Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning

Michael David Abràmoff,[1–3] Yiyue Lou,[4] Ali Erginay,[5] Warren Clarida,[3] Ryan Amelon,[3] James C. Folk,[1,3] and Meindert Niemeijer[3]

[1]Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, Iowa City, Iowa, United States
[2]Iowa City Veterans Affairs Medical Center, Iowa City, Iowa, United States
[3]IDx LLC, Iowa City, Iowa, United States
[4]Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, Iowa, United States
[5]Service d' Ophtalmologie, Hôpital Lariboisière, APHP, Paris, France

Correspondence: Michael David Abràmoff, 11205 PFP, University of Iowa Hospital and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA;
michael-abramoff@uiowa.edu.

Submitted: May 20, 2016
Accepted: August 18, 2016

Citation: Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200–5206. DOI:10.1167/iovs.16-19964

Purpose. To compare performance of a deep-learning enhanced algorithm for automated detection of diabetic retinopathy (DR), to the previously published performance of that algorithm, the Iowa Detection Program (IDP)–without deep learning components–on the same publicly available set of fundus images and previously reported consensus reference standard set, by three US Board certified retinal specialists.

Methods. We used the previously reported consensus reference standard of referable DR (rDR), defined as International Clinical Classification of Diabetic Retinopathy moderate, severe nonproliferative (NPDR), proliferative DR, and/or macular edema (ME). Neither Messidor-2 images, nor the three retinal specialists setting the Messidor-2 reference standard were used for training IDx-DR version X2.1. Sensitivity, specificity, negative predictive value, area under the curve (AUC), and their confidence intervals (CIs) were calculated.

Results. Sensitivity was 96.8% (95% CI: 93.3%–98.8%), specificity was 87.0% (95% CI: 84.2%–89.4%), with 6/874 false negatives, resulting in a negative predictive value of 99.0% (95% CI: 97.8%–99.6%). No cases of severe NPDR, PDR, or ME were missed. The AUC was 0.980 (95% CI: 0.968–0.992). Sensitivity was not statistically different from published IDP sensitivity, which had a CI of 94.4% to 99.3%, but specificity was significantly better than the published IDP specificity CI of 55.7% to 63.0%.

Conclusions. A deep-learning enhanced algorithm for the automated detection of DR, achieves significantly better performance than a previously reported, otherwise essentially identical, algorithm that does not employ deep learning. Deep learning enhanced algorithms have the potential to improve the efficiency of DR screening, and thereby to prevent visual loss and blindness from this devastating disease.

Keywords: diabetic retinopathy, detection, deep learning, algorithm, diabetes

Automation is a prerequisite to improve health care efficiency affordability and accessibility.[1,2] In the management of the estimated 23 million Americans and 59 million Europeans with diabetes, automation of the retinal exam is sorely needed. The reason is that adherence to the regular eye examinations–necessary to diagnose diabetic retinopathy (DR) at an early stage, when it can be treated with the best prognosis and visual loss can be delayed or deferred[3–5]–is frequently less than 60%.[4,6] This leaves millions of people with diabetes at risk for potentially preventable visual loss and blindness. Though OCT and widefield imaging have been proposed to improve screening performance, most present-day DR screening programs use 1 or 2 field retinal color fundus imaging, in order to reach cost-effectiveness.[7–9]

Over the last two decades, the automated analysis of retinal color images for DR has been studied by many groups.[10–12] Though the lack of widely accepted, well characterized, and representative datasets makes comparison difficult, recent studies show that complete DR screening systems using such algorithms achieve adequate safety, as for example the Iowa Detection Program (IDP).[1,13–16] These algorithms are all based on classical expert designed image analysis, using carefully designed transformations including mathematical morphology and wavelet transformations.[17–20] More recently, we used data-driven machine learning to learn the lowest level wavelet transformations from training data, but this resulted in only marginal performance improvements.[21]

Deep learning, where all transformation levels are determined from training data, instead of being designed by experts,[22] has been highly successful in a large number of computer vision and image analysis tasks, substantially outperforming all classical image analysis techniques,[23] and, given the spatial coherence that is characteristic of images, typically implemented as Convolutional Neural Networks (CNN).[23] Indeed, the highest performing algorithms in the recent Kaggle

Investigative Ophthalmology & Visual Science

competition, which completed July 2015, all used CNNs to identify signs of DR in retinal images.[24]

Training CNNs to detect DR directly from complete retinal images may lead to unanticipated, but undetected, associations. A simple example of such unwanted associations, is a CNN trained on data from a subgroup of patients with a high prevalence of DME. If the photographers of this population tend to image patients with the disc slightly temporal, the algorithm will be taught to associate diabetic macular edema (ME) with this temporal location of the disc, which is of course incorrect outside these specific populations.

Alternatively, CNNs can be used as high performing lesion detectors, and thus, IDx-DR X2.1 ('the device') is a hybrid system. It makes use of multiple CNNs, trained and used to detect hemorrhages, exudates, and other lesions, as well as normal retinal anatomy, which are integrated into a classic system that is otherwise very similar to its prototype, IDP. It is thus of broad interest to determine whether the replacement of classical image analysis approaches by dedicated CNNs has a significant effect on performance, by comparing this hybrid system to the published performance of a similar but classical image analysis system (i.e., IDP), as published in 2013,[1] on a standard, publicly available, dataset with a trusted reference standard.

The purpose of the present study is to determine the sensitivity, specificity, and area under the operator receiving characteristics curve (AUC) of the device to referable DR (rDR), defined as moderate or severe nonproliferative DR (NPDR), proliferative DR (PDR), and/or macular edema (ME), and compare these with those published for the IDP at a similarly high sensitivity set point. IDx-DR X2.1 also added a vision-threatening DR (vtDR) output, and so we also report performance of that to detect vision threatening DR, defined as severe NPDR, PDR, and/or ME. We used the same publicly available set of fundus images and the same previously reported consensus reference standard set by three US Board certified retinal specialists for comparison.[1]

Methods

Subjects, Retinal Image Dataset, Reference Standard

For the present study, we used the exact same dataset as in our 2013 publication.[1] In summary, the Messidor-2 dataset[25] consists of the digital retinal color images, one fovea-centered image per eye, of 874 subjects with diabetes, 1748 images.[1] Messidor-2 differs from the original Messidor dataset of 1200 images in that we ensured it has two images for each subject, one for each eye. As reported previously, subjects were pharmacologically dilated at two centers, and were not dilated at the third center, and then imaged using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic fundus camera (Topcon B.V., Capelle A/D IJssel, The Netherlands) with a 45° field of view, centered on the fovea, at 1440*960, 2240*1488, or 2304*1536 pixels.[1] As reported previously, mean age was 57.6 (± 15.9) years and 57% were male.[1] As also reported previously, three board certified retinal specialists independently graded all images from all subjects according to the International Clinical Diabetic Retinopathy severity scale (ICDR; 0–4) and a modified definition of ME (0–1): the presence of exudates, retinal thickening (if visible on nonstereo photographs), or microaneurysms, all within 1 disc diameter of the fovea.[1] In that study, the criterion of one or more microaneurysms was added to prevent that ME would be incorrectly missed, as the isolated presence of one or more microaneurysm(s) can be the only sign of ME visible on nonstereo photographs. Disagreements were adjudicated until consensus. On initial independent grading, κ of expert 1 versus expert 2 (1vs2) was 0.85 (95% confidence interval [CI]: 0.81–0.90), 1vs3 was 0.82 (95% CI: 0.78–0.87), and 2vs3 was 0.79 (95% CI: 0.75–0.84), average κ was 0.822.[1] Because IDx-DR X2.1 requires two images per eye, and Messidor-2 only provides one fovea centered image per eye, all images were duplicated. The Messidor-2 dataset is in the public domain. The corresponding rDR reference standard is available for researchers in the public domain at http://www.medicine.uiowa.edu/eye/abramoff.

Three Categories of Disease

Using the previously reported ICDR and ME gradings,[1] we created four levels of disease for each subject:

- No DR – ICDR level 0 (no DR) or 1 (mild DR), and no ME
- Referable DR – ICDR level 2 (moderate nonproliferative DR), 3 (severe nonproliferative DR), 4 (proliferative DR), or ME
- Vision threatening DR (vtDR) – ICDR level 3 (severe nonproliferative DR), 4 (proliferative DR), or ME. A new disease category for this study, to evaluate the performance on this category of disease.
- Macular edema, the adjudicated reference standard for the presence of ME. A new, separate category for this study, all subjects with ME appear in both vtDR and rDR.

Presence of vtDR thus implies the presence of rDR, and ME implies both rDR and vtDR. Using the consensus reference standard, previously reported rDR prevalence was 21.7% (95% CI: 19.1%–24.7%).[1] Vision threatening DR prevalence was 10.6% (95% CI: 8.7%–12.9%), and ME prevalence was 9.50% (95% CI: 7.64%–11.65%).

IDx-DR X2.1

IDx-DR X2.1 ('the device') is an automated system for the detection of DR. It consists of two components, client software running at the point of care, which is not part of this study, and analysis software that is running on a server maintained and controlled by IDx, and which is evaluated in this study. In standard operation, the camera operator acquires four images, one optic disc and one macula centered for each eye, and submits them to IDx-DR. The analysis software provides four types of outputs:

- Negative: implying no or only mild DR present
- rDR: implying rDR is present
- vtDR: implying vtDR is present
- Low exam quality: implying either protocol errors or low quality of the individual images

The device applies a set of CNN-based detectors to each of the images in the exam. These detectors are trained and optimized to detect normal anatomy, such as optic disc and fovea, as well as the lesions characteristic for DR, such as hemorrhages, exudates, and neovascularization. Though the CNNs are particular to each type of lesion and parameters vary slightly between them, they are inspired by Alexnet[23] (for more limited training sets) and the Oxford Visual Geometry Group[26] (for more extensive training sets) network architectures. They were trained on 10,000 to 1,250,000 unique samples, depending on the lesion to be detected, extracted from images from patients with DR, and manually annotated by one or more experts, as well as positive and negative confounders.[27] The unique samples underwent a variety of augmentations to increase spatial, rotational, and scale variance.

Investigative Ophthalmology & Visual Science

TABLE. IDx-DR X2.1 Sensitivity, Specificity, Negative and Positive Predictive Value, and AUC and Corresponding 95% CIs for rDR Output to Detect rDR, ME, and vtDR, and vtDR Output to Detect vtDR

IDx Output For	Disease Level	Sensitivity (95% CI)	Specificity (95% CI)	Negative Predictive Value (95% CI)	Positive Predictive Value (95% CI)	AUC (95% CI)
rDR	rDR	96.8% (93.3%–98.8%)	87.0% (84.2%–89.4%)	99.0% (97.8%–99.6%)	67.4% (61.5%–72.9%)	0.980 (0.968, 0.992)
rDR	vtDR	100% (96.1%–100%)	N.A.	N.A.	N.A.	N.A.
rDR	ME	100% (95.6%–100%)	N.A.	N.A.	N.A.	N.A.
vtDR	vtDR	100.0% (96.1%–100.0%)	90.8% (88.5%–92.7%)	100.0% (99.5%–100.0%)	56.4% (48.4%–64.1%)	0.989 (0.984, 0.994)

N.A., not calculated.

In the case of anatomy detectors, the CNN detector output represents its location in the image at a particular location, while in the case of lesion detectors, the CNN output represents the likelihood that a particular detection is an actual abnormality. Feature vectors formed from these likelihoods are fed into two fusion algorithms. The result of the first, the rDR index, represents the likelihood that a patient has rDR. The result of the second, the vtDR index, represents the likelihood that a patient has vtDR. The exam quality data is used to determine if the exam quality was sufficient for the system to make a diagnostic decision. These fusion classifiers were implemented as random forests, and trained on 25,000 complete exams of four images per subject annotated for their ICDR level by multiple experts. Obviously, none of the experts annotating the training images were involved in setting the previously reported reference standard for Messidor-2. Once available, the vtDR index is thresholded first. If the index is above or equal to the threshold, a positive output for vtDR is returned. If the vtDR index is below this threshold, the rDR index is thresholded. If the rDR index is above or equal to this latter threshold a positive output for rDR is returned. If it is below the latter threshold an output of "negative" is returned.

Messidor-2 images cannot be used for commercial purposes, and were not used for training or testing the device. The images for training the detectors and classifiers were obtained from the EyeCheck project and the University of Iowa.

Modifications to IDx-DR to Create the IDx-DR X2.1 Device

As explained in the introduction, the IDP used only classical detector algorithms, and CNNs have been added as lesion/anatomy detectors to the IDx-DR implementation studied here. In addition, the device had to be modified to support a retrospective dataset of only fovea centered images. Therefore, the device, as tested here, only allows access to the analysis software, the client software having been turned off. Because the image quality verifies that one disc- and one fovea-centered image are present for both a left and a right eye, and Messidor-2 does not contain the required disc centered images, the image quality algorithm had to be turned off. No other changes were made, nor was any component or CNN retrained for this IDx-DR version or this study. The device was run at the University of Iowa. To estimate the fraction of Messidor-2 subjects that would have received an insufficient quality exam if one disc centered and one fovea centered had been available for both eyes, we used the quality algorithm only, outside of the device.

Statistical Analysis of Performance

Analyses were conducted using SAS 9.4 (SAS Institute, Inc., Cary, NC, USA). The device's rDR and vtDR output sensitivity, specificity, negative and positive predictive value, and their 95% CIs were calculated, based on exact binomial distribution, at set points, 0.37 for rDR and 0.329 for vtDR, chosen as high sensitivity set points, in order to compare with the published IDP,[1] which was also set at high sensitivity. We test the differences in sensitivity and specificity of the IDx-DR X2.1 device and IDP using Fisher's exact test.

Sensitivity for detecting vtDR and ME was also calculated for the device's rDR output. The AUC for rDR and vtDR indices were determined by logistic regression (SAS PROC Logistic) from the adjudicated reference standard.[28,29] We compared rDR AUC with the theoretical maximum AUC of 0.955 (95% CI 0.939–0.972), which can be obtained by a perfect detection program, given the characteristics of the three readers and the prevalence of disease in the Messidor-2 population.[1,18]

RESULTS

For 874 subjects, the sensitivity of the device's rDR output to detect rDR was 96.8% (95% CI: 93.3%–98.8%) and specificity was 87.0% (95% CI: 84.2%–89.4%), with 6/874 false negatives, resulting in a negative predictive value of 99.0% (95% CI: 97.8%–99.6%), and positive predictive value of 67.4% (95% CI 61.5%–72.9%). Sensitivity for the device's rDR output to detect vtDR was 100% (95% CI: 96.1%–100%; i.e., no cases of vtDR were missed), and sensitivity for the device's rDR output to detect ME was also 100% (95% CI: 95.6%–100%; i.e., no cases of ME were missed by the rDR output). The AUC for the device's rDR index to detect rDR was 0.980 (95% CI: 0.968–0.992; Fig. 1; Table).

The device's rDR output sensitivity to detect rDR, of 96.8%, was not statistically different from the previously published IDP sensitivity to detect rDR (P value 0.615), but its specificity, at 87.0%, was significantly better than that of IDP (P value $<$ 0.0001). The 6/874 false negatives are shown in Figure 2.

For the device's vtDR output, sensitivity to detect vtDR was 100.0% (95% CI: 96.1%–100.0%) and specificity was 90.8% (95% CI: 88.5%–92.7%), resulting in a negative predictive value of 100.0% (95% CI 99.5%–100.0%), and positive predictive value of 56.4% (95% CI 48.4%–64.1%). The AUC for the device's vtDR index to detect vtDR was 0.989 (95% CI: 0.984–0.994; Fig. 1).

We previously reported that the theoretical maximum AUC measurable on this specific dataset and reference standard has a 95% CI of 0.939 to 0.972, which thus overlaps with the 95% CI of 0.968 to 0.992 of the measured AUC for the device's rDR index.[1,18]

Thirty-four subjects (4%) had at least one image that was deemed insufficient by the quality algorithm run outside the device.

DISCUSSION

The results show that a device to detect rDR, where standard Gaussian derivatives and similarly designed, lowest level

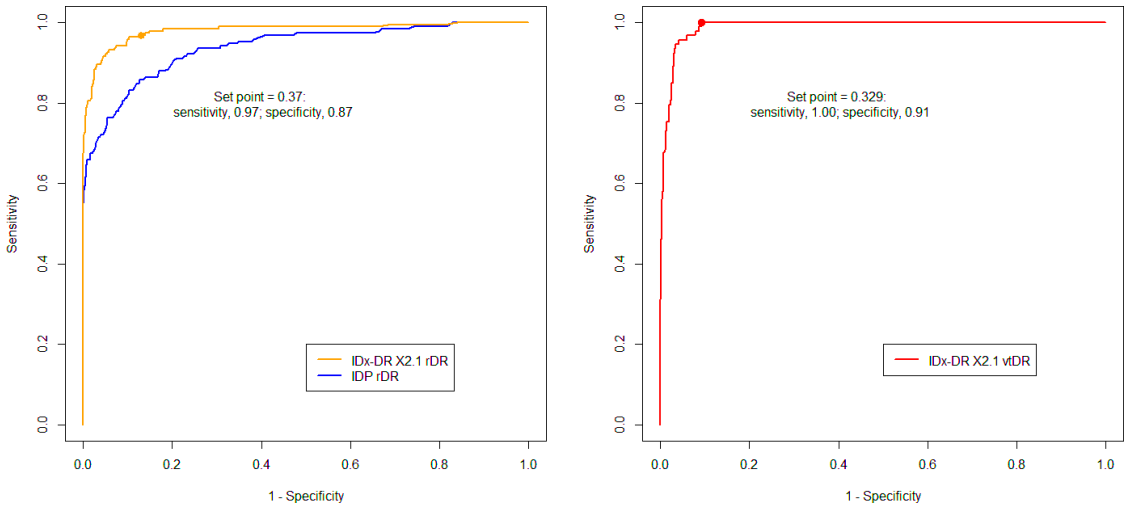

Figure 1. (**Left**) ROC of IDx-DR X2.1 rDR outputs (*orange*), respectively, IDP outputs (*blue*), to detect rDR, defined as more than mild nonproliferative retinopathy and/or macular edema according to ICDR criteria by an adjudicated consensus of three retinal specialists as previously reported.[1] The AUC for rDR output is 0.980, and for IDP was 0.937. (**Right**) ROC of IDx-DR X2.1 vtDR outputs to detect vision threatening retinopathy (vtDR; *red*), defined as moderate or severe nonproliferative retinopathy, proliferative retinopathy, and/or macular edema. AUC for vtDR is 0.989. The *lines* represents sensitivity/specificity pairs for vtDR. Horizontal axes: 1-specificity; vertical axes: sensitivity.

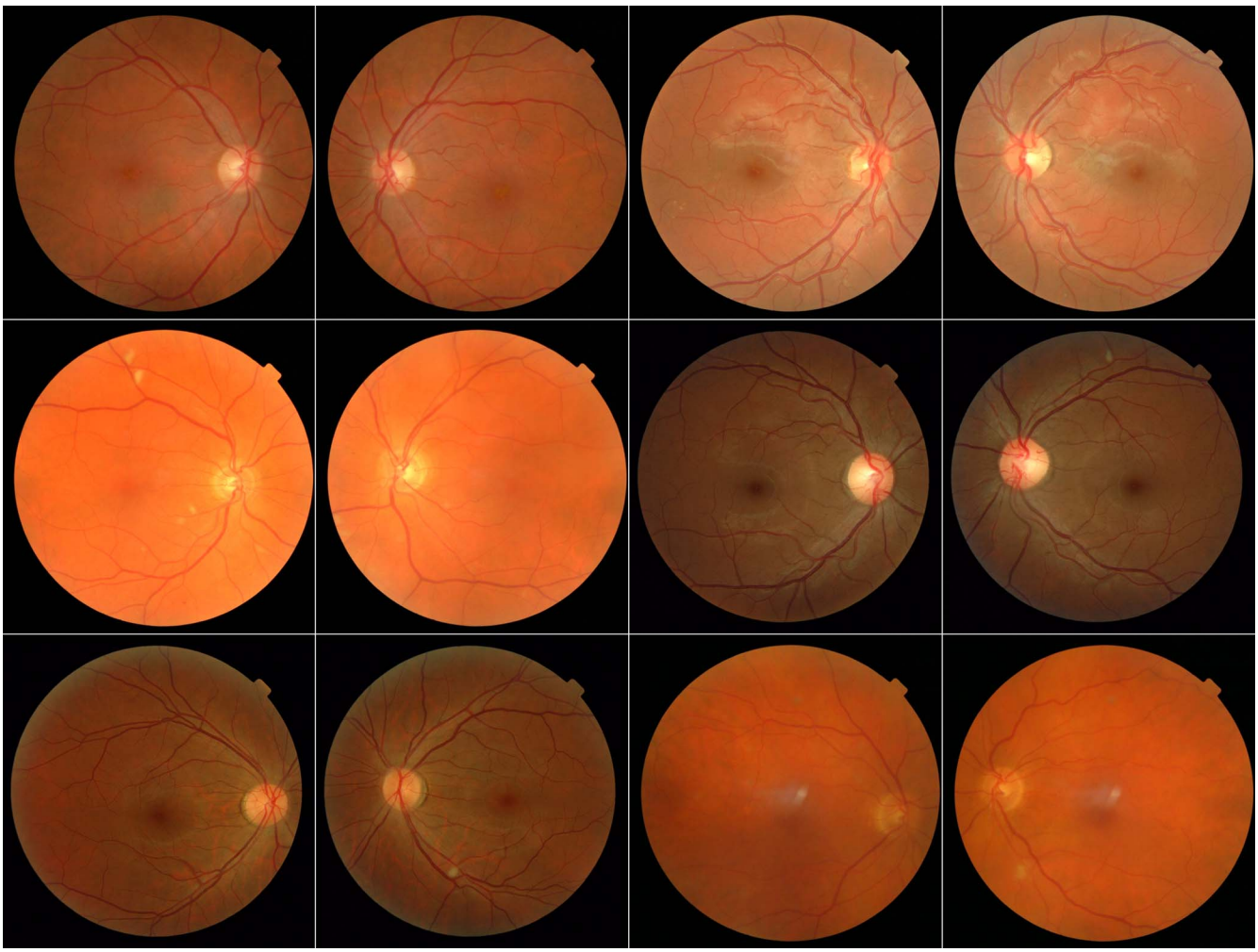

Figure 2. Right and left eye images of the 6/874 subjects diagnosed as having rDR, per the adjudicated consensus of the retinal experts, that were false-negatives for the rDR output (i.e., were missed by the device). All six subjects had consensus grading of moderate DR without ME.

Investigative Ophthalmology & Visual Science

features are replaced by lesion specific trained CNNs, performs significantly better, over a range of performance balances.

The device with CNN based detectors, IDx-DR X2.1, achieved a high performance of 96.8% sensitivity and 87.0% specificity to detect rDR. All false negatives had moderate NPDR according to ICDR (i.e., the lowest level of referable DR), thus the sensitivity to detect vtDR and ME were both 100% (Fig. 2), showing all of the images showing moderate NPDR that were missed by the device's rDR output. The device's AUC 95% CI ranged from 0.968 to 0.992, which overlaps with the 95% CI of 0.939 to 0.972 for achievable AUC given the Messidor-2 dataset and the associated reference standard with three readers.[1,18] Adding additional readers to the reference standard will allow a higher AUC to be measured.[30]

The device's vtDR output had a sensitivity of 100%, and specificity of 91% to detect vtDR, and vtDR AUC was 0.989. Thus, the device's vtDR output may have utility in screening protocols where only vision-threatening DR (i.e., those that may be considered for treatment) are referred.[31,32]

The NPVs of the device are high, 99.0% and 100%, respectively, and thus a negative rDR or vtDR output means that DR is likely absent. The NPV and PPV are dependent on the prevalence of the disease in the population. We and others have found rDR prevalences ranging from 10.4%[33] to 38.1% (Maguire MG, et al. *IOVS* 2015;56:ARVO E-Abstract 2014) and prevalence of vtDR from approximately 1.7%[33] to 19.1% (Maguire et al., *IOVS* 2015;56:ARVO E-Abstract 2014). In Messidor-2, 21.7% had rDR, and 10.6% vtDR, thus within these ranges. The NPV and PPV are thus sensitive to 'age' of a screening program, in other words, how long a specific population has been undergoing screening for DR by either clinicians or a device. The prevalence of vtDR in the first year of running a screening program can be as high as 30% to 40%, as this is primarily prevalence (number of existing cases that had developed over the years without screening), not incidence (number of new cases that develop in a year). After the vtDR cases have been 'caught' in the first year, in the following years the number of cases of vtDR will reflect only the much lower incidence, depending on the presence of risk factors in the population. It is of utmost importance that DR detection algorithms operate well within these ranges of prevalence, including inner city populations at high risk that have never had screening.

Advantages of this study are that it was performed on the publicly available Messidor-2 dataset,[1,23] and a published adjudicated standard of three experts, that is now available in the public domain at http://www.medicine.uiowa.edu/eye/abramoff. For the device, some of the IDP's classic feature based lesion detectors were replaced by CNN based lesion detectors, leading to significantly better performance in a hybrid architecture. The device was never trained on any of the Messidor-2 images. Other groups have used full image-based CNNs (i.e., the CNN is trained with only complete retinal images), lacking explicit lesion detection,[24] and in the Introduction we explained the unanticipated, but undetected, associations that may result in.

The study has potential limitations. The purpose was to determine whether the use of deep learning techniques for DR detection results in performance improvements compared with a high performing classic, nondeep learning algorithm on a standardized dataset, that has been previously published.[1] The purpose was not to determine real world performance of the device or other DR detection algorithms. Real world performance can only be adequately determined from a prospective study on people with diabetes and standardized reading.

Though Messidor-2 contains 93 cases of vtDR according to the present reference standard, of which only eight have neovascularizations on the disc or elsewhere (i.e., PDR),[34] and two of eight have isolated PDR without DME. Thus, this study is underpowered to determine the detection performance of isolated PDR without ME–though none were missed by the device.

The Messidor-2 dataset used for this comparison study consists of high quality retinal images,[9] which are not necessarily a good representation of data from screening programs, generally, and certainly not reflective of the quality of images that are seen in the non-eye care settings where screening algorithms have the potential to deliver their biggest impact.[24] The ability to detect an ungradeable image is an important component when assessing the capabilities of a device for automated detection of diabetic retinopathy in the real world. Because of the relatively high quality of the images in Messidor-2, only a small number (4%) would have had an insufficient image quality output if the protocol had been complete. Thus, while Messidor-2 is a dataset that is useful in measuring performance of an algorithm on high quality exams, or comparing it with other algorithms, as in the present study, it is not sufficient to establish an algorithm's performance in broader clinical use. In addition, Messidor-2 images contain a single image per eye, limiting the area of retina covered. Many screening programs,[9] and algorithms such as the device, are designed with two images per eye, one fovea centered and one disc centered, leading to a larger area of retina examined. Using two or more images per eye, algorithms as well as human experts may find additional cases of DR not visible on the single image,[8] leading to different measured performance. Similarly, in the real world, reference standards often differ, depending on the characteristics of the clinicians reading the images and how many are involved in reading and how consensus is reached. For example, the ME reference standard was graded from the retinal images, which lack stereo, and no optical coherence tomography (OCT) was available. This implies that isolated retinal thickening cannot be appreciated,[7] though human expert detection of ME from exudates only, in single images, may be almost as sensitive as clinical stereo biomicroscopic analysis of retinal thickening.[35] Thus, DR and ME prevalence and severity may be underestimated in this dataset, and a different reference standard could lead to differences in a device's measured algorithmic performance. Finally, we purposely chose the device's rDR and vtDR outputs to have set points that can be expected to result in high sensitivity, to be able to compare performance of the IDP with the device. This indeed resulted in a high sensitivity of 96.8% and specificity of 87%. In the real world, algorithms such as the device can potentially have different set points that allow a more equal balance between sensitivity and specificity, in accordance with the prevalence of vtDR in the population as well as medical and public health objectives for screening.

Whether the resulting sensitivity is 'good enough'–though higher than typical clinicians evaluating images–is beyond the scope of this study, and is constrained by ethical, legal, and financial considerations, best addressed through ethical, legal, and cost-effectiveness analyses. Furthermore, while in the original study we used retinal specialists to set the reference standard–also used in this study and made publicly available–research studies have shown that, over time, clinicians increasingly deviate from methods using reading centers, as defined in the original standards.[36,37] Diabetic retinopathy detection likely also exhibits such a "diagnostic drift" from the original methods used by the reading center–for example, determining whether a red lesion is a microaneurysm or a hemorrhage, which can make the difference between a mild versus moderate level of DR–that were used in the primary

outcome studies that to a great degree still determine the management of DR:Diabetic Retinopathy Study,[38] Early Treatment of Diabetic Retinopathy Study,[39] and Epidemiology of Diabetes Interventions and Complications/Diabetes Control and Complications Trial,[40] and it is thus important to employ methods that are as close as possible to those.

The device detects rDR and vtDR, and, while automated detection of other diseases that manifest in the retina is desirable, the persistence of diabetes as the leading cause of preventable blindness among working age adults in the developed world, coupled with the established cost-effectiveness of DR screening, rank orders DR as the most important target to address through automated detection.[2,41]

In summary, the IDx-DR X2.1 device has achieved significantly better performance to detect referable DR, quantifiable by a 30% increase in specificity at the high reported sensitivity of 0.97 on a public dataset, using deep learning, through the use of CNN-based lesion detectors in a hybrid architecture. Deep-learning enhanced systems for automated detection of DR, thus have the potential to improve the efficiency and accessibility of DR screening, and thereby to prevent visual loss and blindness from this devastating disease.

Acknowledgments

The authors thank J.C. Klein, MD, Pascale Massin, MD, Beatrice Cochener, MD, PhD, and Philippe Gain, MD for organizing the Messidor study and creating the Messidor-2 resource, as well as Dennis P. Han, MD, Jonathan D. Walker, MD and David F. Williams, MD for creating the reference standard for Messidor-2.

Supported by grants from IDx LLC (Iowa City, IA, USA), Research to Prevent Blindness (New York, NY, USA). MDA is recipient of the Robert C. Watzke Endowment. This material is the result of work supported with resources and the use of facilities at the Iowa City Veterans Affairs Medical Center.

Disclosure: **M.D. Abràmoff**, IDx LLC (C, F, I, R, S), Research to Prevent Blindness (F), P; **Y. Lou**, None; **A. Erginay**, None; **W. Clarida**, IDx LLC (E, I); **R. Amelon**, IDx LLC (E, I); **J.C. Folk**, IDx LLC (I), Research to Prevent Blindness (F); **M. Niemeijer**, IDx LLC (E, I)

References

1. Abramoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131:351–357.
2. Helmchen LA, Lehmann HP, Abramoff MD. Automated detection of retinal disease. *Am J Manag Care*. 2014;11: 20(Spec No. 17):eSP48–eSP52.
3. Bragge P, Gruen RL, Chau M, Forbes A, Taylor HR. Screening for presence or absence of diabetic retinopathy: a meta-analysis. *Arch Ophthalmol*. 2011;129:435–444.
4. Hazin R, Colyer M, Lum F, Barazi MK. Revisiting diabetes 2000: challenges in establishing nationwide diabetic retinopathy prevention programs. *Am J Ophthalmology*. 2011;152:723–729.
5. National Health Services Diabetic Eye Screening Programme of the United Kingdom. *National Health Service Diabetic Retinopathy Programme Annual Report, April 2007-March 2008*. Gloucester, England: National Health Services Diabetic Eye Screening Programme of the United Kingdom; 2008.
6. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2008.
7. Wang YT, Tadarati M, Wolfson Y, Bressler SB, Bressler NM. Comparison of prevalence of diabetic macular edema based on monocular fundus photography vs optical coherence tomography. *JAMA Ophthalmol*. 2016;134:222–228.
8. Sun JK, Aiello LP. The future of ultrawide field imaging for diabetic retinopathy: pondering the retinal periphery. *JAMA Ophthalmol*. 2016;134:247–248.
9. Scanlon PH, Malhotra R, Greenwood RH. Comparison of two reference standards in validating two field mydriatic digital photography as a method of screening for diabetic retinopathy. *Br J Ophthalmol*. 2003;87:1258–1263.
10. Abramoff MD, Staal J, Suttorp MSA, Polak BC, Viergever MA. Low level screening of exudates and hemorrhages in background diabetic retinopathy. *Proceedings of the 1st International Workshop on Computer Assisted Fundus Image Analysis*. Copenhagen, Denmark, May 29–30, 2000:15.
11. Hipwell JH, Strachan F, Olson JA, McHardy KC, Sharp PF, Forrester JV. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. *Diabet Med*. 2000;17:588–594.
12. Ege BM, Hejlesen OK, Larsen OV, et al. Screening for diabetic retinopathy using computer based image analysis and statistical classification. *Comput Methods Programs Biomed*. 2000; 62:165–175.
13. Fleming AD, Goatman KA, Philip S, Prescott GJ, Sharp PF, Olson JA. Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts. *Br J Ophthalmol*. 2010;94;1606–1610.
14. Hansen MB, Abramoff MD, Folk JC, Mathenge W, Bastawrous A, Peto T. Results of automated retinal image analysis for detection of diabetic retinopathy from the Nakuru Study, Kenya. *PLoS One*. 2015;10:e0139148.
15. Roychowdhury S, Koozekanani DD, Parhi KK. DREAM: diabetic retinopathy analysis using machine learning. *IEEE J Biomed Health Inform*. 2014;18:1717–1728.
16. Trucco E, Ruggeri A, Karnowski T, et al. Validating retinal fundus image analysis algorithms: issues and a proposal. *Invest Ophthalmol Vis Sci*. 2013;54:3546–3559.
17. Quellec G, Lamard M, Cazuguel G, et al. Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs. *Invest Ophthalmol Vis Sci*. 2011;52:8342–8348.
18. Abramoff MD, Reinhardt JM, Russell SR, et al. Automated early detection of diabetic retinopathy. *Ophthalmology*. 2010;117: 1147–1154.
19. Chaum E, Karnowski TP, Govindasamy VP, Abdelrahman M, Tobin KW. Automated diagnosis of retinopathy by content-based image retrieval. *Retina*. 2008;28:1463–1477.
20. Fleming AD, Goatman KA, Philip S, et al. The role of haemorrhage and exudate detection in automated grading of diabetic retinopathy. *Br J Ophthalmol*. 2010;94:706–711.
21. Quellec G, Russell S, Abramoff MD. Optimal filter framework for automated, instantaneous detection of lesions in retinal images. *IEEE Trans Med Imaging*. 2011;30:523–533.
22. Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*. 1980;36:193–202.
23. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In XX ed(s). *Advances in Neural Information Processing Systems*. Publisher: City, State; 2012:1097–1105.
24. Kaggle, Inc. Diabetic Retinopathy Detection Vol. 2016. 2015. Available at: https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed September 1, 2016.
25. Laboratoire de Traitement de l'Information Médicale (LaTIM - INSERM U650). Messidor-2 dataset (Méthodes d'Evaluation de Systèmes de Segmentation et d'Indexation Dédiées à l'Ophtalmologie Rétinienne). 2011. Available at: http://latim.univ-brest.fr/indexfce0.html. Accessed September 19, 2016.

26. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. 2014; arXiv: 1409.1556.
27. Quellec G, Russell SR, Abramoff MD. Optimal filter framework for automated instantaneous detection of lesions in retinal images. *IEEE Trans Med Imaging*. 2011;30:523–533.
28. Hosmer DW, Lemeshow S. Applied Logistic Regression. Wiley: New York; 2000.
29. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 1994;309:188.
30. Quellec G, Abramoff MD. Estimating maximal measurable performance for automated decision systems from the characteristics of the reference standard. application to diabetic retinopathy screening. *Conf Proc IEEE Eng Med Biol Soc*. 2014;2014:154–157.
31. Mansberger SL, Gleitsmann K, Gardiner S, et al. Comparing the effectiveness of telemedicine and traditional surveillance in providing diabetic retinopathy screening examinations: a randomized controlled trial. *Telemed J E Health*. 2013;19: 942–948.
32. Mansberger SL, Sheppler C, Barker G, et al. Long-term comparative effectiveness of telemedicine in providing diabetic retinopathy screening examinations: a randomized clinical trial. *JAMA Ophthalmol*. 2015;133:518–525.
33. Abramoff MD, Suttorp-Schulten MS. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. *Telemed J E Health*. 2005;11:668–674.
34. Wilkinson CP, Ferris FL III, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110: 1677–1682.
35. Rudnisky CJ, Tennant MT, de Leon AR, Hinz BJ, Greve MD. Benefits of stereopsis when identifying clinically significant macular edema via teleophthalmology. *Can J Ophthalmol*. 2006;41:727–732.
36. Chen JH, Goldstein MK, Asch SM, Altman RB. Dynamically evolving clinical practices and implications for predicting medical decisions. *Pac Symp Biocomput*. 2016;21:195–206.
37. Arbel Y, Qiu F, Bennell MC, et al. Association between publication of appropriate use criteria and the temporal trends in diagnostic angiography in stable coronary artery disease: a population-based study. *Am Heart J*. 2016;175:153–159.
38. Photocoagulation treatment of proliferative diabetic retinopathy: the second report of diabetic retinopathy study findings. *Ophthalmology*. 1978;85:82–106.
39. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. *Ophthalmology*. 1991; 98:823–833.
40. American Diabetes Association. Executive summary: standards of medical care in diabetes–2012. *Diabetes Care*. 2012; 35(suppl 1):S4–S10.
41. Klonoff DC, Schwartz DM. An economic analysis of interventions for diabetes. *Diabetes Care*. 2000;23:390–404.

Investigative Ophthalmology & Visual Science