tvst

Special Issue

Evidence Based Prediction and Progression Monitoring on Retinal Images from Three Nations

Lutfiah Al Turk[1], Su Wang[2], Paul Krause[2], James Wawrzynski[8], George M. Saleh[8], Hend Alsawadi[3], Abdulrahman Zaid Alshamrani[4], Tunde Peto[5], Andrew Bastawrous[6], Jingren Li[7], and Hongying Lilian Tang[2]

[1] Department of Statistics, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
[2] Department of Computer Science, University of Surrey, Guildford, Surrey, UK
[3] Faculty of Medicine, King Abdulaziz University, Saudi Arabia
[4] Ophthalmology Department, Faculty of Medicine, University of Jeddah, Saudi Arabia
[5] Centre for Public Health, Queen's University Belfast, Northern Ireland, UK
[6] International Centre for Eye Health, Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK
[7] 7th Medical Center of PLA General Hospital, Diabetes Professional Committee of China, Geriatric Health Association, P.R. China
[8] NIHR Biomedical Research Centre at Moorfield Eye Hospital and the UCL Institute of Ophthalmology, London, UK

Correspondence: Paul Krause, Department of Computer Science, University of Surrey, Guildford, Surrey, GU2 7XH. e-mail: p.krause@surrey.ac.uk

Received: January 27, 2020
Accepted: June 18, 2020
Published: August 7, 2020

Keywords: diabetic retinopathy; lesion detection; deep learning; AI algorithm; diabetes

Citation: Al Turk L, Wang S, Krause P, Wawrzynski J, Saleh GM, Alsawadi H, Alshamrani AZ, Peto T, Bastawrous A, Li J, Tang HL. Evidence based prediction and progression monitoring on retinal images from three nations. Trans Vis Sci Tech. 2020;9(2):44, https://doi.org/10.1167/tvst.9.2.44

Purpose: The aim of this work is to demonstrate how a retinal image analysis system, DAPHNE, supports the optimization of diabetic retinopathy (DR) screening programs for grading color fundus photography.

Method: Retinal image sets, graded by trained and certified human graders, were acquired from Saudi Arabia, China, and Kenya. Each image was subsequently analyzed by the DAPHNE automated software. The sensitivity, specificity, and positive and negative predictive values for the detection of referable DR or diabetic macular edema were evaluated, taking human grading or clinical assessment outcomes to be the gold standard. The automated software's ability to identify co-pathology and to correctly label DR lesions was also assessed.

Results: In all three datasets the agreement between the automated software and human grading was between 0.84 to 0.88. Sensitivity did not vary significantly between populations (94.28%–97.1%) with specificity ranging between 90.33% to 92.12%. There were excellent negative predictive values above 93% in all image sets. The software was able to monitor DR progression between baseline and follow-up images with the changes visualized. No cases of proliferative DR or DME were missed in the referable recommendations.

Conclusions: The DAPHNE automated software demonstrated its ability not only to grade images but also to reliably monitor and visualize progression. Therefore it has the potential to assist timely image analysis in patients with diabetes in varied populations and also help to discover subtle signs of sight-threatening disease onset.

Translational Relevance: This article takes research on machine vision and evaluates its readiness for clinical use.

Introduction

Diabetic retinopathy (DR) is a common complication of diabetes mellitus. Among patients with diabetes, DR prevalence is approximately 28.5% in the United States,[1] 34.08% in China,[2] and 34.6% in Saudi Arabia.[3] The diagnosis of DR early in the presymptomatic phase through screening is critical to the eventual visual outcome and relies on a detailed analysis of fundus photographs taken regularly (e.g., often annually) within DR screening programs. At present,

photographs are most commonly analyzed by ophthalmologists, optometrists, and professional graders.

Several classification systems have been developed and adopted to guide DR screening frequency and ophthalmic referral based on a population's needs or the resources available in different parts of the world. Two commonly used systems are the International Clinical Diabetic Retinopathy and Diabetic Macular Oedema Severity Scale (ICDRS) and the one defined by the UK National Screening Committee (NSC).[4,5] Guidance on screening intervals, investigation, and treatment are also incorporated into these guidelines.[6–8]

In England and Wales, where more than 80% of those with diabetes undergo DR screening at least annually, DR is no longer the leading cause of blindness in the working-age population.[7] However, the majority of countries around the world have no such established screening program; one of the barriers remains the need to have sufficient trained staff to manually grade every fundus image captured. The availability of automated image grading might become a facilitator to support DR screening services in resource limited settings.

Over the past two decades, automated retinal image analysis for DR detection and grading has been studied extensively. Computer vision and machine learning methods have been proposed.[9–12] The rise of deep learning,[13,14] typically implemented as convolutional neural networks (CNNs),[15–22] has given a significant boost to the field of automated DR detection. This facilitated the usage of large datasets of fundus images to improve the accuracy and scalability of DR recognition and classification.[13,14]

However, such systems continue to suffer from significant limitations:

- The inability of many deep learning systems to provide detail or evidence to support their "black-box" DR classification;
- Absence of the capacity to detect and grade DR within the broader context of other possible diagnoses such as age-related macular degeneration (AMD) and retinal vein occlusion, to minimize false-positive results and indicate the presence of non-DR pathologies.

Hypothesis of the Study

This study aims to evaluate the ability of a retinal image analysis software system to provide effective detection of referable DR and surrogate markers of diabetic macular edema (DME) and monitor the progression of DR and DME based on either the ICDRS or NSC grading criteria. The measure of its performance is based on its agreement with human graders and clinical assessment. Validations were carried out on external testing data collected from three geographic locations, Kenya, Saudi Arabia, and China, with different camera types and settings. The software also provides evidence of its prediction through visualizing relevant lesions and identifies the presence of copathology while not confusing it with DR.

Methods

DAPHNE Automated Software

DAPHNE is an automated system for retinal image analysis developed by the University of Surrey, UK. DAPHNE was originally developed as a software system for diabetic retinopathy filtering of normal images.[23] Over the years it has evolved with the addition of a range of components with the aim of supporting a holistic reading of retinal images on key pathologic manifestations of DR and other disorders. Two major components are evaluated in this work; one is an image-based classifier, the other is an object detector.

The DAPHNE classifier is an end to end CNN architecture with multiple output layers for different classifications based on the training samples annotated on their quality, as well as DR grade in either ICDRS or NSC (see below, in the Retinal Images Datasets section of this article). Images were first preprocessed to subtract local average color to reduce differences in lighting.[24] These were then augmented to increase spatial, rotational, and scale variance. To speed up the "learning" process, batch normalization and pre-initialization were used. Pre-initialization also improved performance of the CNN network. The CNN framework was trained to provide multiple outputs, including (1) quality assessment; (2) ICDRS DR grades (e.g., 0—no DR, 1—mild, 2—moderate, 3—severe, or 4—proliferative); (3) UK NSC DR grades (e.g., R0—no DR, R1—background, R2—preproliferative, or R3—proliferative); and (4) presence of referable DR.

The DAPHNE object detector has a number of components. A set of U-net[25] and CNN-based detectors were trained to detect retinal anatomic structures, such as the optic disc and macula; the DR lesions, such as microaneurysms, hemorrhages, exudates, and other lesions, such as intraretinal microvascular abnormality (IRMA), new vessel on disc, new vessel elsewhere, cotton wool, drusen, and venous beading.

The rest of this section will first describe the prediction workflow after the algorithms are trained, followed by further information on training and validation datasets, prediction output categories as well as how the prediction accuracy is reported.

DAPHNE Workflow

In the first stage of the processing, raw data in any image format are cropped by removing any black mask borders around the retina region then passed through the classifier network to obtain a prediction on both quality and DR scales. To minimize the throwing away of those low readability images caused by the presence of certain pathologies, the probability outputs on both quality measure and disease measures are fed into a logistic regression model parametrized by some samples of images with pure quality issues and those with conditions such as cataract and retinal detachment. This filtering process does not intend to achieve 100% accuracy but aims to pick up some portion of the low quality images caused by different pathological conditions, if any, so they could be processed further. Otherwise those images with low quality scores are marked as ungradable. All the images that are deemed to have a quality score indicating adequate quality, are passed to the next stage.

In the second stage, DAPHNE detectors output the locations of anatomical structures in the fundus image, as well as the likelihood that a certain region is pathologic. This works together with the DAPHNE classifier that outputs the DR severity grading. The detected pathological regions that are consistent with the predicted grading level are visualized as evidence for the predicted grade. With regard to DME analysis, the DAPHNE detector detects and visualizes the location of the optic disc, fovea, and any exudate around the macula region, as well as any appearance of microaneurysms or hemorrhages within 1-disc diameter of the fovea.

Analysis for Progression

DR is a progressive disease, and UK[26] data suggest that it may be possible to stratify patients for risk, using grading outcomes only, into groups with low and high risk of progressing to proliferative DR. Subsequently, screening intervals for such diverse groups of patients could then safely be modified according to their risk stratification. In our work the DR progression monitoring was carried out by combining the results from individual image classification on disease level and then adding in the DAPHNE detectors to visualize the changes between time-points.

When analyzing a set of images taken for DR screening of the same patients at different examination time points, the system first applies detected anatomical structures to register between baseline and follow-up images of the same eye. It then computes any change in the severity of DR by comparing the grading results of these images. After registration, the lesion detectors are used to extract and visualize the following morphologic changes in pathology (see Fig. 1):

- Any new lesion;
- Any disappearing lesion;
- Any change of existing lesions (smaller or bigger compared with baseline images).

Retinal Image Datasets

Training Sets

The development of the DAPHNE classifier was undertaken partly using 35,124 macula-centered retinal fundus images generated by EyePACS and available at Kaggle.[17] These images were already labeled by EyePACS using the ICDRS scheme. We also annotated these images using the NSC grading scheme and used 28,100 of them as part of our training and testing dataset. The remaining 20% (7024 images) served as an internal validation test set. Cameras used to capture the images include the Optovue iCam, Centervue DRS, Topcon NW and Canon CR1/DGi/CR2 using 45° fields-of-view. As with any typical real-world dataset, this dataset included photographs that contained artefacts or could be out of focus, underexposed or overexposed. Another 4980 images were selected from our own collections, from different cameras and ethnic groups, with the annotations in both ICDRS and NSC agreed by three trained graders. Together these form 33,800 training sample images for learning each grading standard (ICDRS and NSC).

For lesion detection in the DAPHNE detector, we first used a public database (DiaRetDB1).[19] All images were obtained using the same 50° field-of-view fundus camera with varying imaging settings. It contains 89 digital retinal images and a human-expert annotated ground truth for several common DR lesions, including microaneurysms, hemorrhages and exudates. We then added a further 1952 images annotated by three trained graders and a medical retina specialist on retinal pathologic regions. After random sampling and preprocessing, 50,000 sub-images in smaller patch sizes are sampled from 2041 (89 + 1952) images for training. These subimages include those regions with or without pixels where the lesions are located to form negative and positive samples. For negative samples, normal patches on various locations of the retina without any pathology can be used. There are many possibilities to sample such subimages with the lesions appearing at

Figure 1. *Top row:* images from the same eye taken on October 26, 2015 (baseline), November 3, 2015 (dot hemorrhage), and March 8, 2016 (MA, hemorrhage and preretinal hemorrhage). *Bottom row:* the comparison of morphological changes for DR signs between a baseline image (October 26, 2015) and a follow-up retinal image (March 8, 2016).

different positions of the patch for positive samples. Positive samples include those features appearing in DR, as well as those non-DR but pathological. Fifty thousand patch samples allowed the algorithm to be trained under different scenarios.

As a general strategy, during the training, all the training data were divided randomly into training and testing sets on the basis of an 80/20 split strategy. These data were not used for any internal or external validation. There was no patient-level overlap in the training and testing sets in either the internal or external validation.

External Validation Test Sets

To test the generalizability and reproducibility of the software, we used three distinct datasets for external validation acquired with varying imaging settings and cameras from the three countries: 15,000 from China, 10,026 from Saudi Arabia, and 24,700 from Kenya.[18]

Table 1. An Overview of the Training and Internal and External Validation Test Sets

	0	1	2	3	4
ICDRS					
Training samples on DAPHNE classifier using 28,100 from Kaggle	20647	1955	4234	698	566
Internal validation (7024 from Kaggle)	5161	488	1058	175	142
External validation datasets					
Kenya	11479	9463	3395	329	34
NSC	R0	R1	R2	R3	—
Additional training samples on DAPHNE classifier (only its distribution in NSC is shown for simplicity)	3659	346	750	224	—
External validation datasets					
NSC	R0	R1	R2	R3	—
China	9986	3279	1240	495	—
Saudi Arabia	7451	1854	582	139	—

0, no DR; 1, mild; 2, moderate; 3, severe; 4, proliferative; R0, no DR; R1; background; R2, preproliferative; R3, proliferative.

China. Images were obtained from DR screening, fully anonymized locally and with appropriate permissions in place. The gold-standard for DR grading using the NSC grading criteria was carried out by trained and certified graders. Two fundus images were taken in each eye; one optic disc centered and the other macula centered. No follow up images or clinical assessments were received.

Saudi Arabia. Images were collected at an Eye Clinic after appropriate approvals were put in place. Zeiss Visucam 500 cameras were used once eyes were dilated using pharmacological dilation. As this was a clinic based population, most patients had eye conditions but not necessarily DR. Multiple fundus images were acquired from the same patients with varied examination intervals between image capture (ranging from one month to one year). The ground truth on images was extracted based on clinical assessment and converted to DR grades using NSC grading including features of DME being noted.

Kenya. Data were from a population-based survey undertaken in 2007 to 2008 in Nakuru district, Kenya, as the baseline using a Topcon NW6S Non Mydriatic camera model (Topcon, Tokyo, Japan), then in 2013 to 2014 as follow-up in the same population using a DRS Digital Fundus Camera (Haag-Streit, Köniz, Switzerland).[18] Two 45° fundus photographs were taken in each eye; one optic disc centered and the other macula centered. The gold-standard for DR grading of DR was carried out using ICDRS grading criteria. In addition, age related macular degeneration (AMD) and optic disc changes based on retinal photographs were completed by trained graders at the Moorfields Eye Hospital Reading Centre, London, UK. This was the most complex image and grading set of the 3 but also the one with the most complete grading.

A public database, Messidor-2[20,21] was also included in the external validation, consisting of 874 subjects with diabetes (1748 digital retinal color images, one fovea-centered image per eye). These subjects were imaged, without pharmacological dilation, using a Topcon TRC NW6 non-mydriatic fundus camera with a 45-degree field of view, centered on the fovea, at varying imaging settings. Two categories of disease have been provided by the medical experts for each image; the ICDR scales and a definition of DME risk based on the distance between macula and hard exudates. Table 1 provides an overview of the training, internal and external validation test sets.

The Grading Categories and Definitions

Different countries adopt different grading schemes depending on their healthcare resources and policies. The DAPHNE software is trained to predict the probabilities of the following categories:

- Image quality (gradable or non-gradable): The quality grading standard is based on using the UK National Screening Programme for Diabetic Retinopathy's guidelines for the definition of acceptable quality.[27]
- Diabetic retinopathy severity: The DAPHNE classifier grades images based on either ICDRS

or UK NSC classification schemes; whichever is suitable for the country's needs.

- A modified definition of DME (0–1): Fundus photography does not reliably identify DME but allows for surrogate markers to be identified. These surrogate markers of edema such as presence of exudates, or microaneurysms within one-disc diameter of the macula,[19] are identified by the DAPHNE software as well. According to the UK NSC guidelines, diabetic maculopathy (M1) is defined as follows: "A group of exudates is an area of exudates that is greater than or equal to half the disc area and this area is all within the macular area," whereas the macula is defined as "that part of the retina which lies within a circle centered on the center of the fovea whose radius is the distance between the center of the fovea and the temporal margin of the disc."[28] The detection of DME markers is carried out by the DAPHNE object detector and subsequently classified as a referable disease.
- Non-referable DR versus referable DR: In ICDRS level 0 or 1 and in UK NSC R0 or R1 are nonreferable DR, whilst ICDRS level 2, 3, 4 and UK NSC R2 or R3 are Referable DR.

Statistical Analysis of Performance

Evaluation was conducted by measuring sensitivity (SN), specificity (SP), positive and negative predictive values (PPV and NPV), and their 95% confidence intervals (CIs). We also calculated the agreement of the DR and DME grading results between the DAPHNE system and human experts by using the quadratic weighted kappa. These analyses were measured through the StatsModels version 0.8.0 and SciPy version 1.0.0 python packages.

Results

The DAPHNE system is being evaluated in its intended stage in the care pathway: reading of referable cases with evidence, noting any progression changes.

On External Validation Datasets

This dataset was graded as 95% gradable by our system. Referable DR prevalence was 42.5% (21,133 images). According to the NSC and ICDRS grading standards, DAPHNE classifies data into referral (R2 and above in NSC or moderate DR and above in ICDRS) and nonreferral cases (R0/R1 in NSC or no apparent retinopathy and Mild NPDR in ICDRS). Because there is no overlapping in images graded in NSC and ICDRS, we report here the combined calculation on referable retinopathy. Any image with detected DME markers is also referable.

The kappa scores to measure the agreement between the ground truth and the software on referable diseases in China, Saudi Arabia and Kenya datasets are as follows: 0.85, 0.88, and 0.84, respectively. The performance of DAPHNE with regard to the detection of referable retinopathy at high sensitivity operating points was as follows: sensitivity, 94.1% (95% CI: 92.3%–95.6%); specificity, 87.0% (95% CI: 84.9%–88.9%); negative predictive value, 93.9% (95% CI: 93.9%–96.3%); and positive predictive value, 85.3% (95% CI: 82.1%–86.1%). At high-specificity operating points, the sensitivity of our system was 88.2% (95% CI: 85.9%–90.3%), specificity was 93.0% (95% CI: 91.4%–94.5%), the negative predictive value was 91.5% (95% CI: 89.9%–92.8%), and the positive predictive value was 90.4% (95% CI: 88.3%–92.1%). Tables 2A through 2C show the detail of the software performance on each of these three populations.

We chose 3548 eyes with baseline and follow-up images. The measure of changes on their disease levels were calculated. The kappa score is 0.827 when comparing those changes assessed by human graders. Figures 2 and 3 show the detected DR signs across baseline and follow-up images using the DAPHNE lesion detector.

We also carried out an evaluation based on the consistency between the detected features by the software and the DR severity level for the whole image annotated by human graders. If the software detects sufficient features that can be mapped to the same level of DR severity graded by human graders, it is considered as an agreement. The DAPHNE system achieved a weighted kappa score of 0.87 on these external validation sets. The software, however, detected some of the other non-DR lesions and individual artefacts as DR related. This showed that further work is still needed to refine the detection. On the other hand, this will aid flaging up any non-DR pathology.

On External Public Dataset Messidor-2

This consisted of 1748 retinal images from 874 subjects. This dataset was assessed as 100% gradable by our system. Two hundred sixty-four images are Referable DR and 125 DME. Table 2D shows the performance of the algorithm for detecting the different levels of diabetic retinopathy.

At the high sensitivity operating point of detecting referable DR levels (according to ICDR scales and

Table 2A. DAPHNE's Performance on External Validations: (a) Sensitivity, Specificity and Corresponding 95% CIs for Referral Level Output to Detect Referral, PDR and DME, and PDR Level Output to Detect PDR on the Kenya Dataset

Disease Level	Daphne Predicted Results	Sensitivity	Specificity
Referral vs Non- Referral	Referral	94.28% (93.1%–95.22%)	92.12% (88.27%–93.33%)
	PDR	100% (95.5%–100%)	—
	DME	—	
PDR vs Non-PDR	PDR	97.35% (92.3%–99.7%)	85.78% (83.2%–87.81%)

Table 2B. DAPHNE's Performance on External Validations: Sensitivity, Specificity and Corresponding 95% CIs for Referral Level Output to Detect Referral, PDR and DME, and PDR Level Output to Detect PDR on the Saudi Arabian Dataset

Disease Level	Daphne Predicted Results	Sensitivity	Specificity
Referral vs. Non- Referral	Referral	97.1% (95.1%–97.25%)	90.33% (85.71%–92.17%)
	PDR	100% (94.5%–100%)	—
	DME	100% (94.5%–100%)	—
PDR vs. Non-PDR	PDR	98.23% (93.3%–99.6%)	83.78% (82.12%–88.87%)

The ground truth of DME was obtained from eye clinic, to assess the detection of DME markers by the DAPHNE detector.

Table 2C. DAPHNE's Performance on External Validations: Sensitivity, Specificity and Corresponding 95% CIs for Referral Level Output to Detect Referral, PDR and DME, and PDR Level Output to Detect PDR on the China Dataset

Disease Level	Daphne Predicted Results	Sensitivity	Specificity
Referral vs. Non- Referral	Referral	95.51% (93.1%–97.50%)	91.11% (85.11%–92.63%)
	PDR	100% (95.8%–100%)	—
	DME	—	
PDR vs. Non-PDR	PDR	97.18% (91.2%–99.6%)	87.77% (85.3%–88.80%)

Table 2D. DAPHNE's Performance on External Validations: Sensitivity, Specificity and Corresponding 95% CIs for Referral Level Output to Detect Referral, PDR and DME, and PDR Level Output to Detect PDR on the Messidor-2 Dataset

Disease Level	Daphne Predicted Results	Sensitivity	Specificity
Referral vs. Nonreferral	Referral	95.8% (94%–97.42%)	91.32% (86.7%–93.53%)
	PDR	100% (96.5%–100%)	—
	DME	100% (95.8%–100%)	—
PDR vs. Non-PDR	PDR	98.55% (91.3%–99.6%)	86.78% (84.2%–88.8%)

DME scales), the sensitivity of our system was 97.1% (95% CI: 94.8%–98.6%) and specificity was 88.3% (95% CI: 86.5%–90.0%), with a negative predictive value of 99.1% (95% CI: 98.5%–99.5%), and positive predictive value of 69.8% (95% CI: 66.6%–72.8%). At the high specificity operating point, the sensitivity of our system was 89.2% (95% CI: 85.7%–92.2%), and specificity was 95.6% (95% CI: 94.4%–96.6%), with a negative predictive value of 97.0% (95% CI: 96.0%–97.7%) and a positive predictive value of 85.0% (95% CI: 81.5%–87.9%).

Sensitivity, based on referral level prediction when proliferative diabetic retinopathy (PDR) cases are in the referable category, was 100% (95% CI: 96.5%–100%, which means no cases of PDR cases were missed), and sensitivity for detecting DME was also

Figure 2. The detected results of DR progression changes over a five-year period: *First and second rows:* from normal images (R0) to background retinopathy (R1); *last row:* from preproliferative retinopathy (R2) to stable treated proliferative retinopathy (R3s). *First column:* baseline images; *second column:* follow-up fundus images.

100% using the referral threshold (95% CI: 95.8%–100%; i.e., no cases of DME were missed). The AUC for detecting the referral DR was 0.983 (95% CI: 0.969%–0.993%). On the other hand, if a threshold is set to separate PDR and non-PDR, there were cases of PDR classified as severe DR but in the referable category as shown in Table 2.

On Internal Validation Dataset

As a part of the internal validation on 7024 images from the Kaggle dataset, two operating points were selected for the detection of referable DR levels (according to the ICDR scales) for fully gradable images; one for high sensitivity and another for high

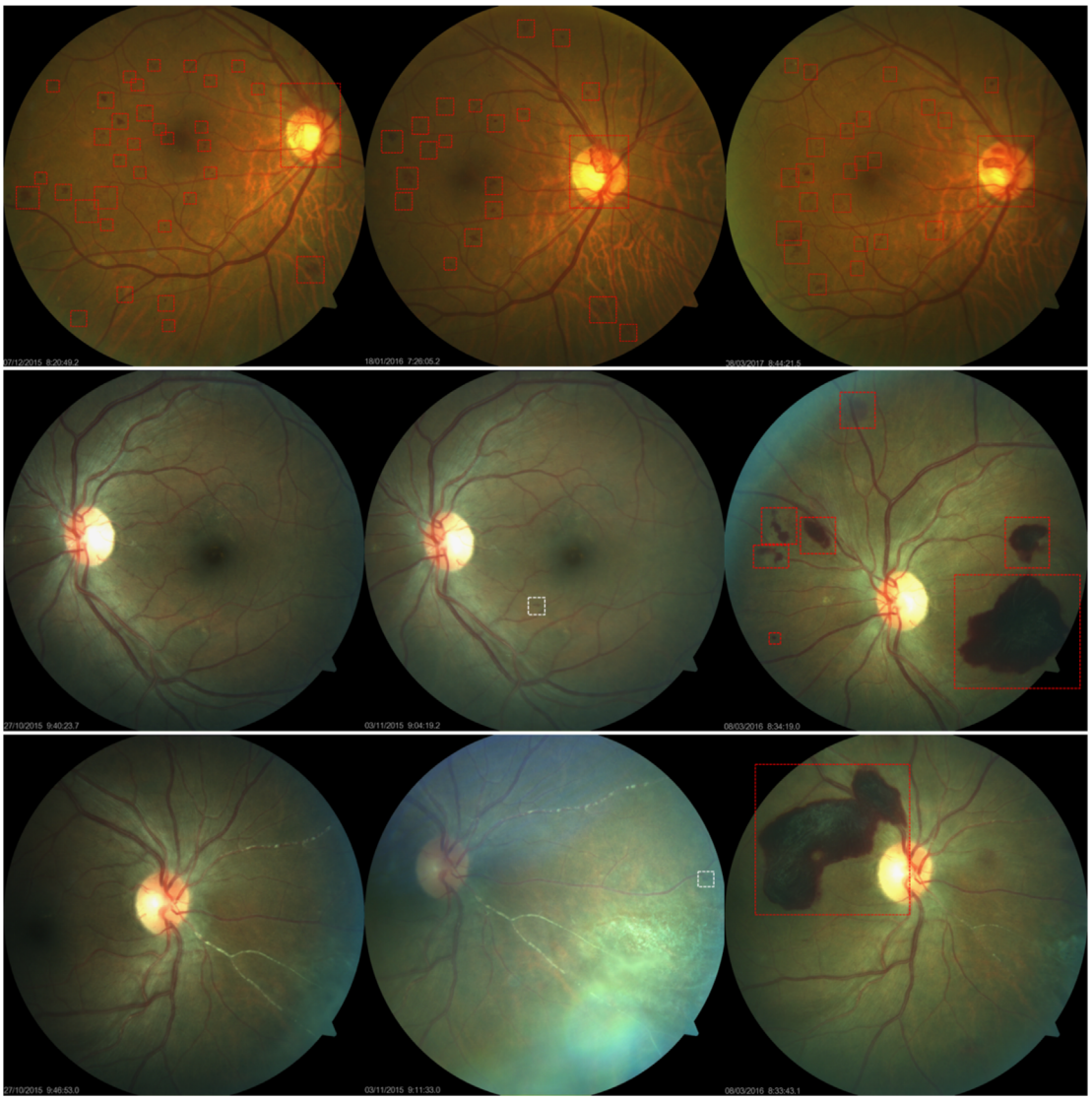

Figure 3. The detected results of DR progression changes within one month (between images columns 1 and 2) and two to four months (between images in columns 2 and 3). Each row shows images from one patient.

specificity. At the high sensitivity operating point, the sensitivity of our system was 94.2% (95% CI: 93.7%–94.7%), and specificity was 76.5% (95% CI: 76.1%–76.9%), a negative predictive value of 98.2% (95% CI: 98.1%–98.4%), and positive predictive value of 48.7% (95% CI: 48.2%–49.1%). At the high-specificity operating point, the sensitivity of our system was 80.1% (95% CI: 80.0%–81.5%) and specificity was 92.6% (95% CI: 92.4%–92.9%), with a negative predictive value of 95.3% (95% CI: 95.1%–95.5%) and positive predictive value of 72.1% (95% CI: 71.4%–72.8%). The AUC for detected referable DR level was 0.985 (95% CI: 0.969%–0.993%). For grading against the ICDR scales, our proposed system obtained a quadratic weighted kappa score of 0.857, which is slightly lower than the winner of the DR competition but higher than other published methods.

Evidence-based Visualization

The current DAPHNE lesion detectors can visualize lesions such as MAs, hemorrhage, exudate, drusen, IRMA, and new vessels explicitly when the prediction probability confidence of these is high. A general category of "abnormal region" is used in the visualization when explicit labeling of the region is of low probability but high as abnormal. Once the CNN outputs any grade indicating the image is not normal

(not 0 in ICDRS or not R0 in NSC), the lesion detectors will visualize at least one of the above pathologic regions that are within the definition of the particular grade. On the other hand, if CNN grades an image as normal, the lesion detectors will search for any missed lesion/abnormal regions using higher prediction probability values.

Discussion

This study demonstrates that all components for the DAPHNE software, such as image quality assessment, image grading, lesion detection, and visualization performed well. These results show good generalizability of the DAPHNE software results to detect gradable quality images, the relevant abnormalities to identify referable DR and DME and to visualize the relevant changes that happened over time. The study was carried out on large sets of images captured from a diverse population with varying camera types and settings. Therefore we believe that so far the evaluation of the DAPHNE software showed sufficiently promising results for it to be useful in the intended stage in the DR screening and imaging analysis care pathway.

In many DR screening programs, human graders do not have access to any other information on the patient but the fundus images they read. Therefore the way the algorithm learns and generates the results needs to mimic the ground truth based on how human graders read the images and come to the conclusion of referral being required. Our purpose was to determine whether the algorithm can learn to perform at an acceptable level of reading fundus images to safely determine quality of the images and then subsequently place patients in the correct referral pathway.

Gradeability of the images can determine the quality of the program and so first of all, DAPHNE looked into any significant impact on software performance when images are from different cameras, with either dilated, or undilated eyes. Testing the data collected from three nations with variation of these factors showed that, as long as the quality assessment is in place in the workflow, the performance of the algorithm is consistent. This is in line with other studies,[29,30] where AI was evaluated in coordination with either assessing the quality and protocol adherence of images, or data imaged with mydriasis through a high-quality imaging platform.

Moreover, an effective algorithm should learn about the true pattern across a large set of images, even if there may be a certain level of noise or variation in some individual samples. Because the DAPHNE software's agreement with human grading was above 84% in all validation datasets, it shows potential for further testing with regard to how it might be incorporated into clinical practice. The software is very sensitive to sight-threatening disease. DAPHNE copes well across different datasets with a varying proportion of normal versus abnormal cases.

When patients do not have DR or DME but do have another pathology, the lesion detectors in DAPHNE are able to recognize them as having abnormal regions. DAPHNE thus has the ability to identify the existence of many common ocular comorbidities in eyes without diabetic retinopathy. The software, however, still needs to be refined to differentiate DR from other visually similar diseases or images with abnormal regions.

The software also shows an ability to monitor DR progression changes between baseline and follow-up images. Because the changes of condition can be measured quantitatively, this progression monitoring may potentially benefit patient care management. Furthermore, it might potentially assist with decision making for optimal screening intervals for patients with diabetes in varied populations.

This automated interpretation addresses only one of the several challenges involved in implementing a successful screening program. This work, however, is not trying to redesign any screening program but rather just to focus on how an automated software system may assist the analysis of the images produced within a functioning screening program. Further work is required to understand a holistic interaction and integration of an AI software into clinical workflows within a given health care system.[31]

Acknowledgments

The authors thank the participants and teams of the Saudi Arabia, China, Kenya Studies. The authors also acknowledge with thanks the Science and Technology Unit, King Abdulaziz University for technical support. The authors thank the Engineering and Physical Sciences Research Council (EPSRC) in the UK for supporting the foundation of this work. We also thank the referees for their comments, which have helped us make some important improvements.

Supported by the National Institute for Health Research (NIHR), Biomedical Research Centre based at Moorfields Eye Hospital, NHS Foundation Trust, UCL Institute of Ophthalmology, and the NSTIP strategic technologies program in the Kingdom of Saudi Arabia (Project No.: 10-INF1262-03).

Disclosure: **L. Al Turk,** None; **S. Wang,** None; **P. Krause,** None; **J. Wawrzynski,** None; **G.M. Saleh,** None; **H. Alsawadi,** None; **A.Z. Alshamrani,** None; **T. Peto,** None; **A. Bastawrous,** None; **J. Li,** None; **H.L. Tang,** None

References

1. Zhang X, Saaddine JB, Chou CF, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. *Jama*. 2010;304:649–656.
2. Liu Y, Song Y, Tao L, et al. Prevalence of diabetic retinopathy among 13473 patients with diabetes mellitus in China: a cross-sectional epidemiological survey in six provinces. *BMJ Open*. 2017;7:e013199.
3. Hajar S, Al Hazmi A, Wasli M, Mousa A, Rabiu M. Prevalence and causes of blindness and diabetic retinopathy in Southern Saudi Arabia. *Saudi Med J*. 2015;36:449.
4. Grading diabetic retinopathy from stereoscopic colour fundus photographs—an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology*. 1991;98:786.
5. Royal College of Ophthalmologists. 1997. *Guidelines for the management of diabetic retinopathy*. London.
6. Chakrabarti R, Harper CA, Keeffe JE. Diabetic retinopathy management guidelines. *Expert Rev Ophthalmol*. 2012;7:417–439.
7. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003–2016. *Acta Diabetologica*. 2017;54:515–525.
8. Peto T, Tadros C. Screening for diabetic retinopathy and diabetic macular edema in the United Kingdom. *Curr Diab Rep*. 2012 Aug;12:338–345, doi:10.1007/s11892-012-0285-4.
9. Mookiah MRK, Acharya UR, Chua CK, Lim CM, Ng EYK, Laude A. Computer-aided diagnosis of diabetic retinopathy: A review. *Comp Biol Med*. 2013;43,:2136–2155.
10. Abràmoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmology*. 2013;131:351–357.
11. Tang HL, Goh J, Peto T, et al. The reading of components of diabetic retinopathy: an evolutionary approach for filtering normal digital fundus imaging in screening and population-based studies. *PloS one*. 2013;8:e66730.
12. Wang S, Tang HL, Hu Y, Sanei S, Saleh GM, Peto T. Localizing microaneurysms in fundus images through singular spectrum analysis. *IEEE Transactions on Biomedical Engineering*, 2017;64:990–1002.
13. 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.
14. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
15. LeCun Y, Bengio Y,, Hinton G. Deep learning. *Nature*. 2015;521:436.
16. Krizhevsky A, Sutskever I,, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012:1097–1105.
17. Kaggle Inc. Diabetic Retinopathy Detection Vol. 2016. 2015. Available at: https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed March 1, 2017.
18. Bastawrous A, Mathenge W, Peto T, et al. The Nakuru eye disease cohort study: methodology & rationale. *BMC Ophthalmol.* 2014;14:60.
19. Kälviäinen RVJPH, Uusitalo H. DIARETDB1 diabetic retinopathy database and evaluation protocol. *Med Image Understanding Anal.* 2007; 2007:61.
20. Quellec G, Lamard M, Josselin PM, Cazuguel G, Cochener B,, Roux C. Optimal wavelet transform for the detection of microaneurysms in retina photographs. *IEEE Trans Med Imaging*, 2008;27:1230–1241.
21. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereolo*. 2014;33:231–234.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014;arXiv preprint arXiv:1409.1556.
23. Hansen MB, Tang HL, Wang S, et al. Automated detection of Diabetic Retinopathy in Three European Populations, Journal of Clinical & Experimental Ophthalmology, cited as: Hansen et al., *J Clin Exp Ophthalmol* 2016;7:4.
24. Graham B. Kaggle Diabetic Retinopathy Detection competition report. Tech. Rep., University of Warwick (2015).
25. Ronneberger O, Fischer P,, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical

image computing and computer-assisted intervention. 2015:234–241. Springer, Cham.
26. Progression of Diabetes Retinal Status Within Community Screening Programs and Potential Implications for Screening Intervals, *Diabetes Care* 2015;38:488–494.
27. Public Health England, Grading Definitions for Referable Disease, NHS Diabetic Eye Screening Programme Technical Report (2017).
28. Scanlon P. Definition of Acceptable Image Quality Version 3, NHS National Screening Programme for Diabetic Retinopathy Technical Report (2007).
29. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;1:39.
30. Olvera-Barrios A, Heeren TF, Balaskas K, et al., Diagnostic accuracy of diabetic retinopathy grading by an artificial intelligence-enabled algorithm compared with a human standard for wide-field true-colour confocal scanning and standard digital retinal images, *Br J Ophthalmol*, doi:10.1136/bjophthalmol-2019-315394.
31. Beede E, Baylo E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020:1–12.