Artificial Intelligence, Machine Learning and Deep Learning in Ophthalmology: Current Clinical Relevance

Daniel S.W. Ting MD PhD[1]
Lily Peng MD PhD[2]
Avinash V Varadarajan MS[2]
Pearse Keane FRCOphth[3]
Phil Burlina PhD[4,5,6]
Michael F. Chiang MD[7]
Leopold Schmetterer PhD [1,8,9,10]
Louis R. Pasquale MD[11]
Neil M. Bressler MD[4]
Dale R Webster PhD[2]
Michael Abramoff MD PhD[12]
Tien Y. Wong MD PhD[1]

1. Singapore Eye Research Institute, Singapore National Eye Center, Duke-NUS Medical School, National University of Singapore
2. Google AI Healthcare, California, USA
3. Moorfields Eye Hospital, London, UK
4. Wilmer Eye Institute, Johns Hopkins University School of Medicine
5. Applied Physics Laboratory, Johns Hopkins University
6. Malone Center for Engineering in Healthcare, Johns Hopkins University
7. Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology, Casey Eye Institute, Oregon Health and Science University
8. Department of Ophthalmology, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
9. Department of Clinical Pharmacology, Medical University of Vienna, Austria
10. Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Austria
11. Department of Ophthalmology, Icahn School of Medicine at Mount Sinai, New York, NY.
12. Department of Ophthalmology and Visual Sciences, University of Iowa Health Care

Financial Disclosure: Drs Daniel SW Ting and TY Wong are the co-inventor of a deep learning system for retinal diseases. Drs L Peng, A Vadarajan and D Webster are the members of the Google AI Healthcare. Dr. Michael F. Chiang is an unpaid member of the Scientific Advisory Board for Clarity Medical Systems (Pleasanton, CA), a Consultant for Novartis (Basel, Switzerland), and an equity owner in Inteleretina, LLC (Honolulu, HI). Dr P Keane is a consultant for Google DeepMind. Dr LR Pasquale is a non-paid consultant for Visulytix and consultant for Google Verily. Dr M.D. Abramoff is the Watzke Professor of Ophthalmology and Visual Sciences, and CEO and Founder, Equity owner, of IDx, Iowa, USA, and is inventor for US patents and patent applications on artificial intelligence, imaging and deep learning, as well as on foreign patents. Drs NM Bressler and P Burlina are the co-inventor and patent holders for a deep learning system for retinal diseases.

Corresponding author:
Daniel SW Ting MD PhD
Assistant Professor in Ophthalmology, Duke-NUS Medical School
Singapore National Eye Center
11 Third Hospital Avenue,
Singapore 168751
Email address: daniel.ting.s.w@singhealth.com.sg

Abstract

With the advent of computer graphic processing units, improvement in mathematical models and availability of big data, artificial intelligence (AI) using machine learning (ML) and deep learning (DL) techniques have achieved robust performance for potential application across many industries, including social-media, the internet of things, the automotive industry and healthcare. DL systems provide capability in image, speech and motion recognition as well as in natural language processing. In medicine, most of the progress of AI, ML and DL systems has been demonstrated in image-centric specialties such as radiology, dermatology and pathology. There is increasing interest in AI in ophthalmology. New studies, including pre-registered prospective clinical trials, have shown DL systems are effective in detecting diabetic retinopathy (DR), glaucoma, age-related macular degeneration, retinopathy of prematurity, refractive error and in identifying cardiovascular risk factors and diseases, using image based data such as fundus photographs and optical coherence tomography. Additionally, the application of ML to Humphrey visual fields may be useful in detecting glaucoma progression. There are fewer studies that incorporate clinical data in AL algorithms and no prospective studies to demonstrate that AI algorithms can predict the development of eye disease. This article describes the current global eye disease burden, clinical unmet needs and selected common ophthalmic conditions of public health importance for which AI and DL systems may be applicable. Technical and clinical aspects to build a DL system to address those gaps, and the potential challenges for clinical adoption are discussed. AI, ML and DL likely will play a crucial role in clinical ophthalmology practice, with implications for screening, diagnosis and follow up of the major causes of vision impairment, in the setting of the ageing population globally.

Introduction

With the advent of graphic processing units (GPUs), advances in mathematical models, the availability of big datasets and low cost sensors, artificial intelligence (AI) using machine learning (ML) techniques initially and deep learning (DL) techniques subsequently, has sparked tremendous interest in many industries.[1] These include application of AI in social-media, the internet of things, finance and banking, the automotive industry and healthcare. AI systems can be designed not only for image,[2,3] speech[4] and motion recognition,[5] but also in natural language processing.[6]

In medicine, the most robust AI algorithms have been demonstrated in image-centric specialties, including radiology, dermatology, pathology and increasingly so in ophthalmology. For example, Lakhani et al demonstrated excellent performance in detecting pulmonary tuberculosis from chest radiographs,[7,8] while Esteva et al was able to differentiate malignant melanoma from benign lesions on skin photographs.[9] In ophthalmology, there have been two major areas in which AI and new DL systems have been applied. First, AI systems have been shown in new studies, including pre-registered prospective clinical trials, to accurately detect diabetic retinopathy (DR),[10-13] glaucoma,[10,14] age-related macular degeneration (AMD),[10,15,16] retinopathy of prematurity (ROP),[17] and refractive error, from digital fundus photographs.[18] A range of cardiovascular risk factors[19] have also been accurately predicted from fundus photographs. Second, several retinal conditions [e.g., neovascular AMD, earlier stages of AMD, and diabetic macular edema (DME)][20] has also be detected accurately using optical coherence tomography (OCT).[21,22] There are relatively fewer AI studies using other data, such as studies which show good performance in detecting glaucoma progression from serial Humphrey visual fields (HVFs).[23] However, there are fewer studies that incorporate clinical and imaging data in AL algorithms, and no prospective studies to demonstrate that AI algorithms can predict the development of eye diseases over time. Furthermore, the implementation and adoption of AI into routine clinical care remains extremely challenging. These remain significant goals of AI research in ophthalmology

This article describes basic concepts of AI, ML and DL and how such systems might address some of the global burdens created by common eye conditions.

Furthermore, the technical and clinical aspects of developing and validating an AI/DL system, potential challenges and future directions are also discussed in this article.

Artificial Intelligence, Machine Learning and Deep Learning

AI was conceptualized in 1956, after a workshop at Dartmouth College **(Figure 1).**[10] In the workshop, many AI groups showed promising results in computer learning of checkers strategies, solving word problems in algebra and proving logical theorems. These tasks involved mostly pattern recognition and computational learning. All AI systems were designed to execute and maximise its chance of 'winning' within a constructed environment. The term 'machine learning' (ML) was subsequently coined by Arthur Samuel in 1959 and stated that "the computer should have the ability to learn using various statistical techniques, without being explicitly programmed".[11] Using ML, the algorithm can learn and make predictions based on the data that has been fed into the training phase, using either a supervised or un-supervised approach. ML has been widely adopted in applications such as computer vision and predictive analytics using complex mathematical models. In supervised learning, the computer is trained with labelled examples, also known as ground truth, whereas for unsupervised learning, no labelling is required for the algorithm to find its own structure in the input. The majority of AI application in biomedical research uses supervised learning.

DL utilizes multiple processing layers to learn representation of data with multiple levels of abstraction.[20] Although some forms of deep neural networks have already been investigated in the past, the advent of graphic processing units (GPU) with improved processing power, larger annotated datasets, and other factors, have recently boosted its diagnostic performance in many domains. Using learning approaches such as backpropagation,[24] a ML or DL system is able to discover intricate structure in large data sets, then changing its internal parameters that are used to compute the representation in each layer from the previous one. These approaches permit the use of regional samples to allow the network to learn to detect biomarkers; furthermore these approaches use complete images, and associate the entire image with a diagnostic output, thereby eliminating the use of "hand-engineered" image features. Given the much improved performance,[11,12] DL has been widely adopted in image recognition, speech recognition and natural language processing.

General Approach in Building a Robust AI system

This section explains some common terminologies, software framework, network architectures, datasets selection, assistive vs. autonomous AI system, consideration factors to ensure the robustness of these algorithms (Table 1).[1,25-30] In order to build a robust DL system, it is important to have 2 main components – the 'brain' (technical networks – Convolutional Neural Network (CNN) and the 'dictionary' (the datasets).

1. *What is a CNN?*

A CNN is a deep neural network consisting of a cascade of processing layers that resemble the biological processes of the animal visual cortex. It transforms the input volume into an output volume via a differentiable function. Inspired by Hubel and Weisel,[31] each neuron in the visual cortex will respond to the stimulus that is specific to a region within an image, similar to how the brain neuron would respond to the visual stimuli, that will activate a particular region of the visual space, known as the receptive field. These receptive fields are tiled together to cover the entire visual field. Two classes of cells are found in this region – simple vs complex cells. The simple cells active when they detect edge-like patterns, while the more complex cells activate when they have a larger receptive field and are invariant to the position of the pattern.

Broadly, the CNN can be divided into the input, hidden (also known as feature-extraction layers) and output layers (Figure 2A). The hidden layers usually consist of convolutional, pooling, fully connected and normalization layers, and the number of hidden layers will differ for different CNNs. The input layer specifies the width, height and the number of channels (usually 3 channels – red, green and blue). The convolutional layer is the core building block of a CNN, transforming the input data by applying a set of filters (also known as kernels) that acts as the feature detectors. The filter will slide over the input image to produce a feature map (as the output). A CNN learns the values of these filters weights on its own during the training process, although the specific parameters such as number of filters, filter size, network architecture still need to be set prior to that. Additional operations called activations (for example ReLU or Rectified Linear Unit) are used after every convolution operation. For pooling, the aim is to reduce the dimensionality of each feature map and make it somewhat spatially invariant, and retain the most important information. Pooling can be divided into different types: maximum, average and minimum. In the case of maximum pooling, the largest element from the rectified feature map will be taken (Figure 2B). The output from the convolutional and pooling layers represent the high-level features of the input image. The purpose of the fully connected layer is to use these high-level features to classify the input image into various classes based on the training dataset. Following which, backpropagation is conducted to compute the network weights and uses the gradient descent to update all filters and parameter values to minimize the output error. This process will be repeated many times during the training process.

2. ***Software frameworks: Keras, TensorFlow, PyTorch***

Deep neural networks are commonly implemented in several popular software frameworks (e.g. Caffe, Tensorflow, PyTorch, etc). Early development in these past 10 years was enabled by the availability of frameworks like Caffe[74] (originally from UC Berkeley), Torch[13] (built on top of Lua) and Theano[75]. These frameworks tend to be less used nowadays, although Caffe2 has been released with both C++ and Python front ends and has features such as ease of deployment for mobile application. More recently, Python-based frameworks such as TensorFlow[76] (from Google) and PyTorch[13] (an evolution of Torch in Python) have gained in popularity. High-level application programming interface (APIs) such as Keras[12] or Lasagne have also made it much easier to develop DL systems, and should be considered the preferred starting point for implementation for new users. In particular, they simplify the reuse of existing networks architectures and pretrained weights, which is convenient for the purposes of transfer learning and fine tuning. Two recent important features of PyTorch are imperative programming (vs. declarative/symbolic for TensorFlow) and the use of dynamic graphs. These features make PyTorch easier to debug and inspect compared to other frameworks where graphs are static (although this feature has also now been made available in TensorFlow).

3. *Popular Network Architectures - AlexNet, VGGNet, Inception, ResNet and DenseNet*

AlexNet, first described in 2012 with 5 convolutional layers, has been the most widely used CNN, after winning the ImageNet Large Scale Visual Competition Recognition (ILSVCR).[32] Following which, more CNNs with deeper layers and unique features were described subsequently. Each CNN can also have different versions and layers, for example VGGNet (16 or 19 layers), Inception V1 to V4 (27 layers), ResNet (18, 50, 152 or even up to 1202 layers with stochastic depth) and DenseNet (40, 100, 121, 169 layers). Compared to AlexNet, the newer networks have unique features to help improve performance, including the addition of more layers, smaller convolutional filters, skip connections, repeated modules with more complex/parallel filters, bottleneck connection and dropout. Although deeper CNNs (e.g. ResNet and DenseNet) have been reported to achieve improved performance, older architectures (e.g. VGGNet and Inception) have consistently shown comparable outcomes in medical imaging analysis. In order to further boost performance, multiple deep neural networks are commonly trained and ensembled. Transfer learning with pretrained weights has also been reported to aid training and performance, especially with smaller datasets.

Rather than training the CNN entirely from scratch (i.e. starting with randomly initialized values), it has been common practice in retinal image and many other DL applications to perform transfer learning. Transfer learning is the process of reusing models developed for other applications (e.g. for performing full image classification from ImageNet images) and further refining these weights for a different target domain (e.g. detection of age-related macular degeneration [AMD] on fundus images). The most popular transfer learning approach has been to use fine-tuning.[10,11,13,17-19,21,67-69] It has been shown that once a network weight is optimized to solve a certain problem, the weights for the resulting model, and especially those corresponding to lower-level layers can be largely reused or slightly modified for solving other tasks. In this approach, called ’fine-tuning’, the original network weights are used as a starting point and further optimized (fine-tuned) to solve another task (such as going from an original domain, i.e. common everyday images found in ImageNet, to retinal imaging). The approach may also involve selectively freezing some of the network layers' weights (e.g. early layers usually encode low level feature computation that are likely to be universally applicable across domains), and selectively fine-tuning other layers (e.g. mid-level convolutional or higher-level fully connected layers, which encode more domain-specific features).

4. ***Dataset splitting and evaluation***

As is usually done in ML, data is split into training, validation and testing datasets. These datasets must not intersect – in other words, an image that is in one of the datasets, must not appear in any of the other datasets. Ideally, this non-intersection should extend to patients. The general class distribution for the targeted condition should be maintained in all these datasets.

Training of deep neural nets is generally done in batches (subsets) randomly sampled from the training dataset. The training dataset is what is used for optimizing the network weights via backpropagation. The validation dataset is used for hyperparameter selection and tuning, and is customarily also used to implement stopping conditions for training.

Finally, the reported performance should be computed exclusively using the selected optimized model weights, on the testing datasets. It is also critical to test the AI system using independent datasets, captured using different devices, population and clinical settings. This will ensure the generalizability of the system in the clinical settings.

5. ***Reporting of the datasets characteristics and diagnostic performance***

For any AI study in medical imaging analysis, it is important to demonstrate the population in which the DL system was developed and tested on. The reporting of dataset characteristics, including age, gender, ethnic groups, imaging platform, size of field of view, reference standard, are important, especially now that we know that DL system can predict additional features that are not discernable to manual inspection like gender and age.[19] These characteristics might be augmented by including the systemic vascular risk factors (e.g. blood pressure, blood sugar level and etc.) for vascular conditions such as diabetic retinopathy (DR).

In order to report the diagnostic performance of an AI system, it is important to first define the gold standard or reference standard (also known as ground truth). In ophthalmology, the reference standard can be the classifications rendered by a reading center, ophthalmologists, professional trained graders, or optometrists. In terms of the performance metrics, the most commonly used one is the area under the receiver's operator characteristics curve (AUC), computed using sensitivity and specificity. In order to ascertain the true performance of an AI system, it is important to report the AUC of testing datasets (locally and externally), using either a pre-set sensitivity or specificity. If the operating threshold is not set suitably, an AI system with good AUC (e.g., >0.90) potentially could have suboptimal sensitivity or specificity, resulting in adverse events within clinical settings.

Apart from the above-mentioned parameters, investigators could consider reporting positive predictive value (also known as precision), negative predictive value or Cohen Kappas. Lastly, many studies utilize accuracy as one of the main measurement outcomes. Similar to AUC, the reporting of accuracy could be potentially 'over-optimistic' given that it takes into account both true positive and true negative as the nominator, with true and false positive, and true and false negative as the denominator. If a dataset contains only a few positive images and the AI system under-detect them, the reported diagnostic accuracy will be high, although the sensitivity will be very poor. Thus, the above-mentioned reasons state the importance of including AUC, sensitivity and specificity as the bare minimum for any AI study for the literature. Directly comparing AUC, sensitivity and specificity between different CNNs is, however, misleading of the data do not stem from the same validation dataset. Usually the AUC is the higher the more severe cases are in the validation dataset.

6. ***Methods to explain the diagnosis***

DL systems are commonly referred to as a 'black-box', and it could be a potential cause of low adoption of such technology within clinical settings. It is important for the patients to be informed of their diagnosis, and why the diagnosis was made. At present, the deep learning community is actively researching ways to rectify this. Highlighting the image features that add diagnostic value to a medical image could provide a relatively novel teaching opportunity in medicine. Visualization of the network workings and activation via *saliency maps* has allowed the generation of overlay highlights that *show* where the network *is looking* when it renders a classification. **Figure 3** shows the example of the heat map detection for referable DR and advanced AMD. Visualization may be achieved using multiple methods, such as occlusion testing, integrated gradients and soft attention. Occlusion testing is performed by sliding an occluding window across an image and checking the resulting effect on output classification. Integrated gradients, on the other hand, perturbs continuously an image from a baseline image to the output image, while monitoring the activation out of the network to characterize the sensitivity of the output to each pixel input. Lastly, for soft attention, the saliency map outputs of convolutional layers are up-scaled via reverse max pooling and passed through additional convolutional layers. This method was used by Poplin et al to identify the locations in the fundus image that were predictive of cardiovascular risks or other information such as gender **(Figure 4).**[19]

AI to Solve Clinical Unmet Need in Ophthalmology

Global Eye Health Burden

By 2050, the world's population aged 60 years and older is estimated to be 2 billion, up from 900 million in 2015, with 80% of whom living in low- and middle-income countries.[33] People are living longer, and the pace of ageing is much faster than in the past.[34] Because of this, there is a need for longer disease surveillance for many ocular and systemic conditions like DR, glaucoma, AMD and cardiovascular conditions **(Table 2)**.[35] Population expansion also creates pressure to screen for important causes of childhood blindness such as retinopathy of prematurity (ROP), refractive error, and amblyopia.[36] In view of these unmet needs, many groups have published the AI system, using retinal images, OCTs and other imaging modalities for glaucoma (e.g. HVF).

1. AI for Diabetic Retinopathy

1.1 Clinical Unmet Need

Diabetes mellitus (DM) is one of the world's fastest growing chronic diseases and a leading cause of acquired vision loss.[37,38] According to the World Health Organization, it is estimated that the total number of people with diabetes will double from 171 million in 2000 to 422 million by 2040 **(Table 2)**.[39] DR, a specific microvascular complication of DM, remains the leading cause of acquired vision loss worldwide in middle-aged and therefore economically active people.[37,40,41] With the increasing number of people with DM, the number of DR and vision-threatening DR (VTDR), which includes severe non-proliferative DR, proliferative DR (PDR) and diabetic macular edema (DME), has been estimated to rise to 191.0 million and 56.3 million respectively by 2030.[42] It is estimated that DR accounts for 4.8% of the number of cases of blindness (37 million) worldwide.[43] A pooled analysis of 22,896 people with DM from 35 population-based studies in the U.S., Australia, Europe and Asia (between 1980-2008) showed that the overall prevalence of any DR was 34.6% (95%CI 34.5-34.8), with 7% (6.9-7.0) suffering from VTDR.[44] Screening for DR, coupled with timely referral and treatment, is a universally accepted strategy for blindness prevention. DR screening programs, however, are challenged by issues related to implementation, availability of human assessors and long-term financial sustainability. Thus, more novel and economical screening technologies are useful to screen for DR.

The idea of automated DR detection is not a new one. In fact, the concept of using software to help with the heavy load of retinal images for DR screening was introduced 20 years ago.[45] Prior to DL, many automated systems were built using ML algorithms operating on hand-crafted "features" to detect DR lesions,[45-50] with overall performance that was lower than manual grading.[51-54] Over the past few years, DL has been shown to greatly improve the performance of automated DR grading systems.[12] In the review of DL papers, there are 3 major areas to consider: (1) What are the inputs and outputs of the model? (2) What is the reference standard? and (3) How well does the algorithm generalize?

1.2 Algorithm Design

Hybrid, or biomarker-based algorithms, use multiple partially dependent detectors for the biomarkers or lesions characteristic for DR, such as microaneurysms, hemorrhages and lipoprotein exudates.[55] The outputs of these are then fused into a disease level output, using a separately trained and validated ML algorithm.[56] The detectors themselves, are independently validated, and can be implemented as multilayer CNNs,[57] wavelet filters, or both. [49,58]

As mentioned above, multilayer CNN,[57] exploiting the spatial coherence that is characteristic of retinal images, and where all transformation levels are determined from training data, instead of being designed by experts,[59] have been highly successful,[11] and have been substantially outperforming classical image analysis techniques in many tasks. [57] Their greatest advantage is that their development only requires a dataset with sufficient quantity and quality of the training data, and not on a mathematical coding representative of DR lesions lesions.[11]

1.3 Inputs of the algorithm for DR

In terms of inputs, most of the published studies describe algorithms that have been trained to take in a single macula-centered or primary field of view 45-degree fundus image and gives an image-level read **(Table 3).** The exception to this is the recent work done by Abramoff et al and Ting et al,[10,60] which actually requires 4 images, two from each eye -- one macula-centered and one disc-centered -- and returns a patient-level read. In terms of the model outputs, there are two major ways that the models have been trained to make predictions -- either as a binary or multi-class classification tasks. While many of the grading scales are usually based on a multi-class clinical grading scale like the International Classification of DR (ICDR) severity scale, a majority of the studies stratify the prediction on a particular severity threshold. Most of these models have been trained to detect referable DR defined as moderate DR or worse and/or DME because it is at this threshold that many guidelines suggest closer follow up (rather than follow up in a year). For example, Abramoff *et al.*[60], Gulshan *et al.*[11], and Ting *et al.*[10] all developed DL systems that were based on detecting referable DR while Gargeya and Leng[13] trained a model to detect any DR.

1.4 Gradeability

One important but often under-appreciated output of the model is gradeability. Since ungradable images will also result in a referral, it is important that the models can accurately identify images that are not gradable. The performance metrics reported by Ting *et al*[10] accounted for gradability issues by default. After accounting for ungradable images, the model in Gulshan *et al*[11] had roughly the same sensitivity (97.5% vs 96.7%) but lower specificity (93.4% vs 84.0%) **(Table 3).** These studies were not preregistered, and so the datasets did not account for every subject the system is intended to screen. In a preregistered, prospective, intention to screen trial, Abramoff et al reported a sensitivity of 87.2% and specificity of 90.7% in detection of referable DR, with reference to reference standard graded by the Wisconsin Reading Center based on a 4 wide-field retinal images, twice the image area than the AI system 'saw', as well as optical coherence tomography in detecting diabetic macular edema. This demonstrates the importance of conducting a real-world clinical trial in testing an AI system, as this similar AI system also had comparable AUC in Messidor-2 dataset, achieving an extremely high AUC of 0.98 in the earlier study.[12]

1.5 Reference standard

These studies have consistently demonstrated that it is possible to train DL algorithms that recapitulate the reference standard with high performance metrics (sensitivity, specificity >90% and/or AUCs >0.95) **(Table 3).** Thus, if an algorithm that was trained to predict the majority decision of ophthalmologists, it will perform that task with high fidelity.[11] Similarly if it were trained to predict the adjudicated grade from a panel of retinal specialists, the model will also recapitulate that classification very well.[61] Finally, if it is designed and built to predict the same for the reference standard that has been used for over 3 decades to evaluate diagnosis and treatment of DR, the Early Treatment of Diabetic Retinopathy Severity Scale, that can be recapitulated well also.[60] Nonetheless, many DR FDA pharmacological trials were performed using the reference standard from an established reading centers (e.g. Wisconsin Reading Center) or comprehensive slit lamp examination, using more fields (>2 fields) or stricter criteria (using OCT as the reference standard to determine whether this is presence of diabetic macular edema, as compared to a 2-field non-stereoscopic fundus photographs). Hence, for reviewing papers about DL, it is very important to be able to clearly identify how the reference standard was established, and some performance may not be an apple-to-apple comparison. Performance numbers from studies with different reference standards may not be comparable.

For example, Abramoff *et al* reported 96.8% sensitivity and 87.0% specificity in detection of referable DR on the Messidor-2 dataset for their DL-based system[12] when using three retinal specialist adjudication on a single macula-centered 45 degree fundus image. Subsequently, the same group reported the performance of a similar algorithm but improved with deep learning based detectors, with a sensitivity of 96.8% (95% CI: 93.3%–98.8%), and specificity of 87.0% (95% CI: 84.2%–89.4%), and AUC of 0.980 (95% CI: 0.968–0.992), against 3 retinal specialists.[12] Finally, in their FDA pivotal trial where the reference standard was based on a four widefield stereoscopic fundus images and OCT (for DME), read by the Wisconsin Reading Center, the standard for FDA drug trials, sensitivity of 87.2% (95% CI, 81.8–91.2%) (>85%), specificity of 90.7% (95% CI, 88.3–92.7%) (>82.5%), and imageability rate of 96.1% (95% CI, 94.6–97.3%) were reported.[60] Similarly, the model in Gulshan *et al* had a sensitivity of 97.5% and 93.4% on the EyePACS-1 primary validation dataset where the reference standard was the majority decision of ophthalmologists.[11] Subsequently, Krause *et al* demonstrated that compared to adjudication of retinal specialists, the majority decision of ophthalmologists had 83.8% sensitivity and 98.1% specificity.[61] Not surprisingly, when DL was used to train a model that recapitulates the adjudication of retinal specialists, the new model performed well, with a sensitivity of 97.1% and specificity of 92.3%. In the study conducted by Ting *et al*, the reference standard varied but generally consisted of at two independent graders, with a 3rd senior grader to adjudicate disagreements.[10]

1.7 Generalizability of the algorithm

One of the most important considerations in training DL models is generalization -- that is how well do the models perform on new data, especially data that is derived from populations that are distinct from the population used for model training. Prior to validation, it is important to pre-set the desired operating threshold on the development dataset based on desired sensitivity and specificity. Subsequent validation at these pre-defined operating points will better represent performance in real-world settings. Validation in the same population as the development set (perhaps at a different time frame) is often called “primary validation.” Validation in a different population is often called “secondary validation.” Ting *et al* developed a model that was validated on 11 independent datasets.[10] On the primary validation set, the model achieved a 90.5% sensitivity and 91.6% specificity. On the largest secondary validation dataset of more than 15,000 images, the algorithm had a sensitivity of 98.7% and specificity of 81.6%. The model also did well on the 9 other secondary validation datasets with high sensitivity (>90%) and acceptable specificity (>70%).

1.8 Future Directions

Despite the high performance metrics reported by numerous studies that leverage DL for DR detection, there is still much work ahead in terms of implementation in clinical practice. First, it would be important for screening programs considering the implementation of these systems to understand the steps to capturing the necessary data for the algorithms to be used. For example, what type of equipment would be required for the system to work? Is the model compatible with multiple imaging cameras? What is the recommended procedure to capture fundus images? How often will dilation be necessary? All of these could have important implications about whether the adoption of these algorithms is even feasible, especially in resource constrained settings.

Next, performance verification of a trained model in the population where it is to be deployed will be required. This is particularly important for models that have been trained utilizing datasets from relatively homogenous populations, without secondary validation. Retinal images that could be used for training and inference can often be quite variable from one screening program to another. Variabilities like field of view, image magnification, image quality and participant ethnicities are all considerations that should be accounted for. Diversification of the training dataset would be critical in addressing this challenge.[10]

Another consideration in the development of AI models for DR screening is how to address non-DR findings. It is common practice that if there are non-DR findings identified during DR screening that these findings are reported back to the clinic. However, there is still some uncertainty and heterogeneity about when these other findings should be considered referable. In addition, there can be substantial grader variability in the manual interpretation of fundus images for other disease. For example, when to refer a suspicious cup-to-disc ratio could vary from one screening program to another. Ting *et al* reported the development of additional models that also could detect AMD and the glaucoma-like disc.[10] There are other publications (covered later in this review) focused on building models that detect non-DR diseases separately. Studies looking at both DR and non-DR findings would be an important area for future development.

In addition to performance, impact on clinical workflow, model explainability may be an important aspect of the adoption of DL systems. Because DL models do not utilize explicit feature engineering, attention techniques can help visualize the regions of the image that is most relevant for the prediction. Large longitudinal clinical trials with AI systems implemented end-to-end with diverse hardware, population characteristics, and local environmental will be critical milestones in evaluating the actual safety and efficacy of AI systems. Furthermore, real-world deployment of these new systems in multiple settings will be critical in understanding the full impact of AI on clinical care. For example, increased number of screenings enabled by automated screening algorithms will increase demand for follow-up and treatment. Healthcare systems will have to adapt so that they can manage this additional volume. Moreover, real time feedback from a model might enable follow-up actions to be initiated at the same visit. If a patient does not need to be referred, this would also be an opportunity to reinforce and commend the patient on efforts in managing their disease and emphasize the need for follow-up. If a patient is found to have referable disease, this allows for timely follow-up appointments to be scheduled before the patient leaves the office. There is limited information available regarding the potential success of such management.

Despite the tremendous progress made in the application of DL for DR screening, there are still many challenges ahead -- from identifying image features that are critical to image classification to large scale implementation. However, the rapid progress and excitement in this field make it fairly clear that DL systems will have a profound impact on DR screening in the coming decades.

2. **AI for Glaucoma and Glaucoma Suspect**

2.1 Clinical Unmet Need

Apart from DR, many screening programs also screen for the referable glaucoma suspect. The World Health Organization has declared Glaucoma to be the second largest cause of blindness worldwide, comprising 15% of the blindness cases globally, or 5.2 million patients. This number is expected to increase up to 111.8 million by 2040 **(Table 2).**[62] As glaucoma is an optic neuropathy, retinal features for referable glaucoma suspect include an increased vertical cup-to-disc ratio (CDR), neuro-retinal rim thinning, presence of optic disc haemorrhages and retinal nerve fibre layer defects. To date, there is no cost-effective screening strategy for detection of the high-risk glaucoma suspect, mainly due to the absence of an appropriate test.[63] This sentiment is aligned with the most recent US Preventive Services Task Force 2013 position statement claiming that current evidence was insufficient to recommend screening for glaucoma in adults; nonetheless, there was a proviso that high-risk groups (e.g. positive family history, African American) might benefit from early screening with their primary care physicians.[64,65]

The success of AI using DL system in glaucoma in the screening or the clinical setting is predicated on an agreed-upon structural and functional definition of the disease. Certainly, glaucoma is a heterogenous condition, especially considering the various anterior segment features that may be present in the disorder, with the convergent feature being a characteristic optic nerve appearance that corresponds to vision loss. One way to characterize this optic neuropathy is to rely on excavation of the neuroretinal rim that can be quantified with the cup-to-disc-ratio (CDR). Since disc size and shape can vary among people in a population and these features also differ across populations, it is problematic to describe a CDR that defines glaucoma. The International Society for Geographical and Epidemiological Ophthalmology (ISGEO) proposes using the upper 97.5th percentile of vertical CDR or of CDR asymmetry as a standard definition of structural glaucomatous damage.[66] This definition is, however, not sufficient for glaucoma diagnosis, because of the large influence of disc size[67] and the issues in patients with abnormal anatomical configuration of the disc. In addition, measurement of CDR is biased by large grader-variability because of a lack of a solid anatomic basis.[68] On OCT retinal nerve fibre layer thickness and ganglion cell complex measurements are used to discriminate glaucoma from healthy.[69] More recently minimum rim width as measured from Bruch's membrane opening has been used as a novel diagnostic tool in glaucoma.[70] A proposed reference standard for functional loss from glaucoma is a glaucoma hemifield test (GHT) outside normal limits and a cluster of 3 contiguous points with assigned probability of 5% or less on the pattern deviation of a Humphrey visual field analyzer. These contiguous points should follow a nerve fiber layer distribution. Comparable functional loss on other visual field (VF) platforms could be considered. Patients with definite glaucoma would meet both structural and functional criteria while suspects might meet only the structural criterion. The ISGEO proposes that patients with disc haemorrhage, IOP at greater than the 97.5Th percentile or subjects with occludable angles but normal optic nerves, visual fields, IOP and no peripheral anterior synechiae also be regarded as suspects. While no definition of glaucoma is ideal, DL systems can potentially be trained to identify these phenotypic attributes.

2.2 Optic Disc Imaging

Using engineered software, researchers have attempted to auto-segment the disc and cup margin using hard-coded algorithms to ascertain the CDR.[71-73] Peripapillary atrophy and vessel obscuration create major challenges to auto-segment the disc and cup margins. The problem is actually particularly pronounced in annotating the normal cup, which is generally small and with high vascular density. Errors in auto-segmenting the disc and cup contours create challenges to accurately identify the glaucoma-like disc as defined by ISGEO criteria. Researchers have successfully circumvented auto-segmentation problems by training neural networks to recognize the disc with user-defined threshold features for glaucoma referral.[10,14] As shown in Ting et al and Li et al,[10,14] clinicians diagnose glaucoma suspect from the optic disc images for CDR and glaucomatous changes in a training set and a neural network can recognize those photos that meet a predefined threshold for cupping associated with glaucoma with >90% accuracy **(Table 4).** In this setting disc images do not require segmentation during an unsupervised assessment of whether or not they are glaucomatous **(Figure 5).**

In 2018, the use of DL to detect a glaucoma suspect and glaucoma has moved beyond the use of optic nerve photos to detect eyes with CDRs above a pre-selected cutoff. Shibota et al., using 3242 fundus images, was able to train a CNN to detect the definitively glaucomatous optic nerve with an AUC of 0.965**.**[74] The CNN was trained to detect focal disc notching, cup excavation, retinal nerve fibre layer atrophy, disc haemorrhage and peripapillary atrophy, all signs which may occur at CDRs below pre-selected criteria. Using 1758 Spectral Domain OCT images Asaoka was able to detect early glaucoma with an AUC of 0.937 (Sensitivity = 82.5% and Specificity = 93.9%).[75] Interestingly ultra-wide scanning laser ophthalmoscopy is gaining popularity in the detection of DR and fine optic disc details are captured in these images. Masumoto et al. used 1379 Optomap images to detect glaucoma overall with 81.3% sensitivity and 80.2% specificity; values were higher for more severe glaucoma **(Table 4).**[76]

2.3 Visual Fields

Relative to optic disc photographs or OCT images, the data contained in visual field (VF) tests have low dimensionality and high noise. Nonetheless VFs represent an important endpoint in glaucoma clinical trials and VF findings will likely influence glaucoma diagnosis and guide clinical care for the foreseeable future. While the Glaucoma Hemifield Test (GHT) on the Humphrey VF represents a supervised algorithm that is useful in defining glaucoma, DL systems would be useful to define and quantify patterns of VF loss so that minimal thresholds for defining glaucoma could be established. Elze et al. developed an unsupervised algorithm termed archetype analysis to identify VF loss patterns that include glaucomatous and non-glaucomatous deficits and provide weighting coefficients for these patterns.[77] This algorithm has been validated[78] and has proven useful in augmenting the GHT for the detection of early functional glaucomatous loss.[79] Using an entirely different strategy, Li et al trained a CNN to learn the Pattern Deviation probability plots of normal and glaucomatous eyes and was able to detect glaucoma with 93.2% sensitivity and 82.6 sensitivity.[80] Yousefi et al. used an alternative Gaussian mixture and expectation maximization method to decompose VFs along different axes to detect VF progression.[23] This approach was as good or superior to current algorithms, including Glaucoma Progression Analysis, Visual field Index and Mean Deviation slope, in detecting VF progression.

2.4 Clinical forecasting

Kalman filtering (KF) is a ML technique that filters out noise in serial measures of a parameter to forecast trends over time. Glaucoma is generally a chronic slowly progressive disease whose trajectory is influenced by serial IOP, as well as changes in functional and structural data. Researchers at University of Michigan used longitudinal data on IOP and VFs to accurately forecast VF progression for participants in the Collaborative Initial Glaucoma Treatment Study.[81] Using a similar approach on a clinical based sample of Japanese normal tension glaucoma patients, KF was better able to predict 2-year MD forecast than linear regression of MD.[82]

2.5 Clinical and technical challenges in translating the technology

Care is needed to implement promising DL algorithms in the clinic setting so as to empower rather than entangle doctors as well as address patient concerns and improve their clinical care experience. In the screening environment there will be concerns about whether DL algorithms will be robust across the various platforms used to acquire information on glaucoma patients. The role of tonometric data and how to acquire data about the filtration apparatus in the screening setting could be important in populations where the prevalence of elevated IOP[83] and angle closure is high.[62] Furthermore, as is true in DR screening, there will be concerns about whether the algorithms will be applicable to the specific patient populations where they are employed. These challenges seem surmountable if careful planning and beta testing is employed.

2.6 Future directions

Currently, much work is needed to improve AI glaucoma detection algorithms. In the area of imaging, OCT technology demonstrates that the disc edge is best defined based on Bruch's membrane opening (BMO) and clinicians are not well trained to find this landmark on fundus photos.[84] Thus validation of DL systems to detect the glaucoma-like disc may require that training sets contain paired OCT images so that proper ground truth regarding disc margin contour be established. This will help establish the most accurate standardized assessment of CDR. DL systems should account for disc color and textural information embedded in pixel-rich fundus images so that they can detect non-glaucomatous optic nerve disease and leverage the fact that nerve fibre layer atrophy accompanies optic nerve degeneration. Rather than detect the disc with arbitrary CDR cutoffs, more work is needed to calibrate DL systems to detect the disc with manifest VF loss is also needed. Finally, more work on incorporating OCT data into DL algorithms to detect pathologic optic nerves as well as progressive structural damage is needed.[85] Algorithms that not only ascertain if there is optic nerve pathology but the regional location of pathology would be widely accepted.

With respect to VFs, more work is needed on unsupervised approaches to detect VF progression, which is important in the clinical setting. In addition, glaucoma forecasting will also be useful in the clinical setting and could be refined by considering factors other than serial IOP and HVFs, like patient demographics, family history, past medical history, genetic risk scores as well as other ocular parameters. Ultimately, we are likely to see hybrid methods that incorporate structure, function, and non-ophthalmic parameters like genetic risk score into algorithms that predict diagnosis, guide treat and offer prognosis for glaucoma patients.

3. **AI for Age-related Macular Degeneration (AMD)**

3.1 Clinical Unmet Need

AMD is another major cause of vision impairment, accounting for 8.7% of all blindness worldwide **(Table 2).**[86-89] The age-related eye disease study (AREDS) classified AMD stages into none, early, intermediate and late AMD. In a systematic review consisting of ~130,000 individuals from 39 studies, the pooled prevalence of any, early and late AMD were 8.69% (95% CI 4.26-17.4), 8.01% (3.98-15.5), 0.37% (0.18-0.77), respectively.[88] It is projected that 288 million may have some forms of AMD by 2040, with approximately 10% having intermediate AMD or worse.[88] The treatment for neovascular AMD patients has been revolutionized with the advent of anti-vascular endothelial growth factors (VEGF),[90,91] with many countries, e.g. US, Australia, reporting a significant drop in incident blindness by >50%. [92,93] With the ageing population, there is an urgent clinical need to have a robust DL system in detection of these patients for further evaluation in the tertiary eye care centers.

The intermediate stage of AMD is often asymptomatic, characterized by numerous medium-sized drusen or at least 1 large druse or geographic atrophy of retina pigment epithelium that does not involve the fovea. Left untreated, the advanced choroidal neovascular form of AMD can lead to substantial central vision loss in most individuals with at least half having fellow eye involvement in the advanced stage within five years of the first eye involvement. The clinical presentation of AMD includes drusen, retinal pigment epithelium (RPE) abnormalities, geographic atrophy

(GA) or choroidal neovascularization (CNV) with subsequent scarring. The Age-related Eye Diseases Study (AREDS) classified AMD stages into early, intermediate and late stages, based on drusen and other characteristics.[94] The American Academy of Ophthalmology (AAO) recommends an examination for those with the intermediate stage of AMD at least every 2 years, as most of these patients are usually visually asymptomatic, but have a higher risk of developing advanced AMD than individuals without the intermediate stage. An automated AMD screening algorithm for detecting cases of AMD that require management, i.e., either the intermediate stage or advanced stages for which follow-up is therefore desirable.

3.2 AREDS dataset

Many of the AI systems for AMD were built using the AREDS dataset,[16,95] while some utilized other datasets (e.g. Singapore Eye Research Institute).[10] The AREDS was a multi-center double-masked clinical trial, involving 4613 participants, recruited across 11 clinical centers designed to assess the clinical course, prognosis, and risk factors for AMD and cataract.[94] A total of 66,943 macula-centered images (baseline and follow-up) were used from this study. Based on the retinal lesions, AREDS classification proposed 2 classification scales: 1) 4-step - none, early, intermediate and late AMD; and 2) 9-step severity scale, which was based on outcome data, provided predictive variables for 4-year risk of developing choroidal neovascularisation, central geography atrophy or both.[96,97] AREDS grade 1 indicates that fundus images with little or no AMD changes, while AREDS grade 2 through 9 represent changes associated with early or intermediate AMD. AREDS grades 10 through 12 represent late-stage AMD, namely GA, neovascular AMD and images with both late-stage forms. Of these, 44.6% had referable AMD, defined as intermediate AMD or worse.

3.3 Fundus images-based DL systems

In the AREDS 1 dataset, Burlina et al reported an AUC between 0.94 and 0.96 with accuracy between 88.4% and 91.6% in detection of referable AMD, using a 5-fold cross validation and pre-trained AlexNet and Overfeat CNNs **(Table 5).** Using the same dataset, they estimated 5-year risk of AMD progression, with weighted k scores of 0.77 for 4-step severity scales and overall mean estimation error between 3.5% and 5.3%.[95]

Similarly, Grassmann et al built a DL system for detection of early and late AMD, using 6 different CNNs.[16] Early AMD was defined as AREDS grade 2 to 9, while late AMD was defined as AREDS grade 10-12 (GA and neovascular AMD). Using the 9-step AREDS severity scale, the authors reported an accuracy of 63.3% in predicting 13 classes in the AREDS test set with a quadratic weighted k of 92% (95% CI: 89%-92%). This algorithm was validated using the Augsburg dataset, consisting of 5,555 fundus images that were collected as part of the collaborative health research in the region of Augsburg, Germany. There was a reduction in weighted and unweighted k values due to the 313 false positive fundus images, that are deemed to have neovascular AMD (AREDS class 11), but that were actually images from healthy individuals, with most showing dominant macula reflexes. These patients were detected as 'abnormal' as the inclusion criteria for AREDS datasets were individuals aged 55 years or older. This, again, shows the importance of selecting the right screening population for the AMD algorithm in the clinical setting.

Developed using VGGNet and Singaporean population-based cohort, Ting et al reported an AUC, sensitivity and specificity of 0.931 (0.928-0.935), 93.2% (91.1%-99.8%) and 88.7% (88.3-89.0), respectively, in detection of referable AMD on the testing dataset, on a 2-year diabetic cohort (2014 and 2015) recruited from Singapore National DR Screening Program.[10] This algorithm, however, was not tested on a white population. Vice versa, the former 2 groups also did not validate their algorithms in Asian populations. In terms of the technical methodologies, Burlina et al[15,95] performed auto-segmentation on the macular region while the latter 2 DL systems analyse the entire retinal image,[10,16] although diagnostic performance were still comparable between the 3 DL systems using the respective testing datasets. Future research is important to evaluate the generalizability and cost-effectiveness of these DL systems in a larger international multi-ethnic cohort.

4. AI for Retinopathy of Prematurity

4.1 Clinical Unmet Need

ROP is a retinal vascular disease affecting premature infants, characterized by abnormal fibrovascular proliferation at the boundary of the vascularized and avascular peripheral developing retina. Globally, it is estimated that 15 million babies are born prematurely each year.[98] In US, the incidence of ROP was 19.9% **(Table 2).**[99] ROP accounts for 6 to 18% of childhood blindness,[100] causing significant psycho-social impact on the child and the family.[101] According to the Early Treatment for ROP (ETROP) trial,[102] early treatment has shown to be beneficial to improve the visual acuity of high-risk ROP patients, although 9% still eventually became blind. Thus, early screening with regular monitoring is crucial.

ROP diagnosis is traditionally performed by indirect ophthalmoscopy at the neonatal intensive care unit (NICU) bedside, and is increasingly being performed by telemedicine interpretation of wide-angle retinal images.[103-105] Clinical diagnosis is based on parameters defined by the international classification of ROP (ICROP): zone, stage, clock hour extent, and plus disease. “Plus disease” is defined as venous dilatation and arteriolar tortuosity in central retinal vessels greater than or equal to that of a standard published photograph.[106-108] The 2005 revised ICROP defined a newer “pre-plus” category as posterior pole vessels that are not normal but with less than the required amount of vascular abnormality.[108] Based on findings from the NIH-sponsored multicenter Cryotherapy for ROP (CRYO-ROP) and Early Treatment for ROP (ETROP) trials, presence of plus disease has been shown to be the key factor in identifying infants with severe treatment-requiring disease at risk for blindness.[102,106] **Figure 6** displays examples of normal vessels, pre-plus disease, and plus disease. Therefore, it is critical to diagnose plus disease accurately and reproducibly.

4.2 Challenges in ROP Diagnosis

There are a number of challenges with the current approach to ROP diagnosis. From a public health perspective, the number of premature infants at risk for ROP is increasing due to a rising number of preterm births and increased neonatal survival, particularly in the developing world.[109] Meanwhile, the supply of clinicians who perform ROP management is limited by logistical challenges of coordinating examination at the NICU bedside, low physician reimbursements, and extensive medicolegal liability. From an educational perspective, training in ROP diagnosis is often inadequate, further limiting the workforce of ophthalmologists trained to manage this disease.[110-113]

In particular regarding clinical care, there are a number of real-world challenges regarding plus disease diagnosis: (1) There is often significant variability in diagnostic classification (plus vs. pre-plus vs. normal), even among experts,[114-118] leading to inconsistent application of evidence-based practice.[119] This has occurred even in NIH-funded multicenter trials. For example, in the CRYO-ROP protocol, confirmation of threshold disease was required by a second unmasked certified examiner performing dilated ophthalmoscopy. In that setting, the second examiner disagreed with the first examiner regarding clinical diagnosis of threshold disease in 12% of cases.[120] Also, in a multi-center study of telemedicine for ROP diagnosis, nearly 25% of examinations by certified study graders required adjudication because the graders disagreed on one of three criteria for clinically-significant ROP.[121] (2) There is significant variability in diagnostic process among experts, who have been shown in observational studies to consider different retinal vascular features during assessment of disease severity.[122] (3) There is evidence that experts frequently deviate from the published definition of plus disease when assessing ROP, for example by considering factors such as venous tortuosity and peripheral retinal vascular features.[122-125] (4) The published standard photograph for plus disease was from the 1980s, and has a much smaller field of view and larger magnification than clinicians are accustomed to seeing during standard examination methods using indirect ophthalmoscopy or wide-angle retinal images. There is evidence that this causes bias and inconsistency in diagnosis.[126] (5) Studies have shown that there is geographical variation in plus disease diagnosis possibly related to differences in training,[119] and that there may be chronological drift showing a tendency to diagnose "plus disease" more frequently than in the past.[127] (6) The multicenter Supplemental Therapeutic Oxygen for Prethreshold ROP (STOP-ROP) study defined that plus disease is present if there is sufficient venous dilation and arterial tortuosity in at least 2 quadrants, and this definition was incorporated into the 2005 revised ICROP.[108,128] However, there is variability in how this definition is interpreted,[115-117,122] and evidence that this variability may lead to clinically-significant differences in diagnosis.[116,129] (7) The ICROP definition of pre-plus disease[108] is somewhat vague. Studies have found significant levels of variability in diagnosis of pre-plus disease among experts.[114,115] (8) Vascular abnormality in ROP reflects a continuous spectrum of disease,[108,130,131] whereas clinical management is based on a discrete classification (e.g. “plus disease” vs. “not plus”) from findings of clinical trials, which requires determining cut-points for abnormality.[106,120] Research suggests that diagnostic discrepancy results from individual clinicians having different cut-points (e.g. “is this plus or pre-plus disease”), despite having better agreement on relative disease severity (e.g. “which retina looks worse”).[132,133]

4.3 Early Approaches to Image Analysis for ROP

Early approaches to computer-based image analysis for plus disease diagnosis were based on quantification of vascular tortuosity and dilation (RetCam; Natus Medical Incorporated, Pleasanton, CA).[134] Three such systems have been developed and validated for wide-angle RetCam images: ROPTool, Retinal Image multiScale Analysis (RISA), and Computer-Assisted Image Analysis of the Retina (CAIAR).[135-137] These systems have been evaluated against expert diagnostic performance, but have not had real-world application because of limitations such as being semi-automated (e.g. requiring manual identification of optic disc or key vascular segments), or having limited correlation with two-level expert diagnosis (plus disease vs. not plus).

More recently, one system (Imaging & Informatics in ROP, i-ROP) was developed based on machine learning methods, in which a vascular metric termed “acceleration” was found to have best diagnostic performance in a 6 disc-diameter circular crop of wide-angle RetCam images considering all retinal vessels combined.[138] This system had 95% accuracy for 3-level plus disease diagnosis (vs. pre-plus or normal) in a test set of 77 images, compared to a reference standard defined by combining ophthalmoscopic examination by 1 expert with image-based examination by 3 experts. For the same test set of 77 images, 3 individual experts had accuracy of 92-96%, and 31 non-experts had mean accuracy of 81%. However, real-world application of this system has been limited by the requirement for manual segmentation of images.[138]

4.4 Deep Learning for ROP

DL has been applied for automated diagnosis of ROP, which could potentially address barriers to ROP screening on a larger scale.[139] Most recently, Brown et al developed and validated a fully-automated DL system (i-ROP DL) for 3-level plus disease diagnosis (plus vs. pre-plus vs. normal) with an area under the ROC curve of 0.98 for plus disease diagnosis compared to a reference standard defined by combining ophthalmoscopic examination by 1 expert with image-based examination by 3 experts. When evaluated in an independent test set of 100 wide-angle RetCam images, the i-ROP DL system achieved 93% sensitivity and 94% specificity for diagnosis of plus disease, and 100% sensitivity and 94% specificity for diagnosis of pre-plus or worse disease. When compared to 8 international ROP experts evaluating the same 100-image test set, the i-ROP DL system agreed with the consensus diagnosis more frequently than 6 of the 8 experts.[17]

4.5 Future Directions

AI has potential to create assistive technologies to improve the accuracy and consistency of ROP diagnosis by clinicians. In the future, this could produce quantitative ROP severity scores to facilitate objective monitoring of disease progression and treatment response. Future automated systems might provide initial readings of images captured by neonatal intensive care unit nurses, thereby reducing the requirement for traditional ophthalmoscopic examinations in the majority of infants without clinically-relevant disease. These methods may be particularly applicable to the developing world, where the availability of ophthalmology and neonatology expertise may be insufficient to manage the number of premature infants at risk for ROP.

5. AI for Cardiovascular Disease

5.1 Clinical Unmet Need

Cardiovascular diseases (CVDs) is the largest cause of non-communicable deaths worldwide. For 2018, World Health Organization (WHO) estimated that 17.9 million people died of CVD worldwide in 2012, accounting for an estimated 31% of global mortality **(Table 2)**.[140] Of those, ischemic heart disease (IHD) and stroke are the top cause of mortality, responsible for approximately 85% CVD deaths, with >75% occurring in low- to middle-income countries.[141,142] To prevent heart attacks, strokes, and other adverse cardiovascular events, it is important to identify the systemic risk factors, that can be divided into non-modifiable (e.g. age, sex) and modifiable factors (e.g. smoking, hypertension, hyperlipidemia). In the routine clinical setting, many physicians utilized risk calculators, such as the Pooled Cohort equations,[143] Framingham[144-146] and SCORE,[147,148] aiming to perform risk assessment and stratify individual risk to inform treatment decisions. If existing CVD risk assessment tools could be improved, more patients can benefit from early treatment, while minimizing the harms of screening.[149]

Given the ageing population, the clinical unmet need will continue to rise over the next few decades. Most screening programs will face shortage of manpower and infrastructure, especially in the low-to-middle income countries. Thus, there is an urgent call for action in exploring novel and economical screening technologies for these conditions. Cardiovascular disease (CVD) risk assessment is a critical first step in managing and preventing heart attacks, strokes, and other adverse cardiovascular events. Clinicians often utilize risk calculators, such as the Pooled Cohort equations,[143] Framingham[144,145,150] and SCORE,[147,151] which is based on various factors from patient history (e.g. age, self-reported sex, smoking status) and blood samples (e.g. lipid panels).[152] Given that obtaining these values require a blood draw and fasting prior to the procedure, some of these parameters such as cholesterol values may be sparsely available[153].

5.2 Manual grading for retinal vascular Imaging to predict CVD risks

There have been many efforts to improve risk prediction, particularly in incorporating phenotypic information to further refine risk prediction such as the addition of coronary artery calcium[154] or retinal imaging. The retina is unique in that it is one of the only places in the body where vascular tissue can be visualized quickly and noninvasively. Conditions associated with CVD, such as hypertensive retinopathy and cholesterol emboli, can often manifest in the eye. Previous studies have shown that various retinal features may be predictive of cardiovascular events, stroke[155] or chronic kidney disease.[156] These features include vessel caliber,[157-159] bifurcation or tortuosity,[160] Currently, the assessment of such features requires expert assessors going through a fairly long and detailed procedure. For example, to measure vessel diameters, expert assessors must segment vessels, identify specific segments and adjudicate variations, a fairly time-consuming process to measure just one feature of the image. While the previous work in this field is promising, the clinical utility of such features still requires further study.

5. CVD risk prediction using Retinal Imaging and DL

One of the key benefits of DL is the ability to learn the appropriate predictive features directly from the raw examples, rather than requiring features to be hand-engineered. In the context of retinal imaging, what this means is that we can provide all the pixel values from a retinal fundus image with minimal processing as the input to a deep learning model, and train it to predict the desired label (for example the ICDR DR Grade) for that image.

In a recent study, Poplin and Varadarjan *et al*[19] used DL to build a model that predicted cardiovascular risk factors using retinal fundus images **(Table 6)** from 48,101 patients from the UK Biobank study[161] and 236,234 from the EyePACS population.[162] The UK Biobank population was predominantly Caucasian without diabetes while the EyePACS patients were predominantly Hispanic with diabetes. These models were then validated using images from 12,026 patients from UK Biobank, 999 patients from EyePACS, and on an independent cohort of Asian patients **(Table 8).**[163] The model was fairly accurate for some predictions such as age, self-reported sex, blood pressure, and smoking status. In addition, the authors also trained a model to predict the onset of major adverse cardiovascular events (MACE) within 5 years using the UK Biobank study. For this, MACE was defined as the presence of billing codes for unstable angina, myocardial infarction, or stroke or death from cardiovascular causes. Participants that had a MACE prior to the retinal imaging were excluded. Because the UK Biobank recruited relatively healthy participants, MACE were rare (631 events occurred within 5 years of retinal imaging--105 of which were in the clinical validation set). Despite the limited number of events the model achieved an AUC of 0.70 (95% CI: 0.65, 0.74) from retinal fundus images alone **(Table 8),** comparable to the AUC of 0.72 (0.67, 0.76) for the European SCORE risk calculator. Because cholesterol levels were not available at the time of the study, BMI was used as a proxy while calculating the SCORE risk.[164-166]

An explanation technique for DL models called soft-attention was used to identify relevant anatomical regions that the model may be using to make its predictions. This generated a heat map showing the most predictive pixels in the image. A representative example of a single retinal fundus image with accompanying attention maps[167] for a few predictions is shown in **Figure 4**.

5.4 Future direction

Despite these promising results, efforts to improve the performance and interpretability of these DL models seems indicated, especially for MACE. First, the Poplin and Varadarajan *et al* study did not include blood tests such as lipid panels in the analysis because it was not available for the study.[19] Future work should include these important clinical factors. A substantially larger dataset or a population with more cardiovascular events may enable more accurate DL models to be trained and evaluated with high confidence. Training with larger datasets and more clinical validation will help determine whether retinal fundus images may be able to augment or replace some of the other markers, such as lipid panels, to yield more accurate predictions.

6. AI for refractive error

6.1 Clinical Unmet Need

Uncorrected refractive error is a major cause of visual impairment that affects a large proportion of the world population.[168] Uncorrected refractive error is commonly defined as visual acuity of less than 6/12 in the better eye with improvement of at least 0.2 logMAR (equivalent to 2 lines) after refraction. According to the global burden of disease study, 101.2 million cases of moderate and severe visual impairment and 6.8 million cases of blindness were due to uncorrected refractive error **(Table 2).**[169] It is estimated that visual impairment secondary to refractive error resulted in close to 269 billion international dollars loss on the potential productivity cost, mainly attributable to lack of awareness and shortage of optometry expertise. Given its implication, the screening and correction of refractive error is important.

6.2 AI in predicting refractive error

As discussed in the previous examples with cardiovascular disease and retinal imaging, DL has also shown great promise in discovering new associations from imaging or quantifying known associations to a high level of accuracy. Another example of this is the recent work done in applying DL for refractive error. While physicians would generally have difficulty predicting refractive error from a retinal fundus image, DL techniques are able to predict this fairly accurately. Varadarajan and Popin *et al*[170] showed that deep learning can be used to train algorithms with a mean absolute error (MAE) of 0.56 D (95% CI: 0.55, 0.56), and R^2 of 0.90 (95% CI: 0.90, 0.91) using images taken with a 45 degree field of view as the input data **(Table 9).** Given this somewhat surprising finding, the authors also went on to leverage attention maps to determine the parts of an image most relevant for the prediction. They found that the attention maps consistently highlighted the fovea as a feature that was important for the prediction **(Figure 7).** The model also frequently highlighted retinal vessels and cracks in retinal pigment. The model seemed to predict only the spherical component of refractive error well **(Table 9).** The accuracy of the refractive error prediction seemed to decrease with a smaller field of view, poorer image quality, and possibly macular lesions.

6.3 Future Directions

The ability to train accurate models without feature engineering combined with explanation techniques make DL an attractive tool for scientific discovery. Improvements in and experimentation with other explanation methods for DL models will help us understand these novel signals. While these heatmaps can serve as starting points, other techniques can be leveraged to further help explain model predictions -- such as selectively including or excluding parts of the images during training to measure the relative importance of each of these regions to the prediction task. The identification of new features creates new research opportunities for better understanding of the development and management of disease. For researchers, instead of first guessing and then testing hypotheses one by one, they could use neural networks to directly make the prediction of interest and then utilize attention techniques to generate targeted hypotheses. For clinicians, this work also suggests that large datasets could be leveraged to fuel the development of new non-invasive imaging biomarkers for a variety of diseases, from ophthalmological to systemic diseases.

7. AI for Optical Coherence Tomography for Retinal Diseases

7.1 Clinical Unmet Need

OCT has established itself as the dominant imaging modality across ophthalmic disciplines, particularly for the diagnosis and management of retinal disease.[171] 30 million ophthalmic OCT procedures are now performed every year, a figure comparable in scale to other medical imaging such as magnetic resonance imaging (MRI) or computed tomography (CT), and which is more than the sum of all other ophthalmic imaging modalities combined.[172] By allowing personalized therapy for just one retinal disease – neovascular AMD – it is estimated that OCT imaging has saved the United States government at least $9 billion.[173] OCT imaging is increasingly being used in other areas of ophthalmology also, with applications for the care of patients with cataract, corneal disease, uveitis, oculoplastics, and glaucoma. On the commercial side, by 2016, the estimated revenues from ophthalmic OCT systems had reached $1 billion per year.[173] Although widely adopted first in ophthalmology, the use of OCT is expanding to other medical specialties, including neurology, dermatology, cardiology, and gastroenterology, as well as non-medical fields such as archeology, art conservation, and industrial non-destructive testing. Since 2013, there has even been an ophthalmic OCT system on the International Space Station (ISS) as part of the NASA Optical Health Study, investigating the visual impairment that commonly occurs in those exposed to microgravity for long duration space flight.[174]

From an early point in its commercial development OCT imaging has been innovative in its incorporation of automated medical image analysis techniques.[175] Unlike other ophthalmic imaging modalities, which have typically only allowed, at best, semi-quantitative assessments of disease, OCT provides automated measurements of retinal thickness for macular disease and of retinal nerve fiber layer (RNFL) thickness for glaucoma. Ophthalmic clinical trials quickly adopted these measures as inclusion- and retreatment-criteria, and their application in applied clinical research has both clarified many aspects of retinal disease pathophysiology and elucidated many hitherto unrecognized disease characteristics. As our knowledge of OCT image analysis has grown, however, it has become increasingly clear that accurate measurements of retinal thickness may fail to predict visual outcomes.[176,177] Much of the focus of recent research has been on the identification of more novel OCT-derived anatomic biomarkers. The discovery of such biomarkers may provide valuable information regarding therapeutic mechanisms of action, pharmacodynamics, and pharmacokinetics for clinical trials.[178,179] If such biomarkers are shown to predict clinical benefit, they could also serve as surrogate endpoints in these trials, potentially leading to increased accuracy, reduced costs, and shortened trial duration. DL systems would appear to have the most potential at present as the tool that can unlock these advances.

While OCT imaging has clearly revolutionized ophthalmology, its widespread usage also presents challenges and risks for healthcare delivery. In large tertiary referral hospital eye services, OCT imaging may be performed 1000s of times per week and the timely review and actioning of scans can be logistically challenging.[180] This is particularly important for diseases such as neovascular AMD where irreversible visual loss may occur if there are delays in the early initiation of treatment (in the United Kingdom alone, nearly 200 people develop the severe forms of AMD every single day; the Royal College of Ophthalmologists recommends that such patients should be seen and treated within two weeks of their referral to an ophthalmologist).[181,182] These pressures are likely to increase as OCT becomes used increasingly in the community and – in the future – potentially even in the home. In May 2017, one of the world's leading optometry chains, Specsavers, announced that they would be rolling out OCT imaging devices across all of their branches.[183] Given that Specsavers have 740 branches in the UK alone, and perform nine million eye examinations per year, this may place additional pressure on an already overstretched healthcare system unless systems can be put in place to appropriately triage those patients with suspected eye disease. Of course, if such systems can be established, it will likely have huge benefits for patients as diseases are picked up at an earlier stage and thus treated promptly. It seems likely that DL systems, along with the implementation of appropriate referral pathways, will be crucial to these changes.

7.2 Lesion Detection and Segmentation

Much of the initial work in the application of DL to OCT image sets has related to lesion detection (the process of starting with an unlabelled OCT B-scan or volume and marking potential abnormalities) and segmentation (the delineation of margins of any structure, abnormal or otherwise). The term "semantic segmentation" is often used to describe the analysis of every data-point (pixel/voxel) in an image set and its assignment to a specific label class. From an early stage in its development, automated segmentation of OCT images – using classical computer vision techniques such as thresholding and graph search – has provided objective, reproducible, and quantitative measurements of central retinal thickness.[175] Unfortunately, however, these approaches are prone to error, particularly in cases of retinal disease where there is severe disruption of the normal retinal morphology, such as neovascular AMD.[184,185] Aside from retinal thickness, the segmentation of more specific disease features requires expensive, time-consuming, manual segmentation of OCT scans in dedicated image reading centers, an approach typically only possible in the context of large-scale clinical trials and not at all in routine patient care. In recent years, the introduction of DL systems for such tasks has shown great promise in addressing this issue.

Generic CNN architectures are now commonly customised for medical image segmentation tasks. One such adaptation, the U-net, has achieved particular success due to its flexibility in input sizes and dimensionality and its ability to produce good results even with relatively small amounts of training data.[186] A number of methods are typically used to assess segmentation accuracy. In the computer vision and machine learning literature, the most common metrics used are the Dice coefficient, which measures the overlap between automated and "gold standard" manual segmentation, or the Jaccard index ("intersection over union"), which measures the similarity between two datasets.[187] By contrast, in the clinical literature, particularly in ophthalmology, the agreement between automated and manual segmentation is most commonly measured using Bland-Altman plots.[188] These approaches assume that a high-quality, ground-truth measurement can be ascertained. In many cases, segmentation of retinal OCT images is challenging, even following an adjudication process between the best human experts. This problem may also be more pronounced when novel biomarkers, without well establishment grading protocols (e.g., subretinal hyperreflective material (SHRM)), are being evaluated. In such situations, it is important that the variability of manual segmentation between experts is well defined as a comparator. Finally, it is important that DL segmentations approaches are externally validated on large, heterogeneous real-world clinical datasets that will be representative of their ultimate clinical use case.

7.3 Retinal and choroidal thickness segmentation

Segmentation of the neurosensory retina and its sublayers is not only important for the assessment of macular diseases but also for patients with glaucoma and other neuro-ophthalmic diseases. In May 2017, Fang et al. described the combination of a DL system with a more traditional graph theory approach for the delineation of nine retinal sublayer boundaries on Spectralis OCT (Heidelberg Engineering, Germany) images.[189] A CNN was used to generate class labels and probability maps for each of the layer boundaries – the graph search approach was then used to create the final boundaries. The model segmentation outputs were compared against a semi-automated gold standard (initial automated segmentation using the existing DOCTRAP software, followed by manual corrections of errors by a human expert). For all layers, the mean difference (in pixels) between the automated and manual segmentation outputs was then calculated – although useful for algorithm development the clinical significance of such a measure is harder to determine. In July 2017, Venhuizen et al. described an approach which removed the need for an additional graph search step, and which had been trained on a larger dataset of patients with more advanced AMD (the authors term their approach "robust" total retinal thickness segmentation as existing automated segmentation systems have often been shown to fail in this setting).[190] Their system utilises a generalized U-net architecture to provide automated segmentation of the inner and outer retinal boundaries and hence allows automated measurements of central macular thickness. The authors report that their DL system accurately estimated macular thickness with an error of 14.0 ± 22.1 μm when compared with manual segmentation (versus 42.9 ± 116.0 μm and 27.1 ± 69.3 μm for existing commercial and research algorithms that provide automated segmentation, respectively).

In July 2018, Hamwood et al. evaluated the effect of retinal image patch size and network architecture on the performance of a model trained to segment retinal sublayer thicknesses.[191] To address the issue of label imbalances (i.e., the predominance of non-lesion pixels/voxels in medical images), image "patches" are commonly used rather than single slices or whole volumes. This has a further advantage of increasing the amount of data available for training, a fundamental requirement for DL systems. Hamwood et al., demonstrated how increasing patch size (65 x 65 versus the 33 x 33 pixel patch described by Fang et al.) can improve the performance of the segmentation.

Commercially available OCT systems – especially those that provide greater depth penetration through the use of longer wavelength swept source lasers - have begun to offer automated segmentation of the choroid, with choroidal thickness maps. Deep learning-based approaches to choroidal segmentation have recently been reported. Chen et al., described the use of two CNNs – one to delineate the inner boundary of the choroid (Bruch membrane) and one to identify the outer boundary (the choroid-sclera interface).[192] They reported good results (a Dice score of 0.82) in a small cohort of patients with AMD. Considerable further work will be required to develop and validate such algorithms for automated choroidal thickness measurement, particularly in diseases where this parameter is likely to be of most direct clinical benefit such as posterior uveitis, central serous chorioretinopathy, and polypoidal choroidal vasculopathy.

7.4 Disease Detection

7.4.1. Macular edema

Segmentation of macular oedema – the accumulation of fluid in the extracellular space of the retina (intraretinal fluid (IRF)) and between the retina and retinal pigment epithelium (subretinal fluid (SRF)) - is likely to have greater importance in retinal diseases such as DME, retinal venous occlusion (RVO), and neovascular (“wet”) AMD. In July 2017, Lee et al., described the use of DL for segmentation of IRF in OCT images **(Table 10).**[193] They collected a large cohort of Heidelberg Spectralis OCT images from eyes most likely to have IRF, including DME, RVO, and AMD. For OCT volume, they selected a central slice for manual segmentation; for this, they defined IRF as “an intraretinal hyporeflective space surrounded by reflective septate”. 934 manually segmented OCT images, divided into 1,919,680 images patches, were then used as a training set on a modified U-net architecture. A final test set of 30 images were segmented by four independent clinicians. Good results were then seen, with Dice coefficients for human interrater reliability and the DL system being 0.750 and 0.729, respectively. Roy et al. have also reported good results using a U-net architecture for IRF segmentation, albeit using a smaller dataset of 110 manually segmented OCT B-scans solely from 10 patients with DME.[194] In April 2018, Venhuizen et al. described use of a U-net architecture to provide automated segmentation of a wide variety of intraretinal cystoid spaces, ranging from small micro-cysts to larger intraretinal cystoid spaces spanning a wide area of any given OCT B-scan.[190] Importantly, they showed that their system was capable of being applied with good results to OCT scans from four OCT systems of four different OCT vendors.

In April 2018, Schlegl et al., reported a DL system capable of providing automated segmentation of both IRF and SRF, across three different macular diseases (neovascular AMD, DME, RVO), and imaged with two commonly used systems (Zeiss Cirrus and Heidelberg Spectralis).[195] They trained a separate model for each OCT system, using 200 cases for the three disease categories, achieving accuracies in the range of interobserver variability reported for experts in the literature. They evaluated both the binary detection of fluid (present versus not present) as well as its quantification. Training cases were selected from clinical trial data at the Vienna Reading Centre rather than real-world clinic data. A limitation of this work is that only scans with a clear consensus annotation between Vienna Reading Centre graders were taken into the sample; furthermore, scans with low image quality were also excluded.

7.4.2 Pigment epithelium detachment, subretinal hyperreflective material, and the retinal pigment epithelium

More recently, some groups have extended their models to perform segmentation of pigment epithelium detachment (PED), the formation of a potential space between the retinal pigment epithelium (RPE) and Bruch's membrane.[196,197] Schmidt-Erfurth et al. have reported the correlation of PED metrics with visual acuity in patients with neovascular AMD using a DL-based system.[198] Detailed description and validation of this PED segmentation approach has not yet been published but it appears to treat PED as a single entity rather than a range of specific subtypes. This single PED entity was not found to significantly affect visual acuity in these cases.

In August 2018, de Fauw et al., reported a DL system which considerable extends the range of clinically relevant retinal parameters from that previously described.[22] Rather than taking as input image patches from single OCT B-scans, this model utilized nine contiguous slices from the OCT volume. A three-dimensional U-net architecture was then used to output automated segmentations across 15 different label classes, and across the three most commonly used OCT platforms (Topcon 3D-OCT, Heidelberg Spectralis, Zeiss Cirrus). These labels encompass a range of novel OCT biomarkers, including three forms of PED (fibrovascular, serous, and drusenoid) and subretinal hyperreflective material (SHRM), an emerging morphologic parameter which may be central to the fibrotic process in neovascular AMD. This model also segments the posterior hyaloid and epiretinal membrane (ERM), to allow enhanced assessment of vitreomacular interface disorders, as well as the RPE, allowing for the quantification of retinal degeneration and atrophic changes **(Figure 8).** The authors also leverage the intermediate tissue representation created to perform image classification tasks across a range of OCT systems from different vendors, described in detail below.

7.4.3 Other Morphologic Parameters and Approaches

A number of other retinal OCT features are in the early stages of exploration using DL systems. These include retinal hyperreflective foci in patients with AMD,[199] a biomarker that is increasingly recognized as important to the progression of AMD, and the photoreceptor ellipsoid zone, a biomarker vital to the assessment of patients with inherited retinal dystrophies and rare conditions such as macular telangiectasia (MacTel) type 2.[200,201] For the photoreceptor ellipsoid zone segmentation, training labels can be generated from manual segmentation of *en face* OCT images, a potentially quicker process than dense segmentation of OCT B-scans.

7.5 Image-based Classification

DL is very well suited to image classification tasks – in a medical imaging application this can refer to multiple domains, including screening, triage, diagnosis, and monitoring of disease activity.

The first application of DL for ophthalmic OCT classification was reported by Lee et al. in 2017.[202] They aggregated 10 years of Heidelberg Spectralis macular OCT scans, acquired during routine clinical practice at their institution, and each consisting of 61 individual OCT B-scans. They then linked these images to clinical data from their electronic health record (EHR) in an automated fashion, before curating a cohort of “normal” and “AMD” scans. A normal patient was defined as having visual acuity >20/30 in both eyes and no ICD-9 recorded disease diagnosis in the EHR. An AMD patient was defined as having an ICD-9 diagnosis of AMD by a retina specialist, at least one intravitreal injection in either eye, and worse than 20/30 vision in the better seeing eye. Of note, patients with other macular pathology by ICD-9 code were excluded. The central 11 OCT B-scans from each macular OCT set were selected, labelled en bloc as either normal or as AMD, and then used independently for development of the classification model. As a result, 48,312 normal OCT B-scans (4392 OCT volumes) and 52,690 AMD B-scans (6364 OCT volumes) were included. Each image was downsized to 192 x 124 pixels due to memory limitations. At the level of each individual image, the authors achieved an 87.6% accuracy with a sensitivity of 84.6% and a specificity of 91.5%. The performance improved when they grouped images from the same OCT volume, and/or the same patient, and averaged the probabilities from each image. The authors also performed occlusion testing –systematically covering every location in the image with a blank 20x20 pixel area and evaluating the effect on model performance – to try to gain some insights into model decision making. Using this approach, the resulting saliency maps identified key areas of interest on the OCT images which corresponded to areas of known pathology.

Several groups have adopted similar methodologies to that pioneered by Lee et al., namely the mass extraction of real-world imaging data with automated labelling from EHRs, and the subsequent training of DL systems on simple, binary classification tasks. In many cases, they employ transfer learning – the use of neural networks that have been pre-trained on millions of non-medical images and which only require training of one additional layer on the medical image classification task. Using this approach, Treder et al., similarly reported good results for the classification between neovascular AMD and normal OCT scans.[203] Using the publicly-available Duke OCT dataset, Karri et al., were able to distinguish between OCT scans with AMD, DME, or with no retinal pathology.[204] In 2018, using OCT scans from a larger dataset of 4686 patients, Kermany et al., similarly used transfer learning to classify images as: 1) normal, 2) CNV, 3) DME, or 4) drusen, as well as making a referral decision.[21] A triage decision could then be made, consisting of: 1) urgent (CNV and DME), 2) routine (drusen), and 3) observation (normal). Prahs et al. did a similar mass extraction of ~30,000 OCT image sets (in their case, radial lines scans) from their EHR and trained a DL system to distinguish between an “injection” and “noninjection” group.[205] The former was defined as the delivery of an intravitreal injection, for any indication, within 21 days of the OCT scan acquisition. They reported good sensitivity and specificity on this task but correctly caution that it would not be possible to use this as a treatment recommendation. Finally, Sonobe et al., developed a DL model to classify between OCT images with epiretinal membrane (ERM) and those without.[206] These preliminary works represent important first steps, especially given that they are typically performed using relatively lightweight compute resources, and – in many cases – by clinicians without extensive DL expertise. However, in real world settings patients often present with more than one pathology at any given time so their ultimate clinical use case is less clear.

To allow true real-world clinical applicability on retinal OCT imaging, in our opinion, DL systems should fulfil a number of criteria. They should be designed with a specific clinical pathway in mind, be trained on large and heterogeneous image sets that are representative of this use case. They should also be capable of providing multi-class classifications to allow for co-existence of multiple retinal pathologies. Most importantly, they should be able to achieve performance on par with retinal specialists as well as being able to provide some measure of classification certainty for challenging and ambiguous cases. In August 2018, Moorfields Eye Hospital and DeepMind (de Fauw et al.) reported a novel DL framework which utilises one CNN for retinal OCT segmentation following by another CNN for OCT classification.[22] In this approach, a three-dimensional U-net is first used to segment a range of 15 different retinal morphologic parameters and OCT image acquisition artefacts. The output of this network is then passed to a classification network which make a referral triage decision from four categories (urgent, semi-urgent, routine, observation), and classifies the presence of 10 different OCT pathologies (choroidal neovascularization (CNV), macular oedema without CNV, drusen, geographic atrophy, epiretinal membrane (ERM), vitreomacular traction (VMT), full-thickness macular hole, partial thickness macular hole, central serous retinopathy, and “normal”. The performance of this system was then evaluated on a retrospective dataset of 1000 patients newly referred to Moorfields Eye Hospital where a macular OCT scan was performed. On the central task of referral triage, the DL system achieved an accuracy rate of 94.5% - a performance on par with experienced retinal specialists at the institution. The generation of an intermediate tissue representation by the segmentation network also brings a number of advantages. Firstly, it provides some element of interpretability to the retinal specialist when he or she reviews the triage decision and makes the final diagnosis **(Figure 9).** For example, if the clinician suspects that the classification network has reached the wrong decision, they may be able to visualize if this has occurred due to a segmentation error. By providing instant feedback, such information may also be useful to accelerate the training of ophthalmologists in OCT image interpretation. Secondly, the segmentation of numerous retinal morphologic parameters provides objective, quantitative information for monitoring of response to treatment in routine clinical practice and in clinical trials. Finally, the generation of an intermediate tissue representation by the segmentation network means that the framework can be generalized across OCT systems from multiple different vendors without prohibitive requirements for retraining (when moving to new system, only the segmentation network requires retraining). The first application of this system will be in the rapid access "virtual" clinics widely used for macular disease triage in the UK,[207] while the longer-term goal will be the use of such a system in community settings. Prior to this, further validation of the system will be required and multi-center prospective clinical trials are planned to begin recruitment in 2019.

7.6 Prediction

Outside ophthalmology, DL is being applied increasingly to predict events such as mortality, sepsis, length of stay in hospital and other clinical parameters.[208] Such tasks may be more challenging than straightforward image segmentation or classification tasks and, therefore, are at a less mature stage in development. This is particularly the case for real world dataset which are constantly evolving and which differ significantly between healthcare institutions. Currently to our knowledge, there are no prospective clinical studies that evaluate DL models in this regard in any field of medicine. Nevertheless, the capability of DL systems to model high-dimensional data – such as an OCT scan – has great potential for retinal disease. For example, it may allow prediction of disease onset or progression in AMD, or provide much better prognostic information at an individual patient-level. Preliminary work has thus far used a combination of deep learning for OCT image segmentation with classical machine learning/statistical techniques for the prediction task.

Schmidt-Erfurth et al., used the HARBOR data to develop ML models to predict visual acuity in patients receiving ranibizumab for neovascular AMD.[198] They began by selecting 70% of the HARBOR dataset for analysis. They next applied automated segmentation algorithms (using both graph-based and deep learning approaches) to the OCT scans, allowing segmentation of total retinal thickness, IRF, SRF, and PED. This allowed them to generate four morphologic maps and thus a wide range of quantitative structural variables. They then used classical ML techniques (random forest regression) to predict visual acuity at baseline and at 12 months. For the latter, they constructed separate models for the visits at baseline and then for months one to three. Of note, the ranibizumab dose and treatment regimens were included in the model as fixed effects. Their study involved 614 eyes. At baseline, the extracted OCT biomarkers – in particular, the extent of IRF – were found to predict the visual acuity with an R^2 of 21% (i.e., these variables accounted only for 21% of the variation in baseline visual acuity). As with previous studies, they found that SRF and PED did not contribute to baseline visual acuity to any meaningful extent. They also predicted visual acuity at 12 months following initiation of therapy. At baseline, their model accounted for 36% of the variation of visual acuity. As expected, the performance of the model improved with each additional month added, so that, by month three, it accounted for 70% of the variation. In other words, patients with good visual acuity at baseline, and then at each follow-up for three months, were likely to have good visual acuity at 12 months.

This work combining DL for partial OCT segmentation, with conventional statistical techniques for prediction, is an important first step to prognosticate the AMD treatment outcome. End-to-end approaches using DL are likely to provide additional insights, particularly if large, well-labelled datasets can be used for training. However, a potential challenge in this regard will likely be the significant compute resources that will be required to train such models using a high-resolution three-dimensional dataset containing OCTs. It will also be important to make sure that the resulting model is clinically meaningful. For example, it may be possible to predict visual outcomes to high accuracy after 12 months of treatment, but this will be less useful for the patient if it involves incorporation of multiple time series data immediately prior to this. It will also be important to determine what balance of sensitivity and specificity is likely to be clinically meaningful and thus potentially actionable (for example, in potential prophylactic treatment of retinal disease prior to onset or progression). Finally, perhaps even more so than with image classification tasks, it will be important to prove that any models produced can be generalized for wide-spread usage, either in clinical trials or in real-world clinical practice.

Implementation of in clinical settings

Given the ageing population and the ever-increasing expenditure for health care there is a need to innovations. Three main areas are the targets for such solutions: To improve the general health of a population, to lower the costs of healthcare, and to improve patient's perception. AI solutions are among the most promising solutions to tackle these issues.

Providing healthcare is logistically complex and solutions differ significantly between different countries. Implementing AI-based solution into such workflow is challenging and requires sufficient connectivity. A concerted efforts from all stakeholders is required including regulators, insurances, hospital managers, IT teams, physicians, and patients. Implementation needs to be easy and straightforward without administrative hurdles to be accepted. Quick dissemination of results is an important aspect in this respect.

Another step for AI being implemented into a clinical setting is a realistic business model that needs to consider the specific interest of the patient, the payer, and the provider. Main factors to be considered in this respect are reimbursement, efficiency, and unmet clinical need. The business model also needs to consider the long-term implications, because continuous connectivity and the capacity to learn is associated with the ability to improve clinical performance over time.

In the recent years AI solutions were focussed mainly towards cancer, neurology, and cardiovascular disease (PMID: 29507784). Given the global burden that arises from these conditions it is that AI solutions are required. In ophthalmology the field is relatively new. Ophthalmology may, however, in some respects be an optimal field for implementing deep learning. Major blinding diseases such as DR, AMD and glaucoma can be treated and the incidence of blindness can be largely reduced when treatment is initiated early. Moreover, ocular imaging is cheap and fast when compared to other imaging modalities such as computer tomography or magnetic resonance imaging. As such cost effectiveness may be reached easier than in other medical disciplines. Moreover, ocular images may contain significant information on systemic disease and will as such most likely spread to other sub-specialities in the next years.

Potential challenges

AI approaches in ocular disease require a large number of images. Data sharing from different centers is an obvious approach to increase the number of input data for network training. However, Increasing the number of data elements does not necessarily enhance the performance of a network. For example, adding large amounts of data from healthy subjects will most likely not improve the classification of disease. Moreover, very large datasets for training may increase the likelihood of making spurious connections.[209] For use of retinal images to predict and classify ocular and systemic disease a clear guideline for the optimal number of cases for training is needed.

When data are to be shared between different centers regulations and state privacy rules need to be considered. These may differ between different countries and while they are aimed to ensure patients' privacy they sometimes form barriers for effective research initiatives and patient's care. Generally, there is agreement that images and all other patient-related data need to be anonymized and patients' consent has to be obtained before sharing is possible. This requires technical solutions including data storage, management, and analysis. The implementation of such solutions is time and cost-intensive. It requires hardware and software investments, expertise and is labor-intensive. Investing on data-sharing is a difficult decision, because the financial requirements are high and the benefit is not immediate. The decision for data sharing can sometimes be influenced by the fear that competitors explore novel results first. This can even occur within an institution and usually it is the weaker members of a collaborative team that fear about their career opportunities. Indeed, key performance indicators as defined by funding bodies or universities including number of publications, impact factor and citation metrics may represent major hurdles for effective data sharing. On an institutional level the filing of collaboration agreements with other partners is a long and labor-intensive procedure that slows down analysis of shared data. Such periods may even be prolonged when IP issues are to be negotiated. Given that these are usually multiple-institution agreements time spans of one year or more are common. This is associated with the risk that other teams are faster and that collaborators loose interest in the topic.

In the training set a large number of images is required that need to be well phenotyped. The performance of the network will depend on the number of images, the quality of the images, and how representative the data are for the entire spectrum of the disease. In addition, the applicability in clinical practice will depend on the quality of the phenotyping system and the ability of the human graders to follow this system. In DR, most of the classification systems rely of fundus photographs except for classification of macular edema. Popular classification systems include the ETDRS severity scales,[210] the Modified Airlie House Classification[211] and the International Clinical Disease Severity Scale for DR[212]. Based on these classification systems DL networks usually aim to detect cases with referable and vision threatening DR, particularly when used in screening situations. In principle DL may be a good approach to define such novel staging approaches, but clinical outcome trials are required to validate such strategies.

During the Early Treatment of DR Study it was recognized that some features of DR such as foveal avascular zone (FAZ), capillary loss, capillary dilatation, arteriolar and RPE abnormalities as well as fluorescein leakage, fluorescein leakage cystoid changes could be detected better with fluorescein angiography[213]. The technique is, however, time-consuming, has a mortality of 1:230,000 and inclusion of dye-based angiography into a DR screening program is not feasible. Recently introduced OCT angiography technology identifies some of these features and is an attractive alternative due to its non-invasive nature[214-216]. In addition, OCT angiography may be able to identify additional features such as deep capillary plexus nonperfusion associated with macular photoreceptor damage[217,218]. This is compatible with studies showing that the long-term recovery of photoreceptor integrity and visual outcome in DME is dependent on perfusion status of the deep capillary plexus[219]. OCT angiography has, however, only recently been commercialized and rapid technology development and limited experience currently hampers inclusion into deep learning networks for DR.

Multiple classification systems have been proposed for AMD including the recent Clinical Classification as worked out by the Beckman Initiative for Macular Research Classification Committee[220] and the Three Continent AMD Consortium Severity Scale[221] developed by harmonizing the grading of three large-scale population-based studies. Significant differences among these grading systems have been reported in distinguishing early from intermediate AMD when classifying according to the defined criteria[222]. DL-based classification systems have been developed for referability[223], severity characterization and estimation of 5-year risk[95] and disease conversion[224]. In addition, a severity classification based on fundus photography was developed[16]. Again, OCT angiography was not included in classification systems as yet, although a wide variety of studies indicate that choroidal perfusion abnormalities are associated with the risk of disease progression[225-229].

The issue is most complicated when DL approaches shall be applied to the classification of glaucoma. This is related to the difficulties in defining and diagnosing early stages of the disease. A clear diagnosis of early cases is often difficult and patients that show signs of structural disease without visual field defects are called glaucoma suspects[230]. Confirmation of the diagnosis is only possible longitudinally when the patient is either developing corresponding functional loss as identified with visual field testing or progression of structural loss that exceeds the age-related loss of tissue over time. Under these circumstances, it is of course difficult to train a glaucoma network for early cases of glaucoma detection. On the other hand, this is also a chance for AI to be implemented into glaucoma care, but strong longitudinal data are required to train the network for correctly identifying those who will develop glaucoma. Obviously, predictions of incidence are more difficult than simple classification or staging. In glaucoma there is an urgent clinical need for such networks because treatment is possible[231] and advanced visual field defect is an important risk factor for transitioning to functional blind[232]. Although progression of glaucoma cannot be halted with current therapeutic interventions slowing down progression is of utmost importance because it can shift the time to blindness beyond the life expectancy of a patient.

In patients with more advanced stages of glaucoma the classification may be an easier task, although the wide inter-individual variability of optic nerve anatomy, particularly in myopic eyes, needs to be considered[233,234]. As such the training data set needs to consist of a large dataset including a wide variety of different anatomical configurations of the optic nerve head. DL may also have applications in glaucoma progression analysis that likely needs to include structure and function. If clinical decision-making is based on artificial network progression analysis the general acceptance will also depend on the availability of outcome data.

Whereas the number of images that are available for diseases such as glaucoma, DR and AMD is sufficient to train networks orphan diseases represent a problem because of the lack of cases. One approach is to create synthetic fundus images that mimic the disease. This is, however, a difficult task and current approaches have not proven to be successful[235,236]. In addition, it is doubtful that competent authorities would approve an approach where data do not stem from real patients. Nevertheless, generation of synthetic images is an interesting approach that may have potential for future applications.

The capabilities of DL should not be construed as competence. What networks can provide is excellent performance in a well-defined task. Networks are able to classify DR and detect risk factors for AMD but they are not a substitute for a retina specialist. As such the inclusion of novel technology into DL systems is difficult, because it will require again a large number of data with this novel technology. As mentioned above this is the reason why networks that include OCT angiography have not yet been realized, although it may potentially increase their performance. Inclusion of novel technology into network based classification systems is a long and costly effort. Given that there are many novel imaging approaches on the horizon including adaptive optics[237,238], adaptive optics OCT[239,240], polarization-sensitive OCT[241], Doppler OCT[242,243], oximetry[244,245], measurement of oxygen extraction[246,247], and detection of apoptosing retinal cells[248], which may have considerable potential for diagnosis, classification and progression analysis, this is an important challenge for the future. For instance, if imaging of single retinal ganglion cells[249,250] becomes clinically available this may prompt a paradigm shift for glaucoma diagnosis. Generally speaking, a diagnostic tool that as a stand-alone technique provides excellent sensitivity and specificity does not require a DL approach.

To ensure that AI provides an acceptable patient safety profile, it must be tested in a relevant clinical context. The patient sample, the operators, the reference standard, and patient recruitment all need to be primary-care-clinic biased to avoid selection and spectrum bias.[251] Recruitment methods, exclusion criteria, and a statistical analysis plan must be documented before the recruitment of the first subject, a design called preregistration. Results must focus on the intent-to-screen population, in which every recruited subject is important, so that opportunistic exclusion of subjects and endpoints can be avoided.[252] Reference standards, also called 'truth', can be, in order of increasing external validity and decreasing intra- and inter-observer variability, created by individual clinicians, aggregated clinician opinion (via adjudication or voting), or reading centers. [253,254]

The reference standard can be derived from the same images- in the case of a clinician based reference standard -, the same modality - i.e. widefield stereo fundus images in the case of a reading center), or different or additional modalities, such as optical coherence tomography.[255] . It is worthwhile to note that when, instead of using the same images, a reference standard derived from the same or different modality – such as wide-field imaging – but different images, will capture diabetic retinopathy lesions outside of the field of view or outside the resolution of the images which the AI uses as input. This will lead to different – typically lower measured sensitivity and specificity characteristics.

In summary, autonomous AI enables the delivery of real-time, point-of-care-diagnosis in primary care clinics. It also diminishes the risk of human interference, which has proved problematic in hybrid assistive AI-specialist models.[256] Strong diagnostic accuracy, easy access to high quality diagnostics and gains in productivity, can best be realized by autonomous AI, but requires rigorous patient safety testing before.

Conclusions

AI using DL system has the potential to revolutionize how we live and practice medicine. It likely will change the field rapidly in the next few decades, although several challenges need to be resolved to increase AI adoption in healthcare. Many techniques have been described in attempt to unravel the ‘black box’ nature of DL systems, but more need to be done. Furthermore, it is also useful to develop more predictive algorithms to better stratify patients into different risks groups and treatment arms, aiming to deliver personalized medicine to the global population.

Acknowledgement

We would also like to acknowledge Dr Stephanie Lynch (Department of Ophthalmology and Visual Sciences, University of Iowa Hospital and Clinics), Miss Xin Qi Lee, Haslina Hamzah (Singapore Eye Research Institute), Dr Gilbert Lim (National University Singapore School of Computing, Singapore), Dr Liu Yong (A*STAR Institute of High Performance Center, Singapore) and Dr Yun Liu (Google AI Health, California) for their contribution onto this article as well.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
2. Zhang X, Zou J, He K, Sun J. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Trans Pattern Anal Mach Intell.* 2016;38(10):1943-1955.
3. 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(5):1285-1298.
4. Hinton GD, L.; Yu, D; Dahl, G.; Mohamed, A.; Jaitly, n.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine.* Vol 292012:82-97.
5. Tompson JJ, A.; LeCun, Y.; Bregler, C. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Advances in Neural Information Processing Systems 27.* 2014:1799-1807.
6. World Economic Forum. The Fourth Industrial Revolution: what it means, how to respond. 2016; https://www.weforum.org/agenda/2016/01/the-fourth-

industrial-revolution-what-it-means-and-how-to-respond/. Accessed 18 August, 2018.

7. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology.* 2017;284(2):574-582.

8. Ting DSW, Yi PH, Hui F. Clinical Applicability of Deep Learning System in Detecting Tuberculosis with Chest Radiography. *Radiology.* 2018;286(2):729-731.

9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115-118.

10. 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(22):2211-2223.

11. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA.* 2016;316(22):2402-2410.

12. Abramoff MD, Lou Y, Erginay A, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest Ophthalmol Vis Sci.* 2016;57(13):5200-5206.

13. Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology.* 2017;124(7):962-969.

14. 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(8):1199-1206.

15. 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 ophthalmology.* 2017;135(11):1170-1176.

16. 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(9):1410-1420.

17. Brown JM, Campbell JP, Beers A, et al. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. *JAMA ophthalmology.* 2018;136(7):803-810.

18. Varadarajan AV, Poplin R, Blumer K, et al. Deep Learning for Predicting Refractive Error From Retinal Fundus Images. *Investigative ophthalmology & visual science.* 2018;59(7):2861-2868.

19. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering.* 2018;2:158-164.

20. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express.* 2017;8(7):3440-3448.

21. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell.* 2018;172(5):1122-1131 e1129.

22. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med.* 2018.

23. Yousefi S, Goldbaum MH, Balasubramanian M, et al. Learning from data: recognizing glaucomatous defect patterns and detecting progression from visual field measurements. *IEEE Trans Biomed Eng.* 2014;61(7):2112-2124.

24. Rumelhart DE, Hinton GE, WIlliams RJ. Learning representations by back-propagating errors. *Nature.* 1986;323:533-536.

25. Bottou L, Cortes C, Denker JS, et al. Comparison of classifier methods: a case study in handwritten digit recognition. Paper presented at: Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on1994.

26. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna ZB. *Rethinking the Inception Architecture for Computer Vision.* 2016.

27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.

28. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. Paper presented at: AAAI2017.

29. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. Paper presented at: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on2017.

30. Chen Y, Li J, Xiao H, Jin X, Yan S, Feng J. Dual path networks. Paper presented at: Advances in Neural Information Processing Systems2017.

31. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *J Physiol.* 1968;195(1):215-243.

32. Krizhevsky A, Sutskever H, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS.* 2012.

33. World Health Organization (WHO). Global Health and Ageing. 2018. URL: http://www.who.int/ageing/publications/global_health.pdf [Accessed on 17th November, 2018].

34. Divo MJ, Martinez CH, Mannino DM. Ageing and the epidemiology of multimorbidity. *Eur Respir J.* 2014;44(4):1055-1068.

35. Chader GJ, Taylor A. Preface: The aging eye: normal changes, age-related diseases, and sight-saving approaches. *Investigative ophthalmology & visual science.* 2013;54(14):ORSF1-4.

36. Wheatley CM, Dickinson JL, Mackey DA, Craig JE, Sale MM. Retinopathy of prematurity: recent advances in our understanding. *Br J Ophthalmol.* 2002;86(6):696-700.

37. Moss SE, Klein R, Klein BE. The 14-year incidence of visual loss in a diabetic population. *Ophthalmology.* 1998;105(6):998-1003.

38. 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(4):260-277.

39. World Health Organisation (WHO). Diabetes Facts. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed on 18th November, 2018].

40. Klein BE. Overview of epidemiologic studies of diabetic retinopathy. *Ophthalmic epidemiology.* 2007;14(4):179-183.

41. Cheung N, Mitchell P, Wong TY. Diabetic retinopathy. *Lancet.* 2010;376(9735):124-136.

42. International Diabetes Federation. Diabetes atlas, 6th ed, Brussels, Belgium. 2015.

43. Resnikoff S, Pascolini D, Etya'ale D, et al. Global data on visual impairment in the year 2002. *Bulletin of the World Health Organization.* 2004;82(11):844-851.

44. Yau JW, Rogers SL, Kawasaki R, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care.* 2012;35(3):556-564.

45. Gardner GG, Keating D, Williamson TH, Elliott AT. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. *The British journal of ophthalmology.* 1996;80(11):940-944.

46. Lee SC, Lee ET, Kingsley RM, et al. Comparison of diagnosis of early retinal lesions of diabetic retinopathy between a computer system and human experts. *Arch Ophthalmol.* 2001;119(4):509-515.

47. Sinthanayothin C, Boyce JF, Williamson TH, et al. Automated detection of diabetic retinopathy on digital fundus images. *Diabet Med.* 2002;19(2):105-112.

48. Niemeijer M, van Ginneken B, Russell SR, Suttorp-Schulten MS, Abramoff MD. Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis. *Invest Ophthalmol Vis Sci.* 2007;48(5):2260-2267.

49. Niemeijer M, van Ginneken B, Staal J, Suttorp-Schulten MS, Abràmoff MD. Automatic detection of red lesions in digital color fundus photographs. *IEEE Trans Med Imaging.* 2005;24(5):584-592.

50. Walter T, Massin P, Erginay A, Ordonez R, Jeulin C, Klein JC. Automatic detection of microaneurysms in color fundus images. *Med Image Anal.* 2007;11(6):555-566.

51. Hansen MB, Abramoff MD, Folk JC, Mathenge W, Bastawrous A, Peto T. Results of Automated Retinal Image Analysis for Detection of Diabetic Retinopathy from the Nakuru Study, Kenya. *PLoS One.* 2015;10(10):e0139148.

52. Abramoff MD, Folk JC, Han DP, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol.* 2013;131(3):351-357.

53. Abramoff MD, Niemeijer M, Suttorp-Schulten MS, Viergever MA, Russell SR, van Ginneken B. Evaluation of a system for automatic detection of diabetic retinopathy from color fundus photographs in a large population of patients with diabetes. *Diabetes Care.* 2008;31(2):193-198.

54. Abramoff MD, Reinhardt JM, Russell SR, et al. Automated early detection of diabetic retinopathy. *Ophthalmology.* 2010;117(6):1147-1154.

55. Friedenwald J, Day R. The vascular lesions of diabetic retinopathy. *Bull Johns Hopkins Hosp.* 1950;86(4):253-254.

56. Niemeijer M, Abramoff MD, van Ginneken B. Information fusion for diabetic retinopathy CAD in digital color fundus photographs. *IEEE Trans Med Imaging.* 2009;28(5):775-785.

57. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.

58. Quellec G, Russell SR, Abramoff MD. Optimal filter framework for automated, instantaneous detection of lesions in retinal images. *Ieee Transactions on Medical Imaging.* 2011;30(2):523-533.

59. Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern.* 1980;36(4):193-202.

60. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine.* 2018;39:1-8.

61. 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.

62. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology.* 2014;121(11):2081-2090.

63. Maul EA, Jampel HD. Glaucoma screening in the real world. *Ophthalmology.* 2010;117(9):1665-1666.

64. Hove MN, Kristensen JK, Lauritzen T, Bek T. The prevalence of retinopathy in an unselected population of type 2 diabetes patients from Arhus County, Denmark. *Acta Ophthalmol Scand.* 2004;82(4):443-448.

65. U.S. Preventive Services Task Force. Understanding Task Force Recommendations. 2018. Accessed 2018 November 5.

66. Foster PJ, Buhrmann R, Quigley HA, Johnson GJ. The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol.* 2002;86(2):238-242.

67. Crowston JG, Hopley CR, Healey PR, Lee A, Mitchell P, Blue Mountains Eye S. The effect of optic disc diameter on vertical cup to disc ratio percentiles in a population based cohort: the Blue Mountains Eye Study. *Br J Ophthalmol.* 2004;88(6):766-770.

68. Chauhan BC, Burgoyne CF. From clinical examination of the optic disc to clinical assessment of the optic nerve head: a paradigm change. *Am J Ophthalmol.* 2013;156(2):218-227 e212.

69. Savini G, Carbonelli M, Barboni P. Spectral-domain optical coherence tomography for the diagnosis and follow-up of glaucoma. *Current opinion in ophthalmology.* 2011;22(2):115-123.

70. Chauhan BC, Danthurebandara VM, Sharpe GP, et al. Bruch's Membrane Opening Minimum Rim Width and Retinal Nerve Fiber Layer Thickness in a Normal White Population: A Multicenter Study. *Ophthalmology.* 2015;122(9):1786-1794.

71. Haleem MS, Han L, Hemert J, et al. Regional Image Features Model for Automatic Classification between Normal and Glaucoma in Fundus and Scanning Laser Ophthalmoscopy (SLO) Images. *J Med Syst.* 2016;40(6):132.

72. Haleem MS, Han L, Hemert JV, et al. A Novel Adaptive Deformable Model for Automated Optic Disc and Cup Segmentation to Aid Glaucoma Diagnosis. *J Med Syst.* 2017;42(1):20.

73. Abramoff MD, Alward WL, Greenlee EC, et al. Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features. *Investigative ophthalmology & visual science.* 2007;48(4):1665-1673.

74. Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep.* 2018;8(1):14665.

75. Asaoka R, Murata H, Hirasawa K, et al. Using Deep Learning and transform learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol.* 2018.

76. Masumoto H, Tabuchi H, Nakakura S, Ishitobi N, Miki M, Enno H. Deep-learning Classifier With an Ultrawide-field Scanning Laser Ophthalmoscope Detects Glaucoma Visual Field Severity. *J Glaucoma.* 2018;27(7):647-652.

77. Elze T, Pasquale LR, Shen LQ, Chen TC, Wiggs JL, Bex PJ. Patterns of functional vision loss in glaucoma determined with archetypal analysis. *J R Soc Interface.* 2015;12(103).

78. Cai S, Elze T, Bex PJ, Wiggs JL, Pasquale LR, Shen LQ. Clinical Correlates of Computationally Derived Visual Field Defect Archetypes in Patients from a Glaucoma Clinic. *Curr Eye Res.* 2017;42(4):568-574.

79. Wang M, Pasquale LR, Shen LQ, et al. Reversal of Glaucoma Hemifield Test Results and Visual Field Features in Glaucoma. *Ophthalmology.* 2018;125(3):352-360.

80. Li F, Wang Z, Qu G, et al. Automatic differentiation of Glaucoma visual field from non-glaucoma visual filed using deep convolutional neural network. *BMC medical imaging.* 2018;18(1):35.

81. Schell GJ, Lavieri MS, Stein JD, Musch DC. Filtering data from the collaborative initial glaucoma treatment study for improved identification of glaucoma progression. *BMC Med Inform Decis Mak.* 2013;13:137.

82. Garcia GP, Nitta K, Lavieri MS, et al. Using Kalman Filtering to Forecast Disease Trajectory for Patients with Normal Tension Glaucoma. *Am J Ophthalmol.* 2018.

83. Hark LA, Katz LJ, Myers JS, et al. Philadelphia Telemedicine Glaucoma Detection and Follow-up Study: Methods and Screening Results. *Am J Ophthalmol.* 2017;181:114-124.

84. Hong SW, Koenigsman H, Ren R, et al. Glaucoma Specialist Optic Disc Margin, Rim Margin, and Rim Width Discordance in Glaucoma and Glaucoma Suspect Eyes. *Am J Ophthalmol.* 2018;192:65-76.

85. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects. *J Glaucoma.* 2017;26(12):1086-1094.

86. Wong TY, Loon SC, Saw SM. The epidemiology of age related eye diseases in Asia. *The British journal of ophthalmology.* 2006;90(4):506-511.

87. Baeza M, Orozco-Beltran D, Gil-Guillen VF, et al. Screening for sight threatening diabetic retinopathy using non-mydriatic retinal camera in a

primary care setting: to dilate or not to dilate? *Int J Clin Pract.* 2009;63(3):433-438.

88. 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(2):e106-116.

89. Bressler NM. Age-related macular degeneration is the leading cause of blindness. *JAMA.* 2004;291(15):1900-1901.

90. Group CR, Martin DF, Maguire MG, et al. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N Engl J Med.* 2011;364(20):1897-1908.

91. Chakravarthy U, Harding SP, Rogers CA, et al. Alternative treatments to inhibit VEGF in age-related choroidal neovascularisation: 2-year findings of the IVAN randomised controlled trial. *Lancet.* 2013;382(9900):1258-1267.

92. Mitchell P, Bressler N, Doan QV, et al. Estimated cases of blindness and visual impairment from neovascular age-related macular degeneration avoided in Australia by ranibizumab treatment. *PLoS One.* 2014;9(6):e101072.

93. Bressler NM, Doan QV, Varma R, et al. Estimated cases of legal blindness and visual impairment avoided using ranibizumab for choroidal neovascularization: non-Hispanic white population in the United States with age-related macular degeneration. *Arch Ophthalmol.* 2011;129(6):709-717.

94. Chew EY, Clemons TE, Agron 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(8):791-801.

95. Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration. *JAMA ophthalmology.* 2018.

96. Age-Related Eye Disease Study Research G. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8. *Arch Ophthalmol.* 2001;119(10):1417-1436.

97. Davis MD, Gangnon RE, Lee LY, et al. The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17. *Arch Ophthalmol.* 2005;123(11):1484-1498.

98. Quinn GE. Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic. *Eye Brain.* 2016;8:31-36.

99. Ludwig CA, Chen TA, Hernandez-Boussard T, Moshfeghi AA, Moshfeghi DM. The Epidemiology of Retinopathy of Prematurity in the United States. *Ophthalmic Surg Lasers Imaging Retina.* 2017;48(7):553-562.

100. Fleck BW, Dangata Y. Causes of visual handicap in the Royal Blind School, Edinburgh, 1991-2. *Br J Ophthalmol.* 1994;78(5):421.

101. Blencowe H, Vos T, Lee AC, et al. Estimates of neonatal morbidities and disabilities at regional and global levels for 2010: introduction, methods overview, and relevant findings from the Global Burden of Disease study. *Pediatr Res.* 2013;74 Suppl 1:4-16.

102. Early Treatment for Retinopathy of Prematurity Cooperative G, Good WV, Hardy RJ, et al. Final visual acuity results in the early treatment for retinopathy of prematurity study. *Arch Ophthalmol.* 2010;128(6):663-671.

103. Fierson WM, American Academy of Pediatrics Section on O, American Academy of O, American Association for Pediatric O, Strabismus, American Association of Certified O. Screening examination of premature infants for retinopathy of prematurity. *Pediatrics.* 2013;131(1):189-195.

104. Chiang MF, Melia M, Buffenn AN, et al. Detection of clinically significant retinopathy of prematurity using wide-angle digital retinal photography: a report by the American Academy of Ophthalmology. *Ophthalmology.* 2012;119(6):1272-1280.

105. Quinn GE, Ying G-s, Daniel E, et al. Validity of a telemedicine system for the evaluation of acute-phase retinopathy of prematurity. *JAMA ophthalmology.* 2014;132(10):1178-1184.

106. Tasman W. Multicenter trial of cryotherapy for retinopathy of prematurity. *Archives of Ophthalmology.* 1988;106(4):463-464.

107. Garner A. An international classification of retinopathy of prematurity. *Archives of ophthalmology.* 1984;102(8):1130-1134.

108. Gole GA, Ells AL, Katz X, et al. The international classification of retinopathy of prematurity revisited. *JAMA Ophthalmology.* 2005;123(7):991-999.

109. Gilbert C, Rahi J, Eckstein M, O'Sullivan J, Foster A. Retinopathy of prematurity in middle-income countries. *Lancet.* 1997;350(9070):12-14.

110. Chan RP, Williams SL, Yonekawa Y, Weissgold DJ, Lee TC, Chiang MF. Accuracy of retinopathy of prematurity diagnosis by retinal fellows. *Retina (Philadelphia, Pa).* 2010;30(6):958.

111. Myung JS, Chan RVP, Espiritu MJ, et al. Accuracy of retinopathy of prematurity image-based diagnosis by pediatric ophthalmology fellows: implications for training. *Journal of American Association for Pediatric Ophthalmology and Strabismus.* 2011;15(6):573-578.

112. Wong RK, Ventura CV, Espiritu MJ, et al. Training fellows for retinopathy of prematurity care: a Web-based survey. *Journal of American Association for Pediatric Ophthalmology and Strabismus.* 2012;16(2):177-181.

113. Nagiel A, Espiritu MJ, Wong RK, et al. Retinopathy of prematurity residency training. *Ophthalmology.* 2012;119(12):2644-2645. e2642.

114. Chiang MF, Jiang L, Gelman R, Du YE, Flynn JT. Interexpert agreement of plus disease diagnosis in retinopathy of prematurity. *Archives of ophthalmology.* 2007;125(7):875-880.

115. Wallace DK, Quinn GE, Freedman SF, Chiang MF. Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity. *Journal of American Association for Pediatric Ophthalmology and Strabismus.* 2008;12(4):352-356.

116. Slidsborg C, Forman JL, Fielder AR, et al. Experts do not agree when to treat retinopathy of prematurity based on plus disease. *British Journal of Ophthalmology.* 2012;96(4):549-553.

117. Gschließer A, Stifter E, Neumayer T, et al. Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity. *American journal of ophthalmology.* 2015;160(3):553-560. e553.

118. Campbell JP, Ryan MC, Lore E, et al. Diagnostic discrepancies in retinopathy of prematurity classification. *Ophthalmology.* 2016;123(8):1795-1801.

119. 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(1):74-80.

120. Reynolds JD, Dobson V, Quinn GE, et al. Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. *Archives of ophthalmology.* 2002;120(11):1470-1476.

121. 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 ophthalmology.* 2015;133(6):675-682.

122. Hewing NJ, Kaufman DR, Chan RP, Chiang MF. Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts. *JAMA ophthalmology.* 2013;131(8):1026-1032.

123. Rao R, Jonsson NJ, Ventura C, et al. Plus disease in retinopathy of prematurity: diagnostic impact of field of view. *Retina (Philadelphia, Pa).* 2012;32(6):1148.

124. Keck KM, Kalpathy-Cramer J, Ataer-Cansizoglu E, You S, Erdogmus D, Chiang MF. Plus disease diagnosis in retinopathy of prematurity: vascular tortuosity as a function of distance from optic disc. *Retina (Philadelphia, Pa).* 2013;33(8):1700.

125. Campbell JP, Ataer-Cansizoglu E, Bolon-Canedo V, et al. Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis. *JAMA ophthalmology.* 2016;134(6):651-657.

126. Gelman SK, Gelman R, Callahan AB, et al. Plus disease in retinopathy of prematurity: quantitative analysis of standard published photograph. *Archives of Ophthalmology.* 2010;128(9):1217-1220.

127. Moleta C, Campbell JP, Kalpathy-Cramer J, et al. Plus disease in retinopathy of prematurity: diagnostic trends in 2016 versus 2007. *American journal of ophthalmology.* 2017;176:70-76.

128. Group S-RMS. Supplemental therapeutic oxygen for prethreshold retinopathy of prematurity (STOP-ROP), a randomized, controlled trial. I: primary outcomes. *Pediatrics.* 2000;105(2):295-310.

129. Kim SJ, Campbell JP, Kalpathy-Cramer J, et al. Plus disease in retinopathy of prematurity: should diagnosis be eye-based or quadrant-based? *Journal of American Association for Pediatric Ophthalmology and Strabismus {JAAPOS}.* 2018;22(4):e78.

130. Wallace DK, Kylstra JA, Chesnutt DA. Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity. *Journal of American Association for Pediatric Ophthalmology and Strabismus.* 2000;4(4):224-229.

131. Wallace DK, Freedman SF, Hartnett ME, Quinn GE. Predictive value of pre-plus disease in retinopathy of prematurity. *Archives of ophthalmology.* 2011;129(5):591-596.

132. Campbell JP, Kalpathy-Cramer J, Erdogmus D, et al. Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as a basis of diagnostic variability. *Ophthalmology.* 2016;123(11):2338-2344.

133. Kalpathy-Cramer J, Campbell JP, Erdogmus D, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology.* 2016;123(11):2345-2351.

134. Wittenberg LA, Jonsson NJ, Chan RP, Chiang MF. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity. *Journal of pediatric ophthalmology and strabismus.* 2012;49(1):11-19.

135. Koreen S, Gelman R, Martinez-Perez ME, et al. Evaluation of a computer-based system for plus disease diagnosis in retinopathy of prematurity. *Ophthalmology.* 2007;114(12):e59-e67.

136. Wilson CM, Wong K, Ng J, Cocker KD, Ells AL, Fielder AR. Digital image analysis in retinopathy of prematurity: a comparison of vessel selection methods. *Journal of American Association for Pediatric Ophthalmology and Strabismus.* 2012;16(3):223-228.

137. Abbey AM, Besirli CG, Musch DC, et al. Evaluation of screening for retinopathy of prematurity by ROPtool or a lay reader. *Ophthalmology.* 2016;123(2):385-390.

138. Ataer-Cansizoglu E, Bolon-Canedo V, Campbell JP, et al. Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the “i-ROP” system and image features associated with expert diagnosis. *Translational vision science & technology.* 2015;4(6):5-5.

139. Worrall DE, Wilson CM, Brostow GJ. Automated retinopathy of prematurity case detection with convolutional neural networks. *Deep Learning and Data Labeling for Medical Applications*: Springer; 2016:68-76.

140. Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol.* 2017;70(1):1-25.

141. Prabhakaran D, Jeemon P, Roy A. Cardiovascular Diseases in India: Current Epidemiology and Future Directions. *Circulation.* 2016;133(16):1605-1620.

142. The Cardiovascular Disease Statistics. . 2018; https://www.who.int/cardiovascular_diseases/en/. Accessed 5 November, 2018.

143. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol.* 2014;63(25 Pt B):2889-2934.

144. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998;97(18):1837-1847.

145. National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation.* 2002;106(25):3143-3421.

146. Brody AM, Flack JM, Ference BA, Levy PD. Utility of Framingham risk score in urban emergency department patients with asymptomatic hypertension. *Crit Pathw Cardiol.* 2014;13(3):114-116.

147. Conroy RM, Pyorala K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J.* 2003;24(11):987-1003.

148. Graham I, Atar D, Borch-Johnsen K, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts). *Eur Heart J.* 2007;28(19):2375-2414.

149. Force USPST, Curry SJ, Krist AH, et al. Screening for Cardiovascular Disease Risk With Electrocardiography: US Preventive Services Task Force Recommendation Statement. *JAMA.* 2018;319(22):2308-2314.

150. D'Agostino RB, Sr., Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation.* 2008;117(6):743-753.

151. Graham I, Atar D, Borch-Johnsen K, et al. European guidelines on cardiovascular disease prevention in clinical practice: full text. Fourth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). *Eur J Cardiovasc Prev Rehabil.* 2007;14 Suppl 2:S1-113.

152. Goff DC, Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation.* 2014;129(25 Suppl 2):S49-73.

153. Hira RS, Kennedy K, Nambi V, et al. Frequency and practice-level variation in inappropriate aspirin use for the primary prevention of cardiovascular disease: insights from the National Cardiovascular Disease Registry's Practice Innovation and Clinical Excellence registry. *J Am Coll Cardiol.* 2015;65(2):111-121.

154. Yeboah J, McClelland RL, Polonsky TS, et al. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. *JAMA.* 2012;308(8):788-795.

155. Cheung CY, Tay WT, Ikram MK, et al. Retinal microvascular changes and risk of stroke: the Singapore Malay Eye Study. *Stroke.* 2013;44(9):2402-2408.

156. Yip W, Ong PG, Teo BW, et al. Retinal Vascular Imaging Markers and Incident Chronic Kidney Disease: A Prospective Cohort Study. *Sci Rep.* 2017;7(1):9374.

157. Wang JJ, Liew G, Wong TY, et al. Retinal vascular calibre and the risk of coronary heart disease-related death. *Heart.* 2006;92(11):1583-1587.

158. Wong TY, Kamineni A, Klein R, et al. Quantitative retinal venular caliber and risk of cardiovascular disease in older persons: the cardiovascular health study. *Arch Intern Med.* 2006;166(21):2388-2394.

159. Seidelmann SB, Claggett B, Bravo PE, et al. Retinal Vessel Calibers in Predicting Long-Term Cardiovascular Outcomes: The Atherosclerosis Risk in Communities Study. *Circulation.* 2016;134(18):1328-1338.

160. Witt N, Wong TY, Hughes AD, et al. Abnormalities of retinal microvascular structure and risk of mortality from ischemic heart disease and stroke. *Hypertension.* 2006;47(5):975-981.

161. About UK Biobank. 2017; http://www.ukbiobank.ac.uk/about-biobank-uk/. Accessed 26 March, 2017.

162. Welcome to EyePACS 2017; http://www.eyepacs.org/. Accessed 31 July, 2017.

163. Ting DS, Wong TY. Eyeing Cardiovascular Risk Factors. *Nature Biomedical Engineering.* 2018;2:140-141.

164. Cardiovascular Disease (10-year risk). 2017; https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/. Accessed 21 June 2017.

165. Cooney MT, Dudina A, De Bacquer D, et al. How much does HDL cholesterol add to risk estimation? A report from the SCORE Investigators. *Eur J Cardiovasc Prev Rehabil.* 2009;16(3):304-314.

166. Dudina A, Cooney MT, Bacquer DD, et al. Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators. *Eur J Cardiovasc Prev Rehabil.* 2011;18(5):731-742.

167. Simonyan K VA, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Computer Vision and Pattern Recognition (csCV).* 2017.

168. Bourne RR, Stevens GA, White RA, et al. Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health.* 2013;1(6):e339-349.

169. Naidoo KS, Leasher J, Bourne RR, et al. Global Vision Impairment and Blindness Due to Uncorrected Refractive Error, 1990-2010. *Optom Vis Sci.* 2016;93(3):227-234.

170. Varadarajan AV PR, Blumer K, Angermueller C, Ledsam J, Chopra R, et al. Deep learning for predicting refractive error from retinal fundus images. *Computer Vision and Pattern Recognition (csCV).* 2017.

171. Keane PA, Sadda SR. Retinal imaging in the twenty-first century: state of the art and future directions. *Ophthalmology.* 2014;121(12):2489-2500.

172. Fujimoto J, Swanson E. The Development, Commercialization, and Impact of Optical Coherence Tomography. *Investigative ophthalmology & visual science.* 2016;57(9):OCT1-OCT13.

173. Windsor MA, Sun SJJ, Frick KD, Swanson EA, Rosenfeld PJ, Huang D. Estimating Public and Patient Savings From Basic Research-A Study of Optical Coherence Tomography in Managing Antiangiogenic Therapy. *Am J Ophthalmol.* 2018;185:115-122.

174. Patel N, Pass A, Mason S, Gibson CR, Otto C. Optical Coherence Tomography Analysis of the Optic Nerve Head and Surrounding Structures in Long-Duration International Space Station Astronauts. *JAMA ophthalmology.* 2018;136(2):193-200.

175. Hee MR, Puliafito CA, Wong C, et al. Quantitative assessment of macular edema with optical coherence tomography. *Arch Ophthalmol.* 1995;113(8):1019-1029.

176. Keane PA, Sadda SR. Predicting visual outcomes for macular disease using optical coherence tomography. *Saudi J Ophthalmol.* 2011;25(2):145-158.

177. Schmidt-Erfurth U, Waldstein SM. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Prog Retin Eye Res.* 2016;50:1-24.

178. Csaky KG, Richman EA, Ferris FL, 3rd. Report from the NEI/FDA Ophthalmic Clinical Trial Design and Endpoints Symposium. *Investigative ophthalmology & visual science.* 2008;49(2):479-489.

179. Keane PA, Patel PJ, Ouyang Y, et al. Effects of retinal morphology on contrast sensitivity and reading ability in neovascular age-related macular degeneration. *Investigative ophthalmology & visual science.* 2010;51(11):5431-5437.

180. De Fauw J, Keane P, Tomasev N, et al. Automated analysis of retinal imaging using machine learning techniques for computer vision. *F1000Res.* 2016;5:1573.

181. Owen CG, Jarrar Z, Wormald R, Cook DG, Fletcher AE, Rudnicka AR. The estimated prevalence and incidence of late stage age related macular degeneration in the UK. *Br J Ophthalmol.* 2012;96(5):752-756.

182. Royal College of Ophthalmology (RCOphth). Age-related Macular Degeneration. https://www.rcophth.ac.uk/standards-publications-research/commissioning-in-ophthalmology/age-related-macular-degeneration/. [Accessed on 17th November, 2018].

183. Optometry Today. OCT Rollout in Every Specsavers Announced. 22 May 2017. https://www.aop.org.uk/ot/industry/high-street/2017/05/22/oct-rollout-in-every-specsavers-announced. [Accessed on 17th November, 2018].

184. Sadda SR, Wu Z, Walsh AC, et al. Errors in retinal thickness measurements obtained by optical coherence tomography. *Ophthalmology.* 2006;113(2):285-293.

185. Keane PA, Mand PS, Liakopoulos S, Walsh AC, Sadda SR. Accuracy of retinal thickness measurements obtained with Cirrus optical coherence tomography. *Br J Ophthalmol.* 2009;93(11):1461-1467.

186. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. . *In Medical Image Computing and Computer-Assisted Intervention – MICCAI* 2015;9351:234-241.

187. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical Image Analysis using Convolutional Neural Networks: A Review. *J Med Syst.* 2018;42(11):226.

188. Bunce C. Correlation, agreement, and Bland-Altman analysis: statistical analysis of method comparison studies. *Am J Ophthalmol.* 2009;148(1):4-6.

189. Fang L, Cunefare D, Wang C, Guymer RH, Li S, Farsiu S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomedical optics express.* 2017;8(5):2732-2744.

190. Venhuizen FG, van Ginneken B, Liefers B, et al. Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks. *Biomedical optics express.* 2017;8(7):3292-3316.

191. Hamwood J, Alonso-Caneiro D, Read SA, Vincent SJ, Collins MJ. Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers. *Biomedical optics express.* 2018;9(7):3049-3066.

192. Chen M, Wang J, Oguz I, VanderBeek BL, Gee JC. Automated segmentation of the choroid in EDI-OCT images with retinal pathology using convolution neural networks. *Fetal, Infant and Ophthalmic Medical Image Analysis: International Workshop, FIFI 2017, and 4th International Workshop, OMIA 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, Proceedings FIFI Workshop.* 2017;10554:177-184.

193. Lee CS, Tyring AJ, Deruyter NP, Wu Y, Rokem A, Lee AY. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomedical optics express.* 2017;8(7):3440-3448.

194. Roy AG, Conjeti S, Karri SPK, et al. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical optics express.* 2017;8(8):3627-3642.

195. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning. *Ophthalmology.* 2018;125(4):549-558.

196. Zayit-Soudry S, Moroz I, Loewenstein A. Retinal pigment epithelial detachment. *Surv Ophthalmol.* 2007;52(3):227-243.

197. Xu Y, Yan K, Kim J, et al. Dual-stage deep learning framework for pigment epithelium detachment segmentation in polypoidal choroidal vasculopathy. *Biomedical optics express.* 2017;8(9):4061-4076.

198. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, et al. Machine Learning to Analyze the Prognostic Value of Current Imaging Biomarkers in Neovascular Age-Related Macular Degeneration. *Ophthalmol Retina.* 2018;2(1):24-30.

199. Schlegl T, Bogunovic H, Klimscha S, et al. Fully Automated Segmentation of Hyperreflective Foci in Optical Coherence Tomography Images. URL: https://arxiv.org/abs/1805.03278. *CVPR.* 2018.

200. Loo J, Fang L, Cunefare D, Jaffe GJ, Farsiu S. Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. *Biomedical optics express.* 2018;9(6):2681-2698.

201. Camino A, Wang Z, Wang J, et al. Deep learning for the segmentation of preserved photoreceptors on en face optical coherence tomography in two inherited retinal diseases. *Biomedical optics express.* 2018;9(7):3092-3105.

202. Lee CS, Baughman DM, Lee AY. Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. *Ophthalmol Retina.* 2017;1(4):322-327.

203. 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(2):259-265.

204. Karri SP, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomedical optics express.* 2017;8(2):579-592.

205. Prahs P, Radeck V, Mayer C, et al. OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications. *Graefes Arch Clin Exp Ophthalmol.* 2018;256(1):91-98.

206. Sonobe T, Tabuchi H, Ohsugi H, et al. Comparison between support vector machine and deep learning, machine-learning technologies for detecting epiretinal membrane using 3D-OCT. *International ophthalmology.* 2018.

207. Buchan JC, Amoaku W, Barnes B, et al. How to defuse a demographic time bomb: the way forward? *Eye.* 2017;31(11):1519-1522.

208. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface.* 2018;15(141).

209. Gomes L. Machine-learning maestro michael jordan on the delusions of big data and other huge engineering efforts. *IEEE Spectrum, Oct.* 2014;20.

210. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology.* 1991;98(5 Suppl):786-806.

211. Diabetic retinopathy study. Report Number 6. Design, methods, and baseline results. Report Number 7. A modification of the Airlie House classification of diabetic retinopathy. Prepared by the Diabetic Retinopathy. *Invest Ophthalmol Vis Sci.* 1981;21(1 Pt 2):1-226.

212. Wilkinson CP, Ferris FL, 3rd, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology.* 2003;110(9):1677-1682.

213. Classification of diabetic retinopathy from fluorescein angiograms. ETDRS report number 11. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology.* 1991;98(5 Suppl):807-822.

214. Ang M, Tan ACS, Cheung CMG, et al. Optical coherence tomography angiography: a review of current and future clinical applications. *Graefe's*

archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie. 2018;256(2):237-245.

215. Spaide RF, Fujimoto JG, Waheed NK, Sadda SR, Staurenghi G. Optical coherence tomography angiography. *Progress in retinal and eye research.* 2018;64:1-55.

216. Kashani AH, Chen CL, Gahm JK, et al. Optical coherence tomography angiography: A comprehensive review of current methods and clinical applications. *Progress in retinal and eye research.* 2017;60:66-100.

217. Scarinci F, Jampol LM, Linsenmeier RA, Fawzi AA. Association of Diabetic Macular Nonperfusion With Outer Retinal Disruption on Optical Coherence Tomography. *JAMA Ophthalmol.* 2015;133(9):1036-1044.

218. Scarinci F, Nesper PL, Fawzi AA. Deep Retinal Capillary Nonperfusion Is Associated With Photoreceptor Disruption in Diabetic Macular Ischemia. *Am J Ophthalmol.* 2016;168:129-138.

219. Byung GM, ; Um, T. W.; Lee, J. Y.; Young, H. Y. Correlation between Deep Capillary Plexus Perfusion and Long-Term Photoreceptor Recovery after Diabetic Macular Edema Treatment. *Ophthalmology Retina.* 2018;2(3):235-243.

220. Ferris FL, 3rd, Wilkinson CP, Bird A, et al. Clinical Classification of Age-related Macular Degeneration. *Ophthalmology.* 2013;120(4):844-851.

221. Klein R, Meuer SM, Myers CE, et al. Harmonizing the classification of age-related macular degeneration in the three-continent AMD consortium. *Ophthalmic epidemiology.* 2014;21(1):14-23.

222. Brandl C, Zimmermann ME, Gunther F, et al. On the impact of different approaches to classify age-related macular degeneration: Results from the German AugUR study. 2018;8(1):8675.

223. Burlina P, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Utility of Deep Learning Methods for Referability Classification of Age-Related Macular Degeneration. *JAMA ophthalmology.* 2018.

224. Schmidt-Erfurth U, Waldstein SM, Klimscha S, et al. Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence. *Investigative ophthalmology & visual science.* 2018;59(8):3199-3208.

225. Lindner M, Kosanetzky S, Pfau M, et al. Local Progression Kinetics of Geographic Atrophy in Age-Related Macular Degeneration Are Associated

With Atrophy Border Morphology. *Invest Ophthalmol Vis Sci.* 2018;59(4):Amd12-amd18.

226. Ferrara D, Silver RE, Louzada RN, Novais EA, Collins GK, Seddon JM. Optical Coherence Tomography Features Preceding the Onset of Advanced Age-Related Macular Degeneration. *Invest Ophthalmol Vis Sci.* 2017;58(9):3519-3529.

227. Zhou Q, Daniel E, Grunwald JE, et al. Association between pseudodrusen and delayed patchy choroidal filling in the comparison of age-related macular degeneration treatments trials. *Acta ophthalmologica.* 2017;95(6):e518-e520.

228. Xu W, Grunwald JE, Metelitsina TI, et al. Association of risk factors for choroidal neovascularization in age-related macular degeneration with decreased foveolar choroidal circulation. *Am J Ophthalmol.* 2010;150(1):40-47.e42.

229. Boltz A, Luksch A, Wimpissinger B, et al. Choroidal blood flow and progression of age-related macular degeneration in the fellow eye in patients with unilateral choroidal neovascularization. *Invest Ophthalmol Vis Sci.* 2010;51(8):4220-4225.

230. Chang RT, Singh K. Glaucoma Suspect: Diagnosis and Management. *Asia-Pacific journal of ophthalmology (Philadelphia, Pa).* 2016;5(1):32-37.

231. Schmidl D, Schmetterer L, Garhofer G, Popa-Cherecheanu A. Pharmacotherapy of glaucoma. *Journal of ocular pharmacology and therapeutics : the official journal of the Association for Ocular Pharmacology and Therapeutics.* 2015;31(2):63-77.

232. Peters D, Bengtsson B, Heijl A. Factors associated with lifetime risk of open-angle glaucoma blindness. *Acta ophthalmologica.* 2014;92(5):421-425.

233. Fledelius HC, Goldschmidt E. Optic disc appearance and retinal temporal vessel arcade geometry in high myopia, as based on follow-up data over 38 years. *Acta ophthalmologica.* 2010;88(5):514-520.

234. Kwon J, Sung KR, Park JM. Myopic glaucomatous eyes with or without optic disc shape alteration: a longitudinal study. *The British journal of ophthalmology.* 2017;101(12):1618-1622.

235. Menti E, Bonaldi L, Ballerini L, Ruggeri A, Trucco E. Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network. 2016; Cham.

236. Fiorini SB, L.; Trucco, E.; Ruggeri, A. Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network. *Procedia Computer Science.* 2016;90:54-60.

237. Georgiou M, Kalitzeos A, Patterson EJ, Dubra A, Carroll J. Adaptive optics imaging of inherited retinal diseases. *The British journal of ophthalmology.* 2018;102(8):1028-1035.

238. Burns SA, Elsner AE, Sapoznik KA, Warner RL, Gast TJ. Adaptive optics imaging of the human retina. *Progress in retinal and eye research.* 2018.

239. Pircher M, Zawadzki RJ. Review of adaptive optics OCT (AO-OCT): principles and applications for retinal imaging [Invited]. *Biomedical optics express.* 2017;8(5):2536-2562.

240. Dong ZM, Wollstein G, Wang B, Schuman JS. Adaptive optics optical coherence tomography in glaucoma. *Progress in retinal and eye research.* 2017;57:76-88.

241. de Boer JF, Hitzenberger CK, Yasuno Y. Polarization sensitive optical coherence tomography - a review [Invited]. *Biomedical optics express.* 2017;8(3):1838-1873.

242. Doblhoff-Dier V, Schmetterer L, Vilser W, et al. Measurement of the total retinal blood flow using dual beam Fourier-domain Doppler optical coherence tomography with orthogonal detection planes. *Biomedical optics express.* 2014;5(2):630-642.

243. Leitgeb RA, Werkmeister RM, Blatter C, Schmetterer L. Doppler optical coherence tomography. *Progress in retinal and eye research.* 2014;41:26-43.

244. Stefansson E, Olafsdottir OB, Einarsdottir AB, et al. Retinal Oximetry Discovers Novel Biomarkers in Retinal and Brain Diseases. *Invest Ophthalmol Vis Sci.* 2017;58(6):Bio227-bio233.

245. Bek T, Stefansson E, Hardarson SH. Retinal oxygen saturation is an independent risk factor for the severity of diabetic retinopathy. *The British journal of ophthalmology.* 2018.

246. Werkmeister RM, Schmidl D, Aschinger G, et al. Retinal oxygen extraction in humans. *Sci Rep.* 2015;5:15763.

247. Fondi K, Wozniak PA, Howorka K, et al. Retinal oxygen extraction in individuals with type 1 diabetes with no or mild diabetic retinopathy. *Diabetologia.* 2017;60(8):1534-1540.

248. Cordeiro MF, Normando EM, Cardoso MJ, et al. Real-time imaging of single neuronal cell apoptosis in patients with glaucoma. *Brain : a journal of neurology.* 2017;140(6):1757-1767.

249. Liu Z, Kurokawa K, Zhang F, Lee JJ, Miller DT. Imaging and quantifying ganglion cells and other transparent neurons in the living human retina. *Proceedings of the National Academy of Sciences of the United States of America.* 2017;114(48):12803-12808.

250. Rossi EA, Granger CE, Sharma R. Imaging individual neurons in the retinal ganglion cell layer of the living eye. 2017;114(3):586-591.

251. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299(17):926-930.

252. Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, van Aert RC, van Assen MA. Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Front Psychol.* 2016;7:1832.

253. Quellec G, Abramoff MD. Estimating maximal measurable performance for automated decision systems from the characteristics of the reference standard. application to diabetic retinopathy screening. *Conf Proc IEEE Eng Med Biol Soc.* 2014;2014:154-157.

254. Wong TY, Bressler NM. Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening. *JAMA.* 2016;316(22):2366-2367.

255. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine.* 2018;1(1):39.

256. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* 2007;356(14):1399-1409.

257. 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(8):1264-1272.

258. Li Z, Keel S, Liu C, et al. An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs. *Diabetes Care.* 2018;41(12):2509-2516.

TABLES

Table 1: Ten steps in building an artificial intelligence system for medical imaging analysis

1. Identify a clinical unmet need or research question
2. Selection of datasets - splitting of training, validation and testing
3. Selection of CNNs (e.g. AlexNet, VGGNet, ResNet, DenseNet, Ensemble)
4. Selection of software to build the DL systems - Keras, Tensorflow, Cafe, Python
5. The use of transfer learning/pre-training on ImageNet
6. The use of backpropagation for tuning and optimization
7. Reporting of the characteristics of datasets - patients' demographics, retinal image and disease characteristics
8. Reporting of the diagnostic performance on local and external validation datasets - area under curve, sensitivity and specificity, accuracy and kappa
9. The use of heat map to explain the diagnosis - different types of heat map (occlusion test, soft attention map, integrated gradient method)
10. Novel methods in retinal imaging - GAN, VAE and its potential clinical applications

*GAN – generative adversarial network; VAE – variational autoencoder

Table 2: Global prevalence of major eye disease burden

Global Eye Health Burden	Number of people (Millions)	Prevalence
Diabetes (>18 years)	422	8.50%
Glaucoma (aged 40-80 years)	111.8	3.54%
Age-related macular degeneration (aged 30-97 years)	288	8.69%
Retinopathy of prematurity	15	30%*[99]
Refractive error	108	1.11%
Cardiovascular Disease[140]	442.7	

***In newborns with a birth weight <1kg**

Table 3: A summary of artificial intelligence systems using deep learning in the detection of referable diabetic retinopathy

DL systems	Year	Development Dataset	CNN	Clinical Validation	Mydriatic or Non-Mydriatic	Granularity	Ground Truth	Total n (including ungradable)	% ungradable	Referable AUC	Referable DR Sensitivity	Referable DR Specificity
Abramoff et al[12]	2016	10,000 to 1,250,000 unique samples of each lesion type graded by one or more experts	Algorithm is hybrid with CNN-based lesion predictors and classical non-deep learning algorithms	Messidor-2	Mydriatic	Patient-level	Adjudication by 3 retinal specialistis until full consensus for all cases using a single 45 degree FOV image	874	4.00%	0.98	96.80%	87.00%
Gulshan et al[11]	2016	128,175 images graded 3-7 times	Inception-V3	EyePACS-1*	Mostly Non-Mydriatic	Image-level	Majority decision of 7 or 8 ophthalmologists for all cases using single macula-centered image with 45 degree FOV	9963	11.60%	0.991 (0.974)*	97.50% (96.7%)*	93.40% (84%)*
				Messidor-2	Mydriatic	Image-level		1748	0.17%	0.94	96.10%	93.90%
Gargeya and Leng[13]	2017	75,137 images from Kaggle competition graded by "a panel of retinal specialists" (with no additonal detail)	Customized CNN	Messidor-2	Mydriatic	Image-level	Not clearly described, likely the lesion-based grading that came with the public datasets using a single 45 degree FOV image	--	--	0.99	--	--
				E-Ophtha	Likely Non-Mydriatic	Image-level		--	--	0.96	--	--
Ting et al[10]	2017	76,370 images from multiple screening program and clinical studies graded by a minimum of 2	VGG-19	SiDRP 14-15*	Mydriatic	Image-level	Two trained graders for all cases, using 45 degree FOV a single image. If there is a disagreement, a	35,948	1.10%	0.94*	90.50%*	91.60%*

		graders, often with a retinal specialist for arbitration					retinal specialist generated final grade					
				Guangdong	Non-mydriatic	Image-level	2 graders; arbitration by 1 retinal specialist	15,798	1.40%	0.949*	98.70%*	81.60%*
				SIMES	Mydriatic	Image-level	1 grader; 1 retinal specialist	3052	1.80%	0.889*	97.10%*	82%*
				SINDI	Mydriatic	Image-level	1 grader; 1 retinal specialist	4512	2.10%	0.917*	99.30%*	73.30%*
				SCES	Mydriatic	Image-level	1 grader; 1 retinal specialist	1936	1.00%	0.919*	100%*	76.30%*
				BES	Mydriatic	Image-level	2 ophthalmologists	1052	0.40%	0.929*	94.40%*	88.50%*
				AFEDS	Mydriatic	Image-level	2 retinal specialists	1968	4.20%	0.98*	98.80%*	86.50%*
				RVEEH	Mydriatic	Image-level	2 graders	2302	10.90%	0.983*	98.90%*	92.20%*
				Mexican	Mydriatic	Image-level	2 retinal specialists	1172	0.50%	0.95*	91.80%*	84.80%*
				CUHK	Mydriatic	Image-level	2 retinal specialists	1254	0.00%	0.948*	99.30%*	83.10%*
				HKU	Mydriatic	Image-level	2 optometrists	7706	0.00%	0.964*	100%*	81.30%*
Krause et al[257]	2018	1.67M images with clinical grades for train set 3,737 fully adjudicated images for tune set	Inception-V3	EyePACS-2*	Mostly Non-Mydriatic	Image-level	Adjudication by 3 retinal specialistis until full consensus for all cases using a single 45 degree FOV image	--	0%	0.986	97.1%	92.3%
Abramoff et al[60]	2018	10,000 to 1,250,000	Customized CNN	FDA Pivotal Trial	23.6% Mydriatic	Patient-level	Reading center grading of	892	8.20%	-	87.2%	90.7%

		unique samples of each lesion type graded by one or more experts					stereoscopic, 4W field equivalent of ETDRS, with OCT for DME				80.70%*	89.80%*
Li et al[258]	2018	58,790 images from ZhongShan Ophthalmic Eye Center	Inception-v3	ZhongShan	Mostly Non-Mydriatic	Image-level	Panel of 21 ophthalmologists, reference standard was made when consistent grading outcomes achieved by 3 graders. VTDR = ≥severe DR and/or DME	8,000	6.10%	0.989	97%	91.4%
				NIEHS	Mostly Non-Mydriatic		2 ophthalmologists	7,181	1.9%**	0.955**	92.5%**	98.5%**
				SIMES	Mydriatic		1 grader; 1 retinal specialist	15,679				
				AusDiab	Mydriatic			12,341				

***The results included the ungradable images (and the performance is often lower compared to those who excluded the ungradable images from the analysis)**

**Combined performance for 3 external validation studies, the individual diagnostic performance was not reported in the study

Table 4: A summary of artificial intelligence system using deep learning for detection of glaucoma suspect and glaucoma

Author	Year	Disease definition	Development Dataset	CNN	Clincial Validation	Mydriatic or non-myd	Granularity	Ground Truth	Imaging Modality	Number of images	AUC	Sensitivity	Specificity
Li et al.[14]	2018	CDR□0.7 and glaucomatous changes	31,745 images (LabelMe)	Inception-V3	8,000 images (LabelMe)	--	Image-level	Panel of 21 ophthalmologists, reference standard was made when consistent grading outcomes achieved by 3 graders	Fundus photos	48,116	0.986	95.60%	92.00%
Ting et al.[10]	2017	CDR□0.8 and glaucomatous changes	125,189 images (SiDRP 10-13, SIMES, SCES, SINDI and SNEC Glaucoma datasets)	VGG-19	71,896 images (SiDRP 14-15)	Mydriatic	Image-level	1 retinal specialist; 2 senior graders	Fundus photos	197,085	0.942	96.40%	87.20%
Shibata et al[74]	2018	Glaucoma	3,150 eyes (Matsue Red Cross Hospital)	ResNet	110 eyes (Matsue Red Cross Hospital)	Non-mydriatic	Eye-level	3 resident ophthalmologists	Fundus photos	3,260	0.965	NR	NR
Asaoka et al.[75]	2018	Early glaucoma	1936 eyes (Pretraining: JAMIGO; Training: Tokyo University Hospital, Tajimi Iwase eye clinic)	Customized CNN	196 eyes (Tokyo University Hospital, Kitasato University Hospital, Tajimi Iwase eye clinic)	Mydriatic	Eye-level	Panel of 3 glaucoma specialists; glaucomatous VF change defined by Anderson Patella Criteria	SD OCT	2,132	0.937	82.50%	93.90%

Masumoto et al.[76]	2018	Glaucoma	1,117 images (Tsukazaki Hospital)	Customized CNN	282 images (Tsukazaki Hospital)	Non-mydriatic	Image-level	2 glaucoma specialists	Optos wide-field fundus photos	1,399	0.872*	81.3%*	80.2%*
Li et al.[80]	2018	Glaucoma	3,712 images (3 ophthalmic centers in China)	VGG-15	300 images	--	Image-level	9 opthalmologists (3 glaucoma experts, 3 attending ophthalmologists, 3 resident opthalmologists)	HVF PD probability plots	4,012	0.966	93.20%	82.60%

Abbreviations used: CDR=cup-disc ratio; AUC=Area under the receiver operator curve; SD OCT= Spectral domain ocular coherence tomography; HVF PD = Humphrey visual field pattern deviation. For definition of glaucoma see source references. Some form of convoluted neural network was used in all of these deep learning algorithms.

*This represents glaucoma overall averaged over mild, moderate and severe cases.

Table 5: A summary of artificial intelligence system using deep learning for detection of age-related macular degeneration (AMD)

Author	Year	Disease	Development Dataset	CNN	Clincial Validation	Mydriatic or non-mydriatic	Granularity	Ground Truth	Number of retinal images	AUC	Sensitivity	Specificity	Remark
Burlina et al[15]	2017	Referable AMD	107057 images (AREDS 1)	AlexNet DCNN/ OverFeat DCNN	26764 images (AREDS 2)	Mydriatic	Image-Level	AREDS photograph reading center (trained and certified graders)	133,821	0.94-0.96	71.00-88.40%	91.40-94.10%	0.764-0.829 (Kappa)
Burlina et al[95]	2018	5-year risk of AMD Progression to Advanced Stage	59313 images (AREDS 1)	ResNet-50	8088 images (AREDS 2)	Mydriatic	Image-Level	AREDS photograph reading center (trained and certified graders)	67,401	-	-	-	Overall mean estimation error = 3.5% to 5.3%
Ting et al[10]	2017	Referable AMD	38185 images (SIDRP 10-13) 2180 images (SNEC AMD Phenotype Study) 16182 images (SCES) 8616 images (SMES) 7447 images (SINDI)	VGG-19	71896 (SiDRP 14-15)	Mydriatic	Image-Level	1 Retinal Specialist	108,558	0.931	93.20%	88.70%	
Grassmann et al[16]	2018	Any AMD	86,770 images (AREDS 1)	7 CNN (AlexNet; GoogLeNet; VGG; Inception-v3; ResNet; Inception-ResNet-v2; Ensemble: random forest)	33886 images (AREDS 2) 5555 images (Kora)	Non-Mydriatic	Image-Level	AREDS photograph reading center (trained and certified graders)	120,656	-	100% (Late Stage AMD)	96.5% (Late Stage AMD)	

Table 6. Baseline characteristics of patients in the development and validation sets using the UK Biobank and EyePACS dataset for the deep learning system in detecting cardiovascular risk factors.[19]

	Development Set		**Clinical Validation Set**	
Characteristics	**UK Biobank**	**EyePACS**	**UK Biobank**	**EyePACS-2K**
Number of Patients	48,101	236,234	12,026	999
Number of Images	96,082	1,682,938	24,008	1,958
Age: Mean, years (SD)	56.8 (8.2)	53.6 (11.6)	56.9 (8.2)	54.9 (10.9)
Self-reported Sex (% male)	44.9	39.2	44.9	39.2
Ethnicity	1.2% Black, 3.4% Asian/PI, 90.6% White, 4.1% Other	4.9% Black, 5.5% Asian/PI, 7.7% White, 58.1% Hispanic, 1.2% Native Am, 1.7% Other	1.3% Black, 3.6% Asian/PI, 90.1% White, 4.2% Other	6.4% Black, 5.7% Asian/PI, 11.3% White, 57.2% Hispanic, 0.7% Native Am, 2% Other
BMI: Mean (SD)	27.31 (4.78)	n/a	27.37 (4.79)	n/a
Systolic BP: Mean, mmHg (SD)	136.82 (18.41)	n/a	136.89 (18.3)	n/a
Diastolic BP: Mean, mmHg (SD)	81.78 (10.08)	n/a	81.76 (9.87)	n/a
HbA1c: Mean, % (SD)	n/a	8.23 (2.14)	n/a	8.2 (2.13)
Current Smoker: %	9.53%	n/a	9.87%	n/a

Table 7. Algorithm performance on predicting cardiovascular risk factors on three validation sets. 95% confidence intervals on the metrics were calculated with 2000 bootstrap samples (Methods). MAE: Mean Absolute Error; R^2: R-squared, AUC: Area under the Receiver Operator Curve (c-statistic). For continuous risk factors (like age), the baseline value is the Mean Absolute Error of predicting the mean value for all patients.[19]

Predicted Risk Factor (Evaluation Metric)	**UK Biobank Validation Set** (n=12,026 patients)	**EyePACS-2K Validation Set** (n=999 patients)	**Independent Validation Set of Asian Patients** (n=239 patients)
Age (MAE in years)	3.26 (3.22-3.31)	3.42 (3.23-3.61)	3.42 (3.06,3.78)
Age (R^2)	0.74 (0.73-0.75)	0.82 (0.79-0.84)	0.79 (0.74-0.83)
Self-reported sex (AUC)	0.97 (0.966-0.971)	0.97 (0.96-0.98)	0.98 (0.96-0.99)
Current Smoker (AUC)	0.71 (0.70-0.73)	n/a	0.79 (0.70-0.88)
HbA1c (MAE in %)	n/a	1.39 (1.29-1.50)	0.92 (0.83-1.00)
HbA1c (R^2)	n/a	0.09 (0.03-0.16)	0.24 (0.10-0.39)
Systolic BP (MAE in mmHg)	11.35 (11.18-11.51)	n/a	14.31 (12.96-15.66)
Systolic BP (R^2)	0.36 (0.35-0.37)	n/a	0.34 (0.25-0.44)
Diastolic BP (MAE in mmHg)	6.42 (6.33-6.52)	n/a	7.93 (7.25-8.61)
Diastolic BP (R^2)	0.32 (0.30-0.33)	n/a	0.36 (0.26-0.45)
BMI (MAE)	3.29 (3.24-3.34)	n/a	3.57 (3.21-3.94)
BMI (R^2)	0.13 (0.11-0.14)	n/a	0.16 (0.06-0.28)

Table 8. Predicting 5-year major adverse cardiovascular events (MACE) on biobank validation set. Of the 12,026 patients in the UK Biobank validation dataset, 91 experience a previous cardiac event prior to retinal imaging and were excluded from the analysis. Of the 11,835 patients in the validation set without a previous cardiac event, 105 patients experienced a MACE within 5 years of retinal imaging. 95% confidence intervals were calculated using 2000 bootstrap samples.

Model	**AUC (95% CI)**
Age	0.66 (0.61-0.71)
Systolic blood pressure (SBP)	0.66 (0.61-0.71)
Body mass index (BMI)	0.62 (0.56-0.67)
Gender	0.57 (0.53-0.62)
Current smoker	0.55 (0.52-0.59)
Algorithm	0.70 (0.65-0.74)
Age + SBP + BMI + gender + current smoker	0.72 (0.68-0.76)
Algorithm + age + SBP + BMI + gender + current smoker	0.73 (0.69-0.77)
Systematic Coronary Risk Evaluation (SCORE)[6,7]	0.72 (0.67-0.76)
Algorithm + SCORE	0.72 (0.67-0.76)

Table 9: Mean absolute error (MAE) and coefficient of determination (R2) of algorithm vs baseline for predicting the refractive error in the UK Biobank dataset. Baseline metrics are calculated by predicting mean values of the validation set. The 95% confidence intervals are shown in square brackets; all the values are in units of diopters.

	MAE		**R2**	
	Model	Baseline	Model	Baseline
Spherical Equivalent	0.56 [0.55, 0.56]	1.81 [1.79-1.84]	0.90 [0.90, 0.91]	0.0 [0.0, 0.0]
Cylindrical Component	0.43 [0.42, 0.43]	0.48 [0.47-0.49]	0.05 [0.04, 0.06]	0.0 [0.0, 0.0]
Spherical Component	0.63 [0.63, 0.64]	1.89 [1.87-1.92]	0.88 [0.88, 0.89]	0.0 [0.0, 0.0]

Table 10: A summary of artificial intelligence system using deep learning for optical coherence tomography for different retinal diseases

DL systems	Year	Disease	OCT machines	Development Dataset	CNN	Test Images	AUC	Accuracy	Sensitivity	Specificity
Disease Detection										
Lee et al[202]	2017	Exudative AMD	Spectralis	80,839 images	VGG-16	20613 images	0.928	87.60%	84.60%	91.50%
Treder et al[203]	2018	Exudative AMD	Spectralis	1,012 images (University of Muenster Medical Center)	Inception-V3	100 images	NR	100%	92%	96%
Kermany et al[21]	2018	CNV DME Drusen	Spectralis	108,312 images	Inception-V3	1,000 images				
		1. Multi-class comparison					0.999	96.60%	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%
		DME vs normal					0.999	98.20%	96.80%	99.60%
		Drusen vs normal					0.999	99.00%	98.00%	99.20%
De Fauw et al[22]	2018	Urgent, semi-urgent, routine, and observation only	Topcon (device 1)	877 manually segmented scans	Segmentation network U-Net	997 scans	0.992 (Urgent referral)	94.50%		
			Spectralis (retrained device 2)	152 manually segmented scans		116 scans	0.999 (Urgent referral)	96.60%		

		Normal, CNV, Macular Edema, FTMH, PTMH, CSR, VMT, GA, Drusen, ERM	Topcon (device 1)	14,884 scans	Classification network using a custom 29 CNN layers with 5 pooling layers					
Disease Prediction Ursula Schmidth[224]	2018	AMD	Spectralis	HARBOR Trial	Other - Random Forest	614 patients	-	Predictive Accuracy of BCVA R2=0.7	-	-

FIGURES

Figure 1: The introduction of artificial intelligence (AI) in 1950's, followed by machine learning in 1980's and deep learning (DL) in 2010's. Machine learning is a subset of AI, involving using statistical techniques to help computers to learn without being explicitly programmed. With the advent of graphic processing unit with much improved processing power, DL is the state-of-art technique that has revolutionized the machine learning field over the past few years. It has now been widely adopted in image recognition, speech recognition and natural language processing domains.

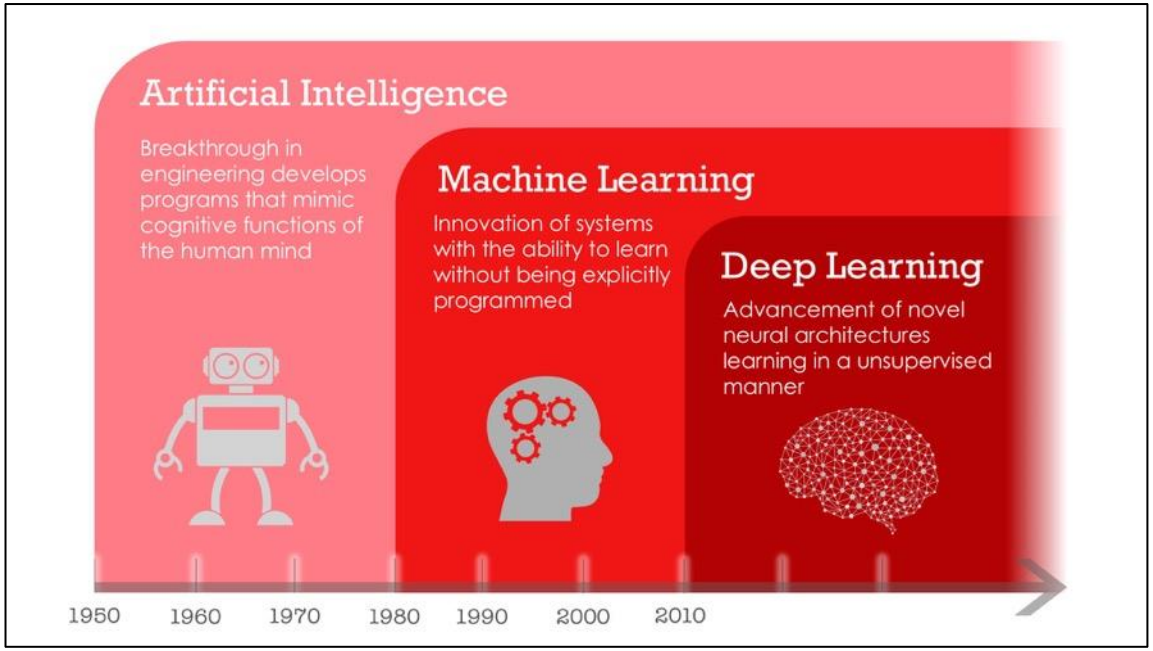

Artificial Intelligence
Breakthrough in engineering develops programs that mimic cognitive functions of the human mind
Machine Learning
Innovation of systems with the ability to learn without being explicitly programmed
Deep Learning
Advancement of novel neural architectures learning in a unsupervised manner
1950
1960
1970
1980
1990
2000
2010

Figure 2A: The input, feature-extraction layers (hidden layer) and classification (output) layers of a convolutional neural network (CNN). The feature extraction layers consist of convolution layer, Rectified Linear Unit (ReLU) layer and Pooling. **Figure 2B:** For max pooling, the largest number within a 2x2 rectified feature map will be chosen to be the representative number on the feature map (output).

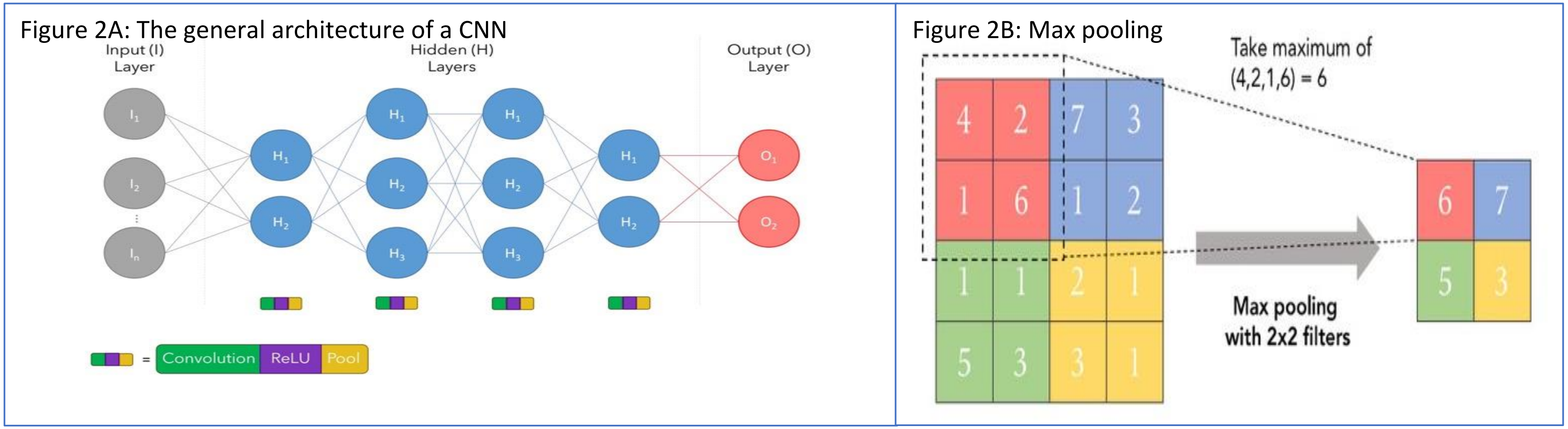


Figure 2A: The general architecture of a CNN

Figure 2B: Max pooling

4	2	7	3
1	6	1	2
1	1	2	1
5	3	3	1

Take maximum of (4,2,1,6) = 6

Max pooling with 2x2 filters

6	7
5	3

Figure 3: The workflow of a deep learning system in detecting referable diabetic retinopathy and age-related macular degeneration, further demonstrated by the heat map

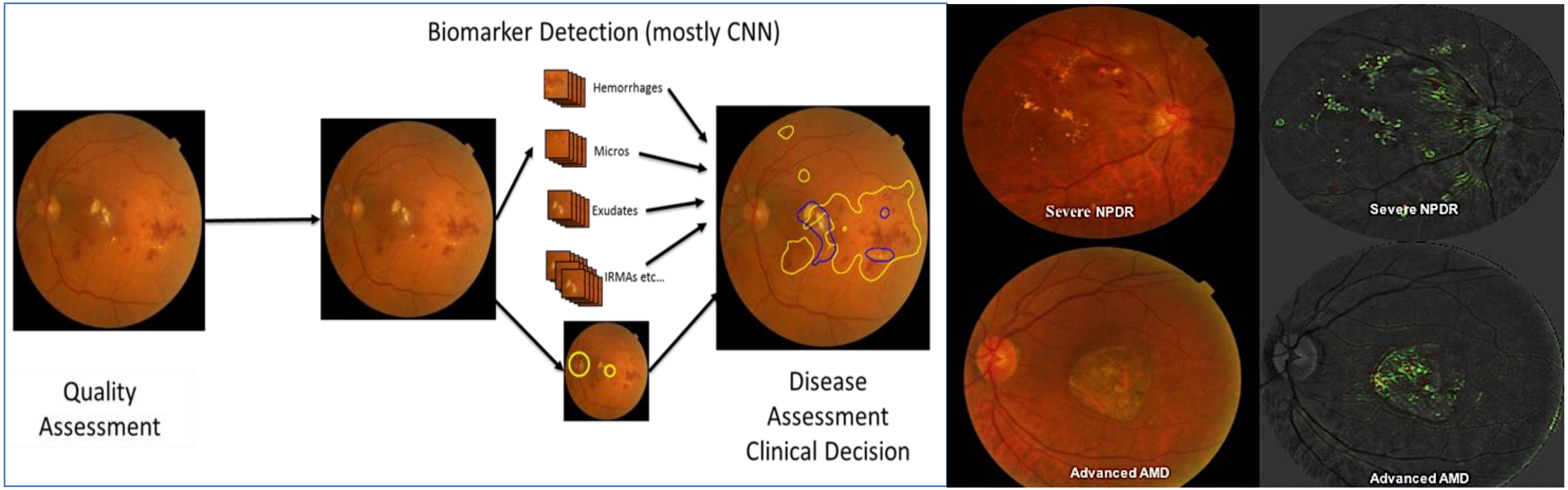


Figure 4: Attention maps for a single retinal fundus image. The left-most image is a sample retinal image in color from the UK Biobank dataset. The remaining images show the same retinal image, but in black and white. The soft attention heat map for each prediction is overlaid in green, indicating the areas of the heat map that the neural-network model is using to make that particular prediction for the image.

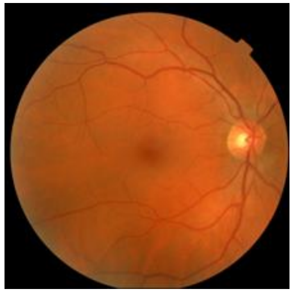
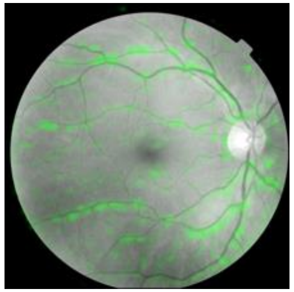
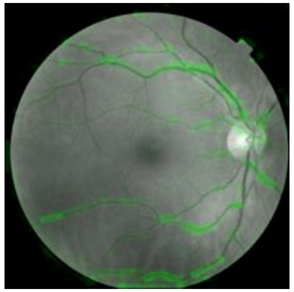
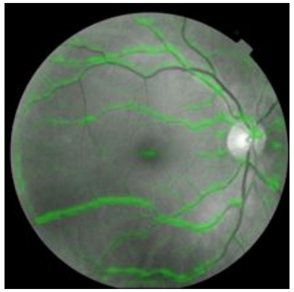
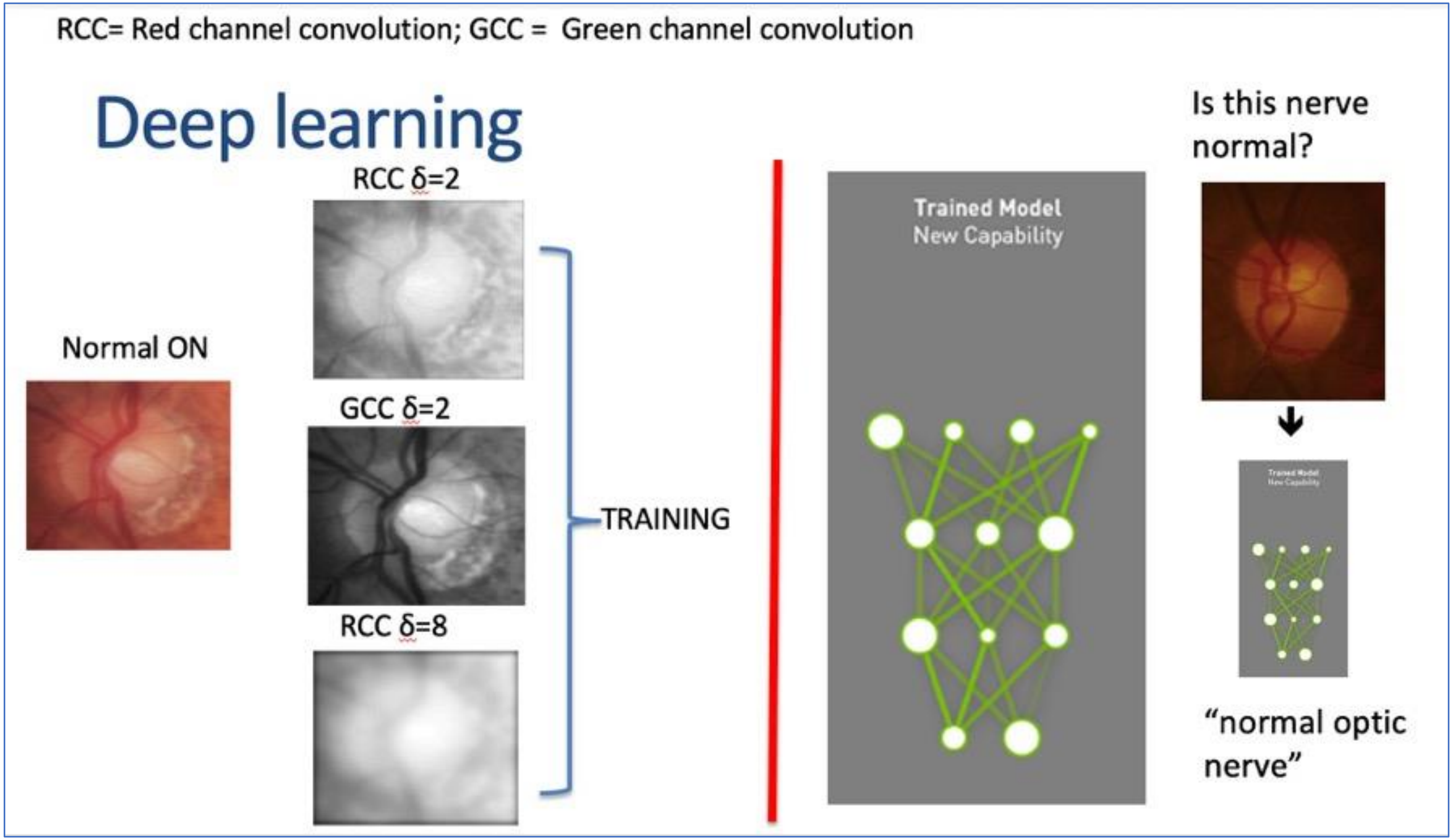
Original	**Age**	**Smoking Status**	**Systolic BP**
			
	Actual: 53.0 years Predicted: 53.8 years	Actual: Nonsmoker Predicted: Nonsmoker	Actual: 128.5 mmHg Predicted: 130.1 mmHg

Figure 5: Deep learning system for detection of glaucomatous optic disc using optic disc imaging



RCC= Red channel convolution; GCC = Green channel convolution

Deep learning

Normal ON

RCC δ=2

GCC δ=2

RCC δ=8

TRAINING

Trained Model
New Capability

Is this nerve normal?

"normal optic nerve"

Figure 6: Continuous spectrum of retinal vascular findings in retinopathy of prematurity (ROP). (A) shows normal posterior retinal vessels. (B) shows pre-plus disease with mild retinal vascular dilation and tortuosity. (C) shows plus disease with significant retinal vascular dilation and tortuosity.

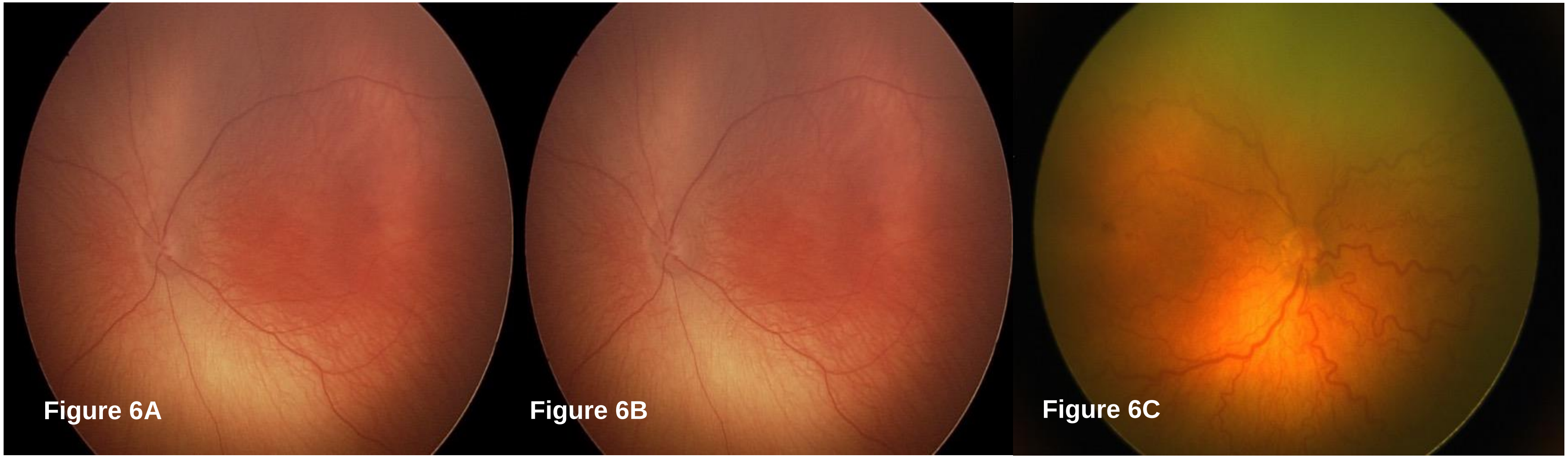


Figure 7. Mean attention map over 1000 images from UK Biobank for severely myopic (SE worse than -6.0), neutral (SE between -0.5 and 0.5), and severely hyperopic (SE worse than 5.0) eyes conditioned on eye position. Scale bar on right denotes attention pixel values, which are between 0 and 1 (exclusive), with the sum of all values equal to one.

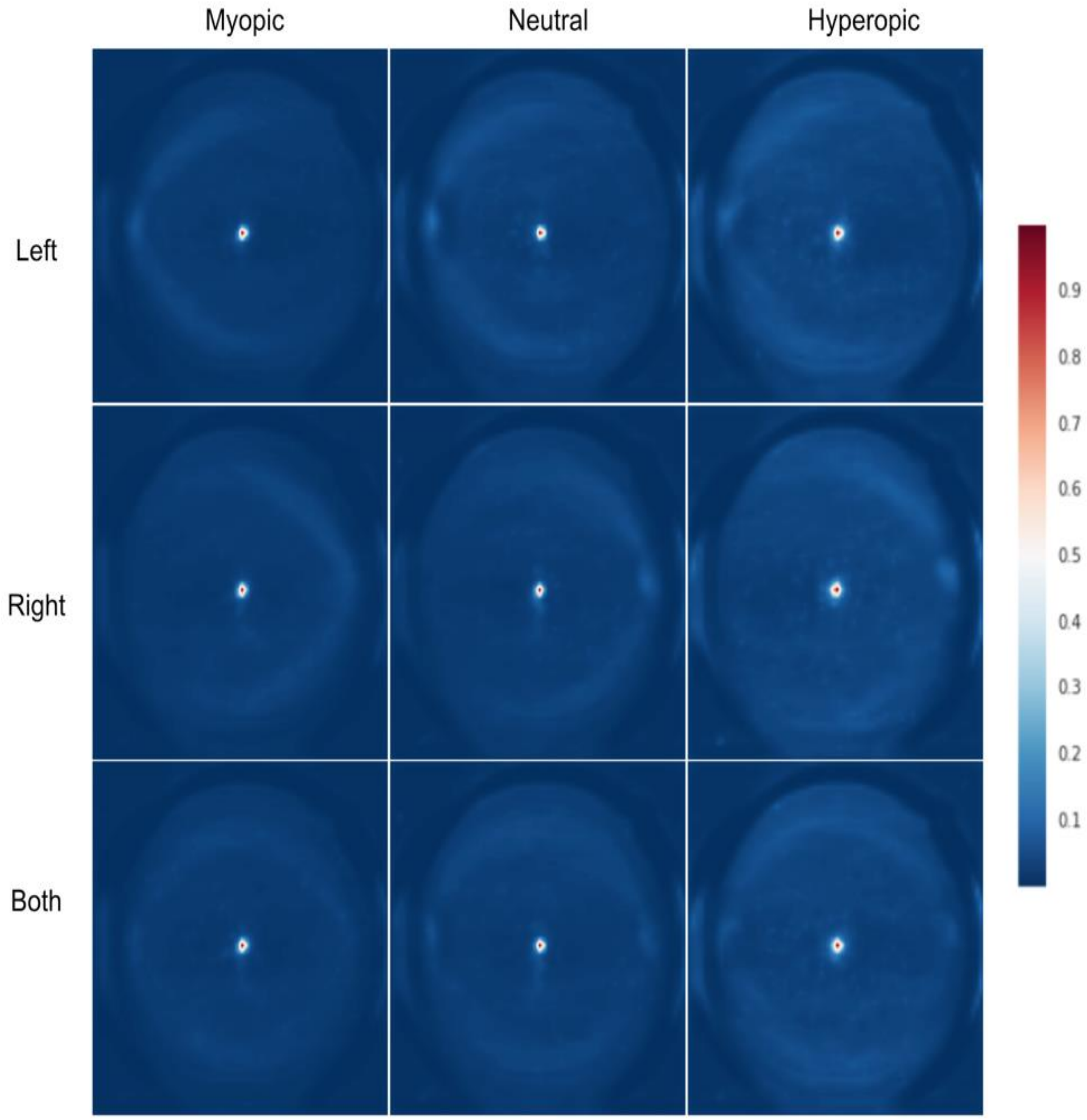


Figure 8: The application of deep learning to the segmentation of retinal optical coherence tomography (OCT) images – the prototype OCT viewer for the Moorfields-DeepMind deep learning system. In this case, the system correctly segments loss of the retinal pigment epithelium (RPE) highlighting an area of geographic atrophy (GA) in age-related macular degeneration (AMD). The GA is surrounded by numerous foci of drusenoid pigment epithelium detachment (PED). The partially detached posterior hyaloid is also clearly delineated.

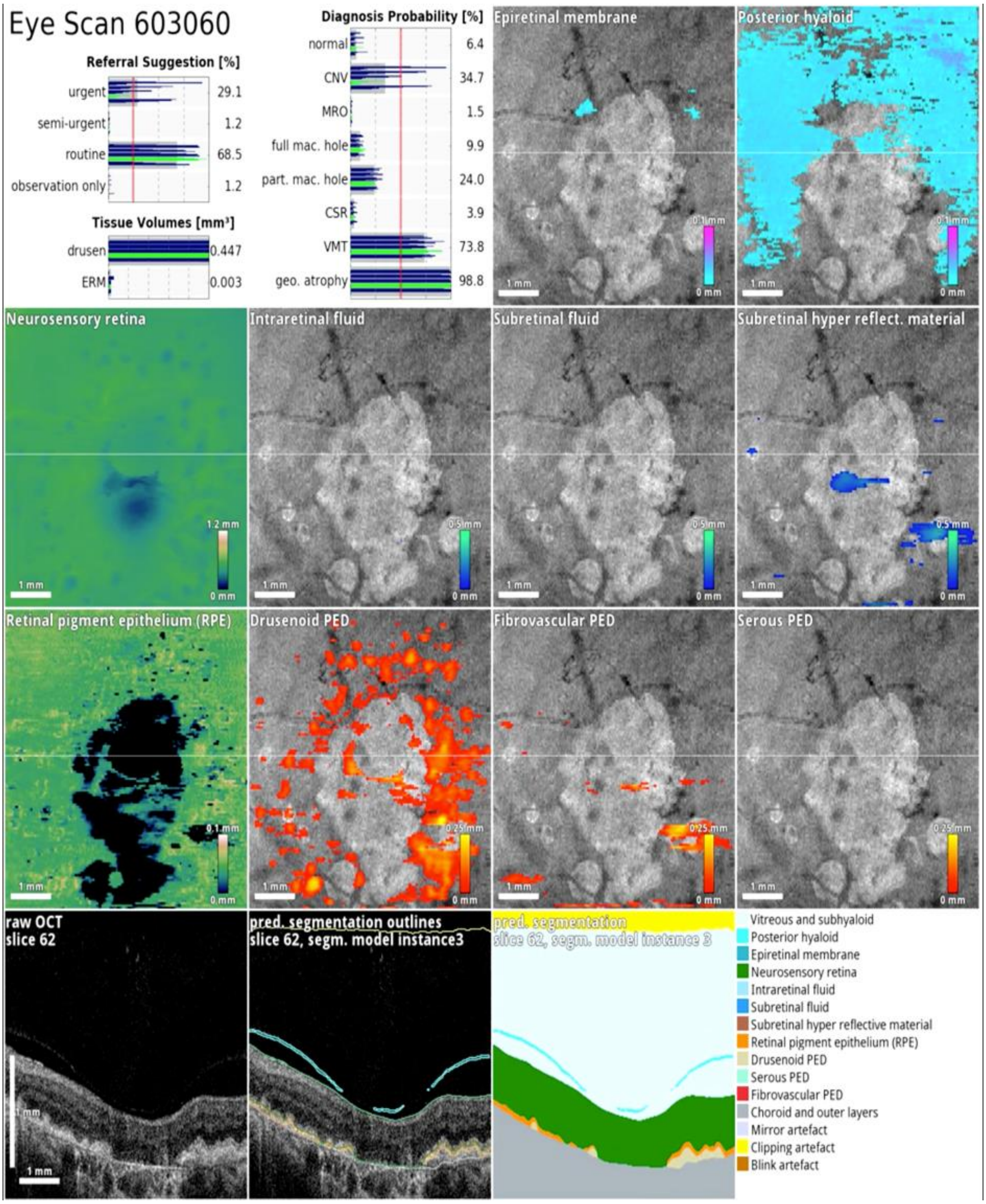


Figure 9: The application of deep learning to the segmentation of retinal optical coherence tomography (OCT) images – the prototype OCT viewer for the Moorfields-DeepMind deep learning system. In this challenging case of retinal angiomatous proliferation (RAP), the system correctly segments an area of intraretinal fluid (IRF) overlying an area of subretinal hyperreflective material (SHRM). It classifies the presence of both macular retinal edema and choroidal neovascularization, but recommends urgent referral to an ophthalmologist. Through the creation of an intermediate tissue representation (seen here as 2D thickness maps for each morphologic parameter), the system provides “interpretability” for the ophthalmologist.