JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

← Editorial

+ Supplemental content

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

EXPOSURE Deep learning–trained algorithm.

MAIN OUTCOMES AND MEASURES The sensitivity and specificity of the algorithm for detecting referable diabetic retinopathy (RDR), defined as moderate and worse diabetic retinopathy, referable diabetic macular edema, or both, were generated based on the reference standard of the majority decision of the ophthalmologist panel. The algorithm was evaluated at 2 operating points selected from the development set, one selected for high specificity and another for high sensitivity.

RESULTS The EyePACS-1 data set consisted of 9963 images from 4997 patients (mean age, 54.4 years; 62.2% women; prevalence of RDR, 683/8878 fully gradable images [7.8%]); the Messidor-2 data set had 1748 images from 874 patients (mean age, 57.6 years; 42.6% women; prevalence of RDR, 254/1745 fully gradable images [14.6%]). For detecting RDR, the algorithm had an area under the receiver operating curve of 0.991 (95% CI, 0.988-0.993) for EyePACS-1 and 0.990 (95% CI, 0.986-0.995) for Messidor-2. Using the first operating cut point with high specificity, for EyePACS-1, the sensitivity was 90.3% (95% CI, 87.5%-92.7%) and the specificity was 98.1% (95% CI, 97.8%-98.5%). For Messidor-2, the sensitivity was 87.0% (95% CI, 81.1%-91.0%) and the specificity was 98.5% (95% CI, 97.7%-99.1%). Using a second operating point with high sensitivity in the development set, for EyePACS-1 the sensitivity was 97.5% and specificity was 93.4% and for Messidor-2 the sensitivity was 96.1% and specificity was 93.9%.

CONCLUSIONS AND RELEVANCE In this evaluation of retinal fundus photographs from adults with diabetes, an algorithm based on deep machine learning had high sensitivity and specificity for detecting referable diabetic retinopathy. Further research is necessary to determine the feasibility of applying this algorithm in the clinical setting and to determine whether use of the algorithm could lead to improved care and outcomes compared with current ophthalmologic assessment.

Author Affiliations: Google Inc, Mountain View, California (Gulshan, Peng, Coram, Stumpe, Wu, Narayanaswamy, Venugopalan, Widner, Madams, Nelson, Webster); Department of Computer Science, University of Texas, Austin (Venugopalan); EyePACS LLC, San Jose, California (Cuadros); School of Optometry, Vision Science Graduate Group, University of California, Berkeley (Cuadros); Aravind Medical Research Foundation, Aravind Eye Care System, Madurai, India (Kim); Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, Tamil Nadu, India (Raman); Verily Life Sciences, Mountain View, California (Mega); Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Mega).

Corresponding Author: Lily Peng, MD, PhD, Google Research, 1600 Amphitheatre Way, Mountain View, CA 94043 (lhpeng@google.com).

Among individuals with diabetes, the prevalence of diabetic retinopathy is approximately 28.5% in the United States[1] and 18% in India.[2] Most guidelines recommend annual screening for those with no retinopathy or mild diabetic retinopathy, repeat examination in 6 months for moderate diabetic retinopathy, and an ophthalmologist referral for treatment evaluation within a few weeks to months for severe or worse diabetic retinopathy or the presence of referable diabetic macular edema, known as clinically significant macular edema.[3] Referable diabetic retinopathy has been defined as moderate or worse diabetic retinopathy or referable diabetic macular edema,[4] given that recommended management changes from yearly screening to closer follow-up at moderate disease severity.[3] Retinal photography with manual interpretation is a widely accepted screening tool for diabetic retinopathy, with performance that can exceed that of in-person dilated eye examinations.[3]

Automated grading of diabetic retinopathy has potential benefits such as increasing efficiency, reproducibility, and coverage of screening programs; reducing barriers to access; and improving patient outcomes by providing early detection and treatment. To maximize the clinical utility of automated grading, an algorithm to detect referable diabetic retinopathy is needed. Machine learning (a discipline within computer science that focuses on teaching machines to detect patterns in data) has been leveraged for a variety of classification tasks including automated classification of diabetic retinopathy. However, much of the work has focused on "feature-engineering," which involves computing explicit features specified by experts, resulting in algorithms designed to detect specific lesions or predicting the presence of any level of diabetic retinopathy.[5] Deep learning[6] is a machine learning technique that avoids such engineering by learning the most predictive features directly from the images given a large data set of labeled examples. This technique uses an optimization algorithm called back-propagation to indicate how a machine should change its internal parameters to best predict the desired output of an image.

In this study, deep learning[7,8] was used to train an algorithm to detect referable diabetic retinopathy and assess the performance of the algorithm in 2 clinical validation sets.

Methods

Data Sets

For algorithm development, macula-centered retinal fundus images were retrospectively obtained from EyePACS in the United States and 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya) among patients presenting for diabetic retinopathy screening. All images were deidentified according to Health Insurance Portability and Accountability Act Safe Harbor prior to transfer to study investigators. Ethics review and institutional review board exemption was obtained using Quorum Review IRB.

Two further data sets were used for clinical validation. The first deidentified data set consisted of a random sample

Key Points

Question How does the performance of an automated deep learning algorithm compare with manual grading by ophthalmologists for identifying diabetic retinopathy in retinal fundus photographs?

Finding In 2 validation sets of 9963 images and 1748 images, at the operating point selected for high specificity, the algorithm had 90.3% and 87.0% sensitivity and 98.1% and 98.5% specificity for detecting referable diabetic retinopathy, defined as moderate or worse diabetic retinopathy or referable macular edema by the majority decision of a panel of at least 7 US board-certified ophthalmologists. At the operating point selected for high sensitivity, the algorithm had 97.5% and 96.1% sensitivity and 93.4% and 93.9% specificity in the 2 validation sets.

Meaning Deep learning algorithms had high sensitivity and specificity for detecting diabetic retinopathy and macular edema in retinal fundus photographs.

of macula-centered images taken at EyePACS screening sites between May 2015 and October 2015. A variety of cameras were used, including Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, and Topcon NW using 45° fields of view. EyePACS images were acquired as a part of routine clinical care for diabetic retinopathy screening, and approximately 40% of the images were acquired with pupil dilation. This data set did not overlap with the EyePACS data used in development. The second data set was the publicly available Messidor-2 data set,[9,10] which has been used by other groups for benchmarking performance of automated detection algorithms for diabetic retinopathy.[11-13] The images were obtained between January 2005 and December 2010 at 3 hospitals in France using a Topcon TRC NW6 nonmydriatic camera and 45° fields of view centered on the fovea. Approximately 44% of images were acquired with pupil dilation.

Grading

All images in the development and clinical validation sets were graded by ophthalmologists for the presence of diabetic retinopathy, diabetic macular edema, and image quality using an annotation tool (eFigures 1 and 2 in the Supplement). Diabetic retinopathy severity (none, mild, moderate, severe, or proliferative) was graded according to the International Clinical Diabetic Retinopathy scale.[14] Referable diabetic macular edema was defined as any hard exudates within 1 disc diameter of the macula,[15] which is a proxy for macular edema when stereoscopic views are not available. Image quality was assessed by graders using the rubric in the "Grading Instructions" section in the Supplement. Images of excellent, good, and adequate quality were considered gradable.

Derivation

The 54 graders for the development set were US-licensed ophthalmologists or ophthalmology trainees in their last year of residency (postgraduate year 4). Each individual graded between 20 and 62 508 images (mean, 9774; median, 2021). Trainee performance was not worse than that of licensed

ophthalmologists, although only 3 trainees graded more than 1000 images. Intergrader reliability was measured for each physician using pairwise comparisons by taking the number of times a grader was in agreement with another grader over the total number of pairwise comparisons. Approximately 10% of the derivation set (128 175 images) were randomly selected to be overread by the same grader to determine intragrader reliability. All graders were paid for their work.

Validation
Graders who were US board-certified ophthalmologists with the highest rate of self-consistency were invited to grade the clinical validation sets, EyePACS-1 (n = 8) and Messidor-2 (n = 7). A simple majority decision (an image was classified as referable if ≥50% of ophthalmologists graded it referable) served as the reference standard for both referability and gradability. Graders were masked to judgments by other graders. (See "Grading Quality Control" section in the Supplement for more details.)

Development of the Algorithm
Deep learning is the process of training a neural network (a large mathematical function with millions of parameters) to perform a given task. The function computes diabetic retinopathy severity from the intensities of the pixels in a fundus image. Creating or "training" this function requires a large set of images for which the diabetic retinopathy severity is already known (training set). During the training process, the parameters of the neural network (mathematical function) are initially set to random values. Then, for each image, the severity grade given by the function is compared with the known grade from the training set, and parameters of the function are then modified slightly to decrease the error on that image. This process is repeated for every image in the training set many times over, and the function "learns" how to accurately compute the diabetic retinopathy severity from the pixel intensities of the image for all images in the training set. With the right training data, the result is a function general enough to compute diabetic retinopathy severity on new images. The network used in this study is a convolutional neural network that uses a function that first combines nearby pixels into local features, then aggregates those into global features. Although the algorithm does not explicitly detect lesions (eg, hemorrhages, microaneurysms), it likely learns to recognize them using the local features. The specific neural network used in this work is the Inception-v3 architecture proposed by Szegedy et al.[8]

Data were preprocessed according to a protocol described in the Supplement. The optimization algorithm used to train the network weights was a distributed stochastic gradient descent implementation by Dean et al.[16] To speed up the training, batch normalization[7] as well as preinitialization using weights from the same network trained to classify objects in the ImageNet data set[17] were used. Preinitialization also improved performance. A single network was trained to make multiple binary predictions, including whether the image was (1) moderate or worse diabetic retinopathy (ie, moderate, severe, or proliferative), (2) severe or worse dia-

betic retinopathy, (3) referable diabetic macular edema, or (4) fully gradable. Referable diabetic retinopathy was defined as any image that fulfilled either criterion 1, criterion 3, or both.

The performance of the algorithm was measured by the area under the receiver operating curve (AUC) generated by plotting sensitivity vs 1 – specificity. Because the network in this study had a large number of parameters (22 million), an early stopping criteria[18] (that stops training when peak AUC is reached on a separate tuning set) was used to terminate training before convergence. The development set was divided into 2 parts: (1) training: 80% of the data was used to optimize the network weights and (2) tuning: 20% of the data was used to optimize hyperparameters (such as early stopping for training, image preprocessing options). An ensemble[19] of 10 networks trained on the same data was used, and the final prediction was computed by a linear average over the predictions of the ensemble.

Evaluating the Algorithm
The trained neural network generates a continuous number between 0 and 1 for referable diabetic retinopathy and other diabetic retinopathy classifications, corresponding to the probability of that condition being present in the image. Receiver operating curves were plotted by varying the operating threshold and 2 operating points for the algorithm were selected from the development set. The first operating point approximated the specificity of the ophthalmologists in the derivation set for detecting referable diabetic retinopathy (approximately 98%) and allowed for better comparison between the algorithm's performance and that of the 7 or 8 ophthalmologists that graded the validation set. The second operating point corresponded to a sensitivity of 97% for detecting referable diabetic retinopathy because a high sensitivity is a prerequisite in a potential screening tool.

Statistical Analysis and Performance Comparison on Clinical Validation Sets
Based on the 2 operating points, 2 × 2 tables were generated to characterize the sensitivity and specificity of the algorithm with respect to the reference standard, which was defined as the majority decision of the ophthalmologists' readings based on all available grades. The 95% confidence intervals for the sensitivity and specificity of the algorithm at the 2 operating points were calculated to be "exact" Clopper-Pearson intervals,[20] which corresponded to separate 2-sided confidence intervals with individual coverage probabilities of $sqrt(0.95) \approx 0.975$. The 95% confidence intervals for the intragrader and intergrader reliabilities are z confidence intervals.

Statistical significance and simultaneous 2-sided confidence intervals were computed using the StatsModels version 0.6.1 and SciPy version 0.15.1 python packages.

Subsampling Experiments
Experiments to understand the relationship between the amount of development data on the performance of the resulting algorithms also were conducted. To understand the effect of reducing the number of images in the training set,

Table. Baseline Characteristics[a]

Characteristics	Development Data Set	EyePACS-1 Validation Data Set	Messidor-2 Validation Data Set
No. of images	128 175	9963	1748
No. of ophthalmologists	54	8	7
No. of grades per image	3-7	8	7
Grades per ophthalmologist, median (interquartile range)	2021 (304-8366)	8906 (8744-9360)	1745 (1742-1748)
Patient demographics			
No. of unique individuals	69 573[b]	4997	874
Age, mean (SD), y	55.1 (11.2)[c]	54.4 (11.3)	57.6 (15.9)
Female, No./total (%) among images for which sex was known	50 769/84 734 (59.9)[c]	5463/8784 (62.2)	743/1745 (42.6)
Image quality distribution			
Fully gradable, No./total (%) among images for which image quality was assessed	52 311/69 598 (75.1)[d]	8788/9946 (88.4)	1745/1748 (99.8)
Disease severity distribution classified by majority decision of ophthalmologists (reference standard)			
Total images for which both diabetic retinopathy and diabetic macular edema were assessed, No. (%)	118 419 (100)	8788 (100)	1745 (100)
No diabetic retinopathy	53 759 (45.4)	7252 (82.5)	1217 (69.7)
Mild diabetic retinopathy	30 637 (25.9)	842 (9.6)	264 (15.1)
Moderate diabetic retinopathy	24 366 (20.6)	545 (6.2)	211 (12.1)
Severe diabetic retinopathy	5298 (4.5)	54 (0.6)	28 (1.6)
Proliferative diabetic retinopathy	4359 (3.7)	95 (1.1)	25 (1.4)
Referable diabetic macular edema	18 224 (15.4)	272 (3.1)	125 (7.2)
Referable diabetic retinopathy[e]	33 246 (28.1)	683 (7.8)	254 (14.6)

[a] Summary of image characteristics and available demographic information in the development and clinical validation data sets (EyePACS-1 and Messidor-2). Abnormal images were oversampled for the development set for algorithm training. The clinical validation sets were not enriched for abnormal images.

[b] Unique patient codes (deidentified) were available for 89.3% of the development set (n = 114 398 images).

[c] Individual-level data including age and sex were available for 66.1% of the development set (n = 84 734 images).

[d] Image quality was assessed for a subset of the development set.

[e] Referable diabetic retinopathy, defined as the presence of moderate and worse diabetic retinopathy and/or referable diabetic macular edema according to the International Clinical Diabetic Retinopathy Scale,[14] was calculated for each ophthalmologist before combining them using a majority decision. The 5-point grades represent the grade that received the highest number of votes for diabetic retinopathy alone. Hence, the sum of moderate, severe, and proliferative diabetic retinopathy for the 5-point grade differs slightly from the count of referable diabetic retinopathy images.

images were sampled at rates of 0.2%, 2%, and N × 10%, in which N ranged from 1 to 10; a new algorithm was trained for each data set; and its performance was measured on a fixed tuning set. To understand the effect of reducing the number of grades per image, 2 experiments were run: (1) training: grades in the training set were subsampled at rates of N × 20%, in which N ranged from 0 to 5, with a restriction that the minimum number of grades to sample per image was 1. A new algorithm was trained for each N, and its performance was measured on a fixed tuning set with all the available grades and (2) tuning: grades in the tuning set (used to measure performance) were sampled using the same procedure as in the training experiment. The training set and the algorithm were fixed and used all available grades.

Results

Patient demographics and image characteristics are summarized in the **Table**. The development set included 128 175 images, of which 118 419 were assessed for referable diabetic retinopathy and 33 246 (28.1%) had referable diabetic reti-

nopathy. Each image was graded by ophthalmologists between 3 and 7 times. The EyePACS-1 and Messidor-2 clinical validation sets consisted of 9963 images (8788 fully gradable; 683 [7.8%] referable) and 1748 images (1745 fully gradable; 254 [14.6%] referable), respectively. Image quality was assessed only for a subset of the development set, and fully gradable images ranged from 75.1% (52 311 of 69 598 images that were assessed for image quality) for the development set to 99.8% (1745 of 1748) for the Messidor-2 validation set (Table and **Figure 1**).

In the development set, intragrader reliability among the ophthalmologists could be assessed among 16 graders who had graded a sufficient volume of repeat images. The mean intragrader reliability for referable diabetic retinopathy of these graders was 94.0% (95% CI, 91.2%-96.8%). Intergrader reliability could be assessed on 26 graders. The mean intergrader reliability for these graders was 95.5% (95% CI, 94.0%-96.9%).

In the validation sets, a total of 8 grades per image were obtained for the EyePACS-1 data set and 7 grades per image for Messidor-2. The mean intragrader reliability for referable diabetic retinopathy for EyePACS-1 was 95.8% (95% CI,

Figure 1. EyePACS-1 and Messidor-2 Clinical Validation Sets for Detection of Diabetic Retinopathy and All-Cause Referable Diabetic Retinopathy
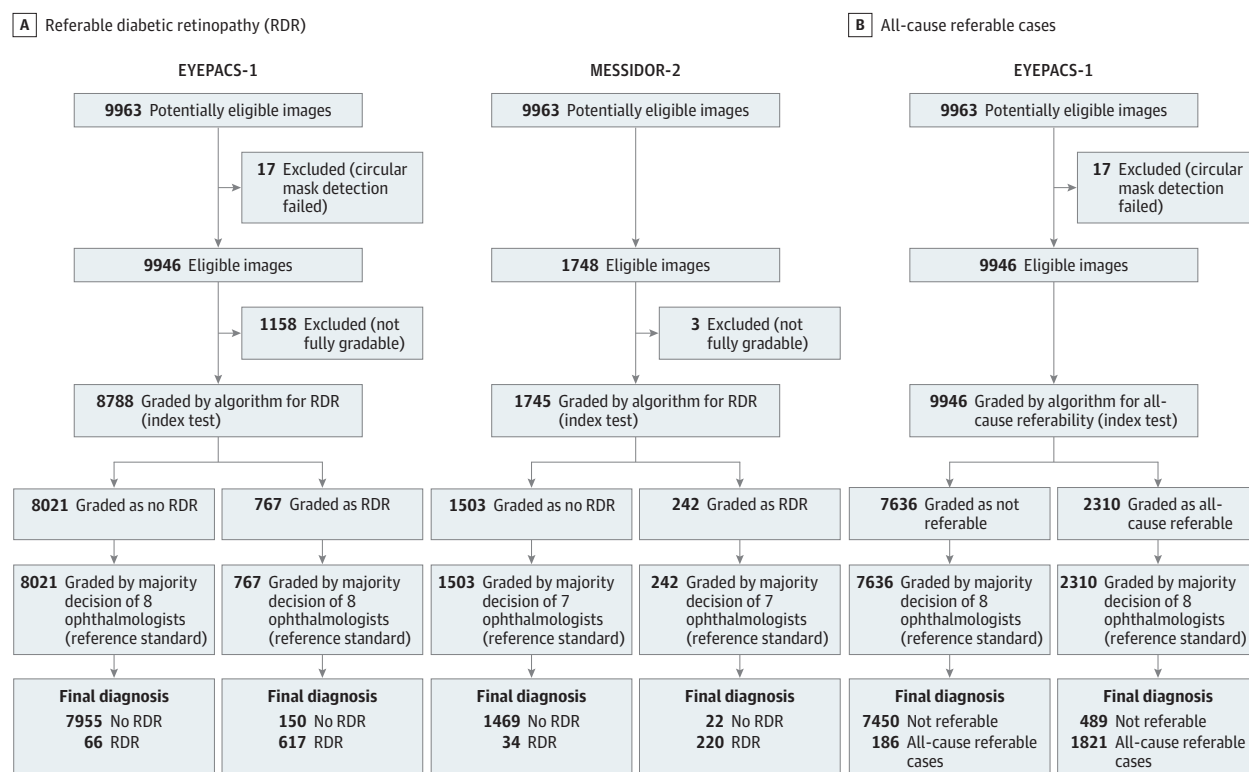
A, Referable diabetic retinopathy, defined as moderate or worse diabetic retinopathy or referable diabetic macular edema. B, All-cause referable cases, defined as moderate or worse diabetic retinopathy, referable diabetic macular edema, or ungradable image quality.

92.8%-98.7%). Intragrader reliability was not assessed for Messidor-2. The mean intergrader reliability was 95.9% (95% CI, 94.0%-97.8%) for EyePACS-1 and 94.6% (95% CI, 93.0%-96.1%) for Messidor-2.

On EyePACS-1, the mean agreement among ophthalmologists on referable diabetic retinopathy images was 77.7% (SD, 16.3%), with complete agreement on 19.6% of the referable cases. On nonreferable images, the average agreement was 97.4% (SD, 7.3%), with complete agreement on 85.6% of the nonreferable cases. On Messidor-2, the average agreement among ophthalmologists on referable diabetic retinopathy images was 82.4% (SD, 16.9%), with complete agreement on 37.8% of the referable cases. On nonreferable images, the average agreement was 96.3% (SD, 9.9%), with complete agreement on 85.1% of the nonreferable cases. The distribution of agreement among ophthalmologists is reported in eFigure 3 in the Supplement.

Figure 2 summarizes the performance of the algorithm in detecting referable diabetic retinopathy in the EyePACS-1 and Messidor-2 validation data sets for fully gradable images. For referable diabetic retinopathy, the algorithm achieved an AUC of 0.991 (95% CI, 0.988-0.993) on EyePACS-1 and an AUC of 0.990 (95% CI, 0.986-0.995) on Messidor-2. Using the first operating cut point with high specificity, approximating the specificity of ophthalmologists in the development set, on EyePACS-1, the algorithm's sensitivity was 90.3% and

specificity was 98.1%. In Messidor-2, the sensitivity was 87.0% and specificity was 98.5%.

A second operating point for the algorithm was evaluated, which had a high sensitivity on the development set, reflecting an output that would be used for a screening tool. Using this operating point, on EyePACS-1, the algorithm had a sensitivity of 97.5% (95% CI, 95.8%-98.7%) and a specificity of 93.4% (95% CI, 92.8%-94.0%). In Messidor-2, the sensitivity was 96.1% (95% CI, 92.4%-98.3%) and the specificity was 93.9% (95% CI, 92.4%-95.3%). Given an approximately 8% prevalence of referable diabetic retinopathy (on a per-image basis [Table]), these findings correspond to a negative predictive value of 99.8% for EyePACS-1 and 99.6% for Messidor-2.

The algorithm performance in making all-cause referable predictions, defined as moderate or worse diabetic retinopathy, referable diabetic macular edema, or ungradable images (Figure 3), was also evaluated using the EyePACS-1 data set. The Messidor-2 data set had only 3 ungradable images, so it was omitted from this analysis. For this task, the algorithm achieved an AUC of 0.974 (95% CI, 0.971-0.978). At the first (high-specificity) operating point, the algorithm had a sensitivity of 90.7% (95% CI, 89.2%-92.1%) and a specificity of 93.8% (95% CI, 93.2%-94.4%). At the second (high-sensitivity) operating point, the algorithm had a sensitivity of 96.7% (95% CI, 95.7%-97.5%) and a specificity of 84.0% (95% CI, 83.1%-85.0%).

Figure 2. Validation Set Performance for Referable Diabetic Retinopathy

A EyePACS-1: AUC, 99.1%; 95% CI, 98.8%-99.3%

B Messidor-2: AUC, 99.0%; 95% CI, 98.6%-99.5%
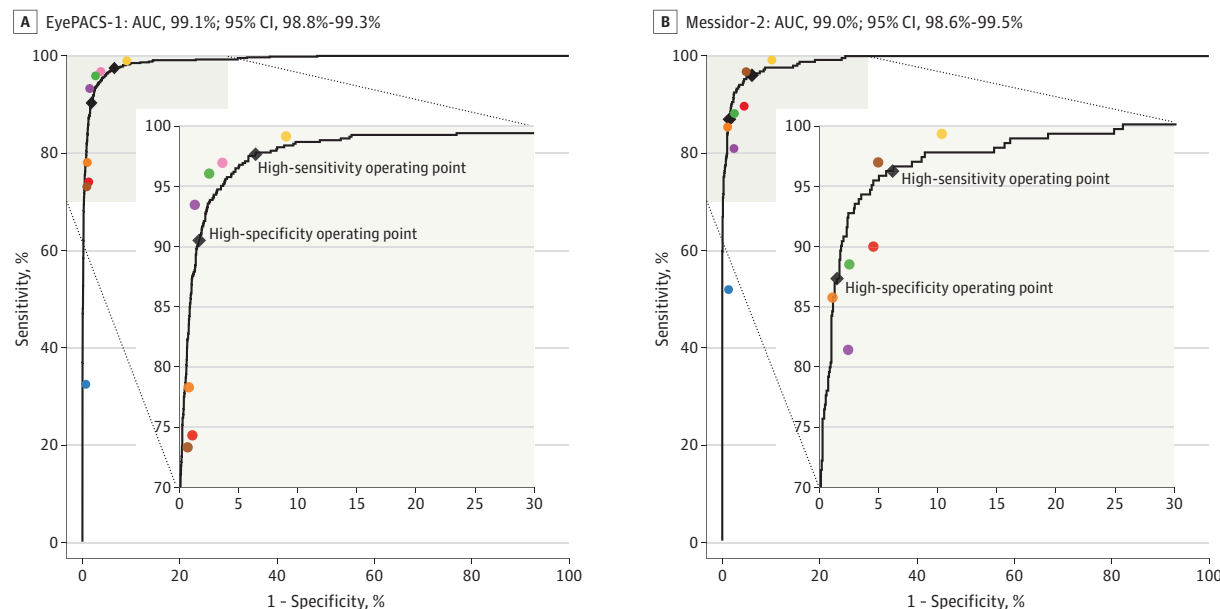
Performance of the algorithm (black curve) and ophthalmologists (colored circles) for the presence of referable diabetic retinopathy (moderate or worse diabetic retinopathy or referable diabetic macular edema) on A, EyePACS-1 (8788 fully gradable images) and B, Messidor-2 (1745 fully gradable images). The black diamonds on the graph correspond to the sensitivity and specificity of the algorithm at the high-sensitivity and high-specificity operating points. In A, for the high-sensitivity operating point, specificity was 93.4% (95% CI, 92.8%-94.0%) and sensitivity was 97.5% (95% CI, 95.8%-98.7%); for the high-specificity operating point, specificity was 98.1% (95% CI, 97.8%-98.5%) and sensitivity was 90.3% (95% CI, 87.5%-92.7%). In B, for the high-sensitivity operating point, specificity was 93.9% (95% CI, 92.4%-95.3%) and sensitivity was 96.1% (95% CI, 92.4%-98.3%); for the high-specificity operating point, specificity was 98.5% (95% CI, 97.7%-99.1%) and sensitivity was 87.0% (95% CI, 81.1%-91.0%). There were 8 ophthalmologists who graded EyePACS-1 and 7 ophthalmologists who graded Messidor-2. AUC indicates area under the receiver operating characteristic curve.

Additional sensitivity analyses were conducted for several subcategories: (1) detecting moderate or worse diabetic retinopathy only; (2) detecting severe or worse diabetic retinopathy only; (3) detecting referable diabetic macular edema only; (4) image quality; and (5) referable diabetic retinopathy on 2 data sets, each restricted to mydriatic and nonmydriatic images, respectively. For each subcategory 1 through 4, the algorithm achieved high sensitivity and specificity (see section on "Performance on Individual Diabetic Retinopathy Subtypes, Image Quality," eTable 1, and eFigure 4 in the Supplement). For example, for the EyePACS-1 data set, at the first operating point for moderate or worse diabetic retinopathy, the algorithm had a sensitivity of 90.1% (95% CI, 87.2%-92.6%) and specificity of 98.2% (95% CI, 97.8%-98.5%). For severe or worse diabetic retinopathy only at the first operating point, the algorithm had a sensitivity of 84.0% (95% CI, 75.3%-90.6%) and specificity of 98.8% (95% CI, 98.5%-99.0%). For diabetic macular edema only, the algorithm's sensitivity was 90.8% (95% CI, 86.1%-94.3%) and specificity was 98.7% (95% CI, 98.4%-99.0%). The algorithm's performance on mydriatic images was very close to its performance on nonmydriatic images (and both were similar to the overall algorithm performance; see eTable 2 in the Supplement).

Multiple networks with varying number of images and grades per image were trained to determine how smaller training data sets related to the performance of the trained algorithms. In the first subsampling experiment (**Figure** 4A), the

effects of data set size on algorithm performance were examined and shown to plateau at around 60 000 images (or approximately 17 000 referable images). In the second experiment (Figure 4B) on subsampling grades, 2 trends emerged: (1) increasing the number of grades per image on the training set did not yield an increase in relative performance (31.6% absolute difference) and (2) using only 1 grade per image on the tuning set led to a decline of 36% in performance compared with using all the available grades on the tuning set (an average of 4.5 grades), and that performance steadily increased as more grades were made available for the tuning set. This suggests that additional grading resources should be devoted to grading the tuning set (on which evaluation is done), which improves the quality of the reference standard and the algorithm performance.

Discussion

These results demonstrate that deep neural networks can be trained, using large data sets and without having to specify lesion-based features, to identify diabetic retinopathy or diabetic macular edema in retinal fundus images with high sensitivity and high specificity. This automated system for the detection of diabetic retinopathy offers several advantages, including consistency of interpretation (because a machine will make the same prediction on a specific image every time), high

sensitivity and specificity, and near instantaneous reporting of results. In addition, because an algorithm can have multiple operating points, its sensitivity and specificity can be tuned to match requirements for specific clinical settings, such as high sensitivity for a screening setting. In this study, sensitivities of 97.5% and 96.1% were achieved.

Automated and semiautomated diabetic retinopathy evaluation has been previously studied by other groups. Abràmoff et al[4] reported a sensitivity of 96.8% at a specificity of 59.4% for detecting referable diabetic retinopathy on the publicly available Messidor-2 data set.[9] Solanki et al[12] reported a sensitivity of 93.8% at a specificity of 72.2% on the same data set. A study by Philip et al[21] reported a sensitivity of 86.2% at a specificity of 76.8% for predicting disease vs no disease on their own data set of 14 406 images. In a recent Kaggle machine-learning competition,[22] deep learning was used to predict the diabetic retinopathy grade only (no diabetic macular edema prediction). The winning entry had a performance that was comparable with an ophthalmologist grading the wrong eye of the same patient[23] and higher than the agreement demonstrated between general physicians and ophthalmologists.[24] Although there are differences in the data set and reference standards compared with the previous studies, the present study extends this body of work by using deep convolutional neural networks and a large data set with multiple grades per image to generate an algorithm with 97.5% sensitivity and 93.4% specificity (at the screening operating point, which had been selected to have high sensitivity). When screening populations with substantial disease, achieving both high sensitivity and high specificity is critical to minimize both false-positive and false-negative results.

In the future, based on the observations from this study, the development of similar high-performing algorithms for medical imaging using deep learning has 2 prerequisites. First,

there must be collection of a large developmental set with tens of thousands of abnormal cases. While performance on the tuning set saturated at 60 000 images, additional gains might be

Figure 3. Validation Set Performance for All-Cause Referable Diabetic Retinopathy in the EyePACS-1 Data Set (9946 Images)
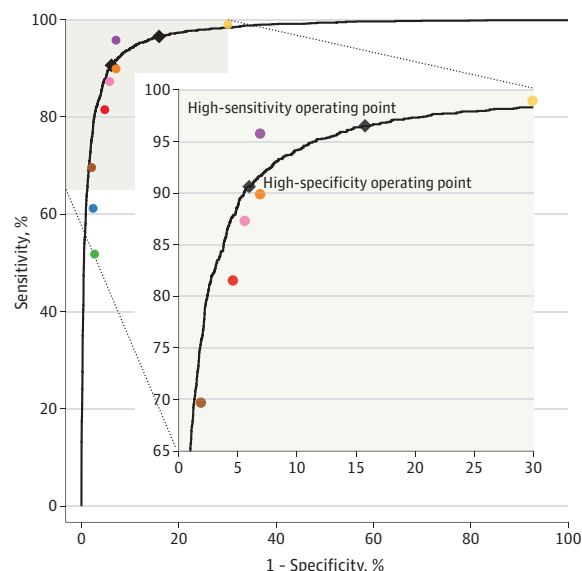
Performance of the algorithm (black curve) and ophthalmologists (colored circles) for all-cause referable diabetic retinopathy, defined as moderate or worse diabetic retinopathy, diabetic macular edema, or ungradable image. The black diamonds highlight the performance of the algorithm at the high-sensitivity and high-specificity operating points. For the high-sensitivity operating point, specificity was 84.0% (95% CI, 83.1%-85.0%) and sensitivity was 96.7% (95% CI, 95.7%-97.5%). For the high-specificity operating point, specificity was 93.8% (95% CI, 93.2%-94.4%) and sensitivity was 90.7% (95% CI, 89.2%-92.1%). There were 8 ophthalmologists who graded EyePACS-1. The area under the receiver operating characteristic curve was 97.4% (95% CI, 97.1%-97.8%).

Figure 4. Model Performance for Detection of Diabetic Retinopathy and All-Cause Referable Diabetic Retinopathy as a Function of the Number of Images and Grades Used in the Development Set
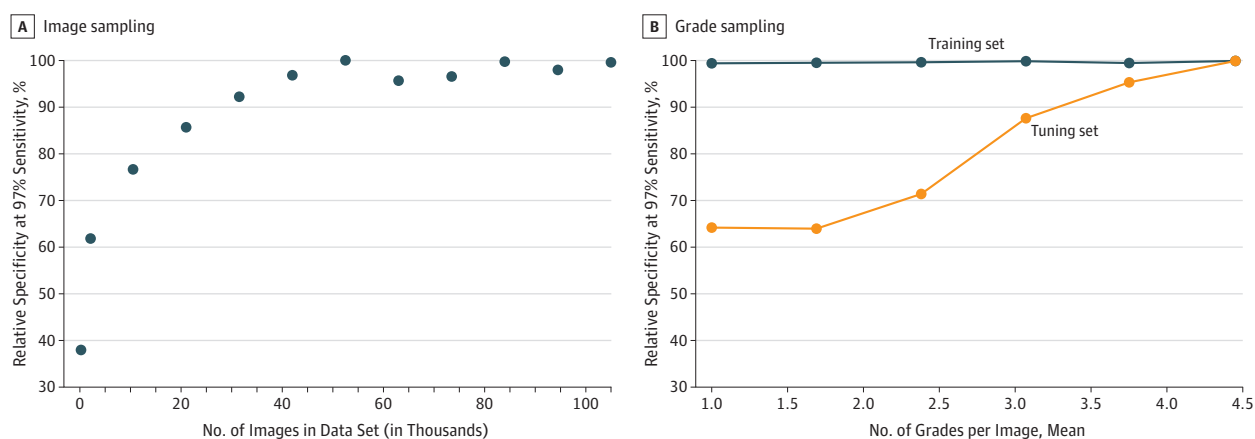
A, Model performance on the tuning set (24 360 images) as a function of number of images used in training (varying between 208 and 103 698). Relative specificity was calculated as specificity at each data set size/specificity at maximum data set size. The training set is the data on which the neural networks are learned, and the tuning set is used to evaluate and optimize the performance. Image subsets were chosen by randomly sampling from the set of all images at rates (determined a priori) of (0.2%, 2%, 10%, 20%, 30%, …, 100%). Each set of images includes all images in the smaller sets. B, Model performance on the tuning set as a function of grades per image. Relative specificity was calculated as specificity at each grade per image/specificity at maximum grade per image.

achieved by increasing the diversity of training data (ie, data from new clinics). Second, data sets used to measure final performance (tuning and the clinical validation data sets) should have multiple grades per image. This provides a more reliable measure of a model's final predictive ability. Although intergrader variability is a well-known issue in many settings in which human interpretation is used as the reference standard (as opposed to "hard" outcomes like mortality), such as in radiology[25] or pathology,[26] diseases with unambiguous interpretations may not require additional grades per image.

There are limitations to this system. The reference standard used for this study was the majority decision of all ophthalmologist graders. This means the algorithm may not perform as well for images with subtle findings that a majority of ophthalmologists would not identify. Another fundamental limitation arises from the nature of deep networks, in which the neural network was provided with only the image and associated grade, without explicit definitions of features (eg, microaneurysms, exudates). Because the network "learned" the features that were most predictive for the referability implicitly, it is possible that the algorithm is using features previously unknown to or ignored by humans. Although this study used images from a variety of clinical settings (hundreds of clinical sites: 3 in India, hundreds in the United States, and 3 in France) with a range of camera types to mitigate the risk that the algorithm is using anomalies in data acquisition to make predictions, the exact features being used are still unknown. Understanding what a deep neural net uses to make predictions is a very active area of research within the larger machine learning community. Another open question is whether the design of the user interface and the online setting for grading used by ophthalmologists has any influence on their performance relative to a clinical setting. This needs further experiments to address. The algorithm has been trained to identify only diabetic retinopathy and diabetic macular edema. It may miss nondiabetic retinopathy lesions that it was not trained to identify. Hence, this algorithm is not a replacement for a comprehensive eye examination, which has many components, such as visual acuity, refraction, slitlamp examination, and eye pressure measurements. In addition, further validation of the algorithm is necessary in a data set in which the gold standard was not a consensus of experts who participated in the derivation of the algorithm.

Conclusions

In this evaluation of retinal fundus photographs from adults with diabetes, an algorithm based on deep machine learning had high sensitivity and specificity for detecting referable diabetic retinopathy. Further research is necessary to determine the feasibility of applying this algorithm in the clinical setting and to determine whether use of the algorithm could lead to improved care and outcomes compared with current ophthalmologic assessment.

REFERENCES

1. Zhang X, Saaddine JB, Chou C-F, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. *JAMA*. 2010;304(6):649-656.

2. Raman R, Rani PK, Reddi Rachepalle S, et al. Prevalence of diabetic retinopathy in India: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study report 2. *Ophthalmology*. 2009;116(2):311-318.

3. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. *Expert Rev Ophthalmol*. 2012;7(5):417-439.

4. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351-357.

5. Mookiah MRK, Acharya UR, Chua CK, Lim CM, Ng EYK, Laude A. Computer-aided diagnosis of diabetic retinopathy: a review. *Comput Biol Med*. 2013;43(12):2136-2155.

6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

7. Ioffe S, Szegedy C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. March 2015. http://arxiv.org/pdf/1502.03167v3.pdf. Accessed November 14, 2016.

8. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. *Rethinking the Inception Architecture for Computer Vision*. December 2015. http://arxiv.org/pdf/1512.00567v3.pdf. Accessed November 14, 2016.

9. Decencière E, Etienne D, Xiwei Z, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereol*. 2014;33(3):231-234. doi:10.5566/ias.1155

10. Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B, Roux C. Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Trans Med Imaging*. 2008;27(9):1230-1241.

11. Giancardo L, Meriaudeau F, Karnowski TP, et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Med Image Anal*. 2012;16(1):216-226.

12. Solanki K, Ramachandra C, Bhat S, Bhaskaranand M, Nittala MG, Sadda SR. EyeArt: automated, high-throughput, image analysis for diabetic retinopathy screening [abstract]. *Invest Ophthalmol Vis Sci*. 2015;56(7):2331130. http://iovs.arvojournals.org/article.aspx?articleid=2331130. Accessed November 14, 2016.

13. Quellec G, Lamard M, Abràmoff MD, et al. A multiple-instance learning framework for diabetic retinopathy screening. *Med Image Anal*. 2012;16(6):1228-1240.

14. American Academy of Ophthalmology. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table*. http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf. Accessed October 14, 2016.

15. Bresnick GH, Mukamel DB, Dickinson JC, Cole DR. A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy. *Ophthalmology*. 2000;107(1):19-24.

16. Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks. *Adv Neural Inf Processing Syst*. 2012:1223-1231.

17. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211-252.

18. Caruana R, Lawrence S, Giles L. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. *Adv Neural Inf Processing Syst*. 2001:402-408.

19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Processing Syst*. 2012:1097-1105.

20. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4):404-413.

21. Philip S, Fleming AD, Goatman KA, et al. The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme. *Br J Ophthalmol*. 2007;91(11):1512-1517.

22. Kaggle Diabetic Retinopathy Detection competition. https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed June 30, 2016.

23. Kaggle diabetic retinopathy competition forum. https://www.kaggle.com/c/diabetic-retinopathy-detection/forums/t/15605/human-performance-on-the-competition-data-set. Accessed June 30, 2016.

24. Verma L, Prakash G, Tewari HK, Gupta SK, Murthy GV, Sharma N. Screening for diabetic retinopathy by non-ophthalmologists: an effective public health tool. *Acta Ophthalmol Scand*. 2003;81(4):373-377.

25. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331(22):1493-1499.

26. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*. 2015;313(11):1122-1132.