Supplementary Online Content

Ting DS, 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*. doi:10.1001/jama.2017.18152

This supplementary material has been provided by the authors to give readers additional information about their work.

eTable 1. Training and Validation Set for Referable Possible Glaucoma and Referable Age-Related Macular Degeneration

Referable Possible Glaucoma	Referable and Non-referable Possible Glaucoma			Referable Possible Glaucoma		
	Total Images	Total Eyes	Total Patients	Total Images	Total Eyes	Total Patients
Training sets						
Singapore Diabetic Retinopathy Screening Program (SiDRP) 2010-13[16]	76,108	38,185	13,099	120	54	47
Singapore Chinese Eye Study[24]	26,731	6,706	3,353	603	182	134
Singapore Malay Eye Study[24]	10,114	6,560	3,280	333	195	150
Singapore Indian Eye Study[24]	10,819	6,800	3,400	157	111	78
Singapore National Eye Center	1417	1,365	846	1,417	1,365	846
Total Possible glaucoma Images	125,189	59,616	23,978	2,630	1,907	1,255
Clinical validation set						
Singapore Diabetic Retinopathy Screening Program (SiDRP) 2014-15[16]	71,896	35,948	14,880	56	28	24
Referable Age-related Macular Degeneration (AMD)	**Referable and Non-referable AMD**			**Referable AMD**		
	Total Images	Total Eyes	Total Patients	Total Images	Total Eyes	Total Patients
Training sets						
Singapore Diabetic Retinopathy Screening Program (SiDRP) 2010-13[16]	38,185	38,185	13,099	1181	771	589
Singapore National Eye Center AMD Phenotype Study[20, 25, 26]	2,180	348	174	1,632	174	174
Singapore Chinese Eye Study[23]	16,182	6,706	3,353	477	328	264
Singapore Malay Eye Study[23]	8,616	6,560	3,280	315	223	179
Singapore Indian Eye Study[23]	7,447	6,800	3,400	410	253	198
Total AMD Images for training	72,610	58,599	23,306	4,015	2,766	1404
Clinical validation set						
Singapore Diabetic Retinopathy Screening Program (SiDRP) 2014-15[16]	35,948	35,948	14,880	773	761	665

Referable possible glaucoma – defined as vertical CDR 0.8 and above, local neuro-retinal rim thinning, focal notching, disc haemorrhage, retinal nerve fibre layer defect; Referable AMD – defined as intermediate AMD and/or advanced AMD (geography atrophy and neovascular AMD).

eTable 2. Demographics, Diabetes History and Systemic Risk Factors of Patients for External Validation Datasets for Referable DR and Training Datasets for Referable Possible Glaucoma and Referable Age-Related Macular Degeneration

	External Validation Sets for referable DR*			Training sets for referable possible glaucoma and AMD			
	SCES	**SIMES**	**SINDI**	**SCES**	**SIMES**	**SINDI**	**AMD Phenotyping Study**
Patients' numbers	484	763	1128	3353	3280	3400 (6800)	174 (248)
Patients' eyes	968	1,526	2,256	6,706	6,560	6,800	248
Age (years) (mean, SD)	63.66 (9.76)	63.07 (9.4)	61.06 (9.9)	59.69 (9.93)	59.19 (11.02)	57.75 (10.06)	69.68 (9.95)
Gender, Male (number, %)	263 (54.34)	330 (43.25)	590 (52.3)	1662 (49.57 %)	1575 (48.02%)	1706 (50.18%)	96 (55.17%)
Ethnicity							
i. Chinese (number, %)	484, 100.0	N/A	N/A	100	N/A	N/A	152 (87.36%)
ii. Indian (number, %)	N/A	N/A	1,128, 100.0	N/A	N/A	100	11 (6.32%)
iii. Malay (number, %)	N/A	763, 100.0	N/A	N/A	100	N/A	11 (6.32%)
Proportion of patients with diabetes (%)	100%	100%	100%	17.66%	32.07%	38.82%	18.97%
Diabetes duration Median, Range (years)	8.08 (0.18-46.5)	6.32 (0.14-46.0)	8.23 (0.14-53.9)	8.15 (0.18-46.5)	6.32 (0.14-46.0)	8.24 (0.14-53.9)	Not Available
HbA1c (%)	7.55 (1.49)	8.42 (2.02)	7.69 (1.68)	6.06 (0.91)	6.45 (1.55)	6.43 (1.38)	6.07 (0.99)
Systemic risk factors							
1. BMI (kg/m^2)	25.17 (3.78)	27.45 (4.81)	26.78 (4.88)	23.69 (3.65)	26.35 (5.11)	26.16 (4.75)	23.12 (3.69)
2. Systolic blood pressure (mmHg)	142.49 (19.83)	154.56 (23.25)	140.47 (19.93)	136.67 (19.58)	147.51 (23.95)	135.88 (20.1)	138.63 (18.53)
3. Diastolic blood pressure (mmHg)	76.15 (9.12)	79.23 (10.99)	76.72 (9.94)	77.57 (9.9)	79.83 (11.33)	77.61 (10.3)	76.86 (10.1)
4. Total cholesterol (mg/dL)	87.84 (20.34)	98.82 (23.4)	86.4 (21.06)	98.1 (19.26)	101.34 (20.88)	93.42 (19.98)	97.92 (21.60)
5. HDL cholesterol (mg/dL)	21.24 (6.12)	23.04 (5.58)	18.72 (5.76)	23.58 (7.2)	24.3 (5.94)	19.26 (5.76)	25.38 (6.48)
6. LDL cholesterol (mg/dL)	50.22 (15.84)	60.12 (18.9)	53.28 (16.92)	59.04 (16.2)	63.9 (18.18)	59.94 (17.1)	64.08 (16.92)
7. Triglycerides (mg/dL)	38.16 (24.66)	34.38 (28.6)	37.26 (22.14)	33.12 (23.04)	28.98 (23.76)	35.28 (20.88)	35.46 (18.72)
8. Creatinine (mg/dL)	0.98 (0.74)	1.2 (0.98)	0.92 (0.42)	0.85 0.43)	1.06 (0.64)	0.88 (0.38)	0.91 (0.31)

SIMES: [18,19,21,22] Singapore Malay Eye Study; SINDI: [18,19,21,22] Singapore Indian Eye Study; SCES -: [18,19,21,22] Singapore Chinese Eye Study; AMD Phenotyping Study [20,25,26] N/A: Not applicable

* Apart from the population-based studies conducted in Singapore, we do not have patients' demographic and systemic vascular risk factors information on all external datasets due to patients' anonymity.

eTable 3. The Area Under Curve (AUC), Sensitivity (%), Specificity (%) of Deep Learning System (DLS) Versus Trained Professional Graders, With Reference to Retinal Specialist's Grading in Unique SiDRP 14-15 Patients

Diagnostic Performance for Referable DR and Vision-threatening DR			
	Deep Learning System	Graders	P value
1. Referable DR*			
Area under curve (95% CI)^	0.879 (0.864, 0.893)		
Sensitivity (95% CI)^	89.56 (85.51, 92.58)	84.84 (81.28, 88.51)	0.04
Specificity (95% CI)^	83.49 (82.68, 84.27)	98.55 (98.27, 98.79)	<0.001
2. Vision-threatening DR			
Area under curve (95% CI)^	0.908 (0.900, 0.915)		
Sensitivity (95% CI)^	100 (90.97 to 100.0)#	89.74 (74.77, 96.27)	0.04
Specificity (95% CI)^	81.4 (80.57, 82.22)	99.09 (98.86, 99.27)	p<0.001

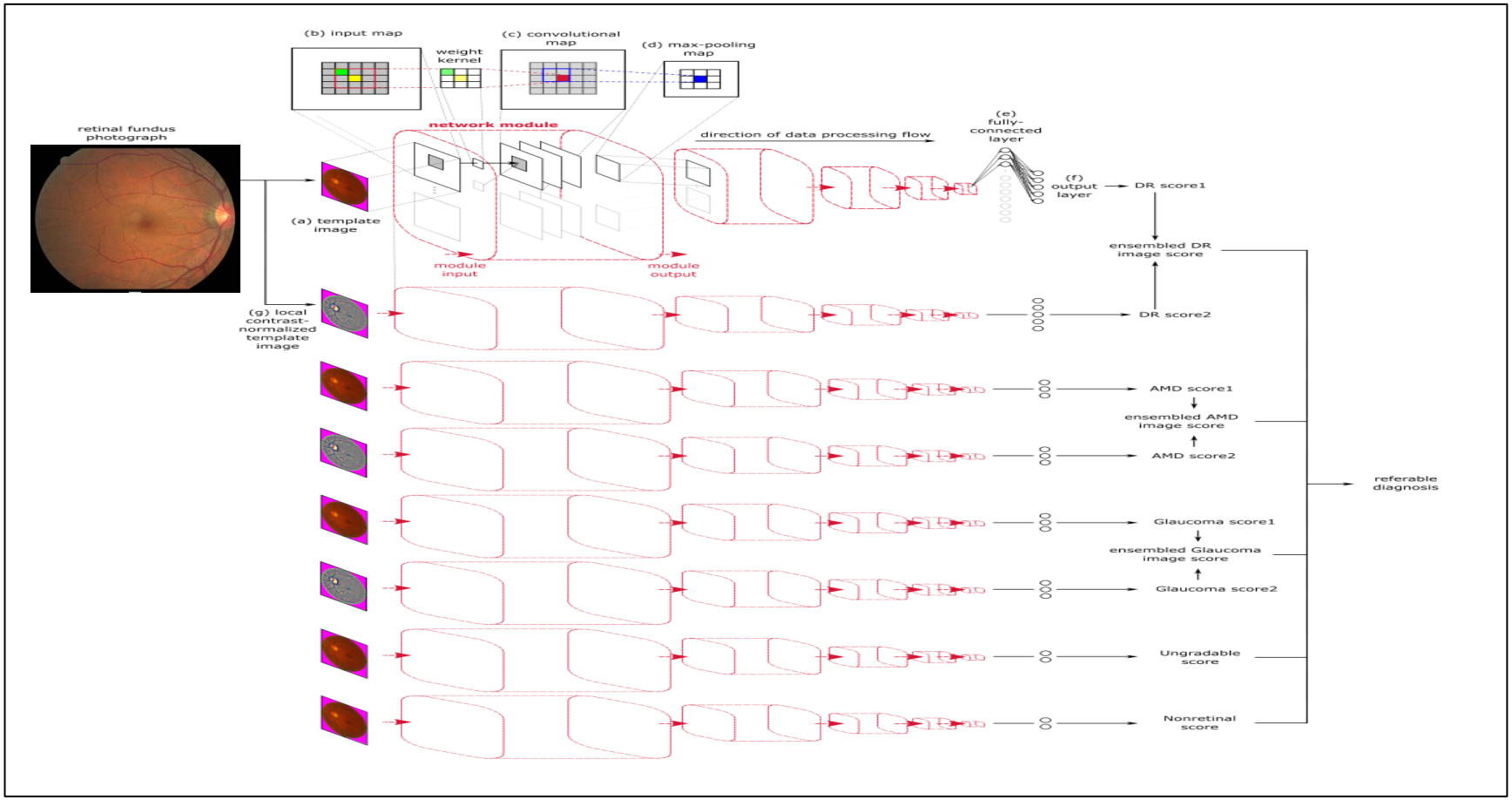
Patients were the units of analysis (n=8,589). These were the new patients who attended SiDRP between 2014 and 15 for the first time and had no previous visit between 2010 and 2013.
DR: Diabetic retinopathy
^ Asymptotic 95% confidence interval was computed for the logit of each proportion
# Exact Clopper-Pearson left-sided 97.5% confidence interval was calculated due to estimate being at the boundary
+ Asymptotic 95% confidence interval was computed for each AUC
* Referable DR was defined as one of the eyes with at least moderate non-proliferative DR (NPDR), severe (NPDR), proliferative DR (PDR) or 'un-gradable'
** Vision-threatening DR was defined as severe non-proliferative DR and PDR
P value was calculated between DLS vs graders using the McNemar's test
Eyes rated 'un-gradable' are treated as referable status
In patients with information in one eye missing, the other eye is used solely to determine referable status
The sensitivity of DLS in detection of DME amongst referable DR eyes = 97.71 (93.04 to 99.27)

eTable 4. The Overall Sensitivity (%), Specificity (%) and Number of Images That Need to Go Through Secondary Grading of 2-Stage Semi-Automated Grading (Deep Learning System as First Stage-Grading, Followed by Manual Grading for Those Test Positive Images), Using a Pre-Set Sensitivity of 90%, 95% and 99% in Detection of Referable Diabetic Retinopathy, Referable Possible Glaucoma and Referable Age-Related Macular Degeneration

Total number of images = 71,896	Semi-automated system with pre-set DLS sensitivity		
	90%	**95%**	**99%**
Overall sensitivity	91.31	95.09	97.05
	(89.67,92.78)^	(93.79,96.19)	(96.00,97.90)
Overall specificity	99.54	99.46	99.38
	(99.45,99.61)	(99.37,99.53)	(99.28,99.46)
Number (%) of images that need to go through secondary grading	18,190	26,601	42,907
	25.30%	37.00%	59.70%

In Singapore, we pre-set the DLS sensitivity at 90%, based on the professional graders' past performances and criteria set by the Ministry of Health, Singapore.
DLS: Deep learning system
^ Asymptotic 95% confidence interval was computed for the logit of each proportion

eFigure 1. Deep Learning System (DLS): The Convolutional Neural Network for Detection of Referable Diabetic Retinopathy (DR), Referable AMD and Referable Possible Glaucoma, Using an Adapted VGGNet Architecture.[1] A colour retinal image will processed as template image (a) to enter an input map (b), a convolutional map (c), a max-pooling map (d), a fully-connected layer (e), an output layer (f) and finally determination of the referable status, taking into account the gradability of the image and presence/absence of referable DR, referable possible glaucoma or referable AMD.



The DLS is composed of eight convolutional neural networks (CNNs), all using an adaptation of the VGGNet architecture:[1] (a) an ensemble of two networks for the classification of DR severity, (ii) an ensemble of two networks for the identification of referable possible glaucoma (iii) an ensemble of two networks for the identification of referable AMD, (iv) one network to assess image quality; and (v) one network to reject invalid non-retinal images. While various network architectures such as Inception and deep residual networks have been employed, VGGNet was employed as it had been demonstrated to produce state-of-the-art performance on the classification of retina images, when our experiments were first conceptualized. The training of a CNN to model DR is achieved by presenting the network with batches of labeled training images. The CNN then incrementally learns the key characteristics of images belonging to each class. We trained multiple CNNs and obtained an image score by assembling the individual CNN scores. Likewise, the eye-level classification is produced using all available images of an eye that are of acceptable quality, and apply score thresholds determined from the training data.

As a preparatory step, each retinal photograph is first automatically segmented to extract only the retina disc. This circular region of interest is then uniformly rescaled to fit a standardized square template of dimension 512x512 pixels **(a)**. The RGB values of the template image are then input as the three channels of the first layer of the relevant CNNs.

A CNN layer consists of many nodes (neurons) that may be arranged in multiple feature maps of the same type (input map, convolutional map, max-pooling). The template image **(a)** will enter the input map **(b),** first layer of the network that directly represent the pixel values of the template image**.** These values are propagated to the convolutional maps **(c)** in the next layer via a convolution operation whereby the value of each node in the source feature map is convolved over a trained weight kernel. We end the series of convolutional maps **(d)** with a 2x2 max-pooling layer that effectively down-samples the feature dimensions by a factor of two**.** These layers form a network module, as enclosed by a red dashed box in eFigure1.

A deep CNN consists of a succession of such network modules where the processing takes place strictly in sequential order where the 2x2 max-pooling layer from an earlier module serve as the inputs to the next module. The series of modules terminates when the features output to the fully-connected layer **(e)** where each circle represents a network node**.** Standard ReLU rectification and dropout layers are then applied, before a final softmax output layer that contains one output node **(f)** for each class trained for**.**

For the classification of DR severity, an ensemble of two convolutional networks was used. One network was provided the original images as input, while the other network was provided locally contrast-normalized images **(g).** The output nodes of each network were indexed according to increasing severity of DR class, from 0 to 4. This allows the predicted DR severity to be represented by a single scalar value, by summing the product of the value of each output node, with its index. The final DR severity score is then the mean of the outputs of the two convolutional networks. Classification of test images is then achieved by thresholding the DR severity score for desired sensitivity/specificity performance, as estimated from the validation set. For the purposes of this paper, a threshold of 0.70 was selected as being adequate for screening purposes. For the classification of AMD and glaucoma severity, a similar procedure was followed, except that each of these conditions admits only three severity classes, from 0 to 2. A threshold of 0.70 was selected for AMD, and 0.70 for glaucoma.

The training procedure for each convolutional network involves repeatedly randomly sampling a batch of images from the training set, together with their ground truth classification. The weight values of the convolutional network are then adjusted by gradient descent, which incrementally improves the general association between images of a certain class, and the value of their corresponding output node. Concurrently, the convolutional network automatically learns useful features at each scale represented by its models, from the smallest-possible pixel level, to scales approaching that of the original input. To expose the convolutional network to additional plausible input feature variations, we apply a limited family of transformations to the input images, involving mirroring, rotation, and scaling by a small degree. Each network was trained approximately to the convergence of its performance, on a small held-out validation set.

Additionally, convolutional networks were trained to reject images for insufficient image quality, as well as for being invalid input (i.e. not being a retinal image). For the latter model, a broad variety of natural images was used as the negative class, in training. Images rejected by either of these models are considered as being recommended for further referral, for the purposes of computing the experimental results. Once an image is analyzed, a report will be generated for the users.

eReference

1. Lim G, Lee ML, Hsu W, Wong TY. Transformed representations for convolutional neural networks in diabetic retinopathy screening. In: *MAIHA, Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014; 34-38.

eFigure 2. Flow Chart of Two Models of the Deep Learning System (DLS) for Diabetic Retinopathy (DR) Screening. Figure 1A shows the fully automated system: All retinal images are analyzed by the DLS. The eye will be considered referable if there is presence of one of the three conditions: referable DR, referable possible glaucoma (GS) and referable age-related macular degeneration (AMD). No human graders are needed. Figure 1B shows the "semi-automated" system: All retinal images are analyzed in initially by DLS, followed by secondary manual grading by a professional grader to reclassify the eyes considered referable.