Artificial intelligence and deep learning in ophthalmology

Daniel Shu Wei Ting,[1] Louis R Pasquale,[2] Lily Peng,[3] John Peter Campbell,[4] Aaron Y Lee,[5] Rajiv Raman,[6] Gavin Siew Wei Tan,[1] Leopold Schmetterer,[1,7,8,9] Pearse A Keane,[10] Tien Yin Wong[1]

For numbered affiliations see end of article.

Correspondence to
Dr Daniel Shu Wei Ting, Assistant Professor in Ophthalmology, Duke-NUS Medical SchoolSingapore National Eye Center, Singapore 168751, Singapore; daniel.ting.s.w@singhealth.com.sg

Received 4 September 2018
Revised 17 September 2018
Accepted 23 September 2018

To cite: Ting DSW, Pasquale LR, Peng L, *et al*. *Br J Ophthalmol* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjophthalmol-2018-313173

ABSTRACT

Artificial intelligence (AI) based on deep learning (DL) has sparked tremendous global interest in recent years. DL has been widely adopted in image recognition, speech recognition and natural language processing, but is only beginning to impact on healthcare. In ophthalmology, DL has been applied to fundus photographs, optical coherence tomography and visual fields, achieving robust classification performance in the detection of diabetic retinopathy and retinopathy of prematurity, the glaucoma-like disc, macular oedema and age-related macular degeneration. DL in ocular imaging may be used in conjunction with telemedicine as a possible solution to screen, diagnose and monitor major eye diseases for patients in primary care and community settings. Nonetheless, there are also potential challenges with DL application in ophthalmology, including clinical and technical challenges, explainability of the algorithm results, medicolegal issues, and physician and patient acceptance of the AI 'black-box' algorithms. DL could potentially revolutionise how ophthalmology is practised in the future. This review provides a summary of the state-of-the-art DL systems described for ophthalmic applications, potential challenges in clinical deployment and the path forward.

INTRODUCTION

Artificial intelligence (AI) is the fourth industrial revolution in mankind's history.[1] Deep learning (DL) is a class of state-of-the-art machine learning techniques that has sparked tremendous global interest in the last few years.[2] DL uses representation-learning methods with multiple levels of abstraction to process input data without the need for manual feature engineering, automatically recognising the intricate structures in high-dimensional data through projection onto a lower dimensional manifold.[2] Compared with conventional techniques, DL has been shown to achieve significantly higher accuracies in many domains, including natural language processing, computer vision[3–5] and voice recognition.[6]

In medicine and healthcare, DL has been primarily applied to medical imaging analysis, in which DL systems have shown robust diagnostic performance in detecting various medical conditions, including tuberculosis from chest X-rays,[7 8] malignant melanoma on skin photographs[9] and lymph node metastases secondary to breast cancer from tissue sections.[10] DL has similarly been applied to ocular imaging, principally fundus photographs and optical coherence tomography (OCT). Major ophthalmic diseases which DL techniques have been used for include diabetic retinopathy (DR),[11–15] glaucoma,[11 16] age-related macular degeneration (AMD)[11 17 18] and retinopathy of prematurity (ROP).[19] DL has also been applied to estimate refractive error and cardiovascular risk factors (eg, age, blood pressure, smoking status and body mass index).[20 21]

A primary benefit of DL in ophthalmology could be in screening, such as for DR and ROP, for which well-established guidelines exist. Other conditions, such as glaucoma and AMD, may also require screening and long-term follow-up. However, screening requires tremendous manpower and financial resources from healthcare systems, in both developed countries and in low-income and middle-income countries. The use of DL, coupled with telemedicine, may be a long-term solution to screen and monitor patients within primary eye care settings. This review summarises new DL systems for ophthalmology applications, potential challenges in clinical deployment and potential paths forward.

DL APPLICATIONS IN OPHTHALMOLOGY

Diabetic retinopathy

Globally, 600 million people will have diabetes by 2040, with a third having DR.[22] A pooled analysis of 22 896 people with diabetes from 35 population-based studies in the USA, Australia, Europe and Asia (between 1980 and 2008) showed that the overall prevalence of any DR (in type 1 and type 2 diabetes) was 34.6%, with 7% vision-threatening diabetic retinopathy.[22] Screening for DR, coupled with timely referral and treatment, is a universally accepted strategy for blindness prevention. DR screening can be performed by different healthcare professionals, including ophthalmologists, optometrists, general practitioners, screening technicians and clinical photographers. The screening methods comprise direct ophthalmoscopy,[23] dilated slit lamp biomicroscopy with a hand-held lens (90 D or 78 D),[24] mydriatic or non-mydriatic retinal photography,[23] teleretinal screening,[25] and retinal video recording.[26] Nonetheless, DR screening programmes are challenged by issues related to implementation, availability of human assessors and long-term financial sustainability.[27]

Over the past few years, DL has revolutionised the diagnostic performance in detecting DR.[2] Using

Table 1 Summary table for the different DL systems in the detection of referable diabetic retinopathy, glaucoma suspect, age-related macular degeneration and retinopathy of prematurity using fundus photographs

DL systems	Year	Test data sets	Test images (n)	CNN	AUC	Sensitivity (%)	Specificity (%)
Referable diabetic retinopathy							
Abràmoff *et al*[14]	2016	Messidor-2	1748	AlexNet/VGG	0.98	96.80	87.00
Gulshan *et al*[12]	2016	Messidor-2	1748	Inception-V3	0.99	87	98.50
						96.10	93.90
		EyePACS-1	9963		0.991	90.30	98.10
						97.50	93.40
Gargeya and Leng[15]	2017	Kaggle images	75 137	Customised CNN	0.97	NA	NA
		E-Ophtha	463		0.96	NA	NA
		Messidor-2	1748		0.94	NA	NA
Ting *et al*[11]	2017	SiDRP 14–15	71 896	VGG-19	0.936	90.50	91.60
		Guangdong	15 798		0.949	98.70	81.60
		SIMES	3052		0.889	97.10	82.00
		SINDI	4512		0.917	99.3	73.3
		SCES	1936		0.919	100	76.30
		BES	1052		0.929	94.40	88.50
		AFEDS	1968		0.98	98.80	86.50
		RVEEH	2302		0.983	98.90	92.20
		Mexican	1172		0.95	91.80	84.80
		CUHK	1254		0.948	99.3	83.10
		HKU	7706		0.964	100	81.30
Abràmoff *et al*[28]	2018	10 primary care practice sites from the USA	892 patients	Alex/VGG	NA	87.2	90.7
Glaucoma suspect*							
Ting *et al*[11]	2017	SiDRP 14–15	71 896	VGG-19	0.942	96.40	93.20
Li *et al*[16]	2018	Guangdong	48 116		0.986	95.60	92.00
Age-related macular degeneration							
Ting *et al*[11]	2017	SiDRP 14–15	35 948	VGG-19	0.932	93.20	88.70
Burlina *et al*[17]	2017	AREDS	120 656	AlexNet, OverFeat	0.940–0.96	NA	NA
Grassmann *et al*[18]	2018	AREDS	120 656	AlexNet, GoogleNet, VGG, Inception-V3, ResNet, Inception-ResNet-V2	NA	84.20	94.30
Retinopathy of prematurity							
Brown *et al*[19]	2018	i-ROP	100	Inception-V1 and U-Net	NA	100	94

The diagnostic performance is not comparable between the different DL systems given the different data sets used in the individual study.
*Definition of glaucoma suspect: (1) Ting *et al*[11]—vertical cup to disc ratio of 0.8 or greater, and any glaucomatous disc changes; (2) Li *et al*[16]—vertical cup to disc ratio of 0.7 or greater, and any glaucomatous disc changes.
AFEDS, African American Eye Disease Study; AREDS, Age-Related Eye Disease Study; AUC, area under the receiver operating characteristic curve; BES, Beijing Eye Study; CNN, convolutional neural network; CUHK, Chinese University Hong Kong; DL, deep learning; SiDRP 14–15, Singapore Integrated Diabetic Retinopathy Screening Programme; HKU, Hong Kong University; NA, not available; RVEEH, Royal Victorian Eye and Ear Hospital; SCES, Singapore Chinese Eye Study; SIMES, Singapore Malay Eye Study; SINDI, Singapore Indian Eye Study.

this technique, many groups have shown excellent diagnostic performance (table 1).[14] Abràmoff *et al*[14] showed that a DL system was able to achieve an area under the receiver operating characteristic curve (AUC) of 0.980, with sensitivity and specificity of 96.8% and 87.0%, respectively, in the detection of referable DR (defined as moderate non-proliferative DR or worse, including diabetic macular oedema (DMO)) on Messidor-2 data set. Similarly, Gargeya and Leng[15] reported an AUC of 0.97 using cross-validation on the same data set, and 0.94 and 0.95 in two independent test sets (Messidor-2 and E-Ophtha).

More recently, Gulshan and colleagues[12] from Google AI Healthcare reported another DL system with excellent diagnostic performance. The DL system was developed using 128 175 retinal images, graded between 3 and 7 times for DR and DMO by a panel of 54 US licensed ophthalmologists and ophthalmology residents between May and December 2015. The test set consisted of approximately 10 000 images retrieved from two publicly available databases (EyePACS-1 and Messidor-2), graded by at least seven US board-certified ophthalmologists with high intragrader consistency. The AUC was 0.991 and 0.990 for EyePACS-1 and Messidor-2, respectively (table 1).

Although a number of groups have demonstrated good results using DL systems on publicly available data sets, the DL systems were not tested in real-world DR screening programmes. In addition, the generalisability of a DL system to populations of different ethnicities, and retinal images captured using different cameras, still remains uncertain. Ting *et al*[11] reported a clinically acceptable diagnostic performance of a DL system, developed and tested using the Singapore Integrated Diabetic Retinopathy Programme over a 5-year period, and 10 external data sets recruited from 6 different countries, including Singapore, China, Hong Kong, Mexico, USA and Australia. The DL system, developed using the DL architecture VGG-19, was reported to have AUC, sensitivity and specificity of 0.936, 90.5% and 91.6% in detecting referable DR. For vision-threatening DR, the corresponding statistics were 0.958, 100% and 91.1%. The

AUC ranged from 0.889 to 0.983 for the 10 external data sets (n=40 752 images). More recently, the DL system, developed by Abramoff *et al*,[28] has obtained a US Food and Drug Administration approval for the diagnosis of DR. It was evaluated in a prospective, although observational setting, achieving 87.2% sensitivity and 90.7% specificity.[28]

Age-related macular degeneration

AMD is a major cause of vision impairment in the elderly population globally. The Age-Related Eye Disease Study (AREDS) classified AMD stages into none, early, intermediate and late AMD.[29] The American Academy of Ophthalmology recommends that people with intermediate AMD should be at least seen once every 2 years. It is projected that 288 million patients may have some forms of AMD by 2040,[30] with approximately 10% having intermediate AMD or worse.[29] With the ageing population, there is an urgent clinical need to have a robust DL system to screen these patients for further evaluation in tertiary eye care centres.

Ting *et al*[11] reported a clinically acceptable DL system diagnostic performance in detecting referable AMD (table 1). Specifically, the DL system was trained and tested using 108 558 retinal images from 38 189 patients. Fovea-centred images without macula segmentation were used in this study. Given that this was the DR screening population, there were relatively few patients with referable AMD. For the other two studies,[17 18] DL systems were developed using the AREDS data set, with a high number of referable AMD (intermediate AMD or worse). Using a fivefold cross-validation, Burlina *et al*[17] reported a diagnostic accuracy of between 88.4% and 91.6%, with an AUC of between 0.94 and 0.96. Unlike Ting *et al*,[11] the authors presegmented the macula region prior to training and testing, with an 80/20 split between the training and testing in each fold. In terms of the DL architecture, both AlexNet and OverFeat have been used, with AlexNet yielding a better performance. Using the same AREDS data set, Grassmann *et al*[18] reported a sensitivity of 84.2% in the detection of any AMD. In this study, the authors used six convolutional neural networks—AlexNet, GoogleNet, VGG, Inception-V3, ResNet and Inception-ResNet-V2—to train different models. Data augmentation was also used to increase the diversity of data set and to reduce the risk of overfitting. For the AREDS data set, all the photographs were captured as analogue photographs and then digitised later. Whether this affects the DL system's performance remains uncertain. In addition, all three abovementioned studies did not have any results for external validation on the individual DL systems.

DM, choroidal neovascularisation and other macular diseases

OCT has had a transformative effect on the management of macular diseases, specifically neovascular AMD and DMO. OCT also provides a near-microscopic view of the retina in vivo with quick acquisition protocols revealing structural detail that cannot be seen using other ophthalmic examination techniques. Thus, the number of macular OCTs has grown from 4.3 million in 2012 to 6.4 million in 2016 in the US Medicare population alone, and will most likely continue to grow worldwide.[31]

From a DL perspective, macular OCTs possess a number of attractive qualities as a modality for DL. First is the explosive growth in the number of macular OCTs that are routinely collected around the world. This large number of OCTs is required to train DL systems where having many training examples can aid in the convergence of many-layered networks with millions of parameters. Second, macular OCTs have dense three-dimensional structural information that is usually consistently captured. Unlike real-world images or even colour fundus photographs, the field of view of the macula and the foveal fixation is usually consistent from one volume scan to another. This lowers the complexity of the computer vision task significantly and allows networks to reach meaningful performance with smaller data sets. Third, OCTs provide structural detail that is not easily visible using conventional imaging techniques and provide an avenue for uncovering novel biomarkers of the disease.

One of the first applications of DL to macular OCTs was in automated classification of AMD. Approximately 100 000 OCT B-scans were used to train a DL classifier based on VGG-16 to achieve an AUC of 0.97 (table 2).[32] Few studies used a technique known as transfer learning, where a neural network is pretrained on ImageNet and subsequently then trained on OCT B-scans for retinal disease classification.[33–35] Of note, these initial studies involve the use of two-dimensional DL models trained on single OCT B-scans rather than three-dimensional models trained on OCT volumes. This may be a barrier to their potential clinical applicability.

DL has also had a transformative impact in boundary and feature-level segmentation using neural networks that have been developed for semantic segmentation such as the U-Net.[36] Specifically, these networks have been trained to segment intraretinal fluid cysts and subretinal fluid on OCT B-scans.[13 37 38] Deep convolutional networks surpassed traditional methods in the quality of segmentation of retinal anatomical boundaries.[39–41] Also similar approaches were used to segment en-face OCTA images to segment the foveal avascular zone.[42]

More recently, DeepMind and the Moorfields Eye Hospital have combined the power of neural networks for both segmentation and classification tasks using a novel AI framework. In this approach, a segmentation network is first used to delineate a range of 15 different retinal morphological features and OCT acquisition artefacts. The output of this network is then passed to a classification network which makes a referral triage decision from four categories (urgent, semiurgent, routine, observation) and classifies the presence of 10 different OCT pathologies (choroidal neovascularisation (CNV), macular oedema without CNV, drusen, geographic atrophy, epiretinal membrane, vitreomacular traction, full-thickness macular hole, partial thickness macular hole, central serous retinopathy and 'normal').[43] Using this approach, the Moorfields-DeepMind system reports a performance on par with experts for these classification tasks (although in a retrospective setting). Moreover, the generation of an intermediate tissue representation by the first, segmentation network means that the framework can be generalised across OCT systems from multiple different vendors without prohibitive requirements for retraining. In the near term, this DL system will be implemented in an existing real-world clinical pathway—the rapid access 'virtual' clinics that are now widely used for triaging of macular disease in the UK.[44] In the longer term, the system could be used in triaging patients outside the hospital setting, particularly as OCT systems are increasingly being adopted by optometrists in the community.[45]

Glaucoma

The global prevalence of glaucoma for people aged 40–80 is 3.4%, and by the year 2040 it is projected there will be approximately 112 million affected individuals worldwide.[46] Clinicians and patients alike would welcome improvements in disease detection, assessment of progressive structural and functional

Table 2 Summary table for the different DL systems in the detection of retinal diseases using OCT

DL systems	Year	Disease	OCT machines	Test images	CNN	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Lee *et al*[13 32]	2017	Exudative AMD	Spectralis	20 613	VGG-16	0.928	87.60	84.60	91.50
Trader *et al*[33]	2018	Exudative AMD	Spectralis	100	Inception-V3	0.980	100	NA	NA
Kermany *et al*[34]	2018	CNV	Spectralis	1000	Inception-V3				
		DMO							
		Drusen							
		1. Multiclass comparison				0.999	96.50	97.80	97.40
		2. Limited model				0.988	93.40	96.60	94.00
		3. Binary model							
		CNV vs normal				1	100	100	100
		DMO vs normal				0.999	98.20	96.80	99.60
		Drusen vs normal				0.999	99	98	99.20
De Fauw *et al*[43]	2018	Urgent, semiurgent, routine and observation only	Topcon	997 patients	1. Deep segmentation network using U-Net	Urgent referral 0.992	94.5		
		Normal, CNV, macular oedema, FTMH, PTMH, CSR, VMT, GA, drusen, ERM	Spectralis	116 patients	2. Deep classification network using a custom 29 CNN layers with 5 pooling layers	Urgent referral 0.999	96.6		

The diagnostic performance is not comparable between the different DL systems given the different data sets used in the individual study. AUC for specific conditions: CNV 0.993; macular oedema 0.990; normal 0.995; FTMH 1.00; PTMH 0.999; CSR 0.995; VMT 0.980; GA 0.990; drusen 0.967; and ERM 0.966.
AMD, age-related macular degeneration; AUC, area under the receiver operating characteristic curve; CNN, convolutional neural network; CNV, choroidal neovascularisation; CSR, central serous chorioretinopathy; DL, deep learning; DMO, diabetic macular oedema; ERM, epiretinal membrane; FTMH, full-thickness macula hole; GA, geographic atrophy; NA, not available; OCT, optical coherence tomography; PTMH, partial thickness macula hole; VMT, vitreomacular traction.

damage, treatment optimisation so as to prevent visual disability, and accurate long-term prognosis.

Glaucoma is an optic nerve disease categorised by excavation and erosion of the neuroretinal rim that clinically manifests itself by increased optic nerve head (ONH) cupping. Yet, because the ONH area varies by fivefold, there is virtually no cup to disc ratio (CDR) that defines pathological cupping, hampering disease detection.[47] Li *et al*[16] and Ting *et al*[11] trained computer algorithms to detect the glaucoma-like disc, defined as a vertical CDR of 0.7 and 0.8, respectively. Investigators have also applied machine learning methods to distinguish glaucomatous nerve fibre layer damage from normal scans on wide-angle OCTs (9×12 mm).[48] Future opportunities include training a neural network to identify the disc that would be associated with manifest visual field (VF) loss across the spectrum of disc size, as our current treatment strategies are aligned with slowing disease detection. Furthermore, DL could be used to detect progressive structural optic nerve changes in glaucoma.

In glaucoma, retinal ganglion cell axons atrophy in a confined space within the ONH and ophthalmologists typically rely on low dimensional psychophysical data to detect the functional consequences of that damage. The outputs from these tests typically provide reliability parameters, age-matched normative comparisons and summary global indices, but more detailed analysis of this functional data is lacking. Elze *et al*[49] developed an unsupervised computer program to analyse VF that recognises clinically relevant VF loss patterns and assigns a weighting coefficient for each of them (figure 1). This method has proven useful in the detection of early VF loss from glaucoma.[50] Furthermore, a myriad of computer programs to detect VF progression exist, ranging from assessment of global indices over time to point-wise analyses, to sectoral VF analysis; however, these approaches are often not aligned with clinical ground truth nor with one another.[51 52] Yousefi *et al*[53] developed a machine-based algorithm that detected VF progression earlier than these conventional strategies. More machine learning algorithms that provide quantitative information about regional VF progression can be expected in the future.

Although intraocular pressure (IOP)-lowering has been shown to be therapeutically effective in delaying glaucoma progression, some demonstrated that disease progression is still inevitable,[54–56] suggesting that we have not arrived at optimised treatment regimens for the various forms of glaucoma. Kazemian *et al*[57] developed a clinical forecasting tool that uses tonometric and VF data to project disease trajectories at different target IOPs. Further refinement of this tool that integrates other ophthalmic and non-ophthalmic data would be useful to establish target IOPs and the best strategies to achieve them on a case-by-case basis. Finally, it is documented that patients with newly diagnosed glaucoma harbour fears of going blind[58]; perhaps, the use of machine learning that incorporates genome-wide data, lifestyle behaviour and medical history into a forecasting algorithm will allow early prognostication regarding the future risk of requiring invasive surgery or losing functional vision from glaucoma.

As machine learning algorithms are revised, the practising ophthalmologist will have a host of tools available to diagnose glaucoma, detect disease progression and identify optimised treatment strategies using a precision medicine approaches. In an ideal future scenario, they may also have clinical forecasting tools that inform patients as to their overall prognosis and expected clinical course with or without treatment.

Retinopathy of prematurity

ROP is a leading cause of childhood blindness worldwide, with an annual incidence of ROP-related blindness of 32 000 worldwide.[59] The regional epidemiology of the disease varies based on a number of factors, including the number of preterm births, neonatal mortality of preterm children and capacity to monitor exposure to oxygen. ROP screening either directly via ophthalmoscopic examination or telemedical evaluation using digital fundus photography can identify the earliest signs of severe

ROP, and with timely treatment can prevent most cases of blindness from ROP.[60 61] Due to the high number of preterm births, reductions in neonatal mortality, and limited capacity for oxygen monitoring and ROP screening, the highest burden of blinding ROP today is in low-income and middle-income countries.[62]

There are two main barriers to effective implementation of ROP screening: (1) the diagnosis of ROP is subjective, with significant interexaminer variability in the diagnosis leading to inconsistent application of evidence-based interventions[63]; and (2) there are too few trained examiners in many regions of the world.[64] Telemedicine has emerged as a viable model to address the latter problem, at least in regions where the cost of a fundus camera is not prohibitive, by allowing a single physician to virtually examine infants over a large geographical area. However, telemedicine itself does not solve the subjectivity problem in ROP diagnosis. Indeed, the acute-phase ROP study found nearly 25% of telemedicine examinations by trained graders required adjudication because the graders disagreed on one of three criteria for clinically significant ROP.[65]

There have been a number of early attempts to use DL for automated diagnosis of ROP,[19 66] which could potentially address both implementation barriers for ROP screening. Most recently, Brown *et al*[19] reported the results of a fully automated DL system that could diagnose plus disease, the most important feature of severe ROP, with an AUC of 0.98 compared with a consensus reference standard diagnosis combining image-based diagnosis and ophthalmoscopy (table 1). When directly compared with the eight international experts in ROP diagnosis, the i-ROP DL system agreed with the consensus diagnosis more frequently than six out of eight experts. Subsequent work found that the i-ROP DL system could also produce a severity score for ROP that demonstrated promise for objective monitoring of disease progression, regression and response to treatment.[67] When compared with the same set of 100 images ranked in order of disease severity by experts, the algorithm had 100% sensitivity an 94% specificity in the detection of pre-plus or worse disease.

Potential challenges

Despite the high level of accuracy of the AI-based models in many of the diseases in ophthalmology, there are still many clinical and technical challenges for clinical implementation and real-time deployment of these models in clinical practice (table 3). These challenges could arise in different stages in both the research and clinical settings. First, many of the studies have used training data sets from relatively homogeneous populations.[12 14 15] AI training and testing using retinal images is often subject to numerous variabilities, including width of field, field

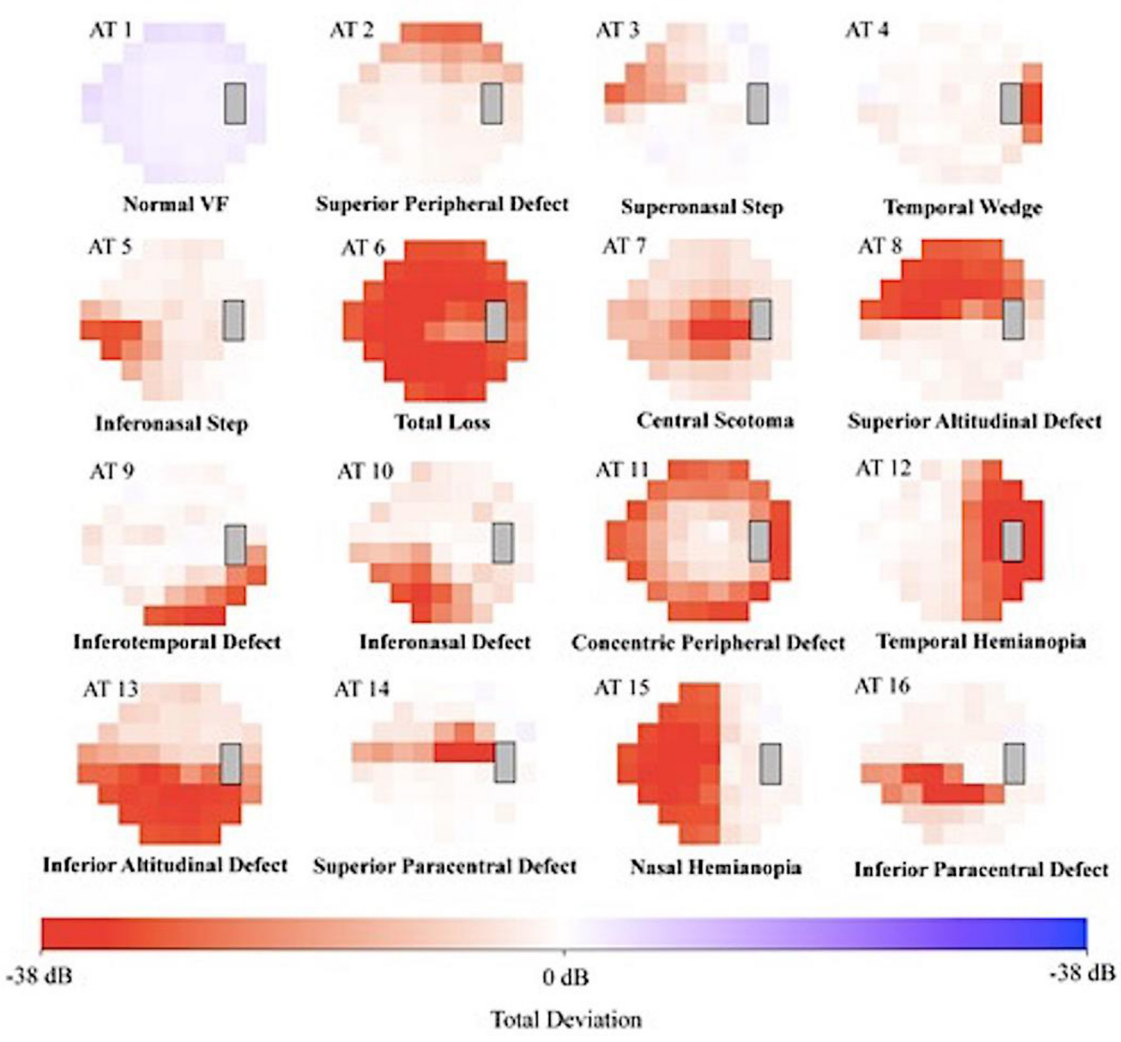

Figure 1 Archetype analysis with 16 visual field (VF) archetypes (ATs) that were derived from an unsupervised computer algorithm described by Elze *et al*.[49]

Table 3 The clinical and technical challenges in building and deploying deep learning (DL) techniques from 'bench to bedside'

Steps	Potential challenges
1. Identification of training data sets	1. Patients' consent and confidentiality issues. 2. Varying standards and regulations between the different institutional review boards. 3. Small training data sets for rare disease (eg, ocular tumours) or common diseases that are not captured in routine (eg, cataracts).
2. Validation and testing data sets	1. Lack of sample size—not sufficiently powered. 2. Lack of generalisability—not tested widely in different populations or on data collected from different devices.
3. Explainability of the results	1. Demonstration of the regions 'deemed' abnormal by DL. 2. Methods to generate heat maps—occlusion tests, class activation, integrated gradient method, soft attention map and so on.
4. Clinical deployment of DL Systems	1. Recommendation of the potential clinical deployment sites. 2. Application of regulatory approval from health authorities (eg, US Food and Drug Administration, Europe CE marking and so on). 3. Conducting prospective clinical trials. 4. Medical rebate scheme and medicolegal requirement. 5. Ethical challenges.

of view, image magnification, image quality and participant ethnicities. Diversifying the data set, in terms of ethnicities, and image-capture hardware could help to address this challenge.[11]

Another challenge in the development of AI models in ophthalmology has been the limited availability of large amounts of data for both the rare diseases (eg, ocular tumours) and for common

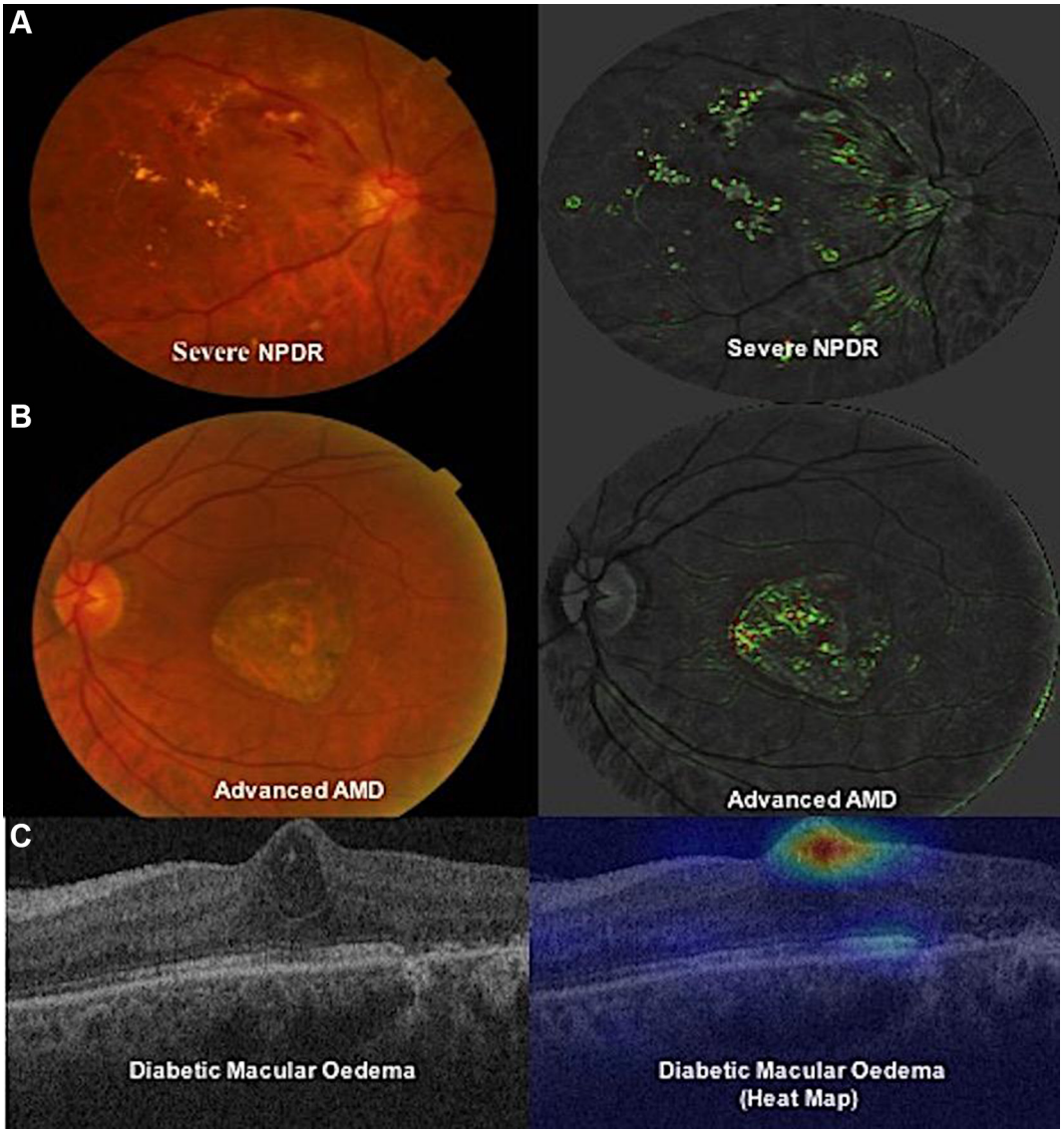

Figure 2 Some examples of heat maps showing the abnormal areas in the retina. (A) Severe non-proliferative diabetic retinopathy (NPDR); (B) geographic atrophy in advanced age-related macular degeneration (AMD) on fundus photographs[11]; and (C) diabetic macular oedema on optical coherence tomography.

diseases which are not imaged routinely in clinical practice such as cataracts. Furthermore, there are diseases such as glaucoma and ROP where there will be disagreement and interobserver variability in the definition of the disease phenotype. The algorithm learns from what they are presented with. The software is unlikely to produce accurate outcomes if the training set of images given to the AI tool is too small or not representative of real patient populations. More evidence on ways of getting high-quality ground-truth labels is required for different imaging tools. Krause *et al*[68] reported that adjudication grades by retina specialists were a more rigorous reference standard, especially to detect artefacts and missed microaneurysms in DR, than a majority decision and improved the algorithm performance.

Second, many AI groups have reported robust diagnostic performance for their DL systems, although some papers did not show how the power calculation was performed for the independent data sets. A power calculation should take the following into consideration: the prevalence of the disease, type 1 and 2 errors, CIs, desired precision and so on. It is important to first preset the desired operating threshold on the training set, followed by analysis of performance metrics such as sensitivity and specificity on the test set to assess calibration of the algorithm.

Third, large-scale adoption of AI in healthcare is still not on the horizon as clinicians and patients are still concerned about AI and DL being 'black-boxes'. In healthcare, it is not only the quantitative algorithmic performance, but the underlying features through which the algorithm classifies disease which is important to improve physician acceptance. Generating heat maps highlighting the regions of influence on the image which contributed to the algorithm conclusion may be a first step (figure 2), although such maps are often challenging to interpret (what does it mean if a map highlights an area of vitreous on an OCT of a patient with drusen?).[69] They may also struggle to deal with negations (what would it mean to highlight the most important part of an ophthalmic image that demonstrates that there is no disease present?).[70 71] An alternative approach has been used for the DL system developed by the Moorfields Eye Hospital and DeepMind—in this system, the generation of an intermediate tissue representation by a segmentation network is used to highlight for the clinician (and quantify) the relevant

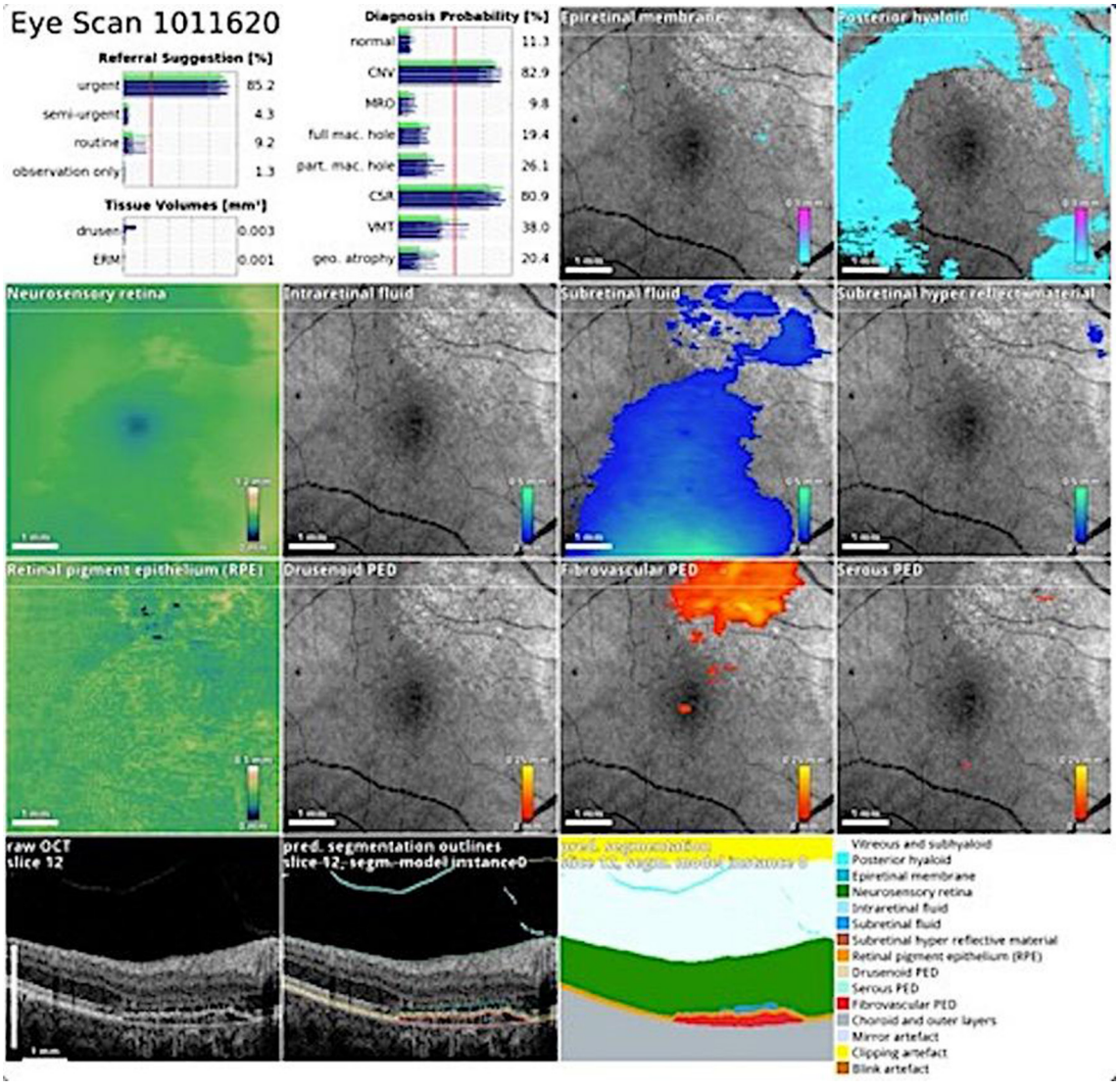

Figure 3 A representative screenshot from the output of the Moorfields-DeepMind deep learning system for optical coherence tomography segmentation and classification. In this case, the system correctly diagnoses a case of central serous retinopathy with secondary choroidal neovascularisation and recommends urgent referral to an ophthalmologist. Through the creation of an intermediate tissue representation (seen here as two-dimensional thickness maps for each morphological parameter), the system provides 'explainability' for the ophthalmologist.

areas of retinal pathology (figure 3).[43] It is also important to highlight that 'interpretability' of DL systems may mean different things to a healthcare professional than to a machine learning expert. Although it seems likely that interpretable algorithms will be more readily accepted by ophthalmologists, future applied clinical research will be necessary to determine whether this is the case and whether it leads to tangible benefits for patients in terms of clinical effectiveness.

Lastly, the current AI screening systems for DR have been developed and validated using two-dimensional images and lack stereoscopic qualities, thus making identification of elevated lesions like retinal tractions challenging. Incorporating the information from multimodal imaging in future AI algorithms may potentially address this challenge. In addition, the medicolegal aspects and the regulatory approvals vary in different countries and settings, and more work will be needed in these areas. An important challenge to the clinical adoption of AI-based technology is how the patients entrust clinical care to machines. Keel *et al*[72] evaluated the patient acceptability of AI-based DR screening within endocrinology outpatient setting and reported that 96% of participants were satisfied or very satisfied with the automated screening model.[72] However, in different populations and settings, the patient's acceptability for AI-based screening may vary and may pose challenge in its implementation.

CONCLUSIONS

DL is the state-of-the-art AI machine learning technique that has revolutionised the AI field. For ophthalmology, DL has shown clinically acceptable diagnostic performance in detecting many retinal diseases, in particular DR and ROP. Future research is crucial in evaluating the clinical deployment and cost-effectiveness of different DL systems in the clinical practice. To improve clinical acceptance of DL systems, it is important to unravel the 'black-box' nature of DL using existing and future methodologies. Although there are challenges ahead, DL will likely impact on the practice of medicine and ophthalmology in the coming decades.

Author affiliations
[1]Singapore Eye Research Institute, Singapore National Eye Center, Duke-NUS Medical School, National University of Singapore, Singapore, Singapore
[2]Department of Ophthalmology, Mt Sinai Hospital, New York City, New York, USA
[3]Google AI Healthcare, Mountain View, California, USA
[4]Casey Eye Institute, Oregon Health and Science University, Portland, Oregon, USA
[5]Department of Ophthalmology, University of Washington, School of Medicine, Seattle, Washington, USA
[6]Vitreo-retinal Department, Sankara Nethralaya, Chennai, Tamil Nadu, India
[7]Department of Ophthalmology, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[8]Department of Clinical Pharmacology, Medical University of Vienna, Vienna, Austria
[9]Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria
[10]Vitreo-retinal Service, Moorfields Eye Hospital, London, UK

Contributors DSWT, LRP, LP, JPC, AYL, RR, GSWT, LS, PAK and TYW have all contributed to manuscript drafting, literature review, critical appraisal and final approval of the manuscript.

Funding This project received funding from the National Medical Research Council (NMRC), Ministry of Health (MOH), Singapore National Health Innovation Center, Innovation to Develop Grant (NHIC-I2D-1409022), SingHealth Foundation Research Grant (SHF/FG648S/2015), and the Tanoto Foundation, and unrestricted donations to the Retina Division, Johns Hopkins University School of Medicine. For the Singapore Epidemiology of Eye Diseases (SEED) study, we received funding from NMRC, MOH (grants 0796/2003, IRG07nov013, IRG09nov014, STaR/0003/2008 and STaR/2013; CG/SERI/2010) and Biomedical Research Council (grants 08/1/35/19/550 and 09/1/35/19/616). The Singapore Integrated Diabetic Retinopathy Programme (SiDRP) received funding from the MOH, Singapore (grants AIC/RPDD/SIDRP/SERI/FY2013/0018 and AIC/HPD/FY2016/0912). In USA, it is supported by the National Institutes of Health (K12 EY027720, R01EY019474, P30EY10572, P41EB015896), by the National Science Foundation (SCH-1622542, SCH-1622536, SCH-1622679) and by unrestricted departmental funding from Research to Prevent Blindness. PAK is supported by a UK National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS--2014-12-023). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Competing interests DSWT and TYW are the coinventors of a deep learning system for retinal diseases. LP is a member of Google AI Healthcare. LRP is a non-paid consultant for Visulytix. PAK is a consultant for DeepMind.

Patient consent Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1. World Economic Forum, 2016. The fourth industrial revolution: what it means, how to respond. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/ (accessed 18 Aug 2018).
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
3. Zhang X, Zou J, He K, *et al*. Accelerating very deep convolutional networks for classification and detection. *IEEE Trans Pattern Anal Mach Intell* 2016;38:1943–55.
4. Shin HC, Roth HR, Gao M, *et al*. Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;35:1285–98.
5. Tompson J, Jain A, LeCun Y. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in Neural Information Processing Systems* 2014;27:1799–807.
6. Hinton G, Deng L, Yu D. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* 2012;29:82–97.
7. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;284:574–82.
8. Ting DSW, Yi PH, Hui F. Clinical applicability of deep learning system in detecting tuberculosis with chest radiography. *Radiology* 2018;286:729–31.
9. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
10. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
11. Ting DSW, Cheung CY, Lim G, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318:2211–23.
12. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
13. Lee CS, Tyring AJ, Deruyter NP, *et al*. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express* 2017;8:3440–8.
14. Abràmoff MD, Lou Y, Erginay A, *et al*. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci* 2016;57:5200–6.
15. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 2017;124:962–9.
16. Li Z, He Y, Keel S, *et al*. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* 2018;125:1199–206.
17. Burlina PM, Joshi N, Pekala M, *et al*. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol* 2017;135:1170–6.
18. Grassmann F, Mengelkamp J, Brandl C, *et al*. 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;125:1410–20.
19. Brown JM, Campbell JP, Beers A, *et al*. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol* 2018;136:803–10.
20. Poplin R, Varadarajan AV, Blumer K, *et al*. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;2:158–64.
21. Varadarajan AV, Poplin R, Blumer K, *et al*. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci* 2018;59:2861–8.
22. Yau JW, Rogers SL, Kawasaki R, *et al*. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* 2012;35:556–64.
23. Hutchinson A, McIntosh A, Peters J, *et al*. Effectiveness of screening and monitoring tests for diabetic retinopathy--a systematic review. *Diabet Med* 2000;17:495–506.
24. Scanlon PH, Malhotra R, Greenwood RH, *et al*. Comparison of two reference standards in validating two field mydriatic digital photography as a method of screening for diabetic retinopathy. *Br J Ophthalmol* 2003;87:1258–63.

25. Murray R, Metcalf SM, Lewis PM. Sustainingremote-area programs: retinal camera use by Aboriginal health workers and nurses in a Kimberley partnership. *MedJ Aust* 2005;182:520–3.
26. Ting DS, Tay-Kearney ML, Constable I, *et al*. Retinal video recording a new way to image and diagnose diabetic retinopathy. *Ophthalmology* 2011;118:1588–93.
27. Ting DS, Cheung GC, Wong TY. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol* 2016;44:260–77.
28 Abramoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 2018;1:1–8.
29. Chew EY, Clemons TE, Agrón E, *et al*. Effect of omega-3 fatty acids, lutein/zeaxanthin, or other nutrient supplementation on cognitive function: the areds2 randomized clinical trial. *JAMA* 2015;314:791–801.
30. Wong WL, Su X, Li X, *et al*. 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.
31. Centers for Medicare & Medicaid Services, 2018. CMS medicare provider utilization and payment data. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html (accessed 4 Sep 2018).
32. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina* 2017;1:322–7.
33. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol* 2018;256:259–65.
34. Kermany DS, Goldbaum M, Cai W. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122–31.
35. Ting DSW, Liu Y, Burlina P, *et al*. AI for medical imaging goes deep. *Nat Med* 2018;24:539–40.
36. Ronneberger O, Fischer P, Brox T. *U-Net: convolutional networks for biomedical image segmentation*. Medical Image Computing and Computer-Assisted Intervention – MICCAI, 2015.
37. Roy AG, Conjeti S, Karri SPK, *et al*. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed Opt Express* 2017;8:3627–42.
38. Venhuizen FG, van Ginneken B, Liefers B, *et al*. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomed Opt Express* 2018;9:1545–69.
39. Hamwood J, Alonso-Caneiro D, Read SA, *et al*. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. *Biomed Opt Express* 2018;9:3049–66.
40. Fang L, Cunefare D, Wang C, *et al*. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed Opt Express* 2017;8:2732–44.
41. Chen M, Wang J, Oguz I, *et al*. Automated segmentation of the choroid in EDI-OCT images with retinal pathology using convolution neural networks. *Fetal Infant Ophthalmic Med Image Anal* 2017;10554:177–84.
42. Prentašic P, Heisler M, Mammo Z, *et al*. Segmentation of the foveal microvasculature using deep learning networks. *J Biomed Opt* 2016;21:75008.
43. De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
44. Buchan JC, Amoaku W, Barnes B, *et al*. How to defuse a demographic time bomb: the way forward? *Eye* 2017;31:1519–22.
45. 2018. OCT rollout in every specsavers announced. https://www.aop.org.uk/ot/industry/high-street/2017/05/22/oct-rollout-in-every-specsavers-announced (accessed 4 Sep 2018).
46. Tham YC, Li X, Wong TY, *et al*. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* 2014;121:2081–90.
47. Jonas JB, Gusek GC, Guggenmoos-Holzmann I, *et al*. Size of the optic nerve scleral canal and comparison with intravital determination of optic disc dimensions. *Graefes Arch Clin Exp Ophthalmol* 1988;226:213–5.
48. Christopher M, Belghith A, Weinreb RN, *et al*. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. *Invest Ophthalmol Vis Sci* 2018;59:2748–56.
49. Elze T, Pasquale LR, Shen LQ, *et al*. Patterns of functional vision loss in glaucoma determined with archetypal analysis. *J R Soc Interface* 2015;12.
50. Wang M, Pasquale LR, Shen LQ, *et al*. Reversal of glaucoma hemifield test results and visual field features in glaucoma. *Ophthalmology* 2018;125:352–60.
51. Tanna AP, Bandi JR, Budenz DL, *et al*. Interobserver agreement and intraobserver reproducibility of the subjective determination of glaucomatous visual field progression. *Ophthalmology* 2011;118:60–5.
52. Viswanathan AC, Crabb DP, McNaught AI, *et al*. Interobserver agreement on visual field progression in glaucoma: a comparison of methods. *Br J Ophthalmol* 2003;87:726–30.
53. Yousefi S, Kiwaki T, Zheng Y, *et al*. Detection of longitudinal visual field progression in glaucoma using machine learning. *Am J Ophthalmol* 2018;193:71–9.
54. Kass MA, Heuer DK, Higginbotham EJ, *et al*. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2002;120:701–13.
55. Heijl A, Leske MC, Bengtsson B, *et al*. Reduction of intraocular pressure and glaucoma progression: results from the Early Manifest Glaucoma Trial. *Arch Ophthalmol* 2002;120:1268–79.
56. Comparison of glaucomatous progression between untreated patients with normal-tension glaucoma and patients with therapeutically reduced intraocular pressures. Collaborative Normal-Tension Glaucoma Study Group. *Am J Ophthalmol* 1998;126:487–97.
57. Kazemian P, Lavieri MS, Van Oyen MP, *et al*. Personalized prediction of glaucoma progression under different target intraocular pressure levels using filtered forecasting methods. *Ophthalmology* 2018;125:569–77.
58. Lichter PR, Musch DC, Gillespie BW, *et al*. Interim clinical outcomes in the Collaborative Initial Glaucoma Treatment Study comparing initial treatment randomized to medications or surgery. *Ophthalmology* 2001;108:1943–53.
59. Blencowe H, Moxon S, Gilbert C. Update on blindness due to retinopathy of prematurity globally and in India. *Indian Pediatr* 2016;53(Suppl 2):S89–92.
60. Robinson R, O'Keefe M. Cryotherapy for retinopathy of prematurity--a prospective study. *Br J Ophthalmol* 1992;76:289–91.
61. Cryotherapy for Retinopathy of Prematurity Cooperative Group. Multicenter trial of cryotherapy for retinopathy of prematurity: ophthalmological outcomes at 10 years. *Arch Ophthalmol* 2001;119:1110–8.
62. Gilbert C, Rahi J, Eckstein M, *et al*. Retinopathy of prematurity in middle-income countries. *Lancet* 1997;350:12–14.
63. Fleck BW, Williams C, Juszczak E, *et al*. An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials. *Eye* 2018;32:74–80.
64. Campbell JP, Swan R, Jonas K, *et al*. Implementation and evaluation of a tele-education system for the diagnosis of ophthalmic disease by international trainees. *AMIA Annu Symp Proc* 2015;2015:366–75.
65. Daniel E, Quinn GE, Hildebrand PL, *et al*. Validated system for centralized grading of retinopathy of prematurity: telemedicine approaches to evaluating acute-phase retinopathy of prematurity (e-ROP) study. *JAMA Ophthalmol* 2015;133:675–82.
66. Worrall D, Wilson CM, Brostow GJ. *Automated retinopathy of prematurity case detection with convolutional neural networks*, 2016.
67. Brown JM, Campbell JP, Beers A. *Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning*. Proceedings Volume 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, 2018.
68. Krause J, Gulshan V, Rahimy E, *et al*. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 2018;125:1264–72.
69. Ramanishka V, Das A, Zhang J, 2016. Top-down visual saliency guided by captions. https://arxiv.org/abs/1612.07360
70. Lam C, Yu C, Huang L, *et al*. Retinal lesion detection with deep learning using image patches. *Invest Ophthalmol Vis Sci* 2018;59:590–6.
71. Quellec G, Charrière K, Boudi Y, *et al*. Deep image mining for diabetic retinopathy screening. *Med Image Anal* 2017;39:178–93.
72. Keel S, Lee PY, Scheetz J, *et al*. 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.