 70% and 80% and specificities more than 97%. Examining the differences between adjudication and reading center grading or between different adjudication groups would be valuable next steps in investigating variability in grading. We also did not study the variability in grading for optometrists or other graders trained to evaluate DR images, which is another area for further exploration.

Other areas of future work include measuring the impact of using adjudicated consensus grades in the training portion of the development dataset on model performance and studying the algorithm's performance on other clinical validation datasets. Although images in the development and clinical validation datasets of our study were taken in nonoverlapping time intervals from the EyePACS population, additional validation on more diverse datasets in terms of cameras, geographies, and patient population would further characterize the algorithm's generalizability.

Despite these limitations, we believe this work provide a basis for further research and begins to lay the groundwork for standards in this field. For example, although adjudication yields a more reliable ground truth, it requires significant time and resources to perform. We demonstrate that by using existing grades and adjudicating a small subset (0.22%) of the training image grades for the "tune" set, we were able to significantly improve model performance without adjudicating the entire training corpus of images. Leveraging these techniques, our resulting model's performance was approximately on par with that of individual ophthalmologists and retinal specialists.

In conclusion, we believe that proper determination of ground truth is a critical aspect of the development of clinically useful machine learning algorithms for use in

Table 8. Comparison of Algorithm Grade versus Adjudicated Grades from Retina Specialists

	Algorithm Grade				
	No	*Mild*	*Moderate*	*Severe*	*Proliferative*
Adjudicated Consensus					
No	1356	74	44	1	3
Mild	7	43	74	0	1
Moderate	3	3	98	36	4
Severe	0	0	2	47	1
Proliferative	0	0	0	4	12

	Algorithm Grade	
	Not Referable	*Referable DME*
Adjudicated Consensus		
Not Referable	1637	97
Referable	4	75

	DR			DME	
	Sensitivity	*Specificity*	*Weighted Kappa*	*Sensitivity*	*Specificity*
Algorithm	97.1%	92.3%	0.84	94.9%	94.4%

DME = diabetic macular edema; DR = diabetic retinopathy.
Confusion matrix for DR and DME where the grade is determined by the algorithm or adjudicated consensus of the retinal specialists. Sensitivity and specificity metrics reported are for moderate or worse DR and referable DME.
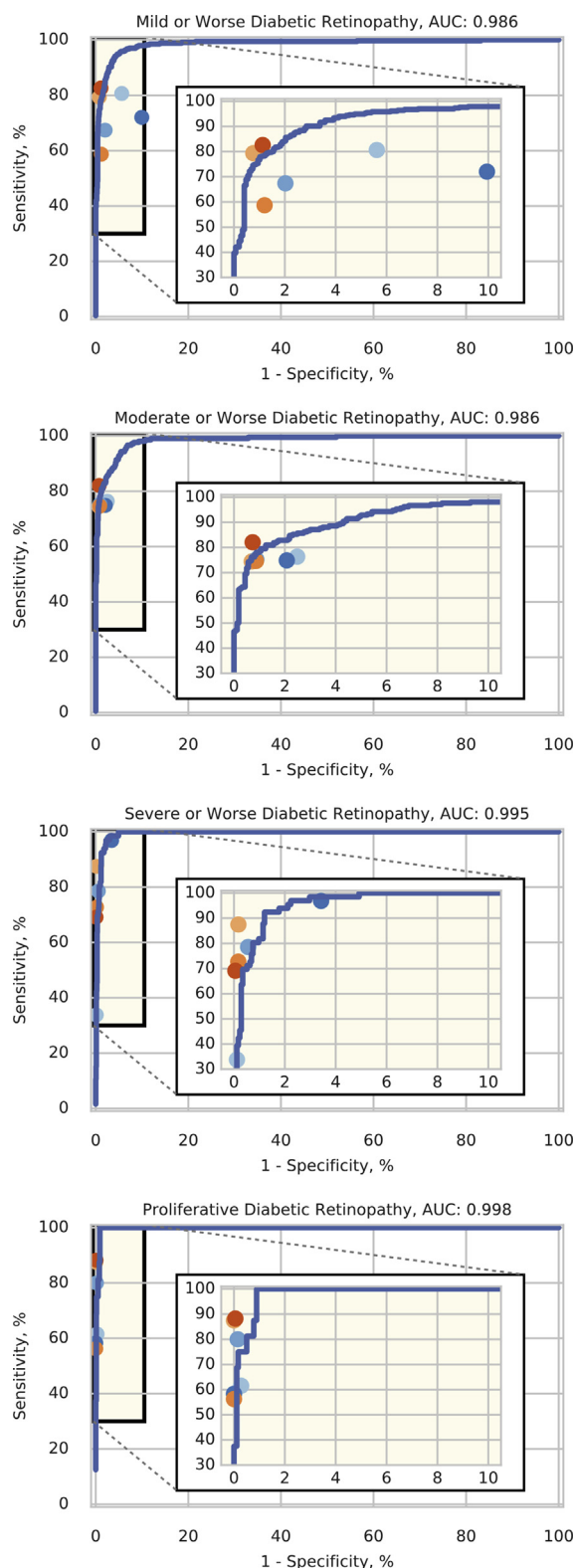
Figure 3. Comparison of the algorithm, ophthalmologists, and retinal specialists using the adjudicated reference standard at various DR severity thresholds. The algorithm's performance is the blue curve. The 3 retina specialists are represented in shades of orange/red, and the 3 ophthalmologists are in shades of blue. N = 1813 fully gradable images. AUC = area under the curve.

detection of or screening for retinal disease. We have shown that our process of live adjudication by multiple subspecialists yielding a consensus grade provides improved algorithmic accuracy and led to improved screening models.

Acknowledgments

References

1. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. *Ophthalmology.* 1991;98:786−806.

2. Diabetic Retinopathy Screening Services in Scotland: A Training Handbook − July 2003: page 17. Available at: http://www.ndrs-wp.scot.nhs.uk/?page_id=1609. Accessed June 21, 2017.

3. American Academy of Ophthalmology. International clinical diabetic retinopathy disease severity scale, detailed table. Available at: http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf. Accessed October 14, 2016.

4. Scott IU, Bressler NM, Bressler SB, et al. Agreement between clinician and reading center gradings of diabetic retinopathy severity level at baseline in a phase 2 study of intravitreal bevacizumab for diabetic macular edema. *Retina.* 2008;28: 36−40.

5. Li HK, Hubbard LD, Danis RP, et al. Digital versus film fundus photography for research grading of diabetic retinopathy severity. *Invest Ophthalmol Vis Sci.* 2010;51: 5846−5852.

6. Gangaputra S, Lovato JF, Hubbard L, et al. Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity. *Retina.* 2013;33: 1393−1399.

7. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K. Thai Screening for Diabetic Retinopathy Study Group. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. *Ophthalmology.* 2006;113:826−832.

8. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretations of mammograms. *N Engl J Med.* 1994;331:1493−1499.

9. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA.* 2015;313:1122−1132.

10. LeCun Y, Yoshua B, Geoffrey H. Deep learning. *Nature.* 2015;521:436−444.

11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542:115−118.

12. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on Gigapixel Pathology Images. arXiv [csCV] 2017. Available at: http://arxiv.org/abs/1703.02442.

13. Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node

metastases in women with breast cancer. *JAMA*. 2017;318: 2199−2210.

14. 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.

15. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124: 962−969.

16. 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.

17. Decencière E, Etienne D, Xiwei Z, et al. Feedback on a publicly distributed image database: the Messidor Database. *Image Anal Stereol*. 2014:0. Available at: http://dx.doi.org/10.5566/ias.1155.

18. Quellec G, Lamard M, Josselin PM, et al. Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Trans Med Imaging*. 2008;27:1230−1241.

19. EyePACS. EyePACS digital retinal image grading protocol narrative. Available at: https://www.eyepacs.org/consultant/Clinical/grading/EyePACS-DIGITAL-RETINAL-IMAGE-GRADING.pdf.

20. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86:2278−2324.

21. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. *J Diabetes Sci Technol*. 2009;3:509−516.

22. Golovin D, Solnik B, Moitra S, et al. Google Vizier: A Service for Black-Box Optimization. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2017:1487−1495. Accessed June 30, 2017.

23. Kolb H. Facts and figures concerning the human retina. In: Kolb H, Fernandez E, Nelson R, eds. *Webvision: The Organization of the Retina and Visual System*. Salt Lake City, UT: University of Utah Health Sciences Center; 2005.

24. A Practical Manual of Diabetic Retinopathy Management. *Clin Exp Optom*. 2009;92:527−528.

25. Szegedy C, Vanhouke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv*. 2015;1502.03167.

26. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016. Thirty-First AAAI Conference on Artificial Intelligence 2016. Available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14806/14311. Accessed December 13, 2017.

27. Cohen JA. Coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37−46.

28. Gonzalez ME, Gonzalez C, Stern MP, et al. Concordance in diagnosis of diabetic retinopathy by fundus photography between retina specialists and a standardized reading center. Mexico City Diabetes Study Retinopathy Group. *Arch Med Res*. 1995;26:127−131.

29. 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.

Footnotes and Financial Disclosures

¹ Google Research, Google Inc., Mountain View, California.

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

³ Oregon Eye Consultants, Eugene, Oregon.

*Equal contribution was provided by L.P. and D.R.W.

Abbreviations and Acronyms:
AUC = area under the curve; **DME** = diabetic macular edema; **DR** = diabetic retinopathy; **ICDR** = International Clinical Diabetic Retinopathy; **MA** = microaneurysm.

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