tvst

Article

In-Person Verification of Deep Learning Algorithm for Diabetic Retinopathy Screening Using Different Techniques Across Fundus Image Devices

Nida Wongchaisuwat[1], Adisak Trinavarat[1], Nuttawut Rodanant[1], Somanus Thoongsuwan[1], Nopasak Phasukkijwatana[1], Supalert Prakhunhungsit[1], Lukana Preechasuk[2], and Papis Wongchaisuwat[3]

[1] Department of Ophthalmology, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand
[2] Siriraj Diabetes Center of Excellence, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand
[3] Department of Industrial Engineering, Kasetsart University, Bangkok, Thailand

Correspondence: Papis Wongchaisuwat, Department of Industrial Engineering, Kasetsart University, 50 Ngamwongwan Road, Chatuchak, Bangkok 10900, Thailand.
e-mail: papis.w@ku.ac.th

Received: May 11, 2021
Accepted: August 4, 2021
Published: November 12, 2021

Keywords: artificial intelligence; deep learning algorithm; diabetic retinopathy screening; retinal fundus photographs

Citation: Wongchaisuwat N, Trinavarat A, Rodanant N, Thoongsuwan S, Phasukkijwatana N, Prakhunhungsit S, Preechasuk L, Wongchaisuwat P. In-person verification of deep learning algorithm for diabetic retinopathy screening using different techniques across fundus image devices. Transl Vis Sci Technol. 2021;10(13):17, https://doi.org/10.1167/tvst.10.13.17

Purpose: To evaluate the clinical performance of an automated diabetic retinopathy (DR) screening model to detect referable cases at Siriraj Hospital, Bangkok, Thailand.

Methods: A retrospective review of two sets of fundus photographs (Eidon and Nidek) was undertaken. The images were classified by DR staging prior to the development of a DR screening model. In a prospective cross-sectional enrollment of patients with diabetes, automated detection of referable DR was compared with the results of the gold standard, a dilated fundus examination.

Results: The study analyzed 2533 Nidek fundus images and 1989 Eidon images. The sensitivities calculated for the Nidek and Eidon images were 0.93 and 0.88 and the specificities were 0.91 and 0.85, respectively. In a clinical verification phase using 982 Nidek and 674 Eidon photographs, the calculated sensitivities and specificities were 0.86 and 0.92 for Nidek along with 0.92 and 0.84 for Eidon, respectively. The 60°-field images from the Eidon yielded a more desirable performance in differentiating referable DR than did the corresponding images from the Nidek.

Conclusions: A conventional fundus examination requires intense healthcare resources. It is time consuming and possibly leads to unavoidable human errors. The deep learning algorithm for the detection of referable DR exhibited a favorable performance and is a promising alternative for DR screening. However, variations in the color and pixels of photographs can cause differences in sensitivity and specificity. The image angle and poor quality of fundus photographs were the main limitations of the automated method.

Translational Relevance: The deep learning algorithm, developed from basic research of image processing, was applied to detect referable DR in a real-word clinical care setting.

Introduction

The prevalence of diabetes has been increasing rapidly, and the World Health Organization predicted that its global prevalence will climb to 750 million by 2030.[1,2] Diabetic retinopathy (DR) is one of the leading causes of preventable blindness. Its early detection can lead to prompt treatment, which effectively prevents irreversible visual loss. The American Diabetes Association and the American Academy of Ophthalmology recommended that an annual fundus examination be conducted for patients with diabetes without retinopathy. Patients who have developed retinopathy with intraretinal hemorrhage or a complication (for example, diabetic macular edema) require

more frequent follow-up visits, such as every 1 to 6 months, depending on the disease severity.[3,4] However, the situation in Thailand is that the majority of patients lack access to healthcare facilities due to the limited number of ophthalmologists, especially in remote areas. In Bangkok, only 38% of patients at Siriraj University Hospital participate in screening programs conducted by ophthalmologists.[5] Although a retinal examination by an ophthalmologist remains the gold standard for screening, many other methods have been developed to improve screening ability. Among those are the use of fundus photographs with trained readers for their interpretation, telemedicine, and automated screening models based on artificial intelligence (AI) technologies.[6,7]

AI with deep learning (DL) technology has been utilized for various medical applications, such as regular chest radiograph screening, DR screening using retinal photographs, and skin lesion detection and diagnosis.[8–15] In 2016, Gulshan and colleagues[16] proposed a novel DR screening software that received desirable results (87%–90% sensitivity and 98% specificity). The excellent performance of their DL algorithm attracted attention in the field of ophthalmology, resulting in numerous studies to determine the accuracy of AI in detecting DR in populations around the world. For example, Ruamviboonsuk and associates[17] applied DL algorithm software developed by Google to the Thai population. Their DR screening sensitivity reached 97%, compared with 74% for manual observation by trained graders. To guarantee a desirable performance, internal and temporal validation processes are essential during the model development.[18] Internal validation, or reproducibility, refers to the performance of the model based on hold-out samples from the model-training step. Temporal validation measures the performance on subsequent samples obtained from the center at which the model was developed. The current study aimed to perform these validation steps to evaluate the performance of our model in preparation for its real-life clinical application.

Materials and Methods

A single-center, cross-sectional study was conducted at the Ophthalmology Outpatient Clinic and the Diabetes Center at Siriraj Hospital, Mahidol University, Bangkok, Thailand, from March 2018 to March 2020. The study adhered to the tenets of the Declaration of Helsinki, and its protocol was approved by the Medical Ethics Committee of Siriraj Hospital, Mahidol University (872/2562 [IRB2]).

Study Participants

The study was divided into two phases: phase 1, DL algorithm software development and verification; and phase 2, clinical verification (in-person study design and interpretation). For phase 1, the inclusion criteria were patients with diabetes with good-quality retinal photographs or images with distinct retinal vessels, optic nerve, and retinal backgrounds. Fair-quality photographs were defined as images with partial visibility of these retinal components. As a result, abnormal lesions were only moderately detectable from the retinal background or the photograph-edge was not clearly visible. Images with blurred components where abnormal lesions could not be distinguished were defined as poor-quality photographs. A retrospective review was conducted of the hospital medical records and of retinal fundus photographs that had been made using two fundus cameras.

Two sets of image data were compiled and analyzed. The first dataset was used to develop DL models, and the second dataset was used to verify the performance of the models (i.e., external validation). This external validation was deemed to represent the temporal validation process referred to earlier. In more detail, all images in the first set of data were randomly divided into three groups using stratified sampling. The proportional distributions of the images were fixed at 80% for DL model training, 10% for model validation, and 10% for model testing. The majority of the images were used for training purposes. The second group of images was intentionally employed to fine-tune the DL model parameters. The resultant version of the model was then tested with the third group of images. This testing was regarded as internal validation of the first set of data. Multiple DL models were experimented with and fine-tuned until at least a 95% sensitivity and 80% specificity were guaranteed. In the next step, the final model developed from the first dataset was applied to the second set of data to verify the performance of the model. As two separate datasets were utilized for development and verification purposes, the second set of data constituted external validation.

Phase 2 was a prospective, cross-sectional study to evaluate the performance of the model in a clinical setting. This phase represented a further temporal validation of the model. For phase 2, the inclusion criteria were patients with diabetes who were 18 years or older and had come for DR screening. Images were excluded if they were of poor quality (due to, for example, unsuccessful picture taking,

dilation failure during the retinal examination, media opacity, or limited patient cooperation). The participants' demographic and clinical data were recorded (age, gender, laterality, quality of images, types of fundus camera, retinal findings, presence of diabetic macula edema, and DR staging). Written informed consent was obtained from each participant.

Fundus Cameras

The images in the Siriraj Hospital database had been taken by two fundus cameras: Nidek (NonMydriatic AFC-330; NIDEK Co., Ltd., Aichi, Japan) and Eidon (Eidon AF TrueColor Confocal Scanner; Centervue, Inc., Fremont, CA). Three, non-mydriatic fundus photographs were specifically obtained from each device. With the Nidek camera, the first image was a single-field, 45°, macula-centered photograph. For the Eidon camera, the first image was a single-field, 60°, macula-centered photograph. The remaining two photographs from each camera were nasal and temporal overlapping images.

Classification of Images

DR staging of the Nidek and Eidon images was thoroughly assessed in accordance with the modified Airlie House classification system developed by the Early Treatment Diabetic Retinopathy Study.[19] Instead of considering all DR stages, our study employed a simplified, binary classification system. Its two categories were (1) non-referable DR (no DR to a mild level of nonproliferative DR [NPDR]), and (2) referable DR and other retinopathies (moderate to severe NPDR, proliferative DR, diabetic macula edema, and other retinopathies).

Phase 1: DL Algorithm Software Development and Verification

One, single-field, 45°, macula-centered, non-mydriatic, fundus photograph from each patient with diabetes at Siriraj Hospital was retrospectively recruited and classified by DR staging by a single retinal expert. The images were classified into the binary groups (non-referable DR or referable DR and other retinopathies) by the same retinal expert. The patients' identifications were masked to the algorithm developer.

Previous work clearly established that DL exhibits promising and acceptable performance in diagnosing DR using retinal fundus images.[16,17] Generally, DL aims to mimic how neurons in the human brain behave using multiple layers of networks. A convolutional neural network (CNN) is a subcategory of DL that has proven to be significantly effective in image classification tasks. CNN was therefore adopted by this work to classify the retinal fundus images, based on the concepts of transfer learning and online active learning for the Nidek and Eidon images, respectively.

There are four main operations in the typical CNN model: convolution, nonlinearity activation, subsampling/pooling, and classification. For the present study, each input image was initially converted into a two-dimensional matrix of pixel values, with three channels representing the colors red, green, and blue. According to the preprocessing step, each image was resized and further normalized with predefined parameters. Prior to training the model, additional data augmentations were utilized if necessary. They included random affine transformation, center cropping, horizontal and vertical image flipping, and changes to the brightness, contrast, saturation, and hue of each image. The convolutional operation extracted features from the input by preserving the spatial relationships between the pixels. An additional rectified linear unit (ReLU) function was applied to feature maps constructed from the convolutional operation to capture the nonlinear relationships within the data. Spatial pooling was used to reduce the dimensionality of the rectified feature map before forming the next layer. Multiple sets of convolutions, ReLU activation, and pooling operations were stacked to construct the overall architecture of the network. Meaningful features were mainly extracted from the original images. The output from the last pooling layer was used as an input for the final classification operation. As the final step, a fully connected layer with a Softmax activation function was utilized to classify the input images into various classes.

Transfer learning is a technique involving the application of the knowledge gained from one task (the "model") to the learning needed for a new task that has limited data. For the purposes of the current research, we retrieved pretrained source models that had been well trained on an image classification task based on a significantly large dataset. These pretrained models were used as the starting point for the training with our fundus images; in other words, the models were fine-tuned for our specific task, which was based on the Nidek images. We drew upon the pretrained models from Kaggle.[20] Three specific CNNs—SE-ResNeXt50_32 × 4d, SE-ResNeXt101_32 × 4d, and SENet154—had initially been trained on a large dataset that was similar to our Nidek images, in terms of both color and characteristics. The weighted average of the three CNN models was further utilized to compute initial results prior to fine-tuning the

models with our Nidek image data set. According to the training configurations, the mean square error was used as the loss function with the RAdam optimizer. CosineAnnealingLR was considered to adjust the learning rate. The predicted class was finally obtained from an ensemble model of the fine-tuned individual models. A grid-search technique was attentively used to find the optimized weights and cut-off probability threshold that provided the best performance. These individual models were equally weighted, and a 0.2 cut-off threshold was selected.

We observed that the colors of the fundus photographs extracted from the Eidon device differed from those of both the Nidek images and the publicly available datasets. In response to this observation, the online active method proposed by Smailagic and coauthors[21] was utilized in addition to a traditional DL. The online active learning, a sequential technique, relies on a relatively small number of required labels for training. In essence, we retrieved the feature embeddings learned from the ResNet baseline classifier on the initial labeled data, and these were further used to query the most informative, unlabeled examples. The selected examples were labeled prior to iteratively retraining the model on the combined labeled data. The hyperparameters used in our model were similar to those stated in Smailagic et al.,[21] such as the number of max entropy samples of 59 and the number of added images per iteration of 20. In accordance with the training configurations, a stochastic gradient descent optimizer with a momentum of 0.9, a learning rate of 0.005, and a weight decay of 0.01 was utilized. The number of epochs, the batch size, and the early stopping patience were set to 50, 32, and 20, respectively. A random search was employed to find the online sample fraction and the cut-off probability threshold hyperparameters; values of 0.875 (online fraction) and 0.1 (cut-off threshold) were obtained.

Phase 2: Clinical Verification (In-Person Study Design and Interpretation)

During DR screening visits by the participants, non-mydriatic retinal photographs were taken by a well-trained investigator and/or two research assistants. After retrieving the required images, mydriatic eye drops were instilled to achieve full pupillary dilatation before the patient underwent a gold-standard retinal examination by a retinal expert. The eight retinal experts at Siriraj Hospital used a biomicroscopic slit lamp with a non-contact lens, and they completed the fundus examinations to zone 3 (i.e., beyond the equatorial zone). All examiners were masked to the DL model output. The trials involved three main components: output from the DL models based only on the central-retinal images; output from the DL models for three consecutive images; and output of the DR stagings obtained from the gold-standard retinal examinations. Only one positive result selected among three images was graded as "referable DR and other retinopathies." The sensitivity and specificity of the algorithm were the primary objectives of the study. The accuracy, receiver operating characteristic (ROC) curve and area under the curve (AUC), and false-positive and false-negative results of the study were also evaluated. The statistical analyses utilized Python packages.

Results

A total of 3515 and 2663 eyes were enrolled for the Nidek and Eidon groups, respectively. The distribution of the DR staging is presented in Table 1. In our proposed model, each fundus image was classified as "non-referable DR" or "referable DR and other retinopathies." Phase 1 considered gradable images, which were separated into two datasets. The first dataset was used for training, fine-tuning of the model hyperparameters, and validation of the model until the desired result (>95% sensitivity and >80% specificity) was achieved. The model specifically trained with the Nidek images reached a sensitivity of 99% and a specificity of 83%, whereas the model trained with the Eidon images achieved an 83% sensitivity and an 81% specificity. The model used for Eidon improved its sensitivity to 96% and its specificity to 85% (Table 2). The second image dataset was used for the external validation of the trained model. A sensitivity of 93% and a specificity of 91% were attained for the Nidek group, while an 88% sensitivity and an 85% specificity were achieved for Eidon. We observed that the model for Eidon yielded a lower sensitivity with the external validation data, but its specificity remained the same as that obtained using the first dataset (Table 2).

In phase 2, the photographs were taken before the patients underwent dilated fundus examinations. In the case of the Nidek group, good- and fair-quality images represented 54% and 22% of the pictures, respectively. However, 318 eyes from the Nidek group (24%) were excluded due to poor-quality images; they mainly resulted from the presence of cataracts, found in 59% of the excluded images. As for the Eidon photographs, 52% and 31% were of good and fair quality, respectively. A total of 140 eyes in the Eidon group (17%) were excluded due to poor image quality, with 39% of the rejections being due to the presence of cataracts.

Table 1. Distribution of DR Gradings in Phases 1 and 2 of the Study

	Phase 1				Phase 2	
	Internal Validation		External Validation		Clinical Verification	
	Nidek	Eidon	Nidek	Eidon	Nidek	Eidon
Train, *n*	1075	898	—	—	—	—
Validate, *n*	124	103	—	—	—	—
Test, *n*	128	134	1206	854	982	674
Classification (%)						
No to mild NPDR	52	78	84	78	70	68
Referable DR	39	14	12	15	21	23
Other retinopathies	9	8	4	7	9	9

Table 2. Accuracy Test Results of DL Algorithm in Detecting Referable Diabetic Retinopathy

Device	*n*	Sensitivity	Specificity	PPV	NPV	Accuracy
Internal validation (phase 1)						
Nidek	128	0.99	0.83	0.89	0.98	0.92
Eidon	134	0.83	0.81	0.61	0.93	0.81
Eidon*	122	0.96	0.85	0.61	0.99	0.87
External validation (phase 1)						
Nidek	1206	0.93	0.91	0.66	0.99	0.91
Eidon*	829	0.88	0.85	0.62	0.96	0.86
Clinical verification (phase 2)						
Single photo						
All						
Nidek	982	0.82	0.92	0.82	0.92	0.89
Eidon*	674	0.89	0.84	0.73	0.94	0.86
Excluding other retinopathies						
Nidek	893	0.86	0.92	0.77	0.96	0.91
Eidon*	612	0.92	0.84	0.66	0.97	0.86
Three photos						
All						
Nidek	964	0.97	0.3	0.37	0.96	0.5
Eidon*	626	0.95	0.66	0.57	0.97	0.76
Excluding other retinopathies						
Nidek	877	0.97	0.3	0.29	0.98	0.46
Eidon*	574	0.97	0.66	0.5	0.99	0.74

NPV, negative predictive value; PPV, positive predictive value.
*New model of Eidon.

Other reasons for inferior-quality images were a small pupil, defocus, and a lack of cooperation by the patient (Fig. 1).

The performance of the DL model in phase 2 contrasted with the first phase (Table 2). The sensitivity for detecting referable DR for Eidon was almost identical to that from the first phase, with an 89% sensitivity and an 84% specificity. A potential improvement to the algorithm was observed when other retinopathies were excluded, which resulted in the sensitivity climbing slightly to 92%. On the other hand, the results of the Nidek group showed a substantial decline in sensitivity, from 93% to 82%, with a specificity of 92%. After excluding other retinopathies, however, the sensitivity rose to 86%. ROC curves giving 0.95 AUC for both Nidek and Eidon are illustrated in Figure 2.

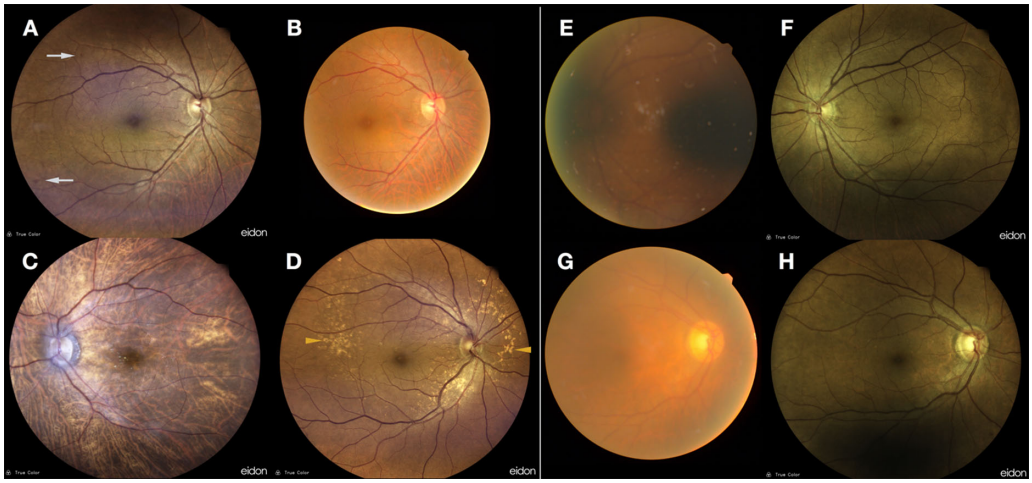

Figure 1. (**A**, **B**) Fundus photographs of a diabetes patient with moderate NPDR in the right eye. An intraretinal hemorrhage in the temporal area was observed in a fundus photograph using a 60° widefield Eidon camera (A, *white arrows*). Photograph of the same eye from the same patient, taken with a Nidek camera (**B**). As the intraretinal hemorrhage could not be detected, the DL algorithm gave a false-negative result for this eye. False-positive results were observed frequently with tigroid appearances of the retinal background, especially with myopia. (**C**) Drusen are presented as a yellowish deposit underneath the retina; this could be misinterpreted as exudate in diabetic retinopathy (**D**, *arrowhead*). Poor-quality images were frequent with small pupils (**E**, **F**) and dense cataracts (**G**, **H**). The Eidon performed better in these conditions, giving better resolution fundus images (**F**, **H**) than the Nidek (**E**, **G**).

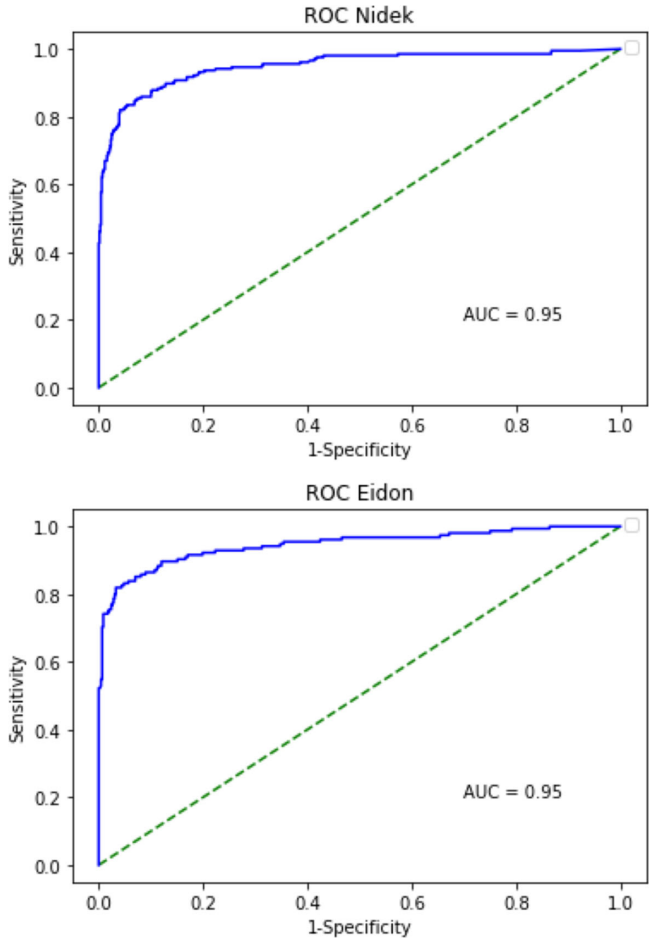

Figure 2. ROC curves of the proposed DL algorithms for the Nidek and Eidon. The AUCs were calculated.

For the three consecutive fundus image results, very high sensitivity with extremely low specificity were observed in both the Nidek and Eidon groups (Table 2). False-negative and false-positive cases were further analyzed (Table 3). Almost every case from both devices was classified as a moderate NPDR, with only a few dots or blot hemorrhages seen. In the Nidek group, 59% of the false-negative cases were diagnosed with peripheral DR beyond the angle of the field of view of the ability of the device to catch an intraretinal hemorrhage or exudate. Although the false-positive cases demonstrated a high proportion of fair-quality images, some artifacts (such as an image glare or a tigroid appearance in a myopic fundus) were abnormally detected as DR. Interestingly, after reviewing the false-positive images, we found that 17% of the false-positive cases in the Eidon group and 2% in the Nidek group derived from human error during the dilated fundus examination, whereas the output from the model gave the correct classification. Examples of false-negative and false-positive cases are presented in Figure 1.

Discussion

The application of DL to the detection of referable DR is promising. Abundant research has already established a good efficacy for DL compared with retinal expert or trained human-grader interpretations

Table 3. False Negatives and False Positives Were Demonstrated

	Nidek	Eidon
False negatives, *n* (%)	29 (3.2)	13 (2.1)
Moderate NPDR, *n*	28	13
Severe NPDR, *n*	1	–
Peripheral DR, *n*	17	3
Proliferative diabetic retinopathy, *n*	–	–
False positives, *n* (%)	53 (5.9)	72 (11.8)
No DR, *n*	42	45
Mild NPDR, *n*	1	6
Human error, *n*	1	12
Other retinopathies, *n*	9	9

of fundus photographs. Our work had the advantage of utilizing a prospective setting. This meant that we could compare the results obtained from DL models used to triage retinal images with those of in-person clinical examinations conducted at the same visit. Hence, we were able to detect some limitations of the clinical use of DL. The study also demonstrated the performance of a new algorithm for widefield confocal scanner imaging photographs.

We conducted the entire process of DL development—from the training steps to validation and finally to verification—in a real clinical setting. The results for the retrospective training dataset (the internal validation process) exhibited very high sensitivity and specificity for both Nidek and Eidon images, with separate models and cut-off thresholds. The external validation process, which used a different dataset, yielded a lower sensitivity and specificity for Eidon due to overfitting. However, contrasting results were presented in the clinical verification phase. To be specific, the sensitivity declined significantly for the Nidek group due to technical problems in taking the fundus photographs. The main problems with the Nidek camera were related to a patient factor and to the ability of the camera to capture small pupil sizes. It is relatively common for elderly diabetic patients to present with early cataracts. The condition typically interferes with the entry of light through the retina. This is significant in that the flashlight source of Nidek is a halogen xenon, which has a poorer ability to transmit light through ocular media opacities than the light-emitting diode of the Eidon light confocal imaging system. As to the technical constraint of the Nidek camera, a small pupil size is likely to make taking photographs more challenging with a Nidek camera than with an Eidon camera. This is because an Eidon camera can take pictures with pupil sizes as small as 2.5 mm. In contrast, a Nidek camera requires at least a 3.3-mm pupil size to achieve an adequate image quality.

Much research has confirmed the outstanding performance of DL in detecting abnormalities in retinal photographs. Some studies found that the ability of DL was superior to that of trained human graders.[8–10,16,22,23] Ruamviboonsuk et al. and Google AI reported a Google DL algorithm performance of over 95% for the detection of referable DR in the Thai population.[17] However, the Google model was derived from another population database while conducting a geographic validation in the Thai population. Our study purposely trained new models suited to the Thai population, with model testing conducted on our own data.

The performance of our DL models in the internal and the external validation processes showed results similar to those reported by previous studies. Our models achieved over 80% sensitivity and 80% specificity, which are highly desirable results for a screening test. Nevertheless, we observed interesting drawbacks in the clinical verification phase. The image quality and the angle of field of view played important roles in the screening. Eidon is an advanced fundus camera that uses light-emitting diodes as its light source. The light penetrates media opacity well, and the camera has an automated operating system. Consequently, it is easy for Eidon to take photographs, especially with small pupil diameters (2.5 mm or wider), and it gives sharp images with differences in color clearly represented. A earlier study reported that the image resolutions and color discriminations obtained with Eidon cameras were superior to those of Nidek cameras (which use a xenon light source) and conventional fundus cameras.[24]

Data regarding the diagnostic accuracy of DL algorithms in widefield confocal scanning photographs (Eidon) to detect DR are still limited. Only a study by

Olvera-Barrios and coauthors reported a similar sensitivity for detecting DR (92%), compared with human graders using the commercially available AI algorithm EyeArt (version 2.1.0; Eyenuk, Inc., Woodland Hills, CA).[25] In contrast, our study used separate models for Nidek and Eidon due to the poor diagnostic accuracy achieved when using the same DR screening model. The difference in the model accuracies of the two fundus cameras might result from differences in the color, image resolution, field of view, and lesion characteristics observable in their respective images.

The image viewing angle had a significant effect on the DR screening. The gold-standard photography method for DR detection is seven standard fields (30°) of stereoscopic color fundus photographs.[19] However, given the uncomfortable and time-consuming process involved, a practical drawback is patient preferences. A report by the American Academy of Ophthalmology showed that screening for referable DR using a single, 45° photograph with 61% to 90% sensitivity is preferable.[26] The advanced technology used in ultra-widefield fundus cameras has demonstrated a favorable performance in distinguishing DR staging relative to seven standard stereoscopic fundus photographs and dilated fundus examinations.[27] Thus, in our study, one single-field, macula-centered fundus photograph and three consecutive photographs were considered. For the three consecutive fundus images, extremely low specificity was observed in both groups due to artifacts arising from the non-mydriatic technique. Only the single-field, macula-centered photographs were considered in our study. After excluding other retinopathies, the ability of the models to detect referable DR decreased by 3.2% and 2.1% for Nidek and Eidon, respectively. The false-negative cases with the Nidek camera primarily resulted from its narrow angle of view, which was limited to 45°. Consequently, 1.9% and 0.5% of the peripheral DR images from Nidek and Eidon, respectively, were missed. The 60°-field images of Eidon provided better performance in distinguishing referable DR. Specifically, the 60°-field images of the camera were able to reveal more peripheral dots, blots, and other abnormalities than the 45°-field images obtained with the Nidek camera. Moreover, it has been suggested that cases of peripheral DR have a 3.2-fold higher risk of DR progression than cases of posterior pole DR.[28] It is apparent that advanced fundus photograph technology does improve image quality, resulting in a high screening sensitivity for DR.[29] Although an ultra-widefield fundus camera is a promising tool for DR screening, there are still insufficient images for the training of a DL model with outstanding performance.

The false-positive cases of both devices mainly derived from fair-quality photographs—for example, those with a mild glare. Nevertheless, the overall quality of the images was still good enough to detect findings in the retinal background. Interestingly, small blot areas in cases of moderate NPDR at an early stage were successfully detected in 12 Eidon eyes and two Nidek eyes, whereas they were missed by the dilated fundus examinations. To address this issue in our future work, an automated system to detect the photograph quality can be added to the pipeline. With this system, the quality of an input fundus photograph would be automatically estimated in real time prior to DR screening without the need for the opinion of a specialist.

Applying the advanced technology to real-world clinical practice has many limitations. Other retinopathies, glaucoma, and optic nerve diseases can be comorbidities in patients with diabetes, and they require proper treatment. DL algorithms are trained by learning from the retinal findings of patients with DR. In view of the current performance levels of automated systems, they might not yet be safe enough to fully replace standard dilated fundus examinations performed by humans. Further development of DL for various diseases is underway.

One limitation of this study was the possibility of spectrum bias from the distribution of the DR staging. The prevalence of DR in our research might differ slightly from the prevalence in the general population. The prevalence of the advanced stages of DR is higher in tertiary-care hospitals, especially proliferative DR. This situation could have affected both the training of the models and the interpretation of the results for the positive predictive value of the DL model. The number of images used for further training was also limited. Although we developed the DL model based on the active learning concept, which requires less training data, Eidon is a new device that currently has limited image resources. An overfitting phenomenon was found in the external validation process, confirming that enhancing the model performance when applied to Eidon cameras requires additional data.

During the clinical verification phase, a difficulty was found with the image preparation procedure and the workflow setting. Taking photographs using both devices could not be done in every case; therefore, the accuracy of the models for the Nidek and Eidon devices could not be directly compared in our study. Only the trends of the data from both devices were demonstrated. The image quality proved to be a critical factor. The good-quality photographs represented approximately 50% of the images from both devices, whereas the poor-quality photographs—which

were excluded from Phase 2—accounted for 24% and 17% of the Nidek and Eidon images, respectively. A model result should be thoroughly interpreted in a real-life clinical application. However, patients with poor-quality photographs will be classified as referable and sent for dilated fundus examinations. Detecting referable DR is a critical input when deciding whether to transfer patients to ophthalmologists. With a better screening ability, algorithms will be able to more successfully distinguish sight-threatening DR requiring urgent referral. As a result, patients will be assessed early and receive prompt treatment before the disease progresses to irreversible blindness. DL models have confirmed their clinical effectiveness in screening for referable DR and in improving visual outcomes. Research is needed to establish the desirable performance of DL models for use with high-definition, widefield fundus photographs. Another direction of future work is to establish standardized protocols for DR screening programs.

Conclusions

Screening tools should be simple, safe, accurate, time saving, and cost effective. Despite some limitations, the use of an automated DL model is a promising alternative approach to distinguishing referable DR in clinical settings. The image color, resolution, viewing angle, and lesion characteristics varied between the Eidon and Nidek cameras; all of these factors play a role in determining the model performance. In addition, more data collection will further improve the performance of DL models for DR screening, which will in turn further enhance human resource substitution.

Acknowledgments

Supported by the Routine-to-Research Unit, Research Department, Siriraj Hospital, Mahidol University, Bangkok, Thailand.

Disclosure: **N. Wongchaisuwat**, None; **A. Trinavarat**, None; **N. Rodanant**, None; **S. Thoongsuwan**, None; **N. Phasukkijwatana**, None; **S. Prakhunhungsit**, None; **L. Preechasuk**, None; **P. Wongchaisuwat**, None

References

1. Roglic G. WHO global report on diabetes: a summary. *Int J Noncommunicable Dis*. 2016;1(1):3.
2. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*. 2004;27(5):1047–1053.
3. American College of Physicians, American Diabetes Association, American Academy of Ophthalmology. Screening guidelines for diabetic retinopathy. *Ann Intern Med.* 1992;116(8):683–685.
4. Fong DS, Aiello L, Gardner TW, et al. Retinopathy in diabetes. *Diabetes Care.* 2004;27(suppl 1):S84–S87.
5. Sriwijitkamol A, Moungngern Y, Vannaseang S. Assessment and prevalences of diabetic complications in 722 Thai type 2 diabetes patients. *J Med Assoc Thai.* 2011;94(2):168.
6. Fenner BJ, Wong RL, Lam WC, Tan GS, Cheung GC. Advances in retinal imaging and applications in diabetic retinopathy screening: a review. *Ophthalmol Ther.* 2018;7(2):333–346.
7. Cicinelli MV, Cavalleri M, Brambati M, Lattanzio R, Bandello F. New imaging systems in diabetic retinopathy. *Acta Diabetol.* 2019;56(9):981–994.
8. 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(13):5200–5206.
9. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology.* 2017;124(7):962–969.
10. Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211–2223.
11. Lakhani P, Prater AB, Hutson RK, et al. Machine learning in radiology: applications beyond image interpretation. *J Am Coll Radiol.* 2018;15(2):350–359.
12. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. *Jpn J Radiol.* 2019;37(1):15–33.
13. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–118.
14. Cui X, Wei R, Gong L, et al. Assessing the effectiveness of artificial intelligence methods for

melanoma: a retrospective review. *J Am Acad Dermatol.* 2019;81(5):1176–1180.
15. Fujisawa Y, Inoue S, Nakamura Y. The possibility of deep learning-based, computer-aided skin tumor classifiers. *Front Med (Lausanne).* 2019;6:191.
16. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–2410.
17. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med.* 2019;2:25.
18. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. *Diagn Progn Res.* 2017;1:12.
19. 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(5):786–806.
20. Kaggle. APTOS 2019 11th place solution. Available at: https://www.kaggle.com/uiiurz1/aptos-2019-11th-place-solution. Accessed August 25, 2021.
21. Smailagic A, Costa P, Gaudio A, et al. O-MedAL: online active deep learning for medical image analysis. *WIREs.* 2020;10(4):e1353.
22. Bellemo V, Lim G, Rim TH, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. *Curr Diab Rep.* 2019;19(9):72.
23. Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol.* 2020;105(5):723–728.
24. Sarao V, Veritti D, Borrelli E, Sadda SVR, Poletti E, Lanzetta P. A comparison between a white LED confocal imaging system and a conventional flash fundus camera using chromaticity analysis. *BMC Ophthalmol.* 2019;19(1):231.
25. 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.* 2020;105(2):265–270.
26. Williams GA, Scott IU, Haller JA, Maguire AM, Marcus D, McDonald HR. Single-field fundus photography for diabetic retinopathy screening: a report by the American Academy of Ophthalmology. *Ophthalmology.* 2004;111(5):1055–1062.
27. Silva PS, Cavallerano JD, Sun JK, Noble J, Aiello LM, Aiello LP. Nonmydriatic ultrawide field retinal imaging compared with dilated standard 7-field 35-mm photography and retinal specialist examination for evaluation of diabetic retinopathy. *Am J Ophthalmol.* 2012;154(3):549–559.
28. Silva PS, Cavallerano JD, Haddad NMN, et al. Peripheral lesions identified on ultrawide field imaging predict increased risk of diabetic retinopathy progression over 4 years. *Ophthalmology.* 2015;122(5):949–956.
29. Rajalakshmi R, Prathiba V, Arulmalar S, Usha M. Review of retinal cameras for global coverage of diabetic retinopathy screening. *Eye (Lond).* 2021;35(1):162–172.