JAMA Ophthalmology | Original Investigation

Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks

James M. Brown, PhD; J. Peter Campbell, MD, MPH; Andrew Beers, BA; Ken Chang, MSE; Susan Ostmo, MS; R. V. Paul Chan, MD; Jennifer Dy, PhD; Deniz Erdogmus, PhD; Stratis Ioannidis, PhD; Jayashree Kalpathy-Cramer, PhD; Michael F. Chiang, MD; for the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium

IMPORTANCE Retinopathy of prematurity (ROP) is a leading cause of childhood blindness worldwide. The decision to treat is primarily based on the presence of plus disease, defined as dilation and tortuosity of retinal vessels. However, clinical diagnosis of plus disease is highly subjective and variable.

OBJECTIVE To implement and validate an algorithm based on deep learning to automatically diagnose plus disease from retinal photographs.

DESIGN, SETTING, AND PARTICIPANTS A deep convolutional neural network was trained using a data set of 5511 retinal photographs. Each image was previously assigned a reference standard diagnosis (RSD) based on consensus of image grading by 3 experts and clinical diagnosis by 1 expert (ie, normal, pre-plus disease, or plus disease). The algorithm was evaluated by 5-fold cross-validation and tested on an independent set of 100 images. Images were collected from 8 academic institutions participating in the Imaging and Informatics in ROP (i-ROP) cohort study. The deep learning algorithm was tested against 8 ROP experts, each of whom had more than 10 years of clinical experience and more than 5 peer-reviewed publications about ROP. Data were collected from July 2011 to December 2016. Data were analyzed from December 2016 to September 2017.

EXPOSURES A deep learning algorithm trained on retinal photographs.

MAIN OUTCOMES AND MEASURES Receiver operating characteristic analysis was performed to evaluate performance of the algorithm against the RSD. Quadratic-weighted κ coefficients were calculated for ternary classification (ie, normal, pre-plus disease, and plus disease) to measure agreement with the RSD and 8 independent experts.

RESULTS Of the 5511 included retinal photographs, 4535 (82.3%) were graded as normal, 805 (14.6%) as pre-plus disease, and 172 (3.1%) as plus disease, based on the RSD. Mean (SD) area under the receiver operating characteristic curve statistics were 0.94 (0.01) for the diagnosis of normal (vs pre-plus disease or plus disease) and 0.98 (0.01) for the diagnosis of plus disease (vs normal or pre-plus disease). For diagnosis of plus disease in an independent test set of 100 retinal images, the algorithm achieved a sensitivity of 93% with 94% specificity. For detection of pre-plus disease or worse, the sensitivity and specificity were 100% and 94%, respectively. On the same test set, the algorithm achieved a quadratic-weighted κ coefficient of 0.92 compared with the RSD, outperforming 6 of 8 ROP experts.

CONCLUSIONS AND RELEVANCE This fully automated algorithm diagnosed plus disease in ROP with comparable or better accuracy than human experts. This has potential applications in disease detection, monitoring, and prognosis in infants at risk of ROP.

JAMA Ophthalmol. 2018;136(7):803-810. doi:10.1001/jamaophthalmol.2018.1934
Published online May 2, 2018.

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Group Information: The members of the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium are listed at the end of this article.

Corresponding Author: Michael F. Chiang, MD, Department of Ophthalmology, Casey Eye Institute, Oregon Health and Science University, 3375 SW Terwilliger Blvd, Portland, OR 97239 (chiangm@ohsu.edu).

Retinopathy of prematurity (ROP) is a proliferative retinal vascular disease that affects approximately two-thirds of premature infants weighing fewer than 1250 g at birth. Most cases of ROP are mild and resolve without intervention within several months. However, 5% to 10% of cases progress to severe ROP, which can lead to retinal detachment and permanent blindness if untreated. A major challenge is that clinical ROP diagnosis is based solely on the appearance of retinal vessels on dilated ophthalmoscopic examination at the neonatal intensive care unit bedside, which is highly subjective and qualitative. The most critical feature of severe, treatment-requiring ROP is the presence of plus disease, which was defined during the 1980s by an international consensus panel as arterial tortuosity and venous dilation of the posterior retinal vessels that is greater than or equal to that found in a standard published retinal photograph.[1,2] In 2005, a revised international consensus panel established a 3-tier grading classification of plus disease (ie, normal, pre-plus disease, and plus disease) to capture an intermediate level of severity as an additional prognostic indicator.[3-5] Several major studies funded by the National Institutes of Health[1,4] and several other clinical trials[6] have shown that severe ROP (characterized by plus disease) may be effectively treated with laser photocoagulation[1,4] or with intravitreal injection of pharmacological agents, such as bevacizumab.[6] Therefore, it is essential to diagnose plus disease in an accurate and timely manner.

Retinopathy of prematurity remains a leading cause of childhood blindness worldwide. There are several challenges to delivery of care: (1) clinical diagnosis is highly variable, and high interobserver inconsistency on plus disease diagnosis, even among ROP experts, has been well-documented[7,8]; (2) the number of ophthalmologists and neonatologists willing and able to manage ROP is insufficient because of logistical difficulties, the extensive training process, time-consuming examination, and significant malpractice liability[9-12]; and (3) the incidence of ROP worldwide is rising because of advances in neonatology.[13] These challenges have stimulated research in developing quantitative and objective approaches to ROP diagnosis using computer-based image analysis (CBIA).[14-18] Although multiple groups have developed CBIA systems for plus disease diagnosis in ROP, no automated systems have demonstrated diagnostic performance equivalent to practicing clinicians.[14] A fully automated, validated CBIA system would improve quality of care by providing diagnostic assistance to clinicians and could improve accessibility of care by creating potential for large-scale automated screening systems.

Deep learning (DL) has become the state-of-the-art solution in a wide range of CBIA problems.[19] Convolutional neural networks (CNNs) have been successfully used for automated diagnosis of skin cancer,[20] glioma,[21] lymph node metastases,[22] macular degeneration,[23-25] and diabetic retinopathy.[26,27] Convolutional neural networks have also been used to predict a range of cardiovascular risk factors from retinal fundus photographs that were previously not thought to be quantifiable.[28] Furthermore, they have shown promising results for 2-level diagnosis of plus disease in ROP.[29] The purpose of this article is to implement and evaluate a CNN-based DL approach for 3-level diagnosis (ie, normal, pre-plus disease, and plus disease) in ROP. We trained CNNs on a large data set of clinical ROP images from 8 different institutions and compared their diagnostic performance with expert human graders.

Key Points

Question Can an algorithm based on deep learning achieve expert-level performance at diagnosing plus disease in retinopathy of prematurity?

Finding In this technology evaluation study including 5511 retinal photographs, using 5-fold cross-validation, the algorithm achieved mean areas under the receiver operating characteristic curve of 0.94 and 0.99 for the diagnoses of normal and plus disease, respectively. On an independent test set of 100 images, the algorithm achieved 91% accuracy and a quadratic-weighted κ coefficient of 0.92, outperforming 6 of 8 retinopathy of prematurity experts.

Meaning These findings suggest the proposed algorithm can objectively diagnose plus disease with a proficiency comparable with human experts.

Methods

This study was approved by the institutional review board at the coordinating center (Oregon Health and Science University, Portland) and at each of 8 study centers (Columbia University, New York, New York; University of Illinois at Chicago; William Beaumont Hospital, Royal Oak, Michigan; Children's Hospital Los Angeles, Los Angeles, California; Cedars-Sinai Medical Center, Los Angeles, California; University of Miami, Miami, Florida; Weill Cornell Medical Center, New York, New York; and Asociacion para Evitar la Ceguera en Mexico, Mexico City, Mexico). This study was conducted in accordance with the Declaration of Helsinki.[30] Written informed consent was obtained from parents of all infants enrolled.

Data Sets

Training, validation, and test data sets were created from a database of almost 6000 deidentified posterior retinal images obtained using a commercially available camera (RetCam; Natus Medical Incorporated) as part of the multicenter Imaging and Informatics in Retinopathy of Prematurity (i-ROP) cohort study. A standard imaging protocol was used by all 8 study centers, and the images were obtained between July 2011 and December 2016. Although images were obtained in 5 standard fields of view (ie, posterior, nasal, temporal, superior, and inferior), only posterior images were used in this analysis.

Image Grading

A reference standard diagnosis (RSD) was assigned to each image using previously published methods[31] based on independent image-based diagnoses by 3 trained graders (2 ophthalmologists and 1 study coordinator) and the clinical diagnosis (obtained by full evaluation, including dilated ophthalmoscopic examination) by an expert ophthalmologist. Images

Table. Breakdown of Training and Validation Data Sets[a]

	No.									
	Training Data Set					Validation Data Set				
Split	Patients	Eyes	Normal	Pre-Plus Disease	Plus Disease	Patients	Eyes	Normal	Pre-Plus Disease	Plus Disease
1	718	1409	3668	653	148	180	353	867	151	24
2	718	1405	3640	673	140	180	357	895	131	32
3	710	1392	3552	628	125	188	370	983	176	47
4	713	1400	3580	597	145	185	362	955	207	27
5	716	1408	3673	642	126	182	354	862	162	46

[a] Each training/validation split constitutes an approximate 80:20 split of the 5511 images, retaining the underlying distribution of plus disease prevalence.

were classified as normal, pre-plus disease, or plus disease. Of the 5511 included retinal photographs, 4535 (82.3%) were graded as normal, 805 (14.6%) as pre-plus disease, and 172 (3.1%) as plus disease, based on the RSD. The RSD was used as the basis for training a CNN. Images were excluded if at least 2 of 3 image graders labeled them as unacceptable for diagnosis or if there was stage 4 or 5 ROP (ie, partial or total retinal detachment). In these advanced stages, diagnosis of plus disease for ROP screening is less relevant, and retinal blood vessels are difficult to visualize.

Algorithm Development

The algorithm used 2 neural network architectures, which are complex functions designed to receive images as input (ie, a grid of pixel intensity values), and were trained to produce some desired output. This training process involved presenting the network with corresponding RSDs, which were used to adjust the network's numerous internal parameters to output the correct diagnoses. Both networks used by our algorithm were CNNs, which are highly specialized for image data. Convolutional neural networks operate by learning and applying a series of filters that emphasize image features that are relevant to the task at hand. The first of the CNNs used by our algorithm was a vessel segmentation network, which was trained to output a new image with pixel intensities ranging between 0 and 1. Each pixel value represents the probability that it belongs to a retinal vessel. This process effectively eliminates variations in pigmentation, illumination, and nonvascular pathology, which are commonly observed in images from patients with ROP. In this work, we used the U-Net architecture[32] (eMethods 1 in the Supplement).

The second CNN was trained to diagnose plus disease from the preprocessed images. Through a series of alternating filtering and down sampling operations, a classification network reduced images to a set of features, which were transformed into 3 values representing the probability of that image corresponding to normal, pre-plus disease, or plus disease. We used the Inception version 1 architecture by Szegedy et al,[33] which was pretrained on the ImageNet database of 1.2 million images from 1000 classes.[34] This process of transfer learning has been shown to improve classification performance because of the network having learned highly generalizable image features from an unrelated but large and highly diverse data set of images (eMethods 2 in the Supplement).[35]

Evaluation

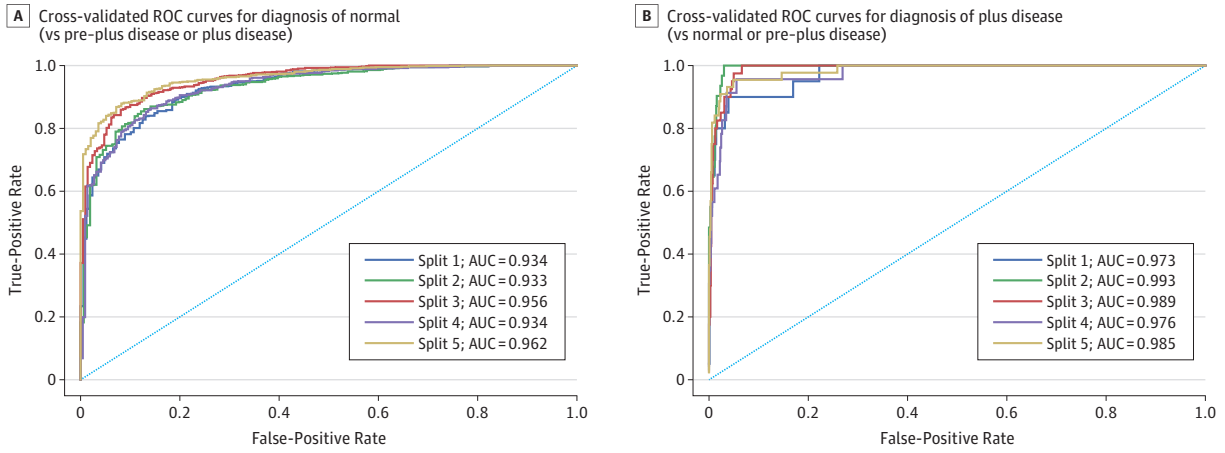
The data set was subdivided into 5 near-equal parts and used to train 5 separate classification CNNs (ie, 5-fold cross-validation). The data were divided to ensure that images acquired from the same patient (eg, left and right eyes or from multiple sessions) were not split across training and validation data. A detailed breakdown of the training and validation sets is provided in the **Table**. Each CNN was trained on 4 splits (80%) and tested on the remaining split (20%) to assess the algorithm's ability to generalize to previously unseen images from different patients. The cross-validated CNNs were evaluated using receiver operator characteristic (ROC) curves. Areas under the ROC curve were used to determine the binary outcomes of normal (vs pre-plus disease/plus disease) and plus disease (vs normal/pre-plus disease) compared with the RSD.

Performance of the best model (based on cross-validation) for plus disease diagnosis was further evaluated against 8 international ROP experts on an independent test set of 100 images, described previously with 54 normal, 31 pre-plus disease, and 15 plus disease images.[8,36] These images were not included in any of the training or validation sets. Each participating expert had more than 10 years of clinical experience in ROP care and had published more than 5 peer-reviewed articles on ROP. Five of 8 experts served as principal investigators for the multicenter Early Treatment for ROP study.[2,4] Interexpert agreement was assessed using quadratic-weighted κ coefficients and interpreted using a commonly accepted scale: 0 to 0.20 indicated slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.0, near-perfect agreement.[37] κ Scores for agreement between the best-performing CNN from cross-validation were calculated for all experts, the 8 expert consensus (mode) diagnosis (there were no ties), and the RSD.

Interpretation of Learned Features

Following training, image features learned by the classification network were extracted for all images in the training set as a high-dimensional vector. Feature vectors were visualized in 2 dimensions using *t*-distributed stochastic neighbor embedding (*t*-SNE), a dimensionality reduction technique that attempts to minimize distances between similar features while maximizing distances between dissimilar features.[38] This *t*-SNE embedding was visualized as a 2-dimensional scatter plot, with

Figure 1. Receiver Operating Characteristic (ROC) Curves for Diagnosis of Plus Disease in Retinopathy of Prematurity



Data were analyzed from 5-fold cross-validation of 5511 retinal images. Mean areas under the ROC curves (AUCs) for the 5 sets were 0.94 for identifying normal images (vs pre-plus disease or plus disease; A) and 0.98 for identifying plus disease images (vs normal or pre-plus disease; B).

each point corresponding to an image in feature space (eMethods 3 in the Supplement).

Results

Automated Diagnosis of Plus Disease Using Deep Learning

Figure 1 displays ROC curves for 5 CNNs produced using 5-fold cross-validation, each of which was evaluated on an independent test data set (mean [SD] retinal photographs, 1113 [70]). The mean (SD) values of the 5 areas under the ROC curve were 0.94 (0.01) for the diagnosis of normal (vs pre-plus disease/plus disease) and 0.98 (0.01) for diagnosis of plus disease (vs normal/pre-plus disease).

Comparison With Expert Diagnosis

Figure 2 summarizes diagnostic performance of the best-performing model from cross-validation (split 3; Figure 1) on 100 images, with diagnoses from 8 international ROP experts. Sensitivity and specificity of the DL algorithm for detecting plus disease were 93% and 94%, respectively. For detection of pre-plus disease or worse, the sensitivity and specificity were 100% and 94%, respectively. As shown in Figure 2A, the DL algorithm diagnosed 91 of 100 images (91.0%) correctly, whereas 8 experts had an average accuracy of 82.0% (range, 77-94).[36] None of the 9 misclassifications resulted in an image with plus disease being identified as normal or vice versa. The quadratic-weighted κ score for the DL algorithm for agreement with the RSD was 0.92, which was better than 6 of 8 experts (mean [range] agreement compared with RSD, 0.85 [0.80-0.95]) (Figure 2B). Receiver operator characteristic analysis (Figure 2C) displays the behavior of the DL algorithm for diagnosis of plus disease as a function of different operating thresholds, with the operating points of each of the 8 experts shown for reference. Most of the experts lie on or near the ROC curve, which suggests the algorithm may be tuned to mimic any individual expert.

Interpretation of Learned Features

The *t*-SNE was used to visualize high-dimensional features learned by the DL algorithm in 2 dimensions (**Figure 3**). Each point on the scatter plot corresponds to an individual retinal image, where similar images (based on their features) appear nearer to one another than dissimilar images. The colored RSD labels are used only for visualization to denote the different clusters. The *t*-SNE demonstrates qualitative separation among different disease grades. Normal and plus disease form 2 distinct clusters with pre-plus disease bridging them, demonstrating a continuum of disease severity.

Discussion

This study presents the results of a DL-based algorithm trained to diagnose plus disease automatically using retinal images from premature infants at risk of ROP. The key findings are (1) this fully automated CBIA system can diagnose plus disease with comparable or better proficiency than ROP experts, and (2) analysis of features using DL provides insight about the diagnostic process used by experts. Evidence-based ROP management guidelines are based on treatment for presence of plus disease to prevent visual loss and blindness,[1,2] yet inconsistency in plus disease diagnosis leads to clinically significant differences in management.[39] In 2007, Chiang et al[7] investigated plus disease diagnosis for 22 ophthalmology experts on a data set of 34 images and found unanimous agreement on plus disease in only 4 of 34 images (12%). Since then, several publications have reported similar results, with mean weighted κ statistics for plus disease diagnosis ranging from fair (0.21-0.40)[36,40] to moderate (0.41-0.60)[7,41] agreement for expert pairs. It had been unclear whether these differences translated to real-world differences in treatment or outcomes, since this problem (systematic bias in plus disease diagnosis) represents an inherent limitation within ROP clinical trials. However, recent secondary analysis from the Benefits of Oxygen

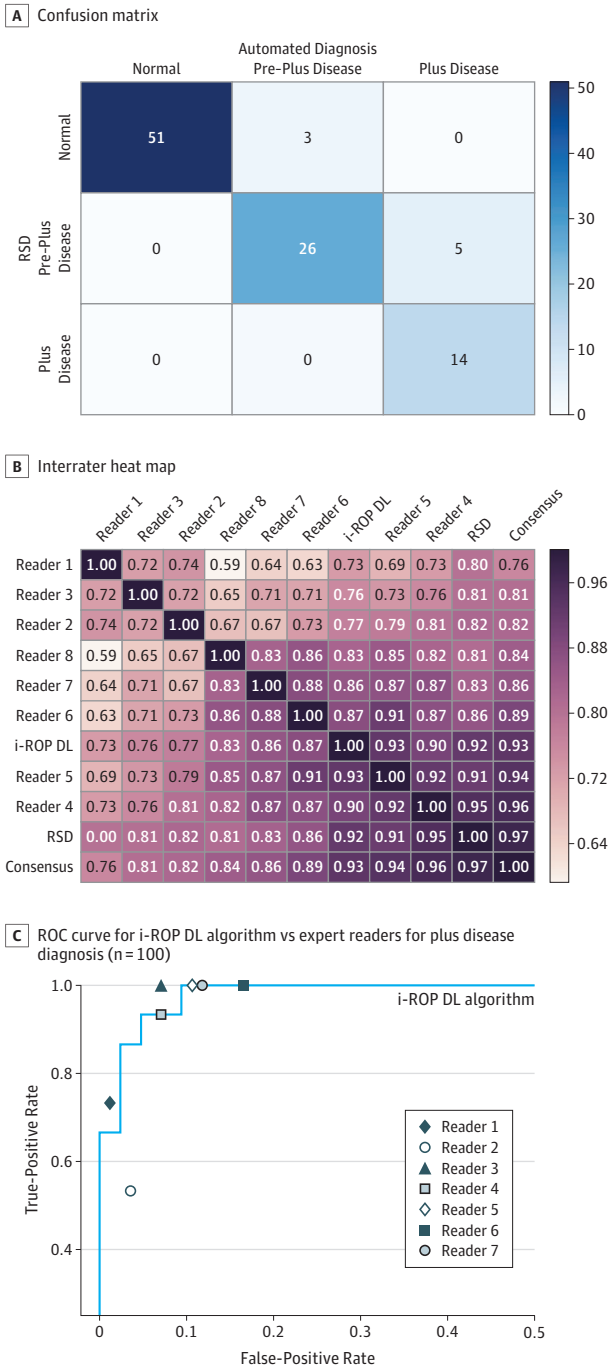
Saturation Targeting-II trials[39] found variation in diagnosis of treatment-requiring ROP among international experts due to differences in plus disease diagnosis. Objective assessment of disease severity, such as with our CNN-based image analysis system, has potential to improve clinical outcomes.

The i-ROP DL algorithm outperformed not only most experts in this study (Figure 2B) but also all prior CBIA systems in ROP.[14] We have previously published results from a different system using machine learning methods rather than CNN-based methods,[42,43] which was able to accurately diagnose pre-plus disease and plus disease but only by using manually segmented images in which retinal vessels were traced by hand and input into the system. Other systems are semiautomated (ie, require manual identification of the optic disc and a few key vessel segments) but have only weakly correlated with 2-level diagnosis (not plus disease vs plus disease).[14,16] In contrast, our current system performed almost perfectly and performed better than most experts on the test set of 100 images at 3-level diagnosis (normal vs pre-plus disease vs plus disease) using raw image files without the need for manual segmentation. We also observed that each of the experts' operating points for sensitivity and specificity of diagnosis of plus disease fell on or near the ROC curve for the DL algorithm (Figure 2C), suggesting that the diagnosis of individual experts may be predicted by tuning the operating point and/or slightly retraining the CNN to better understand that expert's unique biases.[44]

Interpretation of DL outputs may facilitate better understanding of the cognitive processes used to make diagnoses in image-based diagnostic specialties, such as ophthalmology, dermatology, pathology, and radiology.[20,26,45] In ROP, studies have demonstrated that experts deviate from the published and internationally accepted definition of plus disease in several ways[42] and that experts explain their diagnoses using terms such as "experience" and "clinical judgment."[42,43,46] In the same way, neural networks are often regarded as black boxes, since the features used by the multiple layers of the model are not readily interpretable. The *t*-SNE visualization (Figure 3) of image features learned by the algorithm supports the concept of a phenotypic continuum of disease severity, which has been proposed as an explanation for interobserver differences between disease categories.[8,36] This is a particularly interesting finding because the CNN was trained with the categorical labels "normal," "pre-plus disease," and "plus disease" without intrinsic ordering as part of the RSD. In other words, the system learned that normal, pre-plus disease, and plus disease categories reflect a continuum, and our results demonstrate that expert behavior is predictable based on knowing where along that continuum each expert distinguishes between those categories (Figure 1). Further analysis of the CNN using saliency and class activation maps may reveal features and locations that are informative of ROP pathophysiology, potentially serving as an educational tool for clinicians.

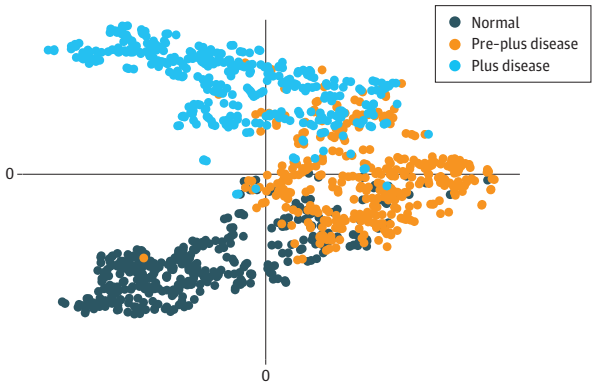
For several reasons, modeling a continuous phenotype using a DL-derived continuous score rather than discrete categories (eg, normal, pre-plus disease, and plus disease) may improve clinical care in ROP. First, a continuous score provides more granularity for determining relative disease

Figure 2. Diagnostic Performance of the Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Deep Learning (DL) Algorithm and 8 ROP Experts Compared With the Reference Standard Diagnosis (RSD) on a Data Set of 100 Images



A, Confusion matrix for the DL algorithm, with numbers of correctly and incorrectly classified images in each class. B, Interrater heat map, with quadratic-weighted κ scores comparing 8 independent experts, the DL algorithm, and the RSD. The consensus diagnosis among 8 experts is also shown, calculated as the most frequent (mode) diagnosis. C, Receiver operating characteristic (ROC) curve for the DL algorithm and performance of 8 experts in terms of true-positive rates (ie, sensitivity) and false-positive rates (ie, 1 − specificity).

Figure 3. *t*-Distributed Stochastic Neighbor Embedding Visualization of Features Extracted From an Intermediate Layer of a Trained Convolutional Neural Network for Plus Disease Diagnosis in Retinopathy of Prematurity



This visualization demonstrates that the convolutional neural network is able to automatically generate features that roughly separate the 3 diagnoses and that they appear to run along a continuum of disease severity.

progression or regression, which may be lost within subjective 2-level or 3-level disease categories because individual eyes may measurably worsen, remain the same, or improve over time. Additionally, physicians are accustomed to incorporating continuous biomarkers (eg, blood pressure) into clinical decision-making. A plus disease score in the upper range may or may not lead to treatment for ROP but could also be put into context of other known risk factors, pace and progression of disease over time, clinical judgment, and published validation studies. Finally, DL-based objective disease metrics may be incorporated into screening strategies to automatically identify clinically relevant disease and initiate appropriate referral. Incorporating DL-based screening into fundus camera systems and telemedicine platforms for ROP and other image-based diseases may improve the objectivity, accuracy, and efficiency of health care delivery.

Limitations

This study has several limitations. Convolutional neural networks are only as robust as the data on which they are trained. In this case, we used nearly 6000 images from 8 different institutions, each with a rigorous RSD, which was itself a consensus diagnosis of 4 separate diagnoses (image-based diagnosis by 3 experts and ophthalmoscopic diagnosis by 1 expert), which should improve the external validity of our system. It is unknown how factors such as image quality, resolution, different camera systems, and field of view may affect the output of the i-ROP DL system.[47] These topics warrant further study. Image preprocessing methods are specific for each data set and CNN, representing a critical step in image classification tasks that eliminates variations, which may introduce bias during model training. In our data set, such variations included differences in retinal pigmentation, brightness, contrast, and textual annotations. Other preprocessing and postprocessing methods, such as binarization and morphological operations, may improve generalizability of our algorithm and could be the subject of future analyses. Our system currently only classifies plus disease, one component of the International Classification of Retinopathy of Prematurity system.[3] Ideally, a fully automated ROP screening platform could classify zone, stage, and overall disease category as well as predict need for treatment. These are the topics of ongoing study.

Conclusions

These results demonstrate that the incorporation of deep neural networks may enable automated screening and diagnosis for ROP with high accuracy and repeatability. These results may change the way ROP is diagnosed in the future and are broadly relevant to other medical fields that rely primarily on subjective image-based diagnostic features. Future work will involve comparison of features learned by the DL algorithm with known morphological features, evaluation of deep neural networks for other components of the ROP clinical examination, and application to other retinal diseases. Incorporation of this technology into fundus cameras or telemedicine systems could provide advice at the point of care and has the potential to improve the quality, accessibility, and cost of ROP screening worldwide.

ARTICLE INFORMATION

Accepted for Publication: April 10, 2018.

Published Online: May 2, 2018. doi:10.1001/jamaophthalmol.2018.1934

Author Affiliations: Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown (Brown, Beers, Chang, Kalpathy-Cramer); Department of Ophthalmology, Casey Eye Institute, Oregon Health and Science University, Portland (Campbell, Ostmo, Chiang); Department of Ophthalmology and Visual Sciences, Illinois Eye and Ear Infirmary, University of Illinois at Chicago (Chan); Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts (Dy, Erdogmus, Ioannidis); Massachusetts General Hospital and Brigham and Women's Hospital Center for Clinical Data Science, Boston (Kalpathy-Cramer); Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland (Chiang).

Author Contributions: Drs Brown and Campbell had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Brown and Campbell contributed equally to this work. Drs Kalpathy-Cramer and Chiang supervised this work equally.
Study concept and design: Brown, Campbell, Beers, Dy, Erdogmus, Ioannidis, Kalpathy-Cramer, Chiang.
Acquisition, analysis, or interpretation of data: Brown, Campbell, Beers, Chang, Ostmo, Chan, Dy, Kalpathy-Cramer, Chiang.
Drafting of the manuscript: Brown, Campbell, Ioannidis, Kalpathy-Cramer.
Critical revision of the manuscript for important intellectual content: Brown, Campbell, Beers, Chang, Ostmo, Chan, Dy, Erdogmus, Kalpathy-Cramer, Chiang.
Statistical analysis: Brown, Campbell, Beers, Kalpathy-Cramer.
Obtained funding: Erdogmus, Ioannidis, Kalpathy-Cramer, Chiang.
Administrative, technical, or material support: Brown, Campbell, Beers, Chang, Ostmo, Chan, Kalpathy-Cramer, Chiang.
Study supervision: Dy, Erdogmus, Ioannidis, Kalpathy-Cramer, Chiang.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for

Disclosure of Potential Conflicts of Interest. Drs Brown, Chang, Dy, Erdogmus, Kalpathy-Cramer, and Chiang received grants from the National Science Foundation and the National Institutes of Health during the conduct of this study. Dr Chan serves on the scientific advisory board of Visunex Medical Systems and serves as a consultant for Alcon, Allergan, and Bausch and Lomb. Dr Ioannidis has received grants from the National Science Foundation and Google and is employed by Yahoo. Dr Chiang serves on the scientific advisory board of Clarity Medical Systems, serves as a consultant for Novartis, and is an initial member of Inteleretina. No other disclosures were reported.

Funding/Support: This work is supported by grants R01EY019474, P30EY10572, and P41EB015896 from the National Institutes of Health, grants SCH-1622542 (Massachusetts General Hospital, Charlestown), SCH-1622536 (Northeastern University, Boston, Massachusetts), and SCH-1622679 (Oregon Health and Science University, Portland) from the National Science Foundation, and training grant T90DA022759/R90DA023427 from the National Institutes of Health Blueprint for Neuroscience Research as well as by unrestricted departmental funding from Research to Prevent Blindness (Oregon Health and Science University, Portland).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Group Members: Members of the Imaging and Informatics in Retinopathy of Prematurity Research Consortium include: *Oregon Health and Science University, Portland*: Michael F. Chiang, MD; Susan Ostmo, MS; Sang Jin Kim, MD, PhD; Kemal Sonmez, PhD; and J. Peter Campbell, MD, MPH; *University of Illinois at Chicago*: R. V. Paul Chan, MD; and Karyn Jonas, RN; *Columbia University, New York, New York*: Jason Horowitz, MD; Osode Coki, RN; Cheryl-Ann Eccles, RN; and Leora Sarna, RN; *Weill Cornell Medical College, New York, New York*: Anton Orlin, MD; *Bascom Palmer Eye Institute, Miami, Florida*: Audina Berrocal, MD; and Catherin Negron, BA; *William Beaumont Hospital, Royal Oak, Michigan*: Kimberly Denser, MD; Kristi Cumming, RN; Tammy Osentoski, RN; Tammy Check, RN; and Mary Zajechowski, RN; *Children's Hospital Los Angeles, Los Angeles, California*: Thomas Lee, MD; Evan Kruger, BA; and Kathryn McGovern, MPH; *Cedars Sinai Hospital, Los Angeles, California*: Charles Simmons, MD; Raghu Murthy, MD; and Sharon Galvis, NNP; *LA Biomedical Research Institute, Los Angeles, California*: Jerome Rotter, MD; Ida Chen, PhD; Xiaohui Li, MD; Kent Taylor, PhD; and Kaye Roll, RN; *Massachusetts General Hospital, Boston*: Jayashree Kalpathy-Cramer, PhD; Ken Chang, BS; Andrew Beers, BS; *Northeastern University, Boston, Massachusetts*: Deniz Erdogmus, PhD; and Stratis Ioannidis, PhD; and *Asociacion para Evitar la Ceguera en Mexico, Mexico City, Mexico*: Maria Ana Martinez-Castellanos, MD; Samantha Salinas-Longoria, MD; Rafael Romero, MD; Andrea Arriola, MD; Francisco Olguin-Manriquez, MD; Miroslava Meraz-Gutierrez, MD; Carlos M. Dulanto-Reinoso, MD; and Cristina Montero-Mendoza, MD.

Disclaimer: The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Meeting Presentation: This article was presented at the 2018 ARVO Annual Meeting; May 2, 2018; Honolulu, Hawaii.

Additional Contributions: We acknowledge the GPU computing resources provided by the Massachusetts General Hospital and Brigham and Women's Hospital Center for Clinical Data Science, Boston.

REFERENCES

1. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: preliminary results. *Arch Ophthalmol*. 1988;106(4):471-479.

2. Early Treatment For Retinopathy Of Prematurity Cooperative Group. Revised indications for the treatment of retinopathy of prematurity: results of the Early Treatment for Retinopathy of Prematurity randomized trial. *Arch Ophthalmol*. 2003;121(12):1684-1694.

3. International Committee for the Classification of Retinopathy of Prematurity. The International Classification of Retinopathy of Prematurity revisited. *Arch Ophthalmol*. 2005;123(7):991-999.

4. Good WV; Early Treatment for Retinopathy of Prematurity Cooperative Group. Final results of the Early Treatment for Retinopathy of Prematurity (ETROP) randomized trial. *Trans Am Ophthalmol Soc*. 2004;102:233-248, 248-250.

5. Reynolds JD, Dobson V, Quinn GE, et al; CRYO-ROP and LIGHT-ROP Cooperative Study Groups. Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. *Arch Ophthalmol*. 2002;120(11):1470-1476.

6. Mintz-Hittner HA, Kennedy KA, Chuang AZ; BEAT-ROP Cooperative Group. Efficacy of intravitreal bevacizumab for stage 3+ retinopathy of prematurity. *N Engl J Med*. 2011;364(7):603-615.

7. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Arch Ophthalmol*. 2007;125(7):875-880.

8. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al; Imaging and Informatics in Retinopathy of Prematurity Research Consortium. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*. 2016;123(11):2345-2351.

9. Braverman RS, Enzenauer RW. Socioeconomics of retinopathy of prematurity in-hospital care. *Arch Ophthalmol*. 2010;128(8):1055-1058.

10. Wallace DK. Fellowship training in retinopathy of prematurity. *J AAPOS*. 2012;16(1):1.

11. Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a web-based survey. *J AAPOS*. 2012;16(2):177-181.

12. Nagiel A, Espiritu MJ, Wong RK, et al. Retinopathy of prematurity residency training. *Ophthalmology*. 2012;119(12):2644-2645.e1, 2.

13. Gilbert C, Rahi J, Eckstein M, O'Sullivan J, Foster A. Retinopathy of prematurity in middle-income countries. *Lancet*. 1997;350(9070):12-14.

14. Wittenberg LA, Jonsson NJ, Chan RVP, Chiang MF. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity. *J Pediatr Ophthalmol Strabismus*. 2012;49(1):11-19.

15. Heneghan C, Flynn J, O'Keefe M, Cahill M. Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis. *Med Image Anal*. 2002;6(4):407-429.

16. Wallace DK, Zhao Z, Freedman SF. A pilot study using "ROPtool" to quantify plus disease in retinopathy of prematurity. *J AAPOS*. 2007;11(4):381-387.

17. Grisan E, Foracchia M, Ruggeri A. A novel method for the automatic grading of retinal vessel tortuosity. *IEEE Trans Med Imaging*. 2008;27(3):310-319.

18. Roth DB, Morales D, Feuer WJ, Hess D, Johnson RA, Flynn JT. Screening for retinopathy of prematurity employing the RetCam 120: sensitivity and specificity. *Arch Ophthalmol*. 2001;119(2):268-272.

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

20. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

21. Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. *AMIA Annu Symp Proc*. 2015;2015:1899-1908.

22. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al; the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199-2210.

23. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1(4):322-327. doi:10.1016/j.oret.2016.12.009

24. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135(11):1170-1176.

25. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172(5):1122-1131.e9.

26. 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(22):2402-2410.

27. 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(22):2211-2223.

28. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0

29. Worrall DE, Wilson CM, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks.

In: Carneiro G, Mateus D, Peter L, et al, eds. *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer International Publishing; 2016:68-76.

30. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-2194. doi:10.1001/jama.2013.281053

31. Ryan MC, Ostmo S, Jonas K, et al. Development and evaluation of reference standards for image-based telemedicine diagnosis and clinical research studies in ophthalmology. *AMIA Annu Symp Proc*. 2014;2014:1902-1910.

32. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Cham, Switzerland: Springer International Publishing; 2015: 234-241.

33. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Paper presented at: 2015 Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition; June 7-12, 2015; Boston, MA.

34. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 Institute of Electrical and Electronics Engineers Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL.

35. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? Paper presented at: 27th International Conference on Neural Information Processing Systems; December 8-13, 2014; Montreal, Québec, Canada.

36. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al; Imaging and Informatics in Retinopathy of Prematurity Research Consortium. Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as a basis of diagnostic variability. *Ophthalmology*. 2016;123(11):2338-2344.

37. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform*. 2002;35(2):99-110.

38. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(Nov):2579-2605.

39. Fleck BW, Williams C, Juszczak E, et al; BOOST II Retinal Image Digital Analysis (RIDA) Group. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. *Eye (Lond)*. 2018;32(1):74-80.

40. Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *Am J Ophthalmol*. 2015;160(3):553-560.e3.

41. Daniel E, Quinn GE, Hildebrand PL, et al; e-ROP Cooperative Group. Validated system for centralized grading of retinopathy of prematurity: telemedicine approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) study. *JAMA Ophthalmol*. 2015;133(6):675-682.

42. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, et al; Imaging and Informatics in ROP (i-ROP) Research Consortium. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. *JAMA Ophthalmol*. 2016;134(6):651-657.

43. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al; i-ROP Research Consortium. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis. *Transl Vis Sci Technol*. 2015;4(6):5.

44. Guan MY, Gulshan V, Dai AM, Hinton GE. Who said what: modeling individual labelers improves classification [published online March 26, 2017]. *Computing Research Repository*.

45. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep*. 2017;7(1):5301.

46. Hewing NJ, Kaufman DR, Chan RV, Chiang MF. Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts. *JAMA Ophthalmol*. 2013;131(8):1026-1032.

47. Coyner A, Swan R, Kalpathy-Cramer J, et al. Automated image quality assessment for fundus images in retinopathy of prematurity. *Invest Ophthalmol Vis Sci*. 2017;58(8):5550.