Automated Early Detection of Diabetic Retinopathy

Michael D. Abràmoff, MD, PhD,[1,2,3] Joseph M. Reinhardt, PhD,[4] Stephen R. Russell, MD,[1,2] James C. Folk, MD,[1,2] Vinit B. Mahajan, MD, PhD,[1,4] Meindert Niemeijer, PhD,[1,3,5] Gwénolé Quellec, PhD[1,6]

Purpose: To compare the performance of automated diabetic retinopathy (DR) detection, using the algorithm that won the 2009 Retinopathy Online Challenge Competition in 2009, the Challenge2009, against that of the one currently used in EyeCheck, a large computer-aided early DR detection project.

Design: Evaluation of diagnostic test or technology.

Participants: Fundus photographic sets, consisting of 2 fundus images from each eye, were evaluated from 16 670 patient visits of 16 670 people with diabetes who had not previously been diagnosed with DR.

Methods: The fundus photographic set from each visit was analyzed by a single retinal expert; 793 of the 16 670 sets were classified as containing more than minimal DR (threshold for referral). The outcomes of the 2 algorithmic detectors were applied separately to the dataset and were compared by standard statistical measures.

Main Outcome Measures: The area under the receiver operating characteristic curve (AUC), a measure of the sensitivity and specificity of DR detection.

Results: Agreement was high, and examination results indicating more than minimal DR were detected with an AUC of 0.839 by the EyeCheck algorithm and an AUC of 0.821 for the Challenge2009 algorithm, a statistically nonsignificant difference (z-score, 1.91). If either of the algorithms detected DR in combination, the AUC for detection was 0.86, the same as the theoretically expected maximum. At 90% sensitivity, the specificity of the EyeCheck algorithm was 47.7% and that of the Challenge2009 algorithm was 43.6%.

Conclusions: Diabetic retinopathy detection algorithms seem to be maturing, and further improvements in detection performance cannot be differentiated from best clinical practices, because the performance of competitive algorithm development now has reached the human intrareader variability limit. Additional validation studies on larger, well-defined, but more diverse populations of patients with diabetes are needed urgently, anticipating cost-effective early detection of DR in millions of people with diabetes to triage those patients who need further care at a time when they have early rather than advanced DR.

Financial Disclosure(s): Proprietary or commercial disclosure may be found after the references.
Ophthalmology 2010;117:1147–1154

Diabetic retinopathy (DR) is the most common cause of blindness in the working population of the United States and of the European Union.[1] Early detection (that is, screening) and timely treatment have been shown to prevent visual loss and blindness in patients with retinal complications of diabetes.[2–4] In the next decade, projections for the United States are that average age will increase, the number of people with diabetes in each age category will increase, and there will be an undersupply of qualified eye care providers, at least in the near term.[5] This so-called perfect storm of healthcare trends will challenge the public health capacity to care for both patients with DR and people with diabetes at risk for this complication.[6] If the previous scenario plays out, it will be necessary either to screen (perform early detection on) large numbers of people with diabetes for DR, to ration access to eye care, or both.

Several European countries successfully have instigated DR early detection programs using digital photography and reading of the images by human experts in their health care systems. In the United Kingdom, 1.7 million people with diabetes were screened for DR in 2007 and 2008.[7] In The Netherlands, more than 30 000 people with diabetes have been screened regularly since 2001 using an early detection project called EyeCheck (www.eyecheck.nl; accessed March 7, 2010).[8] The United States Department of Veterans Affairs has deployed a successful photoscreening program in the Veterans Affairs medical centers, through which 120 883 patients were screened in fiscal year 2008 (Cavallerano A, personal communication, 2009).

Over the last decade, many computer image analysis methods based on image processing and machine learning have been proposed to interpret digital photographs of the retina to increase the efficiency of early detection of DR.[9–23] Few of these methods have been assessed on a large scale in a population with a low incidence of DR that would mimic screening populations.[15,24,25]

The authors have continued to develop new approaches to improve the performance of their algorithms, originally with good success. More recently, they have achieved only limited performance improvements by making the algorithms more sophisticated (Invest Ophthalmol Vis Sci 47[suppl]:ARVO E-Abstract 2735, 2008; Invest Ophthal-

ISSN 0161-6420/10/$–see front matter
doi:10.1016/j.ophtha.2010.03.046

mol Vis Sci 48[suppl]:ARVO E-Abstract 3268, 2009).[26] They expect that given the known intraobserver and interobserver variability in human readers, against which such algorithms are compared, small improvements in performance, even though real, are less and less measurable. Another approach they chose to maximize performance was the organization of a worldwide online DR detection algorithm competition, the Retinopathy Online Challenge (http://roc.healthcare.uiowa.edu; accessed March 7, 2010), to allow investigators to compare their algorithms on a common dataset. The Challenge's intent was to allow the maximum number of research groups and individuals from around the world to participate, and 23 groups did so. The final winners were announced recently, as discussed at a 2009 Association for Research in Vision and Ophthalmology special interest group, and the methods and algorithms were published recently.[27] As organizers, the authors recused their algorithm from participating in this competition. The best performing algorithm in the Challenge was developed by Dr Quellec, then at INSERM 650 in Brest, France, which is termed the Challenge2009 algorithm in this article.[21]

Before translation into clinical use, it is essential to know whether these algorithms approach, or even surpass, the sensitivity and specificity of human detection of DR. This question cannot be answered directly, because there exists no single sensitivity and specificity—these vary for different readers, based on training, background, and other factors. However, whether the algorithms have performance comparable with that of a single reader or a small number of readers can be determined, as well as whether they are mature, that is, whether additional performance improvement can be expected. The authors' hypothesis is that DR detection algorithms are close to the sensitivity and specificity of a single human expert and are mature, that is, close to the measurable performance limit.

The automated algorithms introduced above are optimized to recommend referral for a patient with any form of DR to an ophthalmologist, and they were optimized to detect early DR, because in the authors' opinion, this is the main burden. Nevertheless, they could be modified to diagnose vision-threatening DR, that is, to detect those patients with significant nonproliferative DR, extensive clinically significant macular edema, or any form of proliferative DR. However, this category is small in existing—but not newly started—early detection programs. In addition, the algorithms were limited to detection of so-called red lesions only (microaneurysms and small hemorrhages) to make the comparison more valid, although the authors previously designed and evaluated systems that also detect exudates and cotton-wool spots[26,28] (Invest Ophthalmol Vis Sci 47[suppl]: ARVO E-Abstract 2735, 2008). The automated algorithms thus are optimized to recommend referral for a patient with any form of DR to an ophthalmologist.[29] To test this hypothesis, the EyeCheck algorithm[26] was compared with an independently derived one, the Challenge2009 algorithm, on the same large, single-reader dataset of people with diabetes who were previously known not to have DR.[8,15]

Patients and Methods

Study Population

The study was performed according to the tenets of the Declaration of Helsinki, and institutional review board approval was obtained. The researchers had access only to the deidentified images and their original diagnoses, and the study was Health Insurance Portability and Accountability Act compliant. Because of the retrospective nature of the study and deidentification, informed consent was judged not to be necessary by the institutional review board. From 16 670 people with diabetes who previously were not known to have DR (66 680 color retinal images), 16 670 first-time visits (4 images, 1 centered on the disc and 1 on the fovea for each eye) were selected retrospectively. These images came from the EyeCheck project for online early detection of DR in The Netherlands (www.eyecheck.nl).[8] The examinations were read by 1 of 3 retinal fellowship-trained ophthalmologists (MDA) using a strict protocol for the presence or absence of more than minimal DR,[8,15] the referral threshold for DR, as well as for sufficient quality. If no more than minimal DR was found, examinations also were evaluated for obvious non-DR abnormalities.

Imaging Protocol

As published in detail elsewhere, patients were photographed with nonmydriatic digital retinal cameras by trained technicians at 18 different sites using the Topcon NW 100, the Topcon NW 200, or the Canon CR5-45NM (all Topcon, Tokyo, Japan) cameras.[8] Across sites, 4 different camera settings were used: 640×480 pixels and 45° field of view (FOV), 768×576 pixels and 35° FOV, 1792×1184 pixels and 35° FOV, or 2048×1536 pixels and 35° FOV. Images were JPEG compressed at the minimum compression setting available, resulting in image files of approximately 0.15 to 0.5 MB. After automatic cropping of the black border, all retinal images were automatically resampled to 640×640 pixels.

Training the Algorithms on the Same Training Dataset

Both algorithms were applied to all 4 retinal fundus images. Optimal algorithm performance is reached when they are tuned to the image acquisition protocol(s) used in the early detection programs. To exclude any potential influence on performance from the training data, the same training data for both algorithms were used. One hundred images with red lesions segmented manually by 2 retinal specialists (MDA) were used; these images also were from the EyeCheck project but were not in the dataset used for testing. Fifty-five of these images contained a total of 852 red lesions, consisting of 36 379 pixels for training.

Brief Descriptions of the Algorithms

For a better understanding of how the algorithms work, brief descriptions are given here, and for more detail, the reader is referred to the original publications[15,16,26,28,30–38] (Abramoff MD, et al. Low level screening of exudates and haemorraghes in background DR. Paper presented at: First Computer Aided Fundus Image Analysis Conference, May 8, 2000, Aarhus, Denmark). The original algorithm, EyeCheck, first detects all pixels that appear to be in a red lesion based on pixel feature classification. Clusters of these candidate pixels are clustered in candidate lesions, and features then are extracted from each candidate lesion. These are processed with a kNN classifier to assign it a probability and to indicate the likelihood that it is a red lesion. This algorithm's

performance on a dataset obtained from 7689 retinal examinations as part of a complete DR screening system was published previously and resulted in an area under the receiver operating characteristic (ROC) curve (AUC) of 0.84.[15] More recently, a performance of AUC = 0.88 on a set of 15 000 similar examinations was obtained.[26]

The Challenge2009 algorithm in this study was developed by Dr Quellec and others at Inserm U650, in Brest University Hospital, Brest, France. The algorithm uses a parametric template defined for microaneurysms. Candidate lesions then are searched for in an adapted wavelet domain, where the classification performance of template matching is found to be optimal. Based on the distance to the parametric model, a probability is assigned to each candidate lesion, indicating the likelihood that it is a microaneurysm. A sensitivity and a positive predictive value of 89% were reported for the detection of microaneurysms in 120 retinal images of several imaging methods.[21] Unlike the EyeCheck algorithm, the Challenge2009 algorithm was designed to detect microaneurysms only, although some other red lesions such as pinpoint hemorrhages can be detected. Both of these algorithms calculate a DR risk estimation for an examination consisting of 4 images within minutes on a standard desktop personal computer.

The EyeCheck algorithm is capable of detecting exudates and cotton-wool spots, although this did not result in a substantial performance improvement.[26,28] Because the Challenge2009 algorithm cannot detect these bright lesions, detection of exudates and cotton-wool spots was turned off in the EyeCheck algorithm in this study for comparison purposes. The EyeCheck algorithm also is capable of discarding images of insufficient quality.[26,32] Because the Challenge2009 does not have an image quality component, only images of sufficient quality, as determined by the first reader, were used in this study.

Data Analysis

There is a lack of consensus in the scientific literature on measuring reader agreement and limited guidance on comparing algorithms to human readers.[39] Therefore, 2 commonly used approaches were used. For both algorithms, a probability was assigned to each extracted candidate lesion, reflecting its likelihood of being a microaneurysm or red lesion.[16] Although varying the threshold on the maximum probability for normal or abnormal cutoff, the sensitivity and specificity of each method was compared with that of the human expert as a standard. The resulting multiple sensitivity and specificity pairs are used to create a ROC curve. The ROC curve shows the sensitivity and specificity at different thresholds, and the system can be set for a desired sensitivity or specificity pair simply by selecting a desired threshold. The AUC is considered the most comprehensive measure of system performance, where an area of 1 has sensitivity = specificity = 1 and represents perfect detection, and an area of 0.5 represents a system that essentially performs a coin toss.[40] The level of agreement between the 2 detection methods was also assessed by a κ statistic for several sensitivity settings.[41]

Calculating the Theoretical Limit of the Area under the Receiver Operating Characteristic Curve for a Given Dataset

Early in the analysis, it became clear that using a single human expert as the reference standard implicitly would incorporate an error rate related to the human κ statistic. Therefore, the theoretical limit for AUC was calculated based on the κ statistic. In a previous study, the authors determined the intraobserver and interobserver variability of 3 experts reading the photographs by analyzing the performance of 3 retinal specialists,[15] compared with each expert's first reading, on a different sample of 500 examinations from the EyeCheck dataset. These experts evaluated the 500 examinations for more than minimal DR. Their sensitivity and specificity were found to be 73% and 89%, 62% and 84%, and 85% and 89%, respectively; whereas their κ values were 0.71, 0.55, and 0.41, respectively, and their average κ thus was 0.55. This $\kappa = 0.55$ is within the range of κ (0.34–0.63) in 2 other studies on interobserver agreement of reading images for DR.[42,43] Assume a perfect algorithm, with 0 false-negative results and 0 false-positive results, compared with the true state of disease in the patients. Also assume that this algorithm reads 500 examinations, of which 50 truly have more than minimal DR (a fairly typical 10% of cases). By definition, this algorithm always gives the correct read, and therefore has an AUC of 1.0 (compared with the true state of disease) on this dataset.

Assume human readers with a κ statistic of 0.55 (see above[15]). These human readers, when reading the same dataset of 500 examinations, will make errors compared with the true state of the disease, according to their κ. The resulting number of errors, y, can be calculated from κ and the prevalence of true disease, P, in the dataset, as: $y = n(1–A)(1– \kappa)$, with $A = P^2+(1–P)^2$ and $n = 500$ the size of the dataset.[44] Under the above assumptions, the human reader will make $y = 40$ errors (total of false positives and false negatives).

The performance of any algorithm can be compared only with the human reads, because access to the true state of disease in the dataset is not available. Only access to the human reads is available, which is erroneous to some degree as explained, but which has to be the reference dataset. In the real world, there is no manner in which one can know better what the true state of disease is than this reference dataset.

Potential for Combining Both Algorithms

If both algorithms achieve a similar performance in terms of AUC, but produce different outcomes as measured by κ, the algorithms are complementary. However, if they produce similar outcomes, we would use the method leading to the highest performance. The output of the 2 algorithms was combined by selecting the lesion candidates that were detected by at least 1 of the 2 methods (logical OR combination), which is expected to lead to a system that is more sensitive than either, but at the same specificity.[26]

Results

A single reader identified 690 of the 16 670 subjects in the dataset to have 1 or more images of insufficient quality and found 793 of 15 980 subjects with sufficient quality to have more than minimal DR on 1 examination. Thus, the lower bound of the prevalence of more than minimal DR in this population, assuming none of the patients with insufficient quality images had DR, was 4.7%, whereas the upper bound, assuming all patients with insufficient quality images had DR, was 8.9%. Only the 16 670 examinations of the subjects with sufficient quality images were used as the dataset to compare the algorithms. The reading by the single reader was used as the reference standard for this study to evaluate the performance of the 2 algorithms. The percentage of more than minimal DR was low because some patients were referred not for DR but for other abnormalities. In the dataset, 8192 of 16 670 (51.2%) were women, the average age±standard deviation was 71±7.4 years, and the average hemoglobin A1c±standard deviation within 3 months of the examination was 6.78±1.34%.

Examples of images with more than minimal DR are shown in Figure 1, and the locations where the EyeCheck algorithm and the

Challenge2009 algorithm, or their combination, found microaneurysms are indicated for comparison. The ROC curves obtained for the EyeCheck and Challenge2009 algorithms are shown in Figure 2; the AUC for the EyeCheck algorithm was 0.839, and that for the Challenge2009 algorithm was 0.821, as presented in Table 1. The pairwise z-scores of the differences in AUC between these are reported in Table 2, showing that there was no significant performance difference between the EyeCheck and the Challenge2009. At 90% sensitivity, the specificity of the algorithms for the EyeCheck was 47.7% and that for the Challenge2009 was 43.6%.

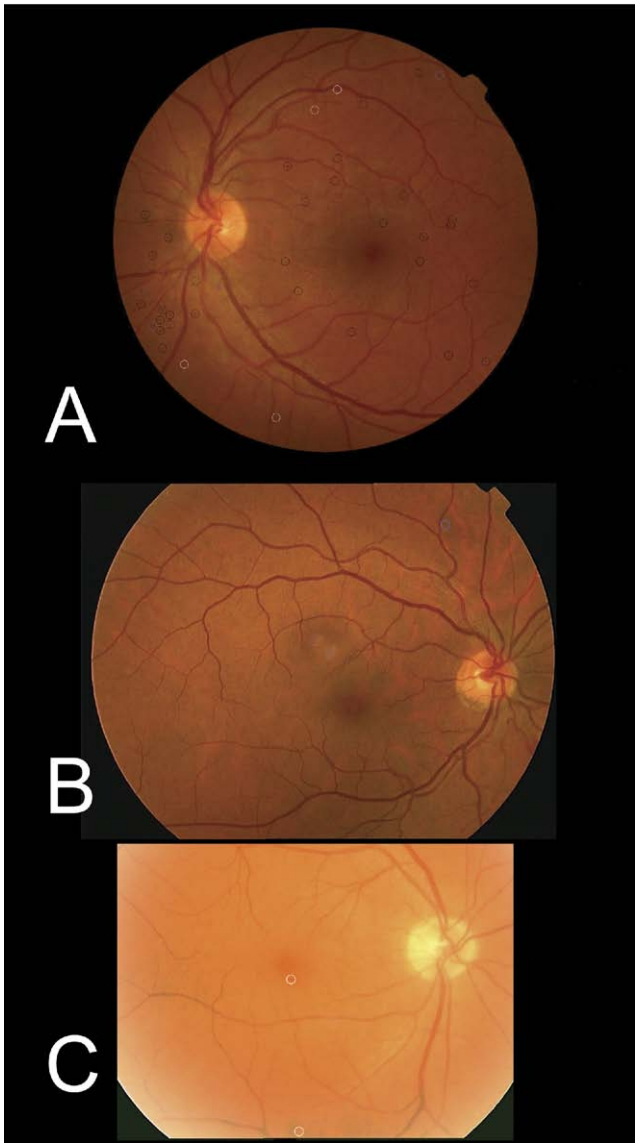

Figure 1. Three fundus photographs showing patients with more than minimal diabetic retinopathy. **A**, EyeCheck algorithm and the Challenge2009 algorithm both agreed that the patient has more than minimal diabetic retinopathy. **B**, Only the EyeCheck algorithm detected more than minimal diabetic retinopathy. **C**, Only the Challenge2009 algorithm detected more than minimal diabetic retinopathy. Black circles indicate lesions detected by both methods, blue circles lesions only detected by the EyeCheck algorithm, and white circles indicate lesions only detected by the Challenge2009 algorithm. Because the dataset is diverse in image size, resolution, and field of view, these are different for all 3 images.

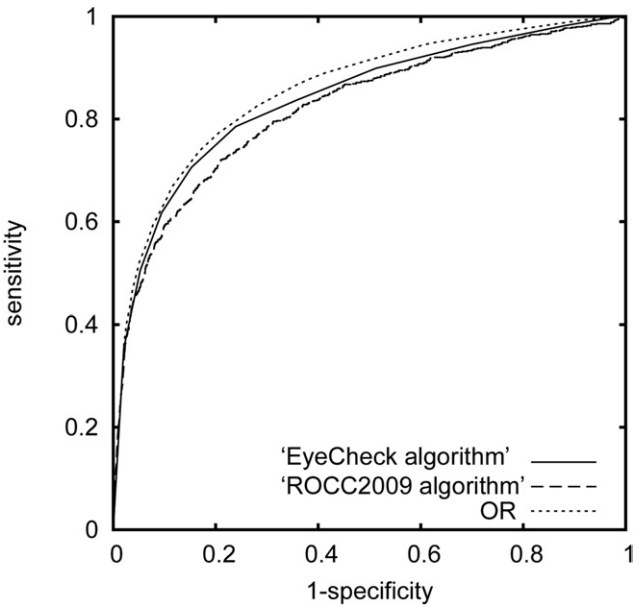

Figure 2. Receiver operating characteristic curves (ROCC) for the EyeCheck algorithm and the Challenge2009 algorithm and their combination (indicated with OR [odds ratio]).

To contrast better these results to human performance and interreader variability, we calculated the κ statistic of agreement between the EyeCheck algorithm and Challenge2009 algorithm for 3 different sensitivity settings, as presented in Table 3. At a sensitivity setting of 83.7% for each algorithm, for example, the κ statistic was 0.304. This sensitivity setting for both algorithms correspondingly missed 129 of 793 abnormal examinations (false negatives); 62 of 793 were missed by both algorithms, 67 of 793 were missed by the EyeCheck but not by Challenge2009, and 67 of 793 were missed by the Challenge2009 but not by the EyeCheck. This difference in output shows that these algorithms truly are different in how they detect DR. The combination of both algorithms (logical OR fusion) gave an AUC of 0.86, which was significantly larger than either the isolated EyeCheck or Challenge2009 algorithm; see Table 2 and Figure 2. At a sensitivity setting of 90%, the specificity of the combined algorithm (logical OR fusion) was 54.7%.

As noted in "Patients and Methods," the performance AUC limit achievable for an algorithm on this dataset was calculated. This performance limit was based on the average κ statistic 0.55 for 3 human readers who had examined all images (not in this study) of the examinations of 500 people with diabetes from the EyeCheck project, and the *P* value (prevalence of true disease) was 0.047. The AUC limit given these numbers is 0.86 for a perfect

Table 1. Area under the Receiver Operating Characteristics Curve and Standard Error Obtained for EyeCheck and Challenge2009 Algorithms and the Combined Algorithm

	EyeCheck	Challenge2009	Combined Algorithm
AUC	0.839	0.821	0.86
SE	0.0089	0.0092	0.0084

AUC = area under the receiver operating characteristic curve; SE = standard error.

Table 2. Z-Test on the Differences between Area under the Receiver Operating Characteristic Curve for the EyeCheck and Challenge2009 Algorithms and the Combined Algorithm

	EyeCheck	Challenge2009	Combined Algorithm
EyeCheck	0	1.91	**2.33**
Challenge2009		0	**4.25**
Combined Algorithm			0

Boldface characters indicate statistical difference with a 95% confidence level.

algorithm. Any AUC over 0.86 is thus not meaningful, and it is not possible to measure any performance improvement over this AUC on this dataset with these expert readings.

Discussion

These results show that the performance of the independently derived Challenge2009 (AUC, 0.82) was not different from that of the EyeCheck algorithm (AUC, 0.84) when tested on the same dataset. The AUC of each of these algorithms and for the combination of both (AUC, 0.86) now is close to or at the mathematical limit of detection for this dataset. Given that equal AUC was reached by 2 totally independently developed algorithms and that this AUC is close to the theoretical limit, it is unlikely that a potential third algorithm can improve on this performance. As noted in "Patients and Methods," the performance AUC limit achievable for an algorithm on this dataset was calculated. This performance limit was based on the average κ statistic 0.55 for 3 human readers who had examined all images (not in this study) of 500 people with diabetes from the EyeCheck project, and the P value (prevalence of true disease) was 0.044. The AUC limit given these numbers was 0.86 (for a perfect algorithm). Any AUC over 0.86 thus is not meaningful, and it is not possible to measure any performance improvement over this AUC.

Although the algorithms led to similar results, they clearly are unique and do not produce the same outcomes at the patient level, as the algorithms' κ statistics in Table 3 show. Combining both methods therefore led to small but significant increases in performance over either system alone. For example, at a sensitivity setting of 90%, the combined algorithm had a specificity of 54.7%, whereas the individual algorithms achieved either 47.7% (EyeCheck algorithm) or 43.6% (Challenge2009 algorithm).

The current best practices standard for the evaluation of DR is 7-field stereo fundus photography read by trained readers in accordance with the Early Treatment Diabetic Retinopathy Study.[45] However, no such best practice standard has ever been available for large populations, which would allow it to be used to evaluate the performance of an automatic detection system.[15] In most large-scale screening systems, only a single expert's reading is available for screened images, as is the case for the patient examinations in this study, and this establishes a statistical limit of measure of performance.[15,24,25]

For the anticipated use of an automated system—for triage, either online, or incorporated into the fundus camera—a high sensitivity is a safety issue and is more important than a high specificity, which is an efficiency issue. Therefore, if the sensitivity of the combined algorithm is adjusted to 90%, which is higher than the 85%, 73%, or 62% sensitivity of the human readers, a specificity of 54.7% is retained, which is lower than the 89% or 84% of the human experts. Based on achieved performance, the algorithm misses fewer patients with DR, but increases the number of false-positive patients. Methods to increase the specificity may include human review of all positives—10% of all examinations, a reduction of 90% compared with human review of all examinations.

In the United States, the impending limitation in the availability of eye care providers coupled with demographically driven increases in age and the number of people with diabetes are starting to exceed the capacity to address this population. The Centers for Disease Control and Prevention indicate that physician undersupply will reach 20% by 2015.[46,47] United States eye care providers examine only approximately 50% of the 23 million people with diabetes.[6] To meet an examination rate of 90% rather than the current 50% would require an estimated additional 10 million annual patient visits. Such a goal would tax the limited resources available for eye care greatly, a problem that would be aggravated with increasing numbers of people with diabetes. Private and public health carriers or state or federal authorities then may consider alternative methods of screening for DR. The adoption of digital camera technology with automated detection systems, such as presented in this study, may fulfill the current and future needs of DR screening.

The application of digital cameras and especially computer reading of these images for early detection, rather than an office visit to an eye care provider, remains controversial.[48] There is a concern about quality of care, because a visit to an eye care provider involves more than the evaluation of the retina for the presence of DR and may result in detection of other pathologic features, such as glaucoma or cataract. Some may be comfortable with digital photography and reading of the images by eye care providers but not by a computer algorithm. Nevertheless, this study shows that computer algorithms seem to be at least as good as a human reader and have the potential to address the needs of the almost 50% of people with diabetes who currently do not undergo regularly any form of dilated eye examination.

This study has several limitations. First, these 2 algorithms were compared with respect to DR detection performance only, and therefore image quality was not considered

Table 3. κ Statistic of Agreement between the EyeCheck and Challenge2009 Algorithms for Specific Sensitivity Settings

Sensitivity	61.7%	70.6%	83.7%
κ	0.374	0.343	0.304

when randomly selecting 16 770 examinations from as many people with diabetes (excluding the training dataset). In practice, a complete automated system would have to ensure adequate image quality before proceeding. The authors and others previously have tested and published image quality assessment algorithms that perform at this level and can be used to exclude insufficient quality images.[15,32] In the authors' experience, approximately 10% of patients have examination results judged to be of insufficient quality by the human reader, which leads to either reimaging or referral to an eye care provider.

Second, the incidence of DR in this population (4.4%) is somewhat low and is most likely the result of excellent metabolic control reflected by an average±standard deviation hemoglobin A1C of 6.78±1.34%.[8] Other populations with less optimal metabolic control are likely to have a higher incidence of DR, and testing the algorithms on such a dataset of meaningful size has not yet been possible.

Third, and most important, a single reader assessment of more than minimal DR is not the comparative (gold) standard, although the examination of a single field retinal photograph per eye has been shown to be sensitive and specific enough to detect early signs of DR.[29] This can be seen from the relatively low κ between experts as reported, and this also limits the ability to measure algorithm performance. This κ was obtained on expert reading from 500 examinations, and not on the entire dataset, because rereading tens of thousands of images was not yet feasible. However, $\kappa = 0.55$ is within the range of κ (0.34–0.63) in a study on interobserver agreement of reading images for DR in a recent paper in *Ophthalmology*,[42] and also in an older study on the same subject.[43] The authors currently are collecting a dataset according to a stricter imaging standard with the goal of increasing the agreement between human readers, as measured by κ. Especially for a system that potentially may be used as a triage system for large populations, in which some images are never read by human experts, validation to an existing standard such as the 7-field stereo photograph evaluation according to the Early Treatment Diabetic Retinopathy Study levels is important.

Fourth, as mentioned at the outset, the algorithms as tested were limited in that they detected only early DR, because in the authors' opinion, this is the main burden.[29] They also are limited in that they detect red lesions only (microaneurysms and small hemorrhages), but not exudates, cotton-wool spots, or isolated retinal thickening without associated exudates, although they are still capable of detecting the vast majority of cases of early DR.[26,28]

In summary, DR detection algorithms achieve comparable performance to a single retinal expert reader and are close to mature, and further measurable improvements in detection performance are unlikely. For translation into clinical practice sooner rather than later, validation on well-defined populations of patients with diabetes, with variable metabolic control and racial and ethnic diversity, are more urgent than further algorithm development. The authors anticipate that automated systems based on algorithms such as those discussed herein will allow cost-effective early detection of DR in millions of people with diabetes and will allow triage of those patients who need further care at a time when they have early rather than advanced DR.

References

1. Klonoff DC, Schwartz DM. An economic analysis of interventions for diabetes. Diabetes Care 2000;23:390–404.
2. 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;107:19–24.
3. Kinyoun JL, Martin DC, Fujimoto WY, Leonetti DL. Ophthalmoscopy versus fundus photographs for detecting and grading diabetic retinopathy. Invest Ophthalmol Vis Sci 1992; 33:1888–93.
4. Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. Ophthalmology 1991;98(suppl):766–85.
5. National Diabetes Statistics, 2007. National Institutes of Health. Available at http://diabetes.niddk.nih.gov/DM/PUBS/statistics/references.htm. Accessed March 10, 2010.
6. Mokdad AH, Bowman BA, Ford ES, et al. The continuing epidemics of obesity and diabetes in the United States. JAMA 2001;286:1195–200.
7. National Health Service. The English Diabetic Retinopathy Programme Annual Report, 1 April 2007–31 March 2008. 2008:8–9. Available at: http://www.retinalscreening.nhs.uk/userFiles/File/Annual%20Report%202007–08%20post-final%20release%202009–03–11.pdf. Accessed March 5, 2010.
8. Abramoff MD, Suttorp-Schulten MS. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemed J E Health 2005;11:668–74.
9. Teng T, Lefley M, Claremont D. Progress towards automated diabetic ocular screening: a review of image analysis and intelligent systems for diabetic retinopathy. Med Biol Eng Comput 2002;40:2–13.
10. Cree MJ, Olson JA, McHardy KC, et al. A fully automated comparative microaneurysm digital detection system. Eye (Lond) 1997;11:622–8.
11. Frame AJ, Undrill PE, Cree MJ, et al. A comparison of computer based classification methods applied to the detection of microaneurysms in ophthalmic fluorescein angiograms. Comput Biol Med 1998;28:225–38.
12. Hipwell JH, Strachan F, Olson JA, et al. Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool. Diabet Med 2000;17: 588–94.
13. Olson JA, Strachan FM, Hipwell JH, et al. A comparative evaluation of digital imaging, retinal photography and optometrist examination in screening for diabetic retinopathy. Diabet Med 2003;20:528–34.
14. Spencer T, Olson JA, McHardy KC, et al. An image-processing strategy for the segmentation and quantification of microaneurysms in fluorescein angiograms of the ocular fundus. Comput Biomed Res 1996;29:284–302.
15. Abràmoff MD, Niemeijer M, Suttorp-Schulten MS, et al. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. Diabetes Care 2008;31: 193–8.

16. Niemeijer M, van Ginneken B, Staal J, et al. Automatic detection of red lesions in digital color fundus photographs. IEEE Trans Med Imaging 2005;24:584–92.
17. Larsen M, Godt J, Larsen N, et al. Automated detection of fundus photographic red lesions in diabetic retinopathy. Invest Ophthalmol Vis Sci 2003;44:761–6.
18. Larsen N, Godt J, Grunkin M, et al. Automated detection of diabetic retinopathy in a fundus photographic screening population. Invest Ophthalmol Vis Sci 2003;44:767–71.
19. Fleming AD, Philip S, Goatman KA, et al. Automated microaneurysm detection using local contrast normalization and local vessel detection. IEEE Trans Med Imaging 2006; 25:1223–32.
20. Walter T, Klein JC, Massin P, Erginay A. A contribution of image processing to the diagnosis of diabetic retinopathy—detection of exudates in color fundus images of the human retina. IEEE Trans Med Imaging 2002;21:1236–43.
21. Quellec G, Lamard M, Josselin PM, et al. Optimal wavelet transform for the detection of microaneurysms in retina photographs. IEEE Trans Med Imaging 2008;27:1230–41.
22. Karnowski TP, Govindasamy V, Tobin KW, et al. Retina lesion and microaneurysm segmentation using morphological reconstruction methods with ground-truth data. Conf Proc IEEE Eng Med Biol Soc 2008;2008:5433–6.
23. Tobin KW, Abramoff MD, Chaum E, et al. Using a patient image archive to diagnose retinopathy. Conf Proc IEEE Eng Med Biol Soc 2008;2008:5441–4.
24. 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;91:1512–7.
25. Scotland GS, McNamee P, Philip S, et al. Cost-effectiveness of implementing automated grading within the National Screening Programme for diabetic retinopathy in Scotland. Br J Ophthalmol 2007;91:1518–23.
26. Niemeijer M, Abramoff M, van Ginneken B. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. IEEE Trans Med Imaging 2009;28:775–85.
27. Niemeijer M, van Ginneken B, Cree MJ, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imaging 2010;29:185–95.
28. Niemeijer M, van Ginneken B, Russell SR, et al. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. Invest Ophthalmol Vis Sci 2007;48: 2260–7.
29. Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. Am J Ophthalmol 2002;134:204–13.
30. Abràmoff MD, Niemeijer M. The automatic detection of the optic disc location in retinal images using optic disc location regression. Conf Proc IEEE Eng Med Biol Soc 2006;1: 4432–5.
31. Sanchez CI, Niemeijer M, Kockelkor T, et al. Active learning approach for detection of hard exudates, cotton wool spots, and drusen in retinal images. In: Karssemeijer N, Giger ML, eds. Medical Imaging 2009: Computer-Aided Diagnosis, 10–12 February 2009, Lake Buena Vista, Florida, United States. Bellingham, WA: SPIE; 2009:72601I. Proceedings of SPIE—the International Society for Optical Engineering, v. 7260.
32. Niemeijer M, Abràmoff MD, van Ginneken B. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Med Image Anal 2006;10:888–98.
33. Niemeijer M, Abràmoff MD, van Ginneken B. Segmentation of the optic disc, macula and vascular arch in fundus photographs. IEEE Trans Med Imaging 2007;26:116–27.
34. Niemeijer M, Staal JS, van Ginneken B, et al. Comparative study of retinal vessel segmentation on a new publicly available database. In: Fitzpatric JM, Sonka M, eds. Medical Imaging 2004: Image Processing, 16–19 February 2004, San Diego, California, United States. Bellingham, WA: SPIE; 2004:648-56. Proceedings of SPIE—the International Society for Optical Engineering, v. 5370.
35. Niemeijer M, Abramoff MD, van Ginneken B. Automated localization of the optic disc and the fovea. Conf Proc IEEE Eng Med Biol Soc 2008;2008:3538–41.
36. Niemeijer M, van Ginneken B, Abramoff MD. Automatic classification of retinal vessels into arteries and veins. In: Karssemeijer N, Giger ML, eds. Medical Imaging 2009: Computer-Aided Diagnosis, 10–12 February 2009, Lake Buena Vista, Florida, United States. Bellingham, WA: SPIE; 2009:72601F. Proceedings of SPIE—the International Society for Optical Engineering, v. 7260.
37. Staal J, Abramoff MD, Niemeijer M, et al. Ridge-based vessel segmentation in color images of the retina. IEEE Trans Med Imaging 2004;23:501–9.
38. Staal J, Kalitzin SN, Abramoff MD, et al. Classifying convex sets for vessel detection in retinal images. In: 2002 IEEE International Symposium on Biomedical Imaging: Proceedings: July 7–10, 2002, Ritz-Carlton Hotel, Washington, D.C., United States. Piscataway, NJ: IEE; 2002:269–72. Available at: http://home.versatel.nl/berendschot/articles/Staa12002.pdf. Accessed March 5, 2010.
39. Hagen MD. Test characteristics: how good is that test? Prim Care 1995;22:213–33.
40. Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997;53:370–82.
41. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. Ophthalmology 1991; 98(suppl):786–806.
42. Ruamviboonsuk P, Teerasuwanajak K, Tiensuwan M, Yuttitham K, Thai Screening for Diabetic Retinopathy Study Group. Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. Ophthalmology 2006;113:826–32.
43. Milton RC, Ganley JP, Lynk RH. Variability in grading diabetic retinopathy from stereo fundus photographs: comparison of physician and lay readers. Br J Ophthalmol 1977;61:192–201.
44. Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. J Clin Pathol 1996;49: 597–9.
45. Early Treatment Diabetic Retinopathy Study Research Group. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Ophthalmology 1991; 98(suppl):823–33.
46. Lee PP, Hoskins HD Jr, Parke DW III. Access to care: eye care provider workforce considerations in 2020. Arch Ophthalmol 2007;125:406–10.
47. Cooper RA, Getzen TE, McKee HJ, Laud P. Economic and demographic trends signal an impending physician shortage. Health Aff (Millwood) 2002;21:140–54.
48. Chew EY. Screening options for diabetic retinopathy. Curr Opin Ophthalmol 2006;17:519–22.

Footnotes and Financial Disclosures

Originally received: June 16, 2009.
Final revision: March 18, 2010.
Accepted: March 19, 2010.
Available online: April 17, 2010. Manuscript no. 2009-816.

[1] Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, Iowa City, Iowa.

[2] Department of Veterans Affairs, Iowa City VA Medical Center, Iowa City, Iowa.

[3] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa.

[4] Omics Laboratory, University of Iowa Hospitals and Clinics, Iowa City, Iowa.

[5] Image Sciences Institute, University of Utrecht, Utrecht, The Netherlands.

[6] Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa.

Financial Disclosure(s):
The author(s) have made the following disclosure(s):
Michael D. Abràmoff - Patents - University of Iowa; Owner - EyeCheck
Meindert Niemeijer - Patents - University of Iowa; Owner - EyeCheck
Gwénolé Quellec - Patents - University of Iowa; DR detection algorithms
Supported by the National Eye Institute, Bethesda, Maryland (grant no.: NEI-EY017066 [MDA]); Research to Prevent Blindness, Inc., New York, New York (SRR, JCF, VBM); the University of Iowa, Iowa City, Iowa; and the Netherlands Organization for Health Related Research (MN), The Hague, Netherlands.

Correspondence:
Michael D. Abràmoff, MD, PhD, Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242.