

ARTICLE OPEN

Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program

Paisan Ruamviboonsuk¹, Jonathan Krause², Peranut Chotcomwongse¹, Rory Sayres², Rajiv Raman³, Kasumi Widner², Bilson J. L. Campana², Sonia Phene², Kornwipa Hemarat⁴, Mongkol Tadarati⁵, Sukhum Silpa-Archa¹, Jirawut Limwattanayingyong¹, Chetan Rao³, Oscar Kuruville⁵, Jesse Jung⁶, Jeffrey Tan⁷, Surapong Orprayoon⁸, Chawawat Kangwanwongpaisan⁹, Ramase Sukumalpaiboon¹⁰, Chainarong Luengchaichawang¹¹, Jitumporn Fuangkaew¹², Pipat Kongsap¹³, Lamyong Chualinpha¹⁴, Sarawuth Saree¹⁵, Srirut Kawinpanitan¹⁶, Korntip Mitvongsa¹⁷, Siriporn Lawanasakol¹⁸, Chaiyasit Thepchatri¹⁹, Lalita Wongpichedchai²⁰, Greg S. Corrado², Lily Peng² and Dale R. Webster²

Deep learning algorithms have been used to detect diabetic retinopathy (DR) with specialist-level accuracy. This study aims to validate one such algorithm on a large-scale clinical population, and compare the algorithm performance with that of human graders. A total of 25,326 gradable retinal images of patients with diabetes from the community-based, nationwide screening program of DR in Thailand were analyzed for DR severity and referable diabetic macular edema (DME). Grades adjudicated by a panel of international retinal specialists served as the reference standard. Relative to human graders, for detecting referable DR (moderate NPDR or worse), the deep learning algorithm had significantly higher sensitivity (0.97 vs. 0.74, $p < 0.001$), and a slightly lower specificity (0.96 vs. 0.98, $p < 0.001$). Higher sensitivity of the algorithm was also observed for each of the categories of severe or worse NPDR, PDR, and DME ($p < 0.001$ for all comparisons). The quadratic-weighted kappa for determination of DR severity levels by the algorithm and human graders was 0.85 and 0.78 respectively ($p < 0.001$ for the difference). Across different severity levels of DR for determining referable disease, deep learning significantly reduced the false negative rate (by 23%) at the cost of slightly higher false positive rates (2%). Deep learning algorithms may serve as a valuable tool for DR screening.

npj Digital Medicine (2019)2:25; <https://doi.org/10.1038/s41746-019-0099-8>

INTRODUCTION

Deep learning (DL) is a field of artificial intelligence which has been applied to develop algorithms for the detection of diabetic retinopathy (DR) with high (>90%) sensitivity and specificity for referable disease (moderate non-proliferative diabetic retinopathy (NPDR) or worse).^{1–3} In addition to high screening accuracy, DL also has advantages in terms of resource consumption, consistency, and scalability, and has the potential to be deployed as an alternative to human graders for classifying or triaging retinal photographs in DR screening programs.

In Thailand, there are 1500 ophthalmologists, including 200 retinal specialists, who provide ophthalmic care to approximately 4.5 million patients with diabetes. Half of the ophthalmologists and retinal specialists practice in Bangkok, the capital of the country, while a majority of patients with diabetes live in areas

100 km or more from provincial hospitals, where ophthalmologists typically practice. The latest Thailand National Survey of Blindness conducted in 2006–2007⁴ showed that 34% of patients with diabetes had low vision or blindness in either eye. DR was and continues to be the most common retinal disease that causes bilateral low vision.^{4,5}

A national screening program for DR was set up by the Ministry of Public Health of Thailand in 2013. The screening was conducted in each of the 13 health regions with an initial target of screening at least 60% of diabetic patients in each region. Unfortunately, Ministry data indicate that less than 50% of the diabetic patients were screened every year since the inception of the program. Because this was in part due to the lack of trained graders, deploying DL in the screening program for DR in Thailand has the

¹Department of Ophthalmology, Rajavithi Hospital, Bangkok, Thailand; ²Google AI, Google, Mountain View, CA, USA; ³Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai, Tamil Nadu, India; ⁴Department of Ophthalmology, Vajira Hospital, Bangkok, Thailand; ⁵Eye and Laser Center, Charlotte, NC, USA; ⁶East Bay Retina Consultants, Oakland, CA; Department of Ophthalmology, University of California San Francisco, San Francisco, CA, USA; ⁷Retina-Vitreous Associates Medical Group, Department of Ophthalmology, University of Southern California, Los Angeles, CA, USA; ⁸Department of Ophthalmology, Lamphun Hospital, Lamphun, Thailand; ⁹Department of Ophthalmology, Somdejphrajaoataksin Maharaj Hospital, Tak, Thailand; ¹⁰Department of Ophthalmology, Sawanpracharak Hospital, Nakhon Sawan, Thailand; ¹¹Department of Ophthalmology, Nakhon Nayok Hospital, Nakhon Nayok, Thailand; ¹²Department of Ophthalmology, Photharam Hospital, Ratchaburi, Thailand; ¹³Department of Ophthalmology, Prapokklao Hospital, Chanthaburi, Thailand; ¹⁴Department of Ophthalmology, Mahasarakham Hospital, Mahasarakham, Thailand; ¹⁵Department of Ophthalmology, Nongbualamphu Hospital, Nongbualamphu, Thailand; ¹⁶Department of Ophthalmology, Pakchongnana Hospital, Nakhon Ratchasima, Thailand; ¹⁷Department of Ophthalmology, Mukdahan Hospital, Mukdahan, Thailand; ¹⁸Department of Ophthalmology, Suratthani Hospital, Suratthani, Thailand; ¹⁹Department of Ophthalmology, Sungaikok Hospital, Narathiwat, Thailand and ²⁰Bangkok Metropolitan Administration Public Health Center 7, Bangkok, Thailand

Correspondence: Lily Peng (lhpeng@google.com)

These authors contributed equally: Paisan Ruamviboonsuk, Jonathan Krause

These authors contributed equally: Peranut Chotcomwongse, Rory Sayres

These authors contributed equally: Lily Peng, Dale R. Webster

Received: 26 October 2018 Accepted: 11 March 2019

Published online: 10 April 2019

potential to solve some of these problems.⁶ Similar issues have been observed in the United Kingdom.⁷

Several DL algorithms for DR have shown promise in populations with multiethnic diabetic patients.^{1–3} However, before the deployment of DL for screening DR, additional large-scale validation on screening populations that are distinct from the original developmental datasets will be critical. In addition, the use of rigorous reference standards that are adjudicated by retinal specialists is important for robust evaluation of the algorithm and human graders.² Lastly, the diagnostic accuracy of DL algorithms should be compared with human graders who routinely grade retinal images in a screening population.

This study was conducted to assess the screening performance of the DL algorithm compared to real-world graders for classifying multiple clinically relevant severity levels of DR in the national screening program for DR in Thailand.

RESULTS

Participant demographics

The characteristics of the images and patients included in this study are described in Table 1. This cohort consisted of 7517 patients, of whom 67.5% were women. The average age was 61.13 (SD = 10.96) years.

Image gradability

Out of 29,943 images, 4595 were deemed not gradable for DR by either the regional grader, the DL algorithm, or both (Supplementary Table 1, Supplementary Figs. 1 and 2). A sample of images where the regional grader disagreed with the DL algorithm on image gradability underwent adjudication and the results of the adjudication are presented in the Supplementary Tables 2 and 3. Adjudicators were approximately 2.5 times more likely to agree with the algorithm than regional graders about the gradability of images for DR. For this difficult image subset, adjudicators agreed with the regional grader 29.0% of the time vs 70.9% of the time with the algorithm. For diabetic macular edema (DME) gradability, they were just as likely to agree with the algorithm as the regional graders.

Manual grading and algorithm performance

A comparison of the performance of regional graders in Thailand and the algorithm compared to the reference standard for all gradable images is summarized in Supplementary Table 4. Out of all the gradable images, the composite sensitivity (i.e., generated by pooling all patients and then computing the metric) of the graders for detecting moderate or worse NPDR was 0.734 (ranging from 0.4071 to 0.914 across regional graders) and the specificity was 0.980 (range: 0.939–1.000). The DL algorithm had a sensitivity of 0.968 (range: 0.893–0.993), specificity of 0.956 (range: 0.983–0.987), and area under the curve (AUC) of 0.987 (range: 0.977–0.995) (Fig. 1). These differences in sensitivity (24% absolute percentage points) and specificity (–2.5%) between the algorithm and regional graders were statistically significant ($p < 0.001$). The algorithm's performance was better than or equal to that of composite grading of regional graders for severe or worse NPDR and proliferative diabetic retinopathy (PDR), with AUC values of 0.991 (range: 0.978–0.997) and 0.993 (range: 0.974–0.995), respectively (Fig. 1). Using moderate or worse NPDR as the referral threshold, regional graders did not refer 73 out of the 602 (12.1%) severe NPDR or worse images and 56 out of the 398 PDR images (14.1%). The algorithm would have missed 20 out of the 602 (3.3%) severe NPDR or worse images and 18 out of 398 PDR images (4.5%).

Results for DME were similar. The sensitivity for regional graders for detecting referable DME was 0.620 (range: 0.450–0.803 across

regions) and the specificity was 0.992 (range: 0.973–0.998). For the DL algorithm, sensitivity was measured at 0.953 (range: 0.859–1.000), specificity was measured at 0.982 (range: 0.944–0.991), and AUC of 0.993 (range: 0.980–0.998; 95% confidence interval (CI) 0.993–0.994).

Our approach to evaluating grader performance involved adjudicating a subset of our image set—cases with disagreement and a sub-sample of cases where the algorithm and human graders agreed. While this is a common reference standard used in previous studies,^{3,8} and allows us to analyze a large representative dataset, it may overestimate the accuracy of both deep learning and human graders. We therefore ran additional analyses to estimate the impact of our adjudication strategy. We subsampled 5% of the cases with agreement between the algorithm and graders randomly for adjudication. We found that the adjudication panel agreed with the algorithm and regional grader 96.2% of the time for moderate or worse NPDR and 87.5% of the time for referable DME. By extrapolating these measured agreement rates to unadjudicated images, we could compute adjusted performance metrics of algorithm and graders as if all images had been adjudicated. For moderate or worse NPDR, the regional grader would have 0.611 sensitivity and 0.959 specificity and the algorithm would have 0.862 sensitivity and 0.934 specificity. For referable DME, the regional grader would have 0.491 sensitivity and 0.989 specificity, and the algorithm would have 0.875 sensitivity and 0.977 specificity.

Because more rapid referral is warranted for cases with severe or worse NPDR and/or DME, the performance of the algorithm for these cases was also examined. At this threshold, regional graders had a sensitivity of 0.635 (range: 0.475–0.831) and specificity of 0.997 (range: 0.967–0.998), while the algorithm had a sensitivity of 0.936 (range: 0.852–0.984) and specificity of 0.982 (range: 0.948–0.993). Using PDR and/or DME as the threshold yields similar performance metrics as severe or worse NPDR and/or DME because the number of DME cases outnumbers that of PDR cases (Figure S1). Additional results for using individual DR severity levels as the threshold are summarized in Supplementary Fig. 3.

The performance of the graders cannot be directly compared to each other because they each graded a different set of images that correspond to their region. However, the algorithm's performance for images from each region could be compared directly to the regional grader from that region (Fig. 2). In nearly all regions, the algorithm's sensitivity was significantly higher than that of the respective regional grader for moderate or worse NPDR and for DME. In one of the regions, the algorithm's sensitivity was lower than that of the regional grader for moderate or worse NPDR, but the difference was not statistically significant ($p = 0.98$).

The overall agreement between the regional graders and algorithm in comparison to the reference standard for DR and DME is shown in a confusion matrix presented in Fig. 3. Furthermore, to compare the agreement for the entire range of DR severities (no/mild, moderate, severe, and proliferative) and for each region, quadratic-weighted Cohen's kappa was used (Supplementary Table 5). Regional graders were measured at 0.773 (range: 0.624–0.875 across regions) and 0.844 (range: 0.736–0.870), $p < 0.001$ for the difference.

In addition to per image metrics, we also calculated the performance of the regional grader and algorithm on a per visit basis (2 images per visit), including both visits, and on a first visit per patient basis. For moderate or worse NPDR, the regional grader sensitivity increased ~3% when going from per image basis to per visit basis, while specificity decreased ~1%. For the algorithm, sensitivity and specificity stayed roughly the same (within ~1%). For referable DME, the regional grader average sensitivity increased ~6% while specificity decreased less than 1%. The algorithm's sensitivity remained roughly unchanged while specificity decreased ~1%. Details of performance by region are shown in Supplementary Tables 4 and 5.

Table 1. Summary of patient characteristics, including breakdowns by region

Region Grader type	1		2		3		4		5		6		7		8		9		10		11		12		13		
	MD	MD	MD	MD	MD	MD	MD	MD	MD	MD	MD	MD	MD	Nurse	MD	MD	Nurse	Nurse	Nurse	Nurse	Nurse	Nurse	Nurse	Nurse	Tech	Tech	
All regions	7517	620	569	440	513	620	440	513	620	680	1005	750	370	500	250	1000	750	370	500	250	1000	750	370	500	250	1000	
Total patients	25985	2467	2256	1760	2051	2424	1760	2051	2424	2720	4020	2989	1582	1986	968	3998	2989	1582	1986	968	3998	2989	1582	1986	968	3998	
Total images	87.83	92.32	94.02	92.95	87.85	88.81	92.95	87.85	88.81	82.85	82.10	89.62	75.24	92.45	86.67	93.81	89.62	75.24	92.45	86.67	93.81	89.62	75.24	92.45	86.67	93.81	
% No/Mild NPDR	9.80	5.65	5.35	5.17	7.40	8.77	5.17	7.40	8.77	12.53	16.21	8.14	20.06	6.15	10.75	5.39	8.14	20.06	6.15	10.75	5.39	8.14	20.06	6.15	10.75	5.39	
% Moderate NPDR	0.81	0.60	0.19	0.77	0.56	0.75	0.77	0.56	0.75	1.38	0.57	1.30	1.16	0.27	1.34	0.40	1.30	1.16	0.27	1.34	0.40	1.30	1.16	0.27	1.34	0.40	
% Severe NPDR	1.57	1.43	0.43	1.12	4.18	1.67	1.12	4.18	1.67	3.24	1.11	0.94	3.55	1.14	1.23	0.40	0.94	3.55	1.14	1.23	0.40	0.94	3.55	1.14	1.23	0.40	
% PDR	6.23	3.08	3.50	3.30	4.00	7.47	3.30	4.00	7.47	8.68	8.81	6.62	15.42	3.83	6.20	2.30	6.62	15.42	3.83	6.20	2.30	6.62	15.42	3.83	6.20	2.30	
% Referable DME	69	61	67	68	65	69	68	65	69	69	77	71	64	75	63	68	71	64	75	63	68	71	64	75	63	68	
% Female	31	39	33	32	35	31	32	35	31	31	23	29	36	25	37	32	29	36	25	37	32	29	36	25	37	32	
% Male	59 (52, 66)	57 (50, 63)	59 (52, 63)	63 (57, 70)	62 (54, 70)	58 (51, 65)	63 (57, 70)	62 (54, 70)	58 (51, 65)	62 (54, 68)	59 (53, 66)	56 (49, 64)	56 (49, 64)	63 (56, 71)	56 (50, 61)	59 (51, 67)	56 (49, 64)	56 (49, 64)	63 (56, 71)	56 (50, 61)	59 (51, 67)	56 (49, 64)	56 (49, 64)	63 (56, 71)	56 (50, 61)	59 (51, 67)	
Age	7.3 (6.5, 8.6)	7.0 (6.4, 8.5)	7.3 (6.4, 8.2)	7.2 (6.4, 8.5)	7.2 (6.5, 8.4)	7.2 (6.3, 8.5)	7.2 (6.4, 8.5)	7.2 (6.5, 8.4)	7.2 (6.3, 8.5)	7.7 (6.8, 8.1)	7.6 (6.8, 9.3)	7.6 (6.5, 9.1)	7.2 (6.3, 8.5)	8.4 (7.3, 9.8)	7.2 (6.5, 8.2)	8.1 (7.1, 9.6)	7.2 (6.3, 8.5)	7.6 (6.8, 9.3)	7.6 (6.5, 9.1)	7.2 (6.3, 8.5)	8.4 (7.3, 9.8)	7.2 (6.3, 8.5)	8.4 (7.3, 9.8)	7.2 (6.5, 8.2)	8.1 (7.1, 9.6)	7.0 (6.3, 8.0)	
HbA1c (%)	139 (118, 169)	130 (110, 172)	136 (118, 168)	133 (115, 166)	133 (114, 156)	150 (126, 181)	133 (115, 166)	133 (114, 156)	150 (126, 181)	140 (118, 199)	144 (121, 172)	133 (107, 170)	149 (122, 188)	131 (115, 154)	149 (130, 186)	136 (118, 163)	133 (107, 170)	144 (121, 172)	133 (107, 170)	149 (122, 188)	131 (115, 154)	149 (130, 186)	136 (118, 163)	131 (115, 154)	149 (130, 186)	136 (118, 163)	
FBS (mg/dL)	105 (83, 130)	117 (107, 147)	102 (79, 124)	101 (81, 128)	94 (75, 120)	103 (80, 129)	101 (81, 128)	94 (75, 120)	103 (80, 129)	102 (93, 122)	109 (86, 132)	107 (86, 131)	104 (85, 124)	96 (73, 119)	118 (96, 142)	108 (88, 132)	107 (86, 131)	104 (85, 124)	96 (73, 119)	118 (96, 142)	108 (88, 132)	107 (86, 131)	104 (85, 124)	96 (73, 119)	118 (96, 142)	108 (88, 132)	
LDL (mg/dL)																											

For blood sample measures and visual acuity, values reflect the distribution across patients at first visit. Numeric values indicate the median across a distribution; values in parentheses indicate the 25th and 75th percentiles. MD ophthalmologist, DME diabetic macular edema, NPDR non-proliferative diabetic retinopathy, PDR proliferative diabetic retinopathy, HbA1c hemoglobin A1c, FBS fasting blood glucose, LDL low-density lipoprotein

Algorithm performance based on confidence score
 While the output of the algorithm is ultimately distilled in a categorical call (e.g., severe NPDR vs PDR), the algorithm originally returns a value between 0 and 1 for each level of DR and DME, indicating a confidence for each severity level. Analysis of the model's performance based on the maximum score of both the DR and DME predictions showed that the algorithm was more sensitive than regional graders at all ranges of confidence score. However, when the algorithm was uncertain (maximum score < 0.7), the specificity of the algorithm was much lower than that of the regional grader at this particular operating point (Figure S2).

DISCUSSION

This study represents one of the largest clinical validations of a deep learning algorithm in a population that is distinct from which the algorithm was trained. In addition, this external validation was conducted in direct comparison with the actual graders in the screening program of the same population. This is pivotal since many countries in the world adopted trained graders for their screening programs for DR. These include the United States,^{9,10} United Kingdom,¹¹ India,¹² Malaysia,¹³ South Africa,¹⁴ Singapore,¹⁵ and China¹⁶ among others. Furthermore, according to the statement by American Academy of Ophthalmology (AAO),^{14,17} there is a strong (level 1) evidence that single-field retinal photography with interpretation by trained graders can serve as a screening tool to identify patients with DR where ophthalmologists are not readily available for referral for further ophthalmic evaluation and management.

Most algorithms in previous studies of DL for DR screening have simplified the various levels of DR into binary predictions, either referable and non-referable or with and without sight-threatening DR. However, in real-world situations, patient management can be different at each level of DR. For example, patients with PDR may require higher urgency for referral for panretinal photocoagulation or intravitreal injections compared to another severity level, such as severe NPDR without DME. On the other hand, patients with moderate NPDR without DME, although labeled as referable, may not require treatment but still require periodic close monitoring by ophthalmologists or retinal specialists. Identifying the referable group of patients requiring treatment accurately may save community resources. Validation of the performance of DL for classifying severity levels of DR would therefore be essential for real-world screening of DR.

Recently, one of the largest head-to-head comparisons between deep learning and human graders was performed by Ting et al.³ The study included 10 secondary validation sets (e.g. validation sets that were drawn from a population that is distinct from the one in which the model was trained), the largest consisting of over 15,000 images. The algorithm validated in this study extends some (but not all) aspects of this body of work. For example, the algorithm, which has been previously validated in populations from the United States and India, showed excellent generalization on a national level in Thailand across different cameras (Supplemental Table 6) and different levels of expertise (ophthalmologists or trained non-ophthalmologists) of graders. It also showed high accuracy when measured on both binary and multi-class tasks. Grading on a more granular 5-point grade is advantageous, especially on a global scale, where follow-up and management guidelines among the many different guideline bodies may vary at each of the 5 levels and in the presence of possible macular edema.¹⁸ For example, while the follow-up recommended by the AAO^{14,17,19} can be up to 12 months for moderate NPDR with no macular edema, that recommendation changes to 3–6 months if there is macular edema and 1 month when there is clinically significant macular edema.¹⁷

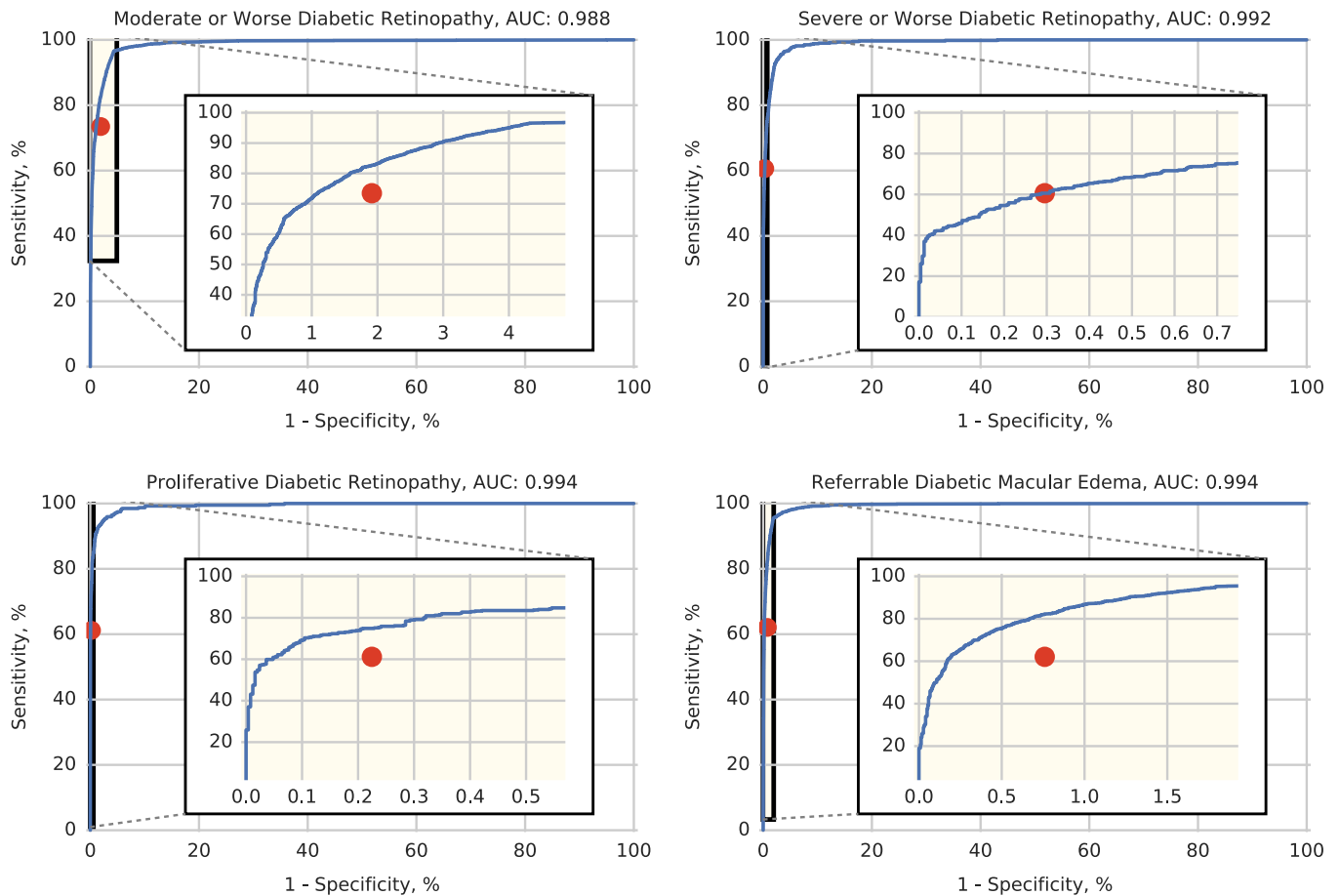


Fig. 1 Comparison of manual grading and algorithm performance. Receiver operating characteristic (ROC) curve of model (blue line) compared to grading by regional graders (red dot) for varying severities of diabetic retinopathy (DR) and diabetic macular edema (DME). The performance represented by the red dot is a combination of all of the grades of the regional graders on all gradable images, since regional graders only graded images from their own region

The threshold level for referral in a screening program for DR may be dependent on the resources of the program. In a lower resource setting where ophthalmologists only see severe cases, the referral threshold may be higher. In a higher resource setting where ophthalmologists prefer to see mild cases, the referral threshold may be lower. There has been some work in the literature to address the importance of this 5-severity levels grading of DR and an adjudicated reference standard.²⁰ However, this work was not validated on a dataset from a different population until this present study. External validation of a deep learning system for accuracy of a 5-point grade should not only give an advantage for selecting an appropriate threshold with acceptable accuracy for a screening system, but also act as a feedback loop to tune the accuracy of the deep learning system itself.

There are some limitations to this study. First, adjudication was performed largely for moderate NPDR or worse cases where there was disagreement between the grader and algorithm. For patients with no retinopathy, screening every 2 years may be appropriate, while those with mild NPDR should be screened once a year.²¹ Therefore, future studies should include adjudication for cases where there is disagreement between no and mild DR.

In addition, the exclusion images with other retinal diseases and images deemed ungradable by either the algorithm or grader may also have inflated algorithm performance. These limitations need to be addressed before deploying the algorithms, particularly in a stand-alone fashion. Future work should include the extension of

the algorithm to detect other common comorbid eye diseases like age-related macular degeneration (AMD) and glaucoma. Compared to the regional graders, the algorithm was more likely to agree with adjudication for gradability. However, more work should be done to study the integration of an image quality algorithm prospectively to see (1) how immediate feedback about image quality may motivate camera operators to re-image the patient for high-quality photos and (2) the disease distribution for images that algorithm considered gradable with no or mild grade but are actually ungradable images with referable disease.

Furthermore, the algorithm's performance could be improved upon for difficult cases. For example, if we used the moderate or worse NPDR operating point, the algorithm would have failed to refer approximate 18 out of 398 cases (4.5%). However, if we used the PDR operating point for referral, the algorithm would have only missed 112 out of 398 cases (28.1%). While the manual grading would have missed even more cases and it is unlikely the PDR operating point would be used for referral instead of the moderate or worse NPDR operating point, there is still room for improvement. Even though the algorithm did miss a few cases with obscure neovascularization picked up by the retinal specialists, it primarily missed cases with inactive fibrous tissue without neovascularization at the optic disc or elsewhere or cases with panretinal photocoagulation scars. Additional training could help improve the algorithm for these cases in the future. Some examples of cases with PDR that were missed by deep learning and the human graders are shown in Figure S3. Given these

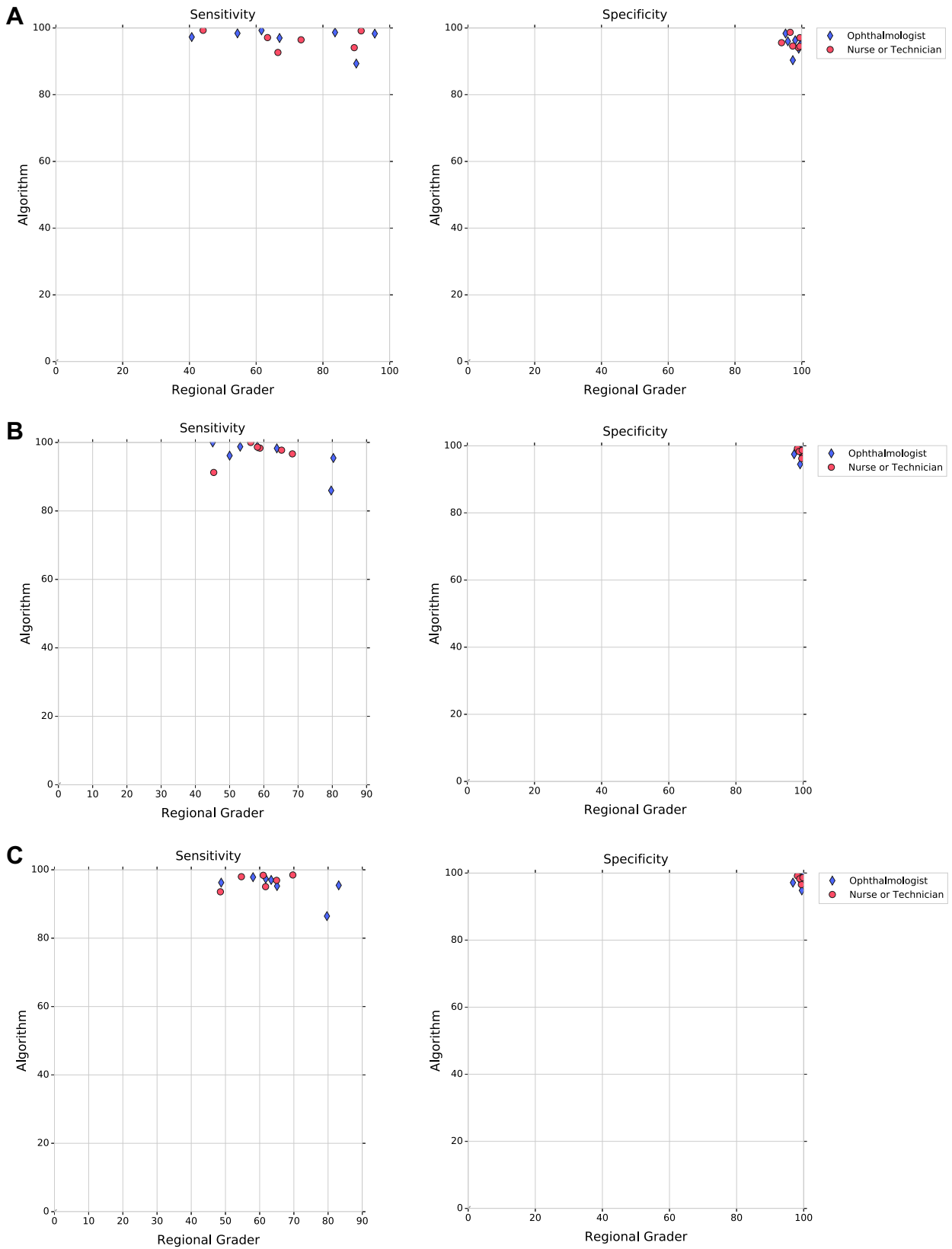


Fig. 2 Comparison of algorithm and individual regional grader performance. Grader performances are represented as blue diamonds (ophthalmologists) and red dots (nurse or technician) for **a** moderate or worse non-proliferative diabetic retinopathy (NPDR), **b** diabetic macular edema (DME), and **c** severe NPDR, proliferative diabetic retinopathy (PDR), and/or DME. Analysis is performed on all gradable images

A

		Regional Grader Label			
		No/Mild NPDR	Moderate NPDR	Severe NPDR	Proliferative DR
Reference Standard	No/Mild NPDR	21,828	380	13	33
	Moderate NPDR	744	1,699	13	14
	Severe NPDR	15	81	100	9
	Proliferative DR	55	85	13	241

Quadratic-Weighted Kappa: 0.774 [0.756-0.791]

B

		Algorithm Label			
		No/Mild NPDR	Moderate NPDR	Severe NPDR	Proliferative DR
Reference Standard	No/Mild NPDR	21,281	919	38	16
	Moderate NPDR	78	1,851	523	18
	Severe NPDR	1	5	195	4
	Proliferative DR	17	17	75	285

Quadratic-Weighted Kappa: 0.844 [0.834-0.855]

C

		Regional Grader Label	
		No DME	DME
Reference Standard	No DME	22,206	170
	DME	699	1,144

D

		Algorithm Label	
		No DME	DME
Reference Standard	No DME	21,975	401
	DME	88	1,755

Fig. 3 Agreement on the image level between the reference standard and regional graders. Comparison of diabetic retinopathy (DR) and diabetic macular edema (DME) performance between the reference standard and **a, c** regional graders or **b, d** the algorithm. Adjudication was performed only for images where either the regional grader or the algorithm identified as moderate and above. Thus, for DR, non-referable cases (no/mild) are combined into a non-referable bucket

findings, if this algorithm were to be deployed in a real-world setting, any image the algorithm grades as moderate or above or severe and above will be immediately referred to manual review to ensure high sensitivity for PDR.

While the regional graders were 80–85% accurate (since regional grader specificity was quite high), the sensitivity of some of the graders was lower than expected for trained graders in a screening program. This not only highlights the necessity of regular audits for human graders, but also an opportunity for algorithms to serve in training, education, and auditing.

In addition, in a real-life screening setting, grading may be performed with a combination of automated and manual grading. A preliminary analysis was performed to look at the relationship between algorithm confidence and performance of both algorithm and manual grading. Future studies could further explore how to combine the algorithm and manual grading to achieve better performance than either alone, while minimizing manual grading workload. In addition, the reference standard for DME in this study was based on monoscopic fundus photos. For DME, optical coherence tomography is now considered the clinical standard, and incorporating this in the reference standard would

be ideal.²² Lastly, DR screening programs generally also refer patients at high suspicion for other non-DR eye diseases such as AMD or glaucoma. The ability to detect other eye diseases would further increase the utility of these algorithms.

This study represents an early milestone in the implementation of a DL algorithm in a large-scale DR screening program. The demonstration of the algorithm's performance and generalizability compared to actual graders in a screening program lays the groundwork for other prospective studies—to further validate the algorithm's performance in real screening workflows and to study its impact on DR screening as a whole. While it is critical that a DL algorithm is accurate, it is equally important to study how the algorithm may affect clinical workflow and outcomes of patients, such as clinician and patient satisfaction and patient adherence to follow-up recommendations, and ultimately impact on disease prevention, progression, and outcomes.

METHODS

This study was approved by the Ethical Review Committee for Research in Human Subjects of the Ministry of Public Health of Thailand and the Ethical Committees of hospitals or health centers from which retinal images of patients with diabetes were used. Patients gave informed consents allowing their retinal images to be used for research. This study was registered in the Thai Clinical Trials Registry, Registration Number TCTR20180716003.

Data acquisition

Diabetic patients were randomly identified from a national registry of diabetic patients, representing hospitals or health centers in each of the 13 health regions in Thailand. Patients were included if they had fundus images of either eye captured using retinal cameras in both the years 2015 and 2017, as part of a 2-year longitudinal study on DR. Retinal images of the patients were single-field, 45-degree field of view, and contained the optic disc and macula, centered on the macula. A variety of cameras were used for image acquisition including ones manufactured by 3nethra, Canon, Kowa, Nidek, Topcon, and Zeiss (Supplementary Table 6). Images were retrieved from the digital archives from retinal cameras utilized in the Thailand DR national screening program. Images were excluded from analysis if they were from patients who had other retinal diseases that precluded classification of severity of DR or DME, such as AMD and other retinal vascular diseases.

Definition of DR severity levels and DME

Severity levels of DR and DME were defined according to the International Clinical Classification of DR (ICDR) disease severity scale.²⁰ In short, DR was classified into no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR. DME was identified as referable DME when hard exudates were found within the distance of 1 disc-diameter from the center of the fovea.^{23,24}

Sample size estimation

According to previous community-based studies of DR in Thailand,²⁵ the prevalence of sight-threatening DR (PDR, severe NPDR, or DME) was approximately 6% of patients with diabetes. With a margin of error of 10%, type 1 error at 0.5 and type 2 error at 0.2, the sample size was estimated at no less than 6112 patients with diabetes. A rate of ungradable images at 20% was estimated, and therefore at least 7450 patients with diabetes were required. The distribution of diabetic patients included from each region was in proportion with the distribution of diabetic patients from each region. The numbers of patients from each of the 13 regions are listed in Table 1.

Deep learning algorithm

The development of the deep learning algorithm for predicting DR and DME is described in detail in Krause et al.² Briefly, a convolutional neural network was trained with an "Inception-v4"²⁶ architecture that predicted a 5-point DR grade, referable DME, gradability of both DR and DME, and an overall image quality score. The input to the neural network was a fundus

image with a resolution of 779 × 779 pixels. Through the use of many stages of computation, parameterized by millions of numbers, the network outputs a real-valued number between 0.0 and 1.0 for each prediction, indicating its confidence. During training, the model was given different images from the training set with a known severity rating for DR, and the model predicted its confidence in each severity level of DR, slowly adjusting its parameters over the course of the training process to increase its accuracy. The model was evaluated on a tuning dataset throughout the training process, which was used to determine model hyperparameters. An "ensemble" of ten individual models was then created to combine their predictions for the final output. To turn the model's confidence-based outputs into discrete predictions, a threshold on the confidence was used for each binary output (DME, DR gradability, and DME gradability), and a cascade of thresholds² was used to output a single DR severity level.

Grading by regional graders

The DL algorithm was compared to 13 human regional graders who actually grade retinal images for the screening program in each of the 13 health regions. In this study, each grader only graded images that have been screened in his or her own region. Some of the graders were general ophthalmologists and others were trained ophthalmic nurses or technicians. Each grader had at least 2 years of experience grading retinal fundus images for DR. Each grader received standard grading protocols for DR and DME, including instructions for the web-based grading tool before the commencement of the study, and each was required to use the same web-based tool for online grading of the retinal images. A tutorial session was conducted for all the graders before the commencement of grading.

Reference standard

There were two groups of retinal specialists who graded retinal images for the reference standard. One group was assigned to grade for DR severity level and another for referable DME.

For gradability, a subset of ~1000 images each for DR and DME where the regional grader disagreed with the DL algorithm on image gradability underwent adjudication. For the remainder of the analysis for both DR and DME, images graded as ungradable by either the algorithm or regional grader were excluded. Additionally, for images which were not adjudicated but the DL algorithm and regional graders were in agreement about the severity of DR, the agreed-upon grade was used as the reference standard. A similar rule was applied for analysis of DME.

For grading DR severity levels the ICDR scale was used. In order to reduce adjudication time for quality of grading, adjudicators were assigned to grade a subset of the images. This subset included all images for which the regional grader and the DL algorithm were in disagreement and at least one graded as moderate NPDR or worse; a random sample of 75 images for which the algorithm and regional grader were in agreement and graded as moderate NPDR or worse; and a random sample of 1175 images for which the algorithm and regional grader both graded as less than moderate NPDR. This random sample represented 5% of all images with agreement between the two modalities. The ratio of images with moderate NPDR or worse to those with less than moderate NPDR in this random sample was proportional to the entire population. For the purposes of this study, no DR and mild NPDR were considered a single category, and only images that both the algorithm and regional grader deemed gradable for DR were adjudicated. For grading referable DME, retinal specialists were assigned to grade all images for which the regional graders and DL were in disagreement about the binary presence or absence of DME, and 5% of the rest of the images were randomly assigned from the subset of those images also determined to be gradable for DME by both the regional grader and the algorithm. Most of the disagreement occurred in cases that were graded as moderate or worse NPDR or referable DME. Further review of the discrepancies revealed that most of the cases were fairly ambiguous and adjudicators tended to err on the side of increased sensitivity for a screening setting. Overall, for moderate or worse NPDR, the regional grader and algorithm grade agreed with adjudication 96.3% of the time. For DME, the agreement rate was 99.1%.

The retinal specialists that served as the reference standard in this study were from Thailand, India, and the United States. There were two retinal specialists per group. Each group graded the images independently, and the same instructions for web-based grading that the regional graders used were issued to them before grading. They were also required to use

the same web-based grading tool as the regional graders. In addition, the group that graded DR severity level had a teleconference and graded a small set of images together to ensure congruence.

The adjudication process was as follows: First, both retinal specialists in a group independently graded each image. Then, until consensus was reached, the retinal specialists took turns revising their grades, each time with access to their previous grade and the other retinal specialist's grade, as well as any additional comments about the case either retinal specialist left. If there was still disagreement after each grader had graded the image three times in this way, then the image's reference standard was determined independently by a separate, senior retinal specialist. For grading DR severity levels, differences between no DR and mild NPDR were not adjudicated in order to focus adjudication time on referable disease.

Statistical analysis

Primary metrics assessed were sensitivity, specificity, and area under the receiver operating characteristic curve. Confidence intervals for sensitivities and specificities were calculated using the Clopper–Pearson interval, and CIs for the quadratic-weighted Cohen's kappa were calculated using a bootstrap, both calculated at the 95% level. All *p* values were calculated using two-sided permutation tests.

DATA AVAILABILITY

The data that support the findings of this study may be available from DR screening programs of Rajivithi Hospital, Lamphun Hospital, Somdejphrajoetaksin Maharaj Hospital, Sawanpracharak Hospital, Nakhon Nayok Hospital, Photharam Hospital, Prapokklao Hospital, Mahasarakham Hospital, Nongbualamphu Hospital, Pakchongnana Hospital, Mukdahan Hospital, Suratthani Hospital, Sungaikolok Hospital, Bangkok Metropolitan Administration Public Health Center 7, but restrictions apply to the availability of these data. These data, or a test subset of them, may be available subject to ethical approvals.

CODE AVAILABILITY

Machine learning models were developed and deployed using standard model libraries and scripts in TensorFlow. Custom code was specific to our computing infrastructure and mainly used for data input/output and parallelization across computers.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the following people for their advice and assistance with this paper: Piyada Pholnonluang, Rada Kanjanapachot, Yothin Jindaluang, Nitaya Suwannaporn, Niteeya Narapisut, Naraporn Sumalgun, Tanawan Sinprasert, Tienchai Methanoppakhun, Ratana Boonyong, Preeyawan Tanomsumpun, Rojsak Phuphiphat, Porntip Nitikarun, Phirun Wongwaithayakonkun, Achareeya Saenmee, Cheunnapa Komkam, Supaporn Numsui, Praween Tantiprapha, Sopon Nilkumhang, Roongruedee Thangwongchai, Supaporn Petcharat, Jansiri Laiprasert, Premrudee Maneechaweng, Sareerah Waesamah, Poowadon Pimpakun, Prapaphan Charncheaw, Ramida Panyarattannakul, Suwannee Saelee, Nutchana Onarun, Yun Liu, Peter Wubbels, Florence Thng, Sunny Virmani, Varun Gulshan, Philip Nelson, David Coz, Derek Wu, Elin Pedersen, William Chen, Jessica Yoshimi, Xiang Ji, Quang Duong, and Brian Basham. This paper will be posted on a preprint server. The results have been presented at the 2018 Asia Pacific Tele-Ophthalmology Society.

AUTHOR CONTRIBUTIONS

Conception or design of the work: P.R., J.K., P.C. R.S., R.R., G.S.C., L.P. and D.R.W. Acquisition, analysis, or interpretation of data: P.R., J.K., P.C., R.S., R.R., L.P., D.R.W, K.W., B.J.L.C., S.P., K.H., M.T., S.S.-A., J.L., C.R., O.K., J.J., J.T., S.O., C.K., R.S., C.L., J.F., P.K., L.C., S. S., S.K., K.M, S.L., C.T. and L.W. Drafting the work or revising it critically: all authors. P.R., J.K. and L.P. take full responsibility for the work as a whole, including the study design, access to data, and the decision to submit and publish the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Digital Medicine* website (<https://doi.org/10.1038/s41746-019-0099-8>).

Competing interests: J.K., R.S., K.W., B.J.L.C., G.S.C., L.P., S.P. and D.R.W. are Google employees and receive salary and stock as a part of the standard compensation package. O.K., J.J. and J.T. are consultants for Google.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
- Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
- Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multi-ethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
- Jenchitr, W. et al. The national survey of blindness low vision and visual impairment in thailand 2006–2007. *Thai. J. Pub. Hlth. Ophthalmol.* **21**, 10 (2007).
- Isipradit, S. et al. The First Rapid Assessment of Avoidable Blindness (RAAB) in Thailand. *PLoS ONE* **9**, e114245 (2014).
- Meenakshi Kumar, et al. Perceived barriers for diabetic retinopathy screening by the staff involved in screening in Thailand.
- Ulrich Freudenstein, J. V. A national screening programme for diabetic retinopathy: needs to learn the lessons of existing screening programmes. *BMJ* **323**, 4 (2001).
- Wang, Y.T., Tadarati, M., Wolfson, Y., Bressler, S. B. & Bressler, N. M. Comparison of prevalence of diabetic macular edema based on monocular fundus photography vs optical coherence tomography. *JAMA Ophthalmol.* **134**, 222–228 (2016).
- Cavallerano, A. A. et al. A telemedicine program for diabetic retinopathy in a Veterans Affairs Medical Center—the Joslin Vision Network Eye Health Care Model. *Am. J. Ophthalmol.* **139**, 597–604 (2005).
- Tsan, G. L. Assessment of diabetic teleretinal imaging program at the Portland Department of Veterans Affairs Medical Center. *J. Rehabil. Res. Dev.* **52**, 193–200 (2015).
- City & Guilds. Diabetic Retinopathy Screening qualifications and training courses. <https://www.cityandguilds.com/qualifications-and-apprenticeships/health-and-social-care/health/7360-diabetic-retinopathy-screening#tab=information> (2018). (Accessed 21 September 2018).
- Namperumalsamy, P., Nirmalan, P. K. & Ramasamy, K. Developing a screening program to detect sight-threatening diabetic retinopathy in South India. *Diabetes Care* **26**, 1831–1835 (2003).
- CPG Secretariat. Screening of Diabetic Retinopathy. *Portal Rasmi Kementerian Kesihatan Malaysia*. www.moh.gov.my/penerbitan/CPG2017/6601.pdf (2001).
- Cook, S. et al. Quality assurance in diabetic retinal screening in South Africa. *S. Afr. Med. J.* **104**, 700–704 (2014).
- Nguyen, H. V. Cost-effectiveness of a National Telemedicine Diabetic Retinopathy Screening Program in Singapore. *Ophthalmology* **123**, 2571–2580 (2016).
- Wong, I. Y. H., Ni, M. Y., Wong, I. O. L., Fong, N. & Leung, G. M. Saving sight in China and beyond: the Lifeline Express model. *BMJ Glob. Health* **3**, e000766 (2018).
- American Academy of Ophthalmology. *Diabetic Retinopathy PPP - Updated 2017* (American Academy of Ophthalmology, 2017). <https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017>.
- Chakrabarti, R., Harper, C. A. & Keefe, J. E. Diabetic retinopathy management guidelines. *Expert Rev. Ophthalmol.* **7**, 417–439 (2012).
- Wong, T. Y. et al. Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. *Ophthalmology* **125**, 1608–1622 (2018).
- American Academy of Ophthalmology. International Council of Ophthalmology: Resources. *International Clinical Diabetic Retinopathy Disease Severity Scale, Detailed Table*. <http://www.icoph.org/resources/45/International-Clinical-Diabetic-Retinopathy-Disease-Severity-Scale-Detailed-Table-.html> (2010).
- Scanlon, P. H. Screening intervals for diabetic retinopathy and implications for care. *Curr. Diab. Rep.* **17**, 96 (2017).

22. Virgili, G. et al. Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy. *Cochrane Database Syst. Rev.* **7**, CD008081 (2011).
23. Bresnick, G. H., Mukamel, D. B., Dickinson, J. C. & Cole, D. R. A screening approach to the surveillance of patients with diabetes for the presence of vision-threatening retinopathy. *Ophthalmology* **107**, 19–24 (2000).
24. Li, H. K. et al. Telehealth practice recommendations for diabetic retinopathy, second edition. *Telemed. e-Health* **17**, 814–837 (2011).
25. Supapluksakul, S., Ruamviboonsuk, P. & Chaowakul, W. The prevalence of diabetic retinopathy in Trang province determined by retinal photography and comprehensive eye examination. *J. Med. Assoc. Thai.* **91**, 716–722 (2008).
26. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Thirty-First AAAI Conference on Artificial Intelligence (AAAI, San Francisco, 2016)*.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019