Author/s:
Keel, S;Li, Z;Scheetz, J;Robman, L;Phung, J;Makeyeva, G;Aung, KZ;Liu, C;Yan, X;Meng, W;Guymer, R;Chang, R;He, M

Title:
Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs

Date:
2019-11-01

Citation:
Keel, S., Li, Z., Scheetz, J., Robman, L., Phung, J., Makeyeva, G., Aung, K. Z., Liu, C., Yan, X., Meng, W., Guymer, R., Chang, R. & He, M. (2019). Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs. Clinical and Experimental Ophthalmology, 47 (8), pp.1009-1018. https://doi.org/10.1111/ceo.13575.

Persistent Link:
https://hdl.handle.net/11343/286203

Keel Stuart (Orcid ID: 0000-0001-6756-348X)
Scheetz Jane (Orcid ID: 0000-0003-0523-1927)

Original Article - Clinical Science

Development and validation of a deep learning algorithm for the detection of neovascular age-related macular degeneration from color fundus photographs

Stuart Keel PhD,[1*] Zhixi Li MD PhD,[2*] Jane Scheetz PhD,[1] Liubov Robman MBBS PhD,[1,3] James Phung BSci(Hons),[3] Galina Makeyeva MBBS PhD,[1] KhinZaw Aung MBBS,[1] Chi Liu MS,[4] Xixi Yan MD,[1] Wei Meng BS,[4] Robyn Guymer PhD FRANZCO,[1] Robert Chang PhD[5] and Mingguang He MD PhD[1,2]

*These two authors contributed equally

[1]Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, University of Melbourne, Melbourne, Australia, [2]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China, [3]Monash University Melbourne, Australia [4]Guangzhou Healgoo Interactive Medical Technology Co.Ltd,[5] Department of Ophthalmology, Byers Eye Institute at Stanford University, Palo Alto, California, USA

Correspondence: Professor Mingguang He, Level 7, Centre for Eye Research Australia, 32 Gisborne Street, East Melbourne, Victoria, Australia 3002
Email: mingguang.he@unimelb.edu.au

Short running title: Automated detection of macular degeneration

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/ceo.13575

Received 22 January 2019; accepted 13 June 2019

Funding sources / Financial disclosure: This research was supported in part by the National Key R&D Program of China (2018YFC0116500), Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (81420108008), Bupa Health Foundation Australia grant, and a MACH 2018 MRFF Rapid Applied Research Translation grant. Prof. Mingguang He receives support from the University of Melbourne at Research Accelerator Program and the CERA Foundation. The Centre for Eye Research Australia receives Operational Infrastructure Support from the Victorian State Government.

Cohort recruitment in the MCCS was funded by VicHealth and The Cancer Council Victoria. Further MCCS funding: the National Health & Medical Research Council of Australia (NHMRC) Program Grant 209057, Capacity Building Grant 251533 and Enabling Grant 396414. The ophthalmic component was funded by the Ophthalmic Research Institute of Australia; American Health Assistance Foundation, Jack Brockhoff Foundation, John Reid Charitable Trust, Perpetual Trustees and Royal Victorian Eye and Ear Hospital.

Conflict of interest: The authors have no financial or other conflicts of interest concerning this study. Mingguang He and Wei Meng report a patent on managing color fundus images using deep learning models. The patent application number was ZL201510758675.5 and patent filing date was May 31, 2017.

Key words: Deep learning algorithm, Age-related macular degeneration (AMD), Retinal-imaging

ABSTRACT

Importance: Detection of early onset neovascular age-related macular degeneration (AMD) is critical to protecting vision.

Background: To describe the development and validation of a deep learning algorithm (DLA) for the detection of neovascular age-related macular degeneration.

Design: Development and validation of a DLA using retrospective datasets.

Participants: We developed and trained the DLA using 56,113 retinal images and an additional 86,162 images from an independent dataset to externally validate the DLA. All images were non-stereoscopic and retrospectively collected.

Methods: The internal validation dataset was derived from real-world clinical settings in China. Gold standard grading was assigned when consensus was reached by 3 individual ophthalmologists. The DLA classified 31,247 images as gradable and 24,866 as ungradable (poor quality or poor field definition). These ungradable images were used to create a classification model for image quality. Efficiency and diagnostic accuracy were tested using 86,162 images derived from the Melbourne Collaborative Cohort Study. Neovascular AMD and/or ungradable outcome in one or both eyes was considered referable.

Main Outcome Measures: Area under the receiver operating characteristic curve (AUC), sensitivity and specificity.

Results: In the internal validation dataset, the AUC, sensitivity and specificity of the DLA for neovascular AMD was 0.995, 96.7%, 96.4%, respectively. Testing against the independent external dataset achieved an AUC, sensitivity and specificity of 0.967, 100% and 93.4%, respectively. More than 60% of false positive cases displayed other macular pathologies. Among the false negative cases (internal validation dataset only), over half (57.2%) proved to be undetected detachment of the neurosensory retina or RPE layer.

Conclusions: This DLA shows robust performance for the detection of neovascular AMD amongst retinal images from a multi-ethnic sample and under different imaging protocols. Further research is warranted to investigate where this technology could be best utilized within screening and research settings.

1. INTRODUCTION

Age-related macular degeneration (AMD) is a leading cause of severe vision loss amongst developed nations, particularly in those aged 65 years and older.[1] With the aging population, it is projected that 288 million people globally will have AMD by 2040.[1] There are two types of advanced AMD, atrophic ('dry') and neovascular ('wet'), with the latter having been responsible for an estimated 90% of severe vision loss cases.[2] Effective therapeutic options are currently only available for neovascular AMD, with anti-vascular endothelial growth factor (VEGF) agents effective in reducing vision loss[8-10] and, in many cases, restoring vision.[3,4]

Detection of early onset neovascular AMD and timely treatment is essential for the protection of visual function. Color fundus photography is a common imaging tool utilised in primary eye care and screening settings and is effective for the diagnosis of AMD.[15] Despite this, accurate interpretation of photographs is dependent on highly trained personnel, limiting its utility in underserved settings, such as developing countries and minority underserved populations. Furthermore, amongst high throughput settings (e.g. screening, epidemiological research), most images captured are normal and therefore manual grading of each image is a cumbersome task requiring an immense number of human graders.

Convolutional neural networks (CNN), a popular deep learning model, have recently been applied to common ophthalmic diseases,[5-7] including AMD,[7-10] with promising results of disease identification from fundus photos. Despite this, most

previously reported systems adopt conservative definitions for referable AMD (e.g. ≥intermediate AMD), which, given the lack of treatment options for atrophic and earlier stages of the disease, may create strain on eye care systems in low resource countries (e.g., China, India, most developing countries). Furthermore, these systems have rarely been validated amongst fundus image datasets that closely resemble real-world screening, where image quality varies considerably with different imaging protocols and retinal pigmentation differs substantially amongst ethnicities. These factors constitute a notable source of potential error for deep learning algorithms (DLA), and must therefore be considered when planning robust evaluations.

The objective of the present study is to describe the development and validation of a DLA for the detection of neovascular AMD using a dataset of over 50,000 retinal photographs, collected from a range of retinal camera models and clinical settings in China. Additionally, we evaluated the accuracy and efficiency of the DLA in a large (>80,000 images), external dataset of non-mydriatic images from a predominantly Caucasian ethnicity.

2. METHODS

This study was approved by the Institutional Review Boards of the Zhongshan Ophthalmic Center, China (2017KYPJ049) the Royal Victorian Eye and Ear Hospital Human Research Ethics Committee (HREC-14/1199H) and conducted in accordance with the Declaration of Helsinki as revised in 2013.

2.1 Development of the DLA

The DLA was developed using 56,113 deidentified, original color fundus photographs acquired from a web-based, cloud sourcing platform (http://www.labelme.org,

Guangzhou, China). These macular and disc centered images were acquired from clinic-based settings and contributed by a total of thirty-six ophthalmology departments, optometry clinics and screening settings in China. All retinal photographs in the training dataset were captured with a variety of common conventional desktop retinal cameras (e.g. Topcon, Canon, Heidelberg and Digital Retinography System), using a variety of imaging protocols. Twenty-one ophthalmologists were enrolled as graders only after meeting a high level of agreement (unweighted Kappa ≥0.70) with an experienced ophthalmologist in a test set of 60 images (20 images of normal fundi, 20 early to late dry AMD and 20 neovascular AMD).

Retinal photographs were graded between October 2016 and March 2017. To ensure an accurate diagnosis of AMD, a multistep method was undertaken. First, images from the total dataset (n=56,113) were randomly assigned to a single ophthalmologist for grading. Once grading was complete, the image was assigned to a separate ophthalmologist until three consistent grading outcomes were achieved for a given image. At this time, a consensus grading outcome was given to the image. Graders were masked to the previous image grading outcomes and a given image could only be assigned to a specific grader once. AMD was graded according to the Beckman clinical classification system that has been described in detail elsewhere.[11] In brief, patients were categorized as no AMD, early, intermediate AMD or late AMD, atrophic or neovascular. An image was defined as 'poor quality' if the vessels within the macular region could not be identified or ≥50% of the macular region was obscured. Images that did not include the macular region were classified as poor field definition. Figure 1 and Table 1 describe the process of image grading and the classification of AMD using the online LabelMe platform.

Figure 1: Image grading process of internal validation dataset. AMD= age related

macular degeneration

Table 1: Classification for age-related macular degeneration on the online system

Classification	Presence of clinical features
Absent	Does not meet any of the following criteria
Early / Intermediate	Any criterion of the following: Drusen > 63 um RPE abnormalities (i.e. hyperpigmentation or depigmentation)
Late dry	Any geographic atrophy
Late wet	Any criterion of the following: Serous detachment of the sensory retina or RPE Sub-RPE retinal hemorrhage Sub-retinal/sub-RPE fibrovascular proliferation Disciform scar
Poor quality	Any criterion of the following: Vessels within the macular area cannot be identified ≥ 50% of the area is obscured
Poor field definition	Not macular-centered photographs

** Lesions were assessed in the circle area within 2-disc diameters of fovea. RPE= retinal pigment epithelium.*

The deep learning classification approach used to train our DLA from image pre-processing to AMD classification is shown in Figure 2. Firstly, several automated pre-processing steps were performed for normalization to control for variations in image size and resolution. This included; (1) applying local space average color for color constancy; (2) downsizing images to a resolution of 299×299 pixels; and (3) online data augmentation to enlarge heterogeneity but keep the prognostic features in the image by a random horizontal shift of 0~3 pixels and 90°, 180° or 270° random rotation. The study included three deep learning models, all using inception-v3

architecture. This included networks for the 1) classification for referable AMD (late-wet AMD), 2) assessment of image quality and 3) assessment of the visual availability of the macular region (field definition).

Figure 2: The deep convolutional neural network used in this study. Data stream is from left to right. A fundus photograph is firstly pre-processed by scaling, subtraction of local space average colour, downsizing the image to a 299 ×299 matrix and data augmentation for normalization. The image is sequentially warped into probability distributions over whether referable AMD is present using Inception-v3 convolutional neural network architecture training from scratch on the training dataset and validation dataset.

2.2 Validation datasets

Using identical pre-processing procedures, the performance of the DLA was internally and externally assessed. Using our local LabelMe dataset using an internal hold-out method a total of 5,554 images from the original dataset of 56,113 we used for internal validation. The consensus grading outcome from the ophthalmologists provided the gold standard for which the DLA was compared to. An experienced ophthalmologist (Z.L) classified false positive and negative images into subgroups.

External validation of the DLA was also assessed on an independent external dataset of 86,162 images of 21,777 participants, derived from the Melbourne Collaborative Cohort Study (MCCS).[12] The participants assessed for AMD in the MCCS in 2003-2007 were aged 47-86 years (mean age 65.3 years, 60.1% female), 14% were born in Southern Europe (Greece, Italy or Malta), with the remaining 86% of Anglo-Celtic origin, born in Australia, the United Kingdom or New Zealand.[12] In the MCCS, digital, non-stereoscopic 45-degree fundus photography of the macular and optic disc

were taken of each eye, using a Canon CR6-45NM non-mydriatic retinal camera with a digital Canon (D60) camera back (Canon Inc., Kanagawa, Japan).[13,14] All images were non mydriatic. Experienced professional graders (K.Z.A and G.M) from the Centre for Eye Research Australia graded each MCCS image and were masked to the identity and clinical characteristics of study participants. Any uncertain cases were adjudicated by a senior retinal specialist (RG) from Australia. For all participants, a single manual grading outcome based on worse affected eye was provided per participant. Overall, 2694 (12.7%) had early stages of AMD (one or more drusen ≥125 microns in size or one or more drusen with 63-124 microns in size with pigmentary abnormalities in a 600-micron diameter grading grid), 122 (0.6%) had late AMD (geographic atrophy or neovascular AMD).[14] Participants with an ungradable manual grading result in both eyes were excluded from the dataset for the purpose of analysis.

DLA grading of MCCS images was undertaken independently of the research team involved in the development of the software. First, images without the manual grading label were transferred by MCCS investigators to the research team, who subsequently processed DLA grading on three computers operating concurrently, using a custom DLA software that allows automated classification consecutively on a set of images. Following this, DLA grading outcomes were transferred to MCCS investigators for preliminary evaluation and identification of discordant cases. Lastly, all discordant cases were assessed by experienced ophthalmologists (X.X.Y and L.R) and any cases identified as suspects for neovascular AMD were presented at a consensus meeting for adjudication by the senior ophthalmologists (R.G and M.H). Given, in the majority of cases, multiple images were available per eye, we adopted the following logic to consolidate a single automated grading result for the right and left eye; 1) positive late-wet AMD = any image for a given eye was found to be positive on automated grading; 2) negative late-wet AMD = no image was found to be positive on automated grading

and at least one image for a given eye was found to be negative; and 3) ungradable = all images for a given eye were ungradable on automated grading. Referable AMD was defined as neovascular AMD and/or ungradable outcome in one or both eyes. The inclusion of ungradable cases as referable is in line with previous reports[7,15] and more closely resembles real-word circumstances, where both positive and ungradable cases go on to manual verification.

2.3 Convolutional Neural Network (CNN) Visualization

To visualize the learning procedure of our networks, we applied an Adaptive Kernel Visualization technique. In brief, this involved applying a sliding window size of 28x28 pixels, with stride of 3 pixels, to crop images into smaller sub-images and produce a (544-28)/3*(544-28)/3) = 172*172 feature map. A random sample of 100 true positive images from the internal validation dataset were selected and utilized as inputs for the trained neovascular AMD deep-learning models. The threshold was set at 0.5 for the late-wet AMD model, meaning that discriminative image regions were highlighted if the classification possibility output of being diagnosed was greater than 50%.

2.4 Statistical Analysis

The sensitivity, specificity, accuracy and area under the curve (AUC) of the DLA in detecting neovascular AMD was performed compared to the reference standard (local validation = retinal specialist; external validation = professional graders) for each participant. The 95% confidence intervals (CIs) were also calculated. STATA version 14.0 (College Station, Texas, USA) was used for all statistical analyses in this study.

3. RESULTS

Each image in the training and internal validation dataset was graded between 3 and 10 times before consensus agreement was reached, with a mean agreement rate of 86.5% (95% CI, 84.5%-88.5%) for the 21 ophthalmologists. Each ophthalmologist graded between 397 and 33,513 (median, 4,135) fundus photographs, with 12 ophthalmologists individually grading more than 5,000 fundus photographs. This considerable variation between the number of images graded existed because ophthalmologists volunteered their time to perform retinal image grading. Of the total 56,113 images, 7,723 (13.8%) were labelled as poor quality and 17,143 (30.5%) labelled as poor field definition (i.e. macular region not in view), leaving 31,247 images with a conclusive AMD classification. Using a simple random sampling method, a total of 27,397 images were assigned to the training dataset and the remaining 3,850 images were held-out for internal validation. A subset of 18,704 images from the entire dataset were used to develop the network for image quality and field definition. Amongst the 27,397 images in the training dataset, 22,553 (82.3%) had no AMD, 1,338 (4.9%) had early or intermediate AMD, 72 (0.3%) had atrophic AMD and 3,434 (12.5%) had neovascular AMD. Investigators purposefully oversampled neovascular AMD cases through targeted image collection in patients undergoing fundus fluorescein angiography in Chinese tertiary hospitals. Table 2 summarises the characteristics of fundus photographs in the training and validation datasets.

	Training set	Validation sample

Table 2: Characteristics of fundus photographs in the training set and randomized validation sample

No AMD	22,553 (82.3)	2,953 (76.7)
Early or intermediate AMD	1,338 (4.9)	180 (4.6)
Late dry AMD	72 (0.3)	11 (0.3)
Late wet AMD	3,434 (12.5)	706 (18.4)
Total	27,397 (100)	3,850 (100)

**AMD= age-related macular degeneration; Data are presented as n (%) unless otherwise indicated.*

3.1 Internal hold-out-validation

In the internal hold out validation dataset (reference to ophthalmologist standard), the AUC, sensitivity, specificity and accuracy of the DLA for neovascular AMD was 0.995 (95% CI, 0.993~0.997), 98.0%, 94.0% and 94.7% respectively (Figure 3). The AUC, sensitivity, specificity and accuracy for image quality and field definition was 0.995 (95% CI, 0.992~0.997), 96.7%, 96.4% 96.5%, respectively. Typical examples of the visualization maps for true positive cases are shown in Figure 4.

Figure 3: The performance of this deep learning algorithm in detection age related macular disease. The area under receiver operating characteristic curve (AUC) of 0.995 (95% confidence interval [CI], 0.993-0.997) was obtained

Figure 4: AMD True Positive. Image A1, B1 & C1 show original images without heat-map. A2 Shows heatmap predominately visualizing the macular, temporal side of the optic nerve head (ONH) and retinal vessels (superiorly). B2 Shows heatmap visualizing the macular region and C2 heatmap shows an area of atrophy superior to the central macular and superior vessels predominantly being visualized. AMD=age-related macular degeneration.

The most common clinical features of false negative cases (n=14) included serous

detachment of the sensory retina or retinal pigment epithelium (RPE) (n=8, 57.2%), followed by sub-retinal/sub-RPE fibrovascular proliferation (n=3, 21.4%). An analysis of false positive cases revealed that 145 (76.3%) had other eye disorders such as diabetic retinopathy (n=38, 20%) and myopic maculopathy (n=22, 11.5%). Normal fundus photographs without artefacts (n=34, 17.9%) and those with artefacts (n=11, 5.8%) made up the remaining false positive cases. Examples of the typical false negative and false positive images can be found in Figure 5 and 6.

Figure 5: Typical cases of false-negative findings in the detection age related macular disease. A, Serous detachment of the sensory retina or RPE; B, sub-retinal/sub-RPE fibrovascular proliferation; C, sub-RPE retinal hemorrhage with serous detachment of the sensory retina or RPE. RPE= retinal pigment epithelium.

Figure 6: Typical cases of false-positive findings. A, Diabetic retinopathy; B, myopic maculopathy; C, early or intermediate AMD; D, choroiditis; E, normal retina; F, normal image with artifacts. AMD= age-related macular degeneration.

3.2 External validation

The external dataset contained 86,202 retinal images of 21,777 participants from the MCCS. Of these, there were 335 (1.54%) participants who had missing or ungradable manual grading outcomes in both eyes that were excluded from analysis. Amongst the remaining 21,327 participants included in the external validation dataset 48 (0.2%) had late-wet AMD.

When adopting the aforementioned definition for referable AMD (i.e. neovascular AMD and/or ungradable outcome in one or both eyes), the AUC, sensitivity and specificity of the DLA in the external validation dataset was 0.967, 100% and 92.6%,

respectively. This consisted of 43 true positives, 322 false positive cases and 1281 ungradable cases in one or both eyes. In total, 5 neovascular AMD cases were labelled as ungradable in DLA analysis. Although not in line with real-world practices, if we were to class these 5 cases as false negatives, the adjusted sensitivity metric would reduce from 100% to 89.6%. Of note, adjudication of discordant manual and DLA grading outcomes via consensus of senior ophthalmologists, revealed that the DLA correctly classified 4 cases as negative that had originally been labelled as positive for neovascular AMD by MCCS manual graders This misclassification was confirmed in the adjudication consensus meeting attended by senior ophthalmologists.

An analysis of false positives revealed that 62.7% (n=202/322) of cases displayed signs of other macular pathology, of which the most common findings were early AMD (n=84/322, 26.1%), atrophic AMD (n=24/322, 7.5%), myopic maculopathy (n=16/322, 5.0%), retinal scarring or irregular macular lesions (n=16/322, 5%) and vitreous opacity (n=5/322, 1.6%). The remaining 37.3% (n=120/322) of false positive cases had no abnormal macular findings, among them, 46.7% (n=56/120) of these displaying image artifacts.

In total, it took about 48 hours (e.g. 6 working days of 8 hours) to complete DLA grading using three computers operating concurrently. To assess repeatability and reliability of the DLA, automated grading was repeated on a random sample of 2000 images from the MCCS dataset. In this subset evaluation, 100% grading consistency was observed.

4. DISCUSSION

This paper describes the development of a DLA for the detection of neovascular AMD based on a large specialist annotated dataset of 56,113 images collected among a

Chinese population. Amongst an independent, local validation dataset of 5,554 images (including 18.4% neovascular AMD cases) derived from a range of providers and camera models, our DLA achieved robust performance for neovascular AMD (AUC = 0.995). Furthermore, the DLA showed excellent efficiency and diagnostic performance (AUC = 0.967) in a large (>80,000 images), external dataset of non-mydriatic images from participants of Southern European and Anglo-Celtic origin.

Over the past decade, automated techniques for the assessment of AMD, via feature extraction from small retinal image datasets (<1,000),[16-18] have been reported with variable accuracy. More recent reports provide novel data on the accuracy of deep learning systems for the detection of AMD.[7-9,19-21] The majority of these studies have utilized the Age-Related Eye Disease Study (AREDS) participants to develop and test the accuracy of their DLAs.[8,9,19-21] The AREDS dataset is useful for training and internally validating DLAs as the images are of high-quality (pupillary dilation ≥5 mm), are collected within research settings and have already been gold standard graded by experts. However, AREDS participants were excluded at recruitment if they had sight threatening disease unrelated. It has been suggested that many CNN's perform sub optimally when tested on unrelated datasets because of technical differences including camera setup illumination and inclusion criteria.[22] In comparison, our DLA was trained using images from multiple clinical settings using various retinal cameras and imaging protocols, and was able to achieve excellent diagnostic accuracy.

Performing external validation is therefore essential given that image quality imperfections are common in real-world screening settings and various camera models are used. The study by Ting et al. (2017)[7] did not perform an external validation for their referable AMD DLA and of those studies that used the AREDS dataset, only Grassman et al. (2018)[19] performed an external validation, albeit on a small (5,555) independent dataset. For the external validation, the algorithm was able to detect 84.2%

of all fundus images with signs of early or late AMD. However, this was only achieved when those aged under 55 years (due to visualization of the macular reflex causing false positives) and images with pathology not related to AMD were excluded. In the current study, only those with missing or ungradable manual grading outcomes in both eyes were excluded from the analysis making the dataset more generalizable to real-world screening settings.

Direct comparison between recent reports relating to AMD detection using DLAs is made difficult due to differing classification criteria. Burlina et al. (2018),[21] Grassman et al (2018)[19] and Burlina et al. (2017)[8] utilised multi step approaches using classifications developed for AREDS. For example, Grassman et al.[19] defined 13 classes, 1 indicating little or no AMD, grades 2-9 representing changes associated with early or intermediate AMD and grades 10-12 covering late-stage AMD such as geographic atrophy and neovascular AMD. Greater numbers of classification groups led to lower kappa scores in the study by Burlina et al. (2018)[21] (0.77 for 4-step approach and 0.74 in the 9-step approach) and an overall accuracy of 63% in the study by Grassman et al. (2018).[19] Similarly, this was shown by Burlina et al. (2017)[8] who obtained accuracy values of 79.4%, 81.5% and 93.4% for 4-class, 3-class and 2-class classifications respectively. Rather than using a multi-step approach, Peng et al (2018)[20] detected individual AMD risk factors including drusen, pigmentary changes and late AMD. Although they used late AMD as a classifier, this included geographic atrophy which our DLA does not classify as referable AMD. A binary classification was used to determine referable AMD (≥intermediate AMD) in the investigations by Ting et al. (2017)[7] and Burlina et al. (2017)[9] who both reported AUC's between 0.93 and 0.96. Whilst these findings are similar to the current study, they are not directly comparable as only those classified as neovascular AMD we considered referable. The choice to only include neovascular AMD as referable was due to the lack of effective treatment options for

atrophic and earlier forms of AMD. Therefore, if deployed in low recourse settings, previously reported DLAs may increase the strain on eye care resources by over-referring many cases.

Taking the current literature into consideration, we developed a DLA to detect neovascular AMD and robustly evaluated its performance in two datasets of images that were taken under varying imaging protocols and across different ethnicities. In order to exhaust all possible variations of neovascular AMD phenotypes, we adopted a training dataset that included a large number (n=3,434) of neovascular AMD cases collected from over 30 clinical settings using different fundus camera and imaging protocols. In the large external validation dataset, the DLA demonstrated reliable automated image analysis under a non-mydriatic retinal protocol, achieving robust diagnostic accuracy (AUC=0.967) and an ungradable image rate of only 5.1%. We hypothesize that the ungradable image rate may be lower if the DLA was deployed within a prospective screening setting, given the software has built in automated classifiers for image quality and field definition that would prompt real-time image re-capture. Our DLA was also successfully validated amongst two ethnic groups (Caucasian AUC = 0.967; Chinese AUC = 0.995) with distinct retinal pigmentation. While we acknowledge that further evaluation is warranted amongst darker skinned ethnic groups, these results provide evidence that the performance of our DLA is likely generalizable to a large number of populations globally.

An important potential benefit of deep learning technology relates to the ability to grade retinal images at an extraordinary speed and scale. In the present study, it took 48 hours to complete DLA grading of the 86,162 images from the MCCS dataset using three computers operating concurrently. In comparison, experienced human graders took several months of dedicated time to complete this task, albeit adopting a greater depth of classification (manual grading = 5-class vs. DLA = 2-class). In total,

the DLA correctly classified 93.8% of participants with healthy fundus images and only 6.8% of participants, including all neovascular AMD cases, were classified as positive (n=43) or ungradable (n=5) by the DLA. This represents a significant potential workload and cost saving if the DLA were deployed as a pre-screening tool in high throughput screening or research settings, where only positive and ungradable cases flagged by automated grading go on to manual verification. In addition, our finding that the DLA correctly classified five cases in the external validation dataset that had been previously misclassified by manual graders as positive for neovascular AMD is noteworthy.

Despite the promising accuracy and efficiency of the DLA, we recognize that one of the major challenges to clinical adoption relates to a major mind-set shift in how clinicians entrust clinical care to machines. To assist in the clinical acceptance,[23] we explored characteristics of misclassification (false negatives and false positives) of the DLA. We found that over 60% of false positive cases in the internal and external validation datasets displayed other macular pathologies, suggesting that most of these false positive cases may have in fact benefited from referral. Among the false negative cases (internal validation dataset only), over half (57.2%) proved to be undetected sensory detachment of neurosensory retina or RPE layer. This is perhaps not a surprising finding given the relative inconspicuous nature of this lesion when compared to other neovascular AMD phenotypes such as fibrovascular or scaring changes. Further training of the DLA with more image examples of this lesion may improve the diagnostic accuracy. In addition to exploring reasons for misclassification, we developed an Adaptive Kernel Visualization method that enables the most discriminative image regions of the DLA to be discerned. This tool offers great potential to alleviate the previously existing tension between accuracy and interpretability of these systems by enabling clinicians to understand important exposure variables in real-time.

The strengths of this study include the utilization of a large training dataset (>50,000 images) of gold standard labelled images, the independent assessment of the DLA on an external dataset, and the robust performance (i.e. accuracy and efficiency) of the DLA on diverse datasets containing images from multiple cameras, under different imaging protocols and across two ethnicities. Some limitations must also be considered. First, there was a relatively small representation of neovascular AMD cases in the external validation dataset (n=48), which may have resulted in an unstable estimate of diagnostic accuracy. Despite this, similar metrics were observed on the internal hold-out validation dataset that contained in excess of 700 neovascular AMD cases. Second, we suspect that a large proportion of DLA screen-positive cases in the internal and external validation datasets would have end-stage lesions (e.g. fibrosis and atrophy), and therefore be ineligible for treatment. Identification of new onset neovascular AMD cases or the existence of choroidal neovascularisation for treatment would rely on confirmation via optical coherence tomography (OCT). Lastly, given that it is not considered cost-effective to screen for AMD in the general population,[24] the usefulness of the evaluated DLA in screening settings may be questioned. It is important to note that we have also developed DLAs to detect referable diabetic retinopathy (DR) and glaucomatous optic neuropathy (GON),[6,25-28] which can be deployed concurrently with the DLA described in the present report. Therefore, given the evaluation of AMD is typically included within manual screening programs amongst the diabetic population, we speculate that, at a minimum, the DLA in question could be included in future AI-based DR screening approaches.

In conclusion, this DLA shows robust performance for the detection of neovascular AMD amongst retinal images from a multi-ethnic sample and under different imaging protocols. The results suggest that this system may provide an efficient and cost-effective 'pre-screener' when applied to large research datasets.

Future efforts will focus on investigating the effectiveness of our DLA for neovascular AMD as a screening tool in high risk populations, when combined with our other fundus-image based DLAs for GON and DR, and what impact the introduction OCT imaging has as a second-line screening tool amongst screen-positive patients.

Acknowledgements

The authors would like to thank the ophthalmologists who volunteered their time to grade fundus images used to train and validate this deep learning algorithm, and the study leads from the Melbourne Collaborative Cohort Study for contributing fundus images for external validation.

REFERENCES

1. Wong WL, Su X, Li X, Cheung CM, Klein R, Cheng CY, Wong TY. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* 2014; **2**: e106-16.
2. Ferris FL, 3rd, Fine SL, Hyman L. Age-related macular degeneration and blindness due to neovascular maculopathy. *Archives of ophthalmology (Chicago, Ill : 1960)* 1984; **102**: 1640-2.
3. Lim LS, Mitchell P, Seddon JM, Holz FG, Wong TY. Age-related macular degeneration. *Lancet* 2012; **379**: 1728-38.
4. Wong TY, Liew G, Mitchell P. Clinical update: new treatments for age-related macular degeneration. *Lancet* 2007; **370**: 204-6.
5. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *Jama* 2016; **316**: 2402-10.
6. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* 2018; **125**: 1199-206.
7. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EYM, Sabanayagam C, Baskaran M, Ibrahim F, Tan NC, Finkelstein EA, Lamoureux EL, Wong IY, Bressler NM, Sivaprasad S, Varma R, Jonas JB, He MG, Cheng CY, Cheung GCM, Aung T, Hsu W, Lee ML, Wong TY. 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**: 2211-23.

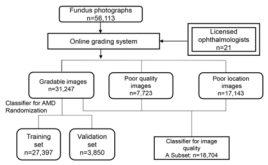
8. Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. *Computers in biology and medicine* 2017; **82**: 80-6.
9. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. *JAMA Ophthalmol* 2017; **135**: 1170-6.
10. Grassmann F, Mengelkamp J, Brandl C, Harsch S, Zimmermann ME, Linkohr B, Peters A, Heid IM, Palm C, Weber BHF. A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography. *Ophthalmology* 2018.
11. Ferris III FL, Wilkinson C, Bird A, Chakravarthy U, Chew E, Csaky K, Sadda SR, Committee BIfMRC. Clinical classification of age-related macular degeneration. *Ophthalmology* 2013; **120**: 844-51.
12. Milne R, Fletcher A, MacInnis R, Hodge A, Hopkins A, Bassett J, Bruinsma F, Lynch B, Dugué P, Jayasekara H. Cohort profile: the melbourne collaborative cohort study (Health 2020). *International journal of epidemiology* 2017; **46**: 1757-i.
13. Aung KZ, Robman L, English DR, Giles GG, Guymer RH. Non-mydriatic digital macular photography: how good is the second eye photograph? *Ophthalmic Epidemiol* 2009; **16**: 254-61.
14. Robman LD, Islam FM, Chong EW, Adams MK, Simpson JA, Aung KZ, Makeyeva GA, Hopper JL, English DR, Giles GG. Age-related macular degeneration in ethnically diverse Australia: Melbourne Collaborative Cohort Study. *Ophthalmic*

Epidemiol 2015; **22**: 75-84.

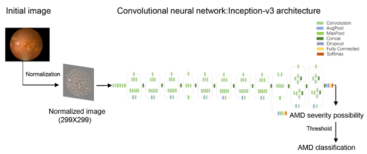
15. Tufail A, Rudisill C, Egan C, Kapetanakis VV, Salas-Vega S, Owen CG, Lee A, Louw V, Anderson J, Liew G, Bolter L, Srinivas S, Nittala M, Sadda S, Taylor P, Rudnicka AR. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology* 2017; **124**: 343-51.
16. Agurto C, Barriga ES, Murray V, Nemeth S, Crammer R, Bauman W, Zamora G, Pattichis MS, Soliz P. Automatic detection of diabetic retinopathy and age-related macular degeneration in digital fundus images. *Invest Ophthalmol Vis Sci* 2011; **52**: 5862-71.
17. Mookiah MR, Acharya UR, Koh JE, Chandran V, Chua CK, Tan JH, Lim CM, Ng EY, Noronha K, Tong L, Laude A. Automated diagnosis of Age-related Macular Degeneration using greyscale features from digital fundus images. *Computers in biology and medicine* 2014; **53**: 55-64.
18. Zheng Y, Hijazi MH, Coenen F. Automated "disease/no disease" grading of age-related macular degeneration by an image mining approach. *Invest Ophthalmol Vis Sci* 2012; **53**: 8310-8.
19. Grassmann F, Mengelkamp J, Brandl C, Harsch S, Zimmermann ME, Linkohr B, Peters A, Heid IM, Palm C, Weber BHJO. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. 2018; **125**: 1410-20.
20. Peng Y, Dharssi S, Chen Q, Keenan TD, Agrón E, Wong WT, Chew EY, Lu ZJO. DeepSeeNet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. 2019; **126**: 565-75.
21. Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NMJJo. Use of

deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. 2018; **136**: 1359-66.

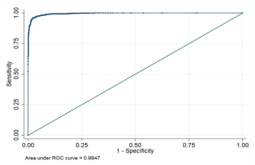
22. Castelvecchi DJNN. Can we open the black box of AI? 2016; **538**: 20.
23. Wong TY, Bressler NM. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *Jama* 2016; **316**: 2366-7.
24. Tamura H, Goto R, Akune Y, Hiratsuka Y, Hiragi S, Yamada M. The Clinical Effectiveness and Cost-Effectiveness of Screening for Age-Related Macular Degeneration in Japan: A Markov Modeling Study. *PLoS One* 2015; **10**: e0133628.
25. Keel S, Lee PY, Scheetz J, Li Z, Kotowicz MA, MacIsaac RJ, He M. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Sci Rep* 2018; **8**: 4330.
26. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, Lee PY, Shaw J, Ting D, Wong T, Taylor H, Chang R, He M. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. *Diabetes care* 2018.
27. Keel S, Lee P, Scheetz J, Le Z, Kotowicz M, MacIsaac R, He M. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. *Scientific Reports* 2018; **Accepted 22nd Feb 2018**.
28. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* 2018.



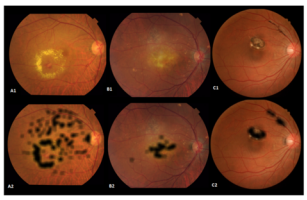
CEO_13575_CEO-19-01-0080 figure 1.tiff



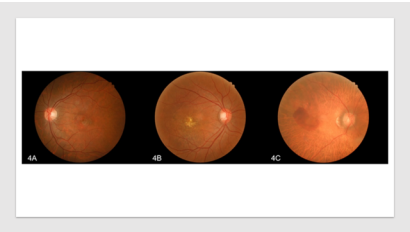
CEO_13575_CEO-19-01-0080 figure 2.tiff



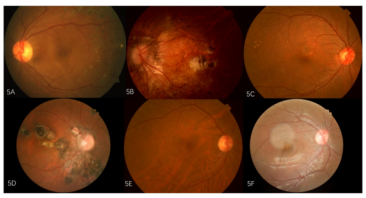
CEO_13575_CEO-19-01-0080 figure 3.tiff



CEO_13575_CEO-19-01-0080 figure 4.tiff



CEO_13575_CEO-19-01-0080 figure 5.tiff



CEO_13575_CEO-19-01-0080 figure 6.tiff