Published in final edited form as:
Ophthalmology. 2010 June ; 117(6): 1147–1154. doi:10.1016/j.ophtha.2010.03.046.

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**[1,3,6], and **Gwénolé Quellec, PhD**[1,5]

[1] Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA [2] Department of Veterans Affairs, Iowa City VA Medical Center, 601 Highway 6 West, Iowa City, IA 55242, USA [3] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA [4] Omics Laboratory, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA [5] Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242, USA [6] Image Sciences Institute, University of Utrecht, Netherlands

Abstract

Purpose—To compare the performance of automated diabetic retinopathy (DR) detection, using the algorithm that won the 2009 Retinopathy Online Challenge Competition in 2009, ('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 two 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 16,770 sets were classified as containing more than minimal DR (threshold for referral). The outcomes of the two algorithmic detectors were applied separately to the dataset and 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 exams containing more than minimal DR were detected with an AUC of 0.839 by the 'Eyecheck' algorithm and an AUC of 0.821 for 'Challenge2009', a

Address for reprints: Department of Ophthalmology and Visual Sciences, University of Iowa Hospitals and Clinics, 200 Hawkins Drive, Iowa City, IA 52242, USA.

Conflicts of Interest

Michael D. Abramoff (P,S) reports that he is named as inventor on a patent, assigned to the University of Iowa, on which the 'EyeCheck' algorithm is based, and that he is named as inventor on additional patent applications, assigned to the University of Iowa, related to DR detection algorithms, and (O) is a shareholder in the EyeCheck project in the Netherlands.
Meindert Niemeijer (P) reports that he is named as inventor on a patent, assigned to the University of Iowa, on which the 'EyeCheck' algorithm is based
Gwenole Quellec (P) reports that he is named as inventor on a patent application, assigned to the University of Iowa, related to DR detection algorithms.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

statistically non-significant difference (z-score 1.91). If either of the algorithms detected DR in combination, 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 the 'Challenge2009' algorithm, 43.6%.

Conclusions—DR detection algorithms appear to be maturing, and further improvements in detection performance cannot be differentiated from best clinical practices, because the performance of competitive algorithm development has now reached the human intra-reader variability limit. Additional validation studies on larger, well-defined, but more diverse populations of patients with diabetes are urgently needed, 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.

Introduction

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 ('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 the 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 "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 to either screen (perform early detection on) large numbers of people with diabetes for DR, ration access to eyecare, or both.

Several European countries have successfully instigated diabetic retinopathy 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 diabetic retinopathy in 2007–2008. [7] In the Netherlands, over 30,000 people with diabetes were screened since 2001 in the same period through an early detection project called EyeCheck (www.eyecheck.nl, last accessed March 7, 2010). [8] The United States (US) Department of Veterans Affairs (VA) has deployed a successful photoscreening program in the VA medical centers, through which 120,883 patients were screened in FY 2008. (A. Cavallerano, personal communication)

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 in order 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 which would mimic screening populations. [15, 24–25]

We have continued to develop new approaches to improve the performance of our algorithms, originally with good success. More recently, we have achieved only limited performance improvements by making the algorithms more sophisticated. (Invest Ophthalmol Vis Sci 47 [Suppl]: ARVO E-Abstract 2735, 2008; Invest Ophthalmol Vis Sci 48 [Suppl]: ARVO E-Abstract 3268, 2009)[26]. We expect that given the known intra- and inter-observer variability in human readers, against which such algorithms are compared, small improvements in performance, even though real, are less and less measurable. Another approach we chose to maximize performance was the organization of a worldwide online DR detection algorithm competition, the Retinopathy Online Challenge (http://roc.healthcare.uiowa.edu, last 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 twenty-three groups were willing to participate. The final winners were announced recently, as discussed at a 2009 Association for Research in Vision and Ophthalmology (ARVO) special interest group, and the methods and algorithms have been accepted for publication[27]. As organizers, we recused our 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 'Challenge2009' algorithm in this paper[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, as there exists no single sensitivity and specificity – these vary for different readers, based on training, background and other factors. However, we can determine whether the algorithms have performance comparable to a single or a small number of readers, and we can also test whether they are mature, in other words, whether additional performance improvement can be expected, or not. Our hypothesis is that DR detection algorithms are close to the sensitivity and specificity of a single human expert, and are mature, i.e., 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 our opinion, this is the main burden. Nevertheless, they could be modified to diagnose vision threatening DR, i.e., detect those patients with significant non proliferative DR, extensive CSME, 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, though we have 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 are thus optimized to recommend referral for a patient with any form of DR to an ophthalmologist[29].

To test this hypothesis, we compared our algorithm[26] ('EyeCheck') to an independently derived one, ('Challenge2009' algorithm), on the same large single reader dataset of people with diabetes who were previously known not to have DR[8, 15].

Methods

Study population

The study was performed according to the tenets of the Declaration of Helsinki, and Institutional Review Board, (IRB), approval was obtained. The researchers had access only to the de-identified images and their original diagnoses, and the study was Health Insurance Portability and Accountability Act (HIPAA) compliant. Because of the retrospective nature of the study and de-identification, informed consent was judged not to be necessary by the IRB. We retrospectively selected 16670 first time visits (four images, one centered on the disc and one on the fovea for each eye) from 16670 people with diabetes who were previously not known to have DR, (66,680 color retinal images). These images came from the EyeCheck project for online early detection of DR in the Netherlands (www.eyecheck.nl)[8]. The exams were read by one of three retinal fellowship trained ophthalmologists utilizing 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, exams were also evaluated for obvious non-DR abnormalities.

Imaging Protocol

As published in detail elsewhere, patients were photographed with “non-mydriatic” digital retinal cameras by trained technicians, at 10 different sites, using either the Topcon NW 100, the Topcon NW 200 (Topcon, Tokyo, Japan), or the Canon CR5-45NM (Topcon, Tokyo, Japan) cameras[8]. Across sites, four different camera settings were used: 640×480 pixels and 45° field of view (FOV), 768×576 pixels and 35° FOV, 1,792×1,184 pixels and 35° FOV or 2,048×1,536 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 four 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, we used the same training data for both algorithms. We used one hundred images with red lesions segmented manually by two retinal specialists, these images were also 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, 30–39] (Abramoff MD, Staal JJ, Suttorp MS, Polak BCP, Viergever MA. Low level screening of exsudates and haemorraghes in background diabetic retinopathy. First Computer Aided Fundus Image Analysis Conference. Aarhus, Denmark, May 2000). Our original algorithm, (‘EyeCheck’ algorithm), first detects all pixels that appear like they might be in a red lesion based on pixel feature classification. Clusters of these candidate pixels are clustered in candidate lesions, and features are then extracted from each candidate lesion. These are processed with a kNN classifier in order to assign it a probability and indicating the likelihood that it is a red lesion.[37] We previously published this algorithm’s performance on a dataset obtained from 7,689 retinal exams as part of a complete DR screening system, obtaining an area under the ROC curve (AUC) of 0.84[15]. More recently, we were able to obtain a performance of AUC=0.88 on a set of 15,000 similar exams[26].

The ‘Challenge2009’ algorithm in this study was developed by Dr. Quellec, *et al*., then at *Inserm U650*, in Brest University Hospital in France, and now at Iowa. The algorithm uses a parametric template defined for microaneurysms. Candidate lesions are then searched for in an adapted wavelet domain, where the classification performance of template matching is found 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 modalities[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 exam consisting of four images within minutes on a standard desktop personal computer.

The ‘EyeCheck’ algorithm is capable of detecting exudates and cotton-wool spots, though this did not result in a substantial performance improvement[26, 28]. Because the ‘Challenge2009’ algorithm cannot detect these bright lesions, we switched detection of exudates and cotton-wool spots off in the 'EyeCheck' algorithm in this study for comparison purposes. The 'EyeCheck' algorithm is also 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[40]. Therefore, we used two commonly used approaches. For both algorithms, a probability was assigned to each extracted candidate lesion, reflecting its likelihood of being a microaneurysm or red lesion[16]. While varying the threshold on the maximum probability for normal/abnormal cutoff, the sensitivity and specificity of each method was compared to the human expert as a standard. The resulting multiple sensitivity/specificity pairs are used to create a receiver operating characteristic (ROC) curve. The ROC curve shows the sensitivity and specificity at different thresholds, and the system can be set for a desired sensitivity/specificity pair by simply selecting a desired threshold. The Area Under the ROC Curve (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[41] The level of agreement between the two detection methods was also assessed by a κ (kappa) statistic for several sensitivity settings [42].

Calculating the theoretical limit of AUC for a given dataset

Early in our analysis we realized that by using a single human expert as our reference standard, would implicitly incorporate an error rate related to the human κ statistic. Therefore we calculated the theoretical limit for AUC based on the κ statistic. In a previous study, we determined the intra- and inter- observer variability of three experts reading the photographs by analyzing the performance of three retinal specialists[15], compared to each expert's first reading, on a different sample of 500 exams from the EyeCheck dataset. These experts evaluated the 500 exams for more than minimal DR. We found their sensitivity/specificity to be 73%/89%, 62%/84%, and 85%/89%, respectively, while their κ were 0.71, 0.55, and 0.41 respectively, and their average κ thus 0.55. This κ =0.55 is within the range of κ (0.34 to 0.63) in two other studies on interobserver agreement of reading images for DR [43–44]. Assume a perfect algorithm, with zero false negatives and zero false positives, compared to the true state of disease in the patients. Also assume that this algorithm reads 500 exams 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 to 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 exams, will make errors compared to 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[45]. 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 only be compared to the human reads, because we do not have access to the true state of disease in the dataset. We only have access to the human reads, 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 we 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. On the other hand, if they produce similar outcomes, then we would simply use the method leading to the highest performance. We combined the output of the two algorithms, by selecting the lesion candidates that were detected by at least one of the two methods (“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 out of the 16,670 subjects in the dataset to have one or more images of insufficient quality, and found 793 of 15980 subjects with sufficient quality to have more than minimal DR on one exam. 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%, while the upper bound, assuming all patients with insufficient quality images had DR, was 8.9%. Only the 16,670 exams 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 two 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, 8,192/16,670 (51.2%) were women, the average age was 71 years ± standard deviation (SD) 7.4 years, and the average HbA1c within 3 months of the exam was 6.78% ± SD 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. ROC curves obtained for ‘EyeCheck’ and ‘Challenge2009’ algorithms are shown in Figure 2; the AUC for the ‘EyeCheck’ algorithm was 0.839, and for the ‘Challenge2009’ algorithm, 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 ‘EyeCheck’ and ‘Challenge2009’. At 90% sensitivity, the specificity of the algorithms for ‘EyeCheck’ was 47.7% and for ‘Challenge2009’, 43.6%.

In order to better contrast these results to human performance and inter-reader variability we calculated the κ statistic of agreement between the ‘EyeCheck’ algorithm and ‘Challenge2009’ algorithm for three 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. At this sensitivity setting both algorithms correspondingly missed 129/793 abnormal exams (false negatives); 62/793 were missed by both algorithms, 67/793 by ‘EyeCheck’ but not by ‘Challenge2009’, and 67/793 by ‘Challenge2009’ but not by ‘EyeCheck’. This difference in output shows that these algorithms are truly different in how they detect DR. The combination of both algorithms (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 (OR fusion) was 54.7%.

As noted in the Methods, we calculated the performance AUC limit achievable for an algorithm on this dataset. We based this performance limit on the average κ statistic 0.55 for three human readers who had examined all images (not in this study), of the exams of 500 people with diabetes from the ‘EyeCheck’ project, and the *P* (prevalence of true disease) =4.7%. The AUC limit given these numbers is 0.86 for a perfect 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 our 'EyeCheck' algorithm (AUC 0.84) when tested on the same dataset. AUC of each of these algorithms and for the combination of both (AUC 0.86) is now close to or at the mathematical limit of detection for this dataset. Given that equal AUC was reached by two totally independently developed algorithms and this AUC is close to the theoretical limit, makes it unlikely that a potential third algorithm can improve on this performance. As noted in the Methods, we calculated the performance AUC limit achievable for an algorithm on this dataset. We based this performance limit on the average κ statistic 0.55 for three human readers who had examined all images (not in this study), of 500 people with diabetes from the 'EyeCheck' project, and the *P* (prevalence of true disease) = 4.4%. The AUC limit given these numbers was 0.86 (for a perfect algorithm). Any AUC over 0.86 is thus not meaningful, and it is not possible to measure any performance improvement over this AUC.

Though the algorithms led to similar results they are clearly unique and do not produce the same outcomes at the patient level, as the algorithms' κ statistic in Table 3 shows. 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%, while 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 seven-field stereo fundus photography read by trained readers in accordance with the Early Treatment Diabetic Retinopathy Study (ETDRS).[46] 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's exams in this study, and this establishes a statistical limit of measure of performance[15, 24–25].

We believe that 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 more important, than a high specificity, which is an *efficiency issue*. Therefore if we adjust the sensitivity of the combined algorithm to 90% which is higher than the 73% or 62% sensitivity of the human readers, we retain a specificity of 54.7%, lower than the 89% or 84% of the human experts. Based on achieved performance, the algorithm would miss fewer patients with DR but increase the number of false positive patients. Methods to increase the specificity might include human review of all positives – 10% of all exams, a reduction of 90% compared to human review of all exams.

In the US, 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 take care of this population. The Centers for Disease Control and Prevention indicate that physician undersupply will reach 20% by 2015[47–48]. U.S. eye care providers only examine about 50% of the 23 million people with diabetes. [6] In order to meet an examination rate of 90% rather than the current 50%, would required 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 may then consider alternative methods of screening for DR. We propose that 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 of rather than an office visit to an eye care provider remains controversial. [49] 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 pathology, 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 appear to be at least as good as a human reader and have potential to address the needs of the almost 50% of people of diabetes who currently do not undergo regular any form of dilated eye exam.

This study has several limitations. First, we wanted to compare these two algorithms on DR detection performance only, and therefore did not consider image quality when randomly selecting 17,877 exams 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. We and others have previously tested and published image quality assessment algorithms that perform at this level and can be used to exclude insufficient quality images [15, 32]. In our experience, approximately 10% of patients have exams judged to be of insufficient quality by the human reader, which leads to either re-imaging or referral to an eye care provider.

Second, the incidence of DR in this population (4.4%) is somewhat low and is most likely due to excellent metabolic control reflected by an average HbA1C of 6.78% ±1.34%[8]. Other populations with less optimal metabolic control are likely to have a higher incidence of DR, and we have not yet been able to test the algorithms on such a dataset of meaningful size.

Third, and most important, a single reader assessment of more than minimal DR is not the comparative (“gold”) standard, even though 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 our ability to measure algorithm performance. We obtained this κ on expert reading from 500 exams, 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 to 0.63) in a study on interobserver agreement of reading images for DR in a recent paper in this journal [43], and also in an older study on the same subject[44]. We are currently 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 may potentially 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 ETDRS levels is important.

Fourth, as mentioned in the Introduction, the algorithms as tested were limited in that they only detected early DR, because in our opinion, this is the main burden[29]. They are also 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, though still capable of detecting the vast majority 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. We anticipate that automated systems based on algorithms such as discussed here, will allow cost-effective early detection of DR in millions of people with diabetes, and allow triage of those patients who need further care at a time when they have early rather than advanced DR.

Acknowledgments

Financial support: Abràmoff: NEI-EY017066; Research to Prevent Blindness (RPB); NY, NY; University of Iowa; Netherlands Organization for Health Related Research (NWO)

Niemeijer: NWO

Russell, Folk, Mahajan: RPB

Dr. Folk is the Judith Gardner and Donald H. Beisner, M.D., Professor of Vitreoretinal Diseases and Surgery

References

1. Klonoff DC, Schwartz DM. An economic analysis of interventions for diabetes. Diabetes Care 2000;23:390–404. [PubMed: 10868871]
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. [PubMed: 10647713]
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. [PubMed: 1582794]
4. Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. Ophthalmology 1991;98(suppl):766–85. [PubMed: 2062512]
5. National Diabetes Statistics. National Institutes of Health; 2007 [last accessed March 10, 2010]. AQ: vague and unverifiable citation; please go to http://diabetes.niddk.nih.gov/DM/PUBS/statistics/ and find specific source such that readers can see from where you are pulling your data http://diabetes.niddk.nih.gov/DM/PUBS/statistics/references.htm
6. PLease replace the citation, which is too similar to the previous one (5), to this one: Mokdad AH, Bowman BA, Ford ES, et al. The continuing epidemics of obesity and diabetes in the United States. JAMA 2001;286:1195–1200. [PubMed: 11559264]
7. National Health Service. The English Diabetic Retinopathy Programme Annual Report, 1 April 2007-31 March 2008. 2008 [Accessed March 5, 2010]. p. 8-9.Available at: http://www.retinalscreening.nhs.uk/userFiles/File/Annual%20Report%202007-08%20post-final%20release%202009-03-11.pdfI meant this report,and page 8 is where discussion of these numbers starts
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. [PubMed: 16430386]
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. [PubMed: 11954703]
10. Cree MJ, Olson JA, McHardy KC, et al. A fully automated comparative microaneurysm digital detection system. Eye (Lond) 1997;11:622–8. [PubMed: 9474307]
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. [PubMed: 9784961]
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. [PubMed: 11073180]
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. [PubMed: 12823232]
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. [PubMed: 8812075]

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. [PubMed: 18024852]

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. [PubMed: 15889546]

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. [PubMed: 12556411]

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. [PubMed: 12556412]

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. [PubMed: 16967807]

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. [PubMed: 12585705]

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. [PubMed: 18779064]

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. [PubMed: 19163946]

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. [PubMed: 19163948]

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. [PubMed: 17504851]

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. [PubMed: 17585001]

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. [PubMed: 19150786]

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. [PubMed: 19822469]

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. [PubMed: 17460289]

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. [PubMed: 12140027]

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. [PubMed: 17947087]

31. Sanchez, CI.; Niemeijer, M.; Kockelkor, T., et al. In: Karssemeijer, N.; Giger, ML., editors. Active learning approach for detection of hard exudates, cotton wool spots, and drusen in retinal images; Medical Imaging 2009: Computer-aided diagnosis; 10–12 February 2009; Lake Buena Vista, Florida, United States. Bellingham, WA: SPIE; 2009. p. 72601IProceedings of SPIE--the International Society for Optical Engineering;

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. [PubMed: 17138215]

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. [PubMed: 17243590]

34. Niemeijer, M.; Staal, JS.; van Ginneken, B., et al. In: Fitzpatric, JM.; Sonka, M., editors. Comparative study of retinal vessel segmentation on a new publicly available database; Medical Imaging 2004: Image processing; 16–19 February 2004; San Diego, California, USA. Bellingham, WA: SPIE; 2004. p. 648-56.Proceedings of SPIE--the International Society for Optical Engineering;

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. [PubMed: 19163472]

36. Niemeijer, M.; van Ginneken, B.; Abramoff, MD. In: Karssemeijer, N.; Giger, ML., editors. Automatic classification of retinal vessels into arteries and veins; 10–12 February 2009; Lake Buena Vista, Florida, United States. Proceedings of SPIE--the International Society for Optical Engineering; Bellingham, WA: SPIE; 2009. p. 72601FMedical Imaging 2009: Computer-aided diagnosis

37. Niemeijer, M.; van Ginneken, B.; Abràmoff, MD. Automatic detection and classification of microaneurysms and small hemorrhages in color fundus photographs. Proc Computer Aided Fundus Imaging and Analysis (CAFIA); 2003.

38. 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. [PubMed: 15084075]

39. Staal, J.; Kalitzin, SN.; Abramoff, MD., et al. Classifying convex sets for vessel detection in retinal images. 2002 IEEE International Symposium on Biomedical Imaging: proceedings; July 7–10, 2002; Ritz-Carlton Hotel, Washington, D.C., USA. Piscataway, NJ: IEE; 2002 [Accessed March 5, 2010.]. p. 269-72.Available at: http://home.versatel.nl/berendschot/articles/Staal2002.pdf

40. Hagen MD. Test characteristics: how good is that test? Prim Care 1995;22:213–33. [PubMed: 7617782]

41. Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997;53:370–82. [PubMed: 9147602]

42. 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. [PubMed: 2062513]

43. 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. [PubMed: 16650679]

44. 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. [PubMed: 851521]

45. Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. J Clin Pathol 1996;49:597–9. [PubMed: 8813964]

46. 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. [PubMed: 2062515]

47. Lee PP, Hoskins HD Jr, Parke DW III. Access to care: eye care provider workforce considerations in 2020. Arch Ophthalmol 2007;125:406–10. [PubMed: 17353416]

48. Cooper RA, Getzen TE, McKee HJ, Laud P. Economic and demographic trends signal an impending physician shortage. Health Aff (Millwood) 2002;21:140–54. [PubMed: 11900066]

49. Chew EY. Screening options for diabetic retinopathy. Curr Opin Ophthalmol 2006;17:519–22. [PubMed: 17065919]

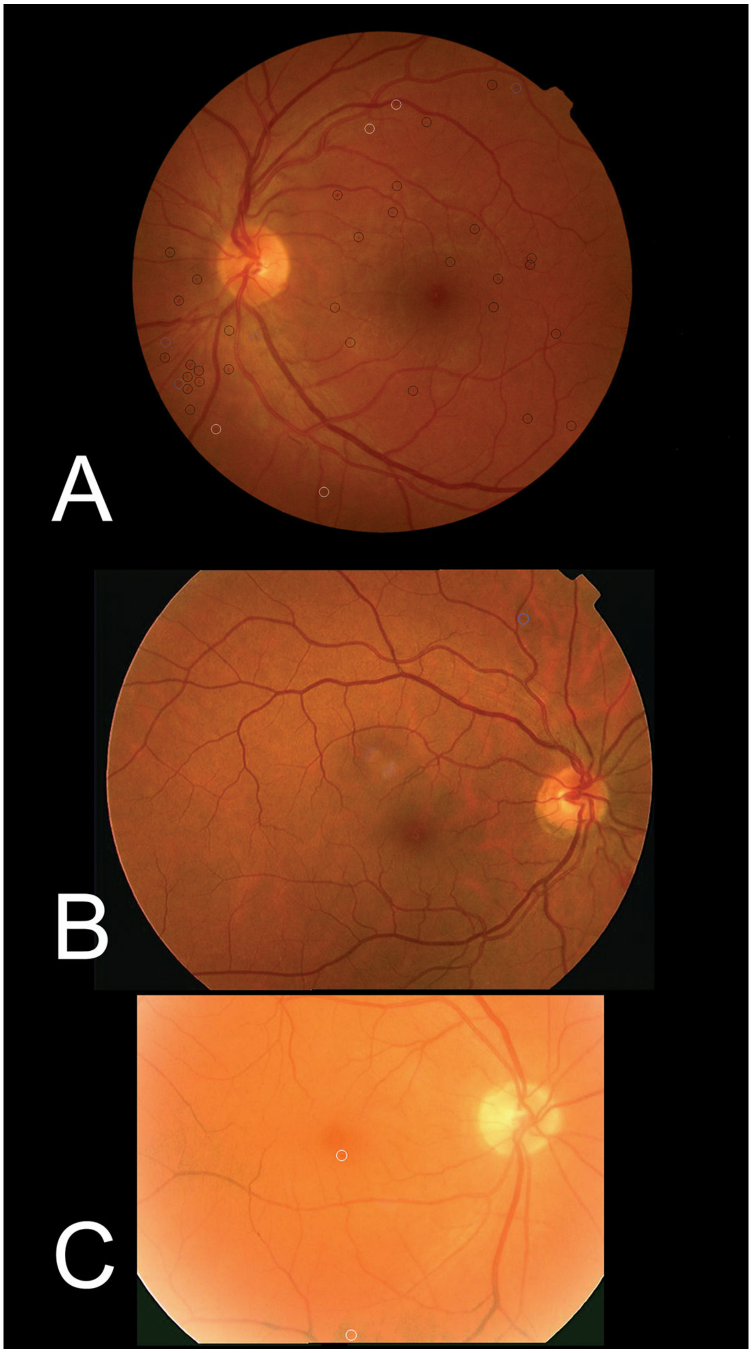

Figure 1.

Three images of patients with more than minimal diabetic retinopathy. In image (a), the 'EyeCheck algorithm' and the 'Challenge2009 algorithm' both agreed that the patient has more than minimal diabetic retinopathy. In image (b) only the 'EyeCheck algorithm', and in (c) only 'Challenge2009' 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 three images.

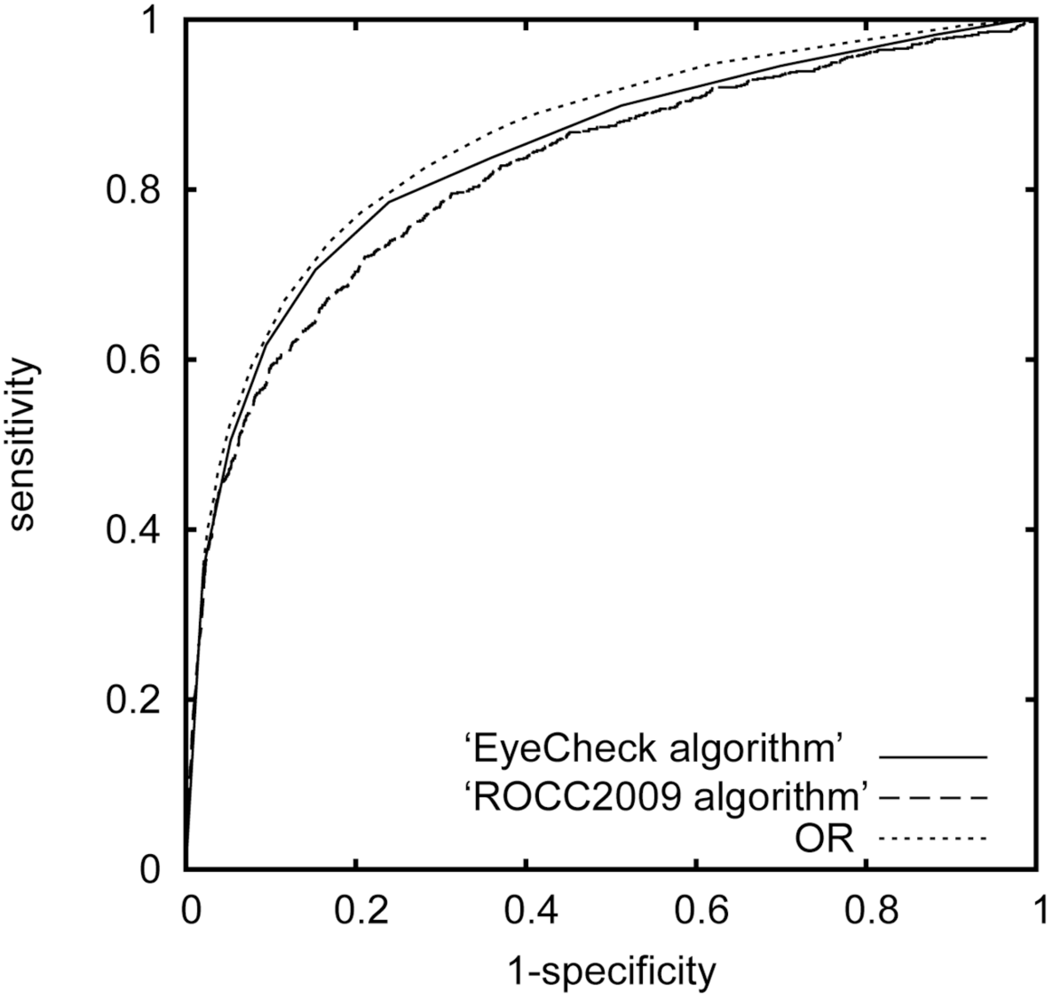

Figure 2.

Receiver Operating Characteristic curves for the 'EyeCheck algorithm' and the 'Challenge2009 algorithm', and their combination (indicated with 'OR').

Computer algorithms that detect diabetic retinopathy from retinal color images have reached a performance comparable to retinal specialists, allowing cost-effective triage of millions of people with diabetes currently not receiving standard of care eye exams.

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 characteristics curve; SE=standard error

Table 2

Z-test on the differences between Area under the Receiver Operating Characteristics Curve for the 'EyeCheck' and 'Challenge2009' algorithms and the combined algorithm.

	EyeCheck	**Challenge2009**	**Combined algorithm**
Algorithm 1	0	1.91	***2.33***
Algorithm 2		0	***4.25***
Combined algorithm			0

Bold italic characters indicate statistical difference with a 95% confidence level.

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