

# Accepted Manuscript

Deep learning in ophthalmology: The technical and clinical considerations

Daniel S.W. Ting, Lily Peng, Avinash V. Varadarajan, Pearse A. Keane, Phil Burlina, Michael F. Chiang, Leopold Schmetterer, Louis R. Pasquale, Neil M. Bressler, Dale R. Webster, Michael Abramoff, Tien Y. Wong



PII: S1350-9462(18)30090-9

DOI: <https://doi.org/10.1016/j.preteyeres.2019.04.003>

Reference: JPRR 759

To appear in: *Progress in Retinal and Eye Research*

Received Date: 23 December 2018

Revised Date: 21 April 2019

Accepted Date: 23 April 2019

Please cite this article as: Ting, D.S.W., Peng, L., Varadarajan, A.V., Keane, P.A., Burlina, P., Chiang, M.F., Schmetterer, L., Pasquale, L.R., Bressler, N.M., Webster, D.R., Abramoff, M., Wong, T.Y., Deep learning in ophthalmology: The technical and clinical considerations, *Progress in Retinal and Eye Research* (2019), doi: <https://doi.org/10.1016/j.preteyeres.2019.04.003>.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Deep Learning in Ophthalmology: The Technical and Clinical Considerations**Daniel S.W. Ting MD PhD<sup>1</sup>Lily Peng MD PhD<sup>2</sup>Avinash V. Varadarajan MS<sup>2</sup>Pearse A. Keane FRCOphth<sup>3</sup>Phil Burlina PhD<sup>4,5,6</sup>Michael F. Chiang MD<sup>7</sup>Leopold Schmetterer PhD<sup>1, 8,9,10</sup>Louis R. Pasquale MD<sup>11</sup>Neil M. Bressler MD<sup>4</sup>Dale R Webster PhD<sup>2</sup>Michael Abramoff MD PhD<sup>12</sup>Tien Y. Wong MD PhD<sup>1</sup>

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 Varadarajan 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 Intelere retina, 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](mailto:daniel.ting.s.w@singhealth.com.sg)

**Abstract**

The advent of computer graphic processing units, improvement in mathematical models and availability of big data has allowed artificial intelligence (AI) using machine learning (ML) and deep learning (DL) techniques to achieve robust performance for broad applications in social-media, the internet of things, the automotive industry and healthcare. DL systems in particular provide improved capability in image, speech and motion recognition as well as in natural language processing. In medicine, significant progress of AI and DL systems has been demonstrated in image-centric specialties such as radiology, dermatology, pathology and ophthalmology. New studies, including pre-registered prospective clinical trials, have shown DL systems are accurate and effective in detecting diabetic retinopathy (DR), glaucoma, age-related macular degeneration (AMD), retinopathy of prematurity, refractive error and in identifying cardiovascular risk factors and diseases, from digital fundus photographs. There is also increasing attention on the use of AI and DL systems in identifying disease features, progression and treatment response for retinal diseases such as neovascular AMD and diabetic macular edema using optical coherence tomography (OCT). Additionally, the application of ML to visual fields may be useful in detecting glaucoma progression. There are limited studies that incorporate clinical data including electronic health records, in AI and DL algorithms, and no prospective studies to demonstrate that AI and DL algorithms can predict the development of clinical eye disease. This article describes global eye disease burden, unmet needs and common 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 needs, and the potential challenges for clinical adoption are discussed. AI, ML and DL will likely 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 ageing populations globally.

## List of Contents

1. Introduction
2. Development of deep learning algorithms: technical considerations
  - i. Fundamentals of a convolutional neural network
  - ii. Software framework
  - iii. Common network architectures and transfer learning
  - iv. Pre-processing and gradability
  - v. Training, validation and testing datasets
  - vi. Datasets characteristics
  - vii. Reference standard
  - viii. Performance metrics
  - ix. Methods to explain diagnosis
3. Deployment of DL algorithms: clinical considerations
  - i. Diabetic retinopathy
    - i. DR classifications
    - ii. Reference standard
    - iii. Fundus Imaging
    - iv. Detection of non-DR findings
    - v. Models of care
    - vi. Screening workflow
    - vii. Future directions
  - ii. Glaucoma and glaucoma suspect
    - i. Challenges in glaucoma and glaucoma suspect definitions
    - ii. Optic disc imaging
    - iii. Visual field
    - iv. Clinical forecasting
    - v. Potential challenges
    - vi. Future directions
  - iii. Age-related macular degeneration
    - i. Different AMD classifications
    - ii. Fundus-based DL algorithms
    - iii. OCT-based DL algorithms
    - iv. OCT segmentation of retinal changes
    - v. OCT algorithm to detect neovascular AMD

- vi. OCT algorithm to triage referral urgency
  - vii. OCT algorithm to predict treatment outcome
  - viii. Future directions
- iv. Retinopathy of prematurity
  - i. Current challenges with ROP diagnosis
  - ii. DL algorithms on Retcam imaging
  - iii. Future directions
- v. Miscellaneous conditions
  - i. Cardiovascular disease
  - ii. Retina is the window to the cardiovascular health
  - iii. AI to predict systemic cardiovascular risk factors
  - iv. AI for refractive error
- 4. Potential challenges
- 5. Conclusions

## 1. Introduction

Artificial intelligence (AI) was conceptualized in 1956, after a workshop at Dartmouth College (**Figure 1**). (McCarthy, Minsky et al. 1955) 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”. (Samuel 1959) 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. With the advent of graphic processing units (GPUs), advances in mathematical models, the availability of big datasets and low cost sensors, deep learning (DL) techniques subsequently, has sparked tremendous interest and been applied in many industries. (LeCun, Bengio et al. 2015) DL utilizes multiple processing layers to learn representation of data with multiple levels of abstraction. (Lee, Tying et al. 2017) DL approaches use complete images, and associate the entire image with a diagnostic output, thereby eliminating the use of “hand-engineered” image features. With improved performance, (Abramoff, Lou et al. 2016, Gulshan, Peng et al. 2016) DL is now widely adopted in image recognition, speech recognition and natural language processing.

In medicine, the most robust AI algorithms have been demonstrated in image-centric specialties, including radiology, dermatology, pathology and increasingly so in ophthalmology. (Schmidt-Erfurth, Sadeghipour et al. 2018, Ting, Pasquale et al. 2018) DL algorithms were found to be effective in detecting pulmonary tuberculosis from chest radiographs, (Lakhani and Sundaram 2017, Hwang, Park et al. 2018) and to differentiate malignant melanoma from benign lesions on digital skin photographs. (Esteva, Kuprel et al. 2017) In ophthalmology, there have been two major areas in which DL systems have been applied. First, DL systems have been shown to accurately detect diabetic retinopathy (DR), (Abramoff, Lou et al. 2016, Gulshan, Peng et al. 2016, Gargeya and Leng 2017, Ting, Cheung et al. 2017) glaucoma, (Ting, Cheung et al. 2017, Li, He et al. 2018) age-related macular degeneration (AMD), (Burlina, Joshi et al. 2017, Ting, Cheung et al. 2017, Grassmann, Mengelkamp et al. 2018) retinopathy of prematurity (ROP), (Brown,

Campbell et al. 2018) and refractive error, (Varadarajan, Poplin et al. 2018) from digital fundus photographs. Cardiovascular risk factors such as blood pressure have also been accurately predicted from fundus photographs. (Poplin, Varadarajan et al. 2018, Ting, Cheung et al. 2019) Second, there are new studies that show several retinal conditions [e.g., choroidal neovascular membrane [CNV], earlier stages of AMD, and diabetic macular edema (DME)] (Lee, Tying et al. 2017) can also be detected accurately with AL algorithms applied on optical coherence tomography (OCT) images. (De Fauw, Ledsam et al. 2018, Kermany, Goldbaum et al. 2018)

To date, several AI review articles have been published thus far, summarizing the deep learning technologies in Ophthalmology. Nevertheless, none of which have focused on the technical and clinical considerations in building deep learning (DL) algorithms for fundus photographs and OCTs. This objective of article, therefore, is to describe the important technical and considerations of building DL algorithms in the research setting, as well as the deployment of these algorithms in the clinical settings.

## 2. Development of DL Algorithms: Technical Consideration

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). This section will focus on the technical aspects of building a DL algorithm, including the understanding of the fundamental of a CNN, software framework, network architectures, datasets selection, characteristics, performance metrics, reference standard and the visualization techniques to improve the explainability of the algorithms (**Table 1**).

### a) *Fundamentals of 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, (Hubel and Wiesel 1968) 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.

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.

#### *b) Software frameworks*

CNNs are commonly implemented in several popular software frameworks. Early development in these past 10 years was enabled by the availability of frameworks like Caffe,<sup>74</sup> Torch,<sup>13</sup> and Theano<sup>75</sup>. More recently, Python-based frameworks such as TensorFlow<sup>76</sup> and PyTorch<sup>13</sup> have gained more popularity among the deep

learning community. High-level application programming interface (APIs) such as Keras<sup>12</sup> or Lasagne have also made it much easier to develop DL systems, by simplifying the existing networks architectures and pretrained weights. Given that this is more convenient for the purposes of transfer learning and fine tuning, they could be considered as starting points for implementation for new users.

*c) Common network architectures and transfer learning*

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.(Krizhevsky, Sutskever et al. 2012) 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. Apart from the classification tasks, U-net, first described by Olaf Ronneberger, has achieved particular success in performing segmentation tasks on optical coherence tomography (OCT), given its flexibility in input sizes and dimensionality.(Ronneberger, Fischer et al. 2015)

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. 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 AMD on fundus images). 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).

*d) Pre-processing and gradability*

A pre-processing algorithm is crucial in standardizing the input of a retinal image, given that different retinal cameras may have different characteristics (e.g. a black border surrounding the retinal image, circular vs rectangular image and etc.). The standardization of the input images (contrast adjustment and auto-cropping of the image borders) may help to optimize the training and testing of a DL algorithm.

It is important for a DL algorithm to assess the image quality of a retinal image using the gradability algorithm, given that a suboptimal retinal image may affect the diagnostic outcome. This is especially important when a DL algorithm is being deployed in the real-world settings, where the patients may be uncooperative, have small pupils or cataracts. On the other hand, one needs to be also cautious in setting the appropriate threshold of this algorithm, as the ungradable retinal images may result in the direct referrals to the tertiary eye care settings. If the criteria for gradability is too stringent, this may result in many unnecessary referrals.

*e) Training, validation and testing datasets*

The training and development phase usually is split into training, validation and testing datasets. These datasets must not intersect; an image that is in one of the datasets (e.g., training) must not be used in any of the other datasets (e.g., validation). Ideally, this non-intersection should extend to patients. The general class distribution for the targeted condition should be maintained in all these datasets.

Training dataset: 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.

Validation dataset: Validation is used for parameter selection and tuning, and is customarily also used to implement stopping conditions for training.

Testing dataset: Finally, the reported performance of the AI algorithm should be computed exclusively using the selected optimized model weights on the testing datasets. It is important 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.

*f) Datasets characteristics*

For any AI study, particularly imaging studies, it is important to demonstrate the population in which the DL system was developed and tested on. The reporting of dataset characteristics, including basic demographics (e.g., age, gender, ethnicity) and imaging data platform, size of field of view, reference standard, are important. This is especially so because DL systems can predict additional features that are not discernable to manual inspection like age and gender.(Poplin, Varadarajan et al. 2018) These characteristics might be augmented by including the systemic factors (e.g. blood pressure, blood sugar level etc.) for vascular conditions such as DR. 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.(Wicherts, Veldkamp et al. 2016) 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.(Quellec and Abramoff 2014, Wong and Bressler 2016)

*g) Reference standard*

In order to report the diagnostic performance of an AI system, gold standard or reference standard (also known as ground truth) plays a pivotal role. In

ophthalmology, the reference standard/s are usually ophthalmologists, reading center graders, non-physician professional trained technicians, or optometrists. In terms of examination methods, it could be done as clinic-based examinations, or image-based examination. When examining the outcome metrics, it is also important to evaluate the design and technical method of a DL algorithm, versus the reference standards. For example, a DL algorithm, if developed using 1-field fundus photograph, will underperform when it compares against the reference standard that uses wide field fundus photography or 7-field 30 degrees retinal photography. Lastly, many conditions have different classifications and it is important to standardize these gradings prior to the training or testing of the DL algorithms. More details will be discussed under the clinical considerations sections.

#### *h) Performance metrics*

In terms of the performance metrics, the most commonly used is the area under the receiver's operator characteristics curve (AUC), computed using sensitivity (also known as recall) 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 a pre-set operating threshold (i.e., 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 and compromising patients' safety. Apart from AUC, other parameters should include 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, for these reasons, the AI study should state AUC, sensitivity and specificity as the bare minimum. For assessment of segmentation accuracy, the dice coefficient is commonly used by the ML community. It measures the overlap between automated and "gold standard" manual

segmentation, or the Jaccard index (“intersection over union”).(Anwar, Majid et al. 2018) In the clinical literature, the agreement between automated and manual segmentation is most commonly measured using Bland-Altman plots.(Bunce 2009)

#### *i) Methods to explain the diagnosis*

After creating a robust DL algorithm using the above approaches, it is important for the DL algorithms to then explain its rationale for diagnosis, in order to assist the physicians to highlight the abnormal areas on the images, and to educate/counsel the patients about their diagnosis (**Figure 3**). DL systems are commonly referred to as a ‘black-box’,(Carin and Pencina 2018) impacting the adoption of such technology within clinical settings. Nevertheless, the recent technical advancement in the visualization maps may provide a solution to this. Visualization of the network workings and activation can be achieved using several methods, for example occlusion testing, integrated gradients and soft attention. It allows the generation of overlay highlights that show where the network is looking when it renders a classification. **Figure 3 and 4** demonstrates some of the visualization techniques used by different AI groups to highlight the abnormal areas in the retinal images.(Poplin, Varadarajan et al. 2018)

### **3. Deployment of DL algorithms: clinical considerations**

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. People are living longer, and the pace of ageing is much faster than in the past.(Divo, Martinez et al. 2014) In a systematic review,(Bourne, Flaxman et al. 2017) the number of people with visual impairment and blindness are growing, given the ageing population and growth of the population. Of these, DR, glaucoma, AMD are found to be the major causes for moderate to severe vision loss.(Flaxman, Bourne et al. 2017) Population expansion also creates pressure to screen for important causes of childhood blindness such as retinopathy of prematurity (ROP), refractive error, and amblyopia.(Wheatley, Dickinson et al. 2002) In order to rectify the manpower and expertise shortage, DL algorithms may be utilized as alternative screening tools. Nevertheless, it is important to consider the various clinical factors associated with

each eye conditions, in order to ensure appropriate deployment and implementation of these DL algorithms within the clinical practice.

#### a) **Diabetic retinopathy**

Over the past 24 months, many AI groups have published various studies in DR screening (**Table 2**), with most reporting robust diagnostic performance in either detecting referable DR or any DR. It is important to understand the clinical implications with respect to the technical design of the individual DL algorithms in DR screening.

##### i. DR classifications

For DR model outputs, it can be either a binary or multi-class classification tasks. Most of the models have been trained to detect referable DR defined as moderate non-proliferative DR (NPDR) 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). At present, there are several DR classifications with variable definitions for DR severity levels. These include the International Clinical Diabetic Retinopathy Severity Scales (ICDRSS), International Clinical Diabetic Macular Edema Severity Scales (ICDMESS), National Health Service (NHS) DR Guidelines, Early Treatment Diabetic Retinopathy Study (ETDRS) classification and etc. For example, the moderate NPDR in ICDRSS is different from the moderate NPDR in ETDRS and R2 (pre-proliferative retinopathy) based on the NHS guidelines. Hence, it is important to understand the differences between these classifications prior to the training and testing of the DL algorithms.

##### ii. Reference standard

Different reference standards were utilized in many published papers in DR screening thus far, including retinal specialists, ophthalmologists, graders from the reading center (e.g. Wisconsin Reading center). In a pre-registered US FDA clinical trial, Abramoff et al reported sensitivity of 87.2% (>85%), specificity of 90.7% (>82.5%) in detection of referable DR (worse than mild DR), and gradability rate of 96.1% were reported, (Abramoff, Lavin et al. 2018) with reference to grading performed at the Wisconsin Reading Center. Notably, the grading was performed



using stricter criteria - 4 fields retinal examination with OCT diagnosis of presence/absence of DME, although the DL algorithm was developed using 2-field and 2-D fundus photographs.

The reference standard varies between different AI studies and thus, it may be challenging to compare one AI system to the other. Moreover, it is important to test an AI system on the independent datasets using a pre-fixed operating threshold. For example, Ting *et al* developed a DL algorithm that was tested on 11 independent datasets.(Ting, Cheung et al. 2017) Using a pre-fixed operating threshold, the DL algorithm achieved a 90.5% sensitivity and 91.6% specificity on a primary testing dataset, and AUC of >0.90, sensitivity (>90%) and specificity (>70%) on the 10 independent datasets, consisting of multi-ethnic population from Singapore, China, Hong Kong, USA, Mexico and Australia.

### iii. Fundus Imaging

Given that many DR screening programs worldwide are performing 2-field retinal still photography, many DL algorithms were trained to detect analyse the optic disc- and macula-centered retinal images.(Ting, Cheung et al. 2017, Abramoff, Lavin et al. 2018, Li, Keel et al. 2018) In the low resource countries, it may be less labour and time consuming to perform 1-field retinal photography. Using the enhanced DL algorithms developed by Gulshan et al,(Gulshan, Peng et al. 2016) the Google AI has reported clinically acceptable diagnostic performance for Thailand population with diabetes.(Raumviboonsuk, Krause et al. 2019) This is an efficient automated DR screening method, as most DR changes usually occur in the posterior pole, although some may occur occasionally at the nasal retina that may not be able to be detected by the 1-field DL algorithm.

### iv. Detection of non-DR findings

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.(Ting, Cheung et al. 2017) 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.

#### v. Models of care

Several models of care can be considered to implement AI for DR screening in the clinical practice. It can be either deployed as cloud-based, office-based or retinal camera-based settings. For cloud-based setting, this requires a tele-ophthalmology platform to enable the AI analysis of the retinal images. This is a suitable model for countries (e.g. Singapore, United Kingdom or United States) that have existing tele-retinal DR screening programs. The AI can be integrated into this information technology (IT) platform to help analyse the retinal images. Should the tele-communication be challenging, the alternative clinical model is to deploy the AI in an application programming interface (API) using tablets, laptops or desktops, in an office-based setting. This may be a more suitable model for the low resource countries where there is suboptimal internet bandwidth. Lastly, it is also possible to build the AI algorithm into the retinal camera, providing an instantaneous diagnosis after the images are captured. However, this approach may potentially limit the use of such DL algorithm for other retinal cameras that are currently available in the market.

#### vi. Screening workflow

AI system can be deployed as a stand-alone fully automated system or an assistive semi-automated model. This is an important factor to decide on the operating threshold of a DL algorithm. For the fully automated system, it is important to take both sensitivity and specificity into account when the operating threshold is set. While attempting to aim for high sensitivity, one also needs to ensure that the specificity is not being highly compromised, resulting in many unnecessary false positive referrals to the healthcare settings. On the other hand, for countries with

existing DR screening programs with manual graders, the assistive semi-automated model may be an excellent alternative approach to reduce the manpower requirement. The DL algorithm can be set at a high-sensitivity threshold to filter the normal or non-referable retinal images, while the manual graders can perform a secondary grading on those retinal images that are deemed referable. This hybrid approach not only can aim for an overall excellent sensitivity and specificity, but also could potentially reduce the manpower headcount for manual grading. With some of the visualization techniques discussed in the previous technical section, this may be a good model for the secondary graders to look for the disease lesions from the abnormal scans for confirmation.

#### vii. Future directions

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 and medicolegal implication.

#### b) **Glaucoma and Glaucoma Suspect**

Apart from DR, many screening programs suggest screening for referable glaucoma suspects. In a systematic review, glaucoma was shown to be leading causes of

blindness worldwide,(Flaxman, Bourne et al. 2017) accounting for 2.9 million patients worldwide. The number with glaucoma is expected to increase up 111.8 million by 2040.(Tham, Li et al. 2014) For glaucoma, AI plays a pivotal role in screening, diagnosis and surveillance of the disease.

#### A. Challenges in glaucoma and glaucoma suspect definitions

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 optic nerve head 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 97.5<sup>th</sup> percentile of vertical CDR or of CDR asymmetry as a standard definition of structural glaucomatous damage.(Foster, Buhrmann et al. 2002) This definition is, however, not sufficient for glaucoma diagnosis, because of the large influence of disc size(Crowston, Hopley et al. 2004) 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.(Chauhan and Burgoyne 2013)

On OCT retinal nerve fibre layer thickness and ganglion cell complex measurements are used to discriminate glaucoma from healthy.(Savini, Carbonelli et al. 2011) More recently minimum rim width as measured from Bruch's membrane opening has been used as a novel diagnostic tool in glaucoma.(Chauhan, Danthurebandara et al. 2015) 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 97.5<sup>th</sup> 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.

## B. Optic Disc Imaging

Optic disc fundus imaging is the least expensive imaging modality to conduct structural assessment of the optic nerve, although the sensitivity and specificity in detecting glaucoma suspect or glaucoma are not comparable to the combined structural and functional assessment using more sophisticated imaging devices such as optical coherence tomography or Humphrey visual fields.

Given that the optic disc fundus imaging is commonly taken and analyzed as part of the DR screening exercises, it is important to have a good DL algorithm in detecting glaucoma +/- glaucoma suspect from the colour retinal images (**Figure 4**). To date, most DL algorithms for disc suspect are developed using large number of retinal images collected from DR screening programs (e.g. Ting et al and Li et al). (Ting, Cheung et al. 2017, Li, He et al. 2018) In these 2 studies, the DL algorithms for detection of glaucoma suspect were developed from the optic disc images (defined as CDR 0.8 or worse and/or glaucomatous changes), with excellent diagnostic outcome of >90% accuracy (**Table 3**). These retinal images, however, were graded and assessed in a 2-dimensional manner without a thorough clinical evaluation with measurement of intraocular pressure, structural or functional confirmation of the diagnosis.

Using 3242 fundus images, Shibota et al developed a DL algorithm that is trained and tested with the eyes with confirmed glaucoma, reporting an excellent AUC of 0.965. (Shibata, Tanito et al. 2018) 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 also able to detect early glaucoma with

an AUC of 0.937 (Sensitivity = 82.5% and Specificity = 93.9%).(Asaoka, Murata et al. 2018) 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 3**). (Masumoto, Tabuchi et al. 2018)

### *C. Visual Field*

Relative to optic disc photographs or OCT images, the data contained in 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 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. (Elze, Pasquale et al. 2015) This algorithm has been validated (Cai, Elze et al. 2017) and has proven useful in augmenting the GHT for the detection of early functional glaucomatous loss. (Wang, Pasquale et al. 2018) 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. (Li, Wang et al. 2018) Yousefi et al. used an alternative Gaussian mixture and expectation maximization method to decompose VFs along different axes to detect VF progression. (Yousefi, Goldbaum et al. 2014) 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.

### *D. 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.(Schell, Lavieri et al. 2013) 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.(Garcia, Nitta et al. 2018)

#### E. Potential challenges

For glaucoma, the issue is most complicated when DL approaches shall be applied to the classification. 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(Chang and Singh 2016). 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(Schmidl, Schmetterer et al. 2015) and advanced visual field defect is an important risk factor for transitioning to functional blind(Peters, Bengtsson et al. 2014). 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(Fledelius and Goldschmidt 2010, Kwon, Sung et al. 2017). 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.

#### *F. 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.(Hong, Koenigsman et al. 2018) 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.(Muhammad, Fuchs et al. 2017) Algorithms that not only ascertain if there is optic nerve pathology but the regional location of pathology would be widely accepted.

#### **c) Age-related Macular Degeneration (AMD)**

AMD is another major cause of vision impairment, accounting for 8.7% of all blindness worldwide.(Bressler 2004, Wong, Loon et al. 2006, Baeza, Orozco-Beltran et al. 2009, Wong, Su et al. 2014) It is projected that 288 million may have some forms of AMD by 2040, with approximately 10% having intermediate AMD or worse.(Wong, Su et al. 2014) The treatment for neovascular AMD patients has been revolutionized with the advent of anti-vascular endothelial growth factors (VEGF),(Group, Martin et al. 2011, Chakravarthy, Harding et al. 2013) with many countries, e.g. US, Australia, reporting a significant drop in incident blindness by >50%. (Bressler, Doan et al. 2011, Mitchell, Bressler et al. 2014) The American Academy of Ophthalmology 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. These patients will require a referral to the tertiary eye care setting for further clinical evaluation and investigations (e.g. OCT and fundus fluorescein angiogram). With ageing population, DL algorithms could be utilized as alternative tools to aid screening, diagnosis, prognostication and disease surveillance.

#### A. Different AMD classifications

For AMD, multiple classification systems have been proposed apart from AREDS, including the recent Clinical Classification as worked out by the Beckman Initiative for Macular Research Classification Committee (Ferris, Wilkinson et al. 2013) and the Three Continent AMD Consortium Severity Scale (Klein, Meuer et al. 2014) 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 (Brandl, Zimmermann et al. 2018). DL-based classification systems have been developed for referability (Burlina, Joshi et al. 2018), severity characterization and estimation of 5-year risk (Burlina, Joshi et al. 2018) and disease conversion (Schmidt-Erfurth, Waldstein et al. 2018).

#### B. Fundus-based DL algorithms

Many of the AI systems for AMD were built using the age-related eye disease study (AREDS) dataset (Burlina, Joshi et al. 2018, Grassmann, Mengelkamp et al. 2018) while some utilized other datasets (Ting, Cheung et al. 2017). Similar to DR and glaucoma, most DL algorithms reported robust diagnostic performance in detecting referable AMD (defined as intermediate AMD or worse) (**Table 4**). Furthermore, using the AREDS dataset, Burlina et al 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% (Burlina, Joshi et al. 2018). Similarly, Grassmann et al built a DL system for detection of early and late AMD (Grassmann, Mengelkamp et al. 2018) developed using AREDS dataset and tested 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. Given that the AREDS dataset mostly consist of patients aged >55 years old, this DL algorithm mis-treated many



dominant macula reflexes as neovascular AMD. Again, this highlights the importance of training DL algorithms with diverse clinical datasets, consisting of a wide range of disease phenotypes and patients' characteristics.

### C. OCT-based DL algorithms

Increasingly, OCT plays a major role in disease detection, prognostication and surveillance in all AMD patients, especially those wet AMD requiring anti-vascular endothelial growth factor (anti-VEGF). OCT has established itself as the dominant imaging modality, particularly for the diagnosis and management of AMD.(Keane and Sadda 2014) Thirty 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.(Fujimoto and Swanson 2016) 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.(Windsor, Sun et al. 2018)

The OCT DL algorithms can be broadly divided into segmentation and classification tasks. With appropriate segmentation, the DL algorithm can also delineate the abnormal areas on the OCT scans, providing the surface areas or volume of the abnormal regions. 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).

### D. OCT segmentation of retinal changes

Lee et al. described the use of DL for segmentation of intraretinal fluid in OCT images (**Table 5**).(Lee, Tying et al. 2017) Using Spectralis OCT images that have intraretinal fluid, including DME, RVO, and AMD, they selected 934 manually segmented central subfoveal scans for manual segmentation and a modified U-net for training and testing. Intraretinal fluid was defined as “an intraretinal hyporeflective space surrounded by reflective septate”. The DL algorithm showed good performance for human interrater reliability and the DL system, with dice coefficients of 0.750 and 0.729, respectively.

Few groups have extended their models to perform segmentation of pigment epithelium detachment (PED), the formation of a potential space between the RPE and Bruch's membrane. (Zayit-Soudry, Moroz et al. 2007, Xu, Yan et al. 2017) Schmidt-Erfurth et al. have reported the correlation of PED metrics with visual acuity in patients with neovascular AMD using a DL-based system. (Schmidt-Erfurth, Bogunovic et al. 2018) 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.

#### E. OCT algorithm to detect neovascular AMD

Using close to 100,000 OCT B-scans (50% normal and 50% AMD scans), Lee et al reported an accuracy of 87.6% with 84.6% sensitivity and 91.5% specificity. (Lee, Baughman et al. 2017) This DL algorithm was developed using the OCT scans identified via the clinical data from the electronic health record (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. They also adopted occlusion testing to highlight the abnormal OCT areas by using a blank 20x20 pixel area.

#### F. OCT algorithm to triage referral urgency

Using 14,884 OCT scans, De Fauw et al. showed that the DL algorithm was able to detect those who require urgent referrals with excellent performance (AUC of >0.90), using 2 different OCT systems (Topcon and Spectralis). (De Fauw, Ledsam et al. 2018) This DL algorithm utilized nine contiguous OCT scans, a three-dimensional U-net architecture and intermediate tissue representation to output automated segmentations across 15 different label classes. These labels encompass a range of novel OCT biomarkers, including three forms of PED (fibrovascular, serous, and drusenoid) and subretinal hyperreflective material. This model segments the posterior hyaloid and epiretinal membrane (ERM), to allow enhanced assessment of

vitreomacular interface disorders, and the RPE, allowing for the quantification of retinal degeneration and atrophic changes (**Figure 5 and 6**). The authors also highlighted the need to perform domain adaptation to fine tune the DL algorithm that was developed using a completely different device. Prior to re-training for the new device, the total error rate for referral suggestions were as high as 46.6%. Nevertheless, by adding an additional 152 scans (527 manually segmented slices in total) from the new device, the error rate was brought down to 3.4% (4 out of 116).

#### G. OCT algorithm to predict treatment outcome

Schmidt-Erfurth et al., used the HARBOR data to develop ML models to predict visual acuity in patients receiving ranibizumab for neovascular AMD. (Schmidt-Erfurth, Bogunovic et al. 2018) They began by selecting 70% of the HARBOR dataset for analysis. They next applied automated segmentation algorithms (using both graph-based and DL 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 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.

#### H. Future Directions

Future research is important to evaluate the generalizability and cost-effectiveness of these DL systems in a larger international multi-ethnic cohort. Apart from screening

purposes, it will be of great value to generate new algorithms to predict and prognosticate the functional, structural and treatment outcome for AMD patients, with appropriate stratification of the risk profiles. Ideally, the development of the algorithm should incorporate multi-modal approach – clinical data (functional, structural, treatment outcome), fundus photographs and OCT imaging. who are likely to progress in the long run, coupled with clinical data and treatment outcome.

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.

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.

#### **d) Retinopathy of Prematurity**

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 (**Figure 7**). Globally, it is estimated that 15 million babies are born prematurely each year.(Quinn 2016) In US, the incidence of ROP was 19.9%.(Ludwig, Chen et al. 2017) ROP accounts for 6 to 18% of childhood blindness,(Fleck and Dangata 1994) causing significant psycho-social impact on the child and the family.(Blencowe, Vos et al. 2013) According to the Early Treatment for ROP (ETROP) trial,(Early Treatment for Retinopathy of Prematurity Cooperative, Good et al. 2010) 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.

#### A. Current challenges with 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.(Gilbert, Rahi et al. 1997) Meanwhile, the supply of clinicians who perform ROP management is limited by logistical challenges of coordinating examination at the neonatal intensive care unit 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.(Chan, Williams et al. 2010, Myung, Chan et al. 2011, Nagiel, Espiritu et al. 2012, Wong, Ventura et al. 2012)

In particular regarding clinical care, there are a number of real-world challenges regarding plus disease diagnosis:

- i. There is often significant variability in diagnostic classification (plus vs. pre-plus vs. normal), even among experts,(Chiang, Jiang et al. 2007, Wallace, Quinn et al. 2008, Slidsborg, Forman et al. 2012, Gschließer, Stifter et al. 2015, Campbell, Ryan et al. 2016) leading to inconsistent application of evidence-based practice.(Fleck, Williams et al. 2018) 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.(Reynolds, Dobson et al. 2002) 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.(Daniel, Quinn et al. 2015)

- ii. 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.(Hewing, Kaufman et al. 2013)
- iii. 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.(Rao, Jonsson et al. 2012, Hewing, Kaufman et al. 2013, Keck, Kalpathy-Cramer et al. 2013, Campbell, Ataer-Cansizoglu et al. 2016)
- iv. 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.(Gelman, Gelman et al. 2010)
- v. Studies have shown that there is a geographical variation in plus disease diagnosis possibly related to differences in training,(Fleck, Williams et al. 2018) and that there may be chronological drift showing a tendency to diagnose “plus disease” more frequently than in the past.(Moleta, Campbell et al. 2017)
- vi. 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.(Group 2000, Gole, Ells et al. 2005) However, there is variability in how this definition is interpreted,(Wallace, Quinn et al. 2008, Slidsborg, Forman et al. 2012, Hewing, Kaufman et al. 2013, Gschließer, Stifter et al. 2015) and evidence that this variability may lead to clinically-significant differences in diagnosis.(Slidsborg, Forman et al. 2012, Kim, Campbell et al. 2018)
- vii. The ICROP definition of pre-plus disease(Gole, Ells et al. 2005) is somewhat vague. Studies have found significant levels of variability in diagnosis of pre-



plus disease among experts.(Chiang, Jiang et al. 2007, Wallace, Quinn et al. 2008)

- viii. Vascular abnormality in ROP reflects a continuous spectrum of disease,(Wallace, Kylstra et al. 2000, Gole, Ells et al. 2005, Wallace, Freedman et al. 2011) 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.(Tasman 1988, Reynolds, Dobson et al. 2002) 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”).(Campbell, Kalpathy-Cramer et al. 2016, Kalpathy-Cramer, Campbell et al. 2016)

#### B. DL algorithms on Retcam Imaging

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).(Wittenberg, Jonsson et al. 2012) 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).(Koreen, Gelman et al. 2007, Wilson, Wong et al. 2012, Abbey, Besirli et al. 2016) 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 ML 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.(Ataer-Cansizoglu, Bolon-Canedo et al. 2015) 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.(Ataer-Cansizoglu, Bolon-Canedo et al. 2015)

DL has been applied for automated diagnosis of ROP, which could potentially address barriers to ROP screening on a larger scale.(Worrall, Wilson et al. 2016) 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.(Brown, Campbell et al. 2018)

### C. 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.

### **e) Miscellaneous conditions**

#### **A. Cardiovascular disease**



Cardiovascular diseases (CVDs) is the largest cause of non-communicable deaths worldwide. For 2018, WHO estimated that 17.9 million people died of CVD worldwide in 2012, accounting for an estimated 31% of global mortality.(Roth, Johnson et al. 2017) 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. 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,(Stone, Robinson et al. 2014) Framingham(Wilson, D'Agostino et al. 1998, National Cholesterol Education Program Expert Panel on Detection and Treatment of High Blood Cholesterol in 2002, D'Agostino, Vasan et al. 2008) and SCORE,(Conroy, Pyorala et al. 2003, Graham, Atar et al. 2007) which is based on various factors from patient history (e.g. age, self-reported sex, smoking status) and blood samples (e.g. lipid panels).(Goff, Lloyd-Jones et al. 2014) 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(Hira, Kennedy et al. 2015).

#### **B. Retina is the window to the cardiovascular health**

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(Yeboah, McClelland et al. 2012) 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(Cheung, Tay et al. 2013) or chronic kidney disease.(Yip, Ong et al. 2017) These features include vessel caliber,(Wang, Liew et al. 2006, Wong, Kamineni et al. 2006, Seidemann, Claggett et al. 2016) bifurcation or tortuosity,(Witt, Wong et al. 2006) 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.

### C. AI to predict systemic cardiovascular risk factors

In a recent study, Poplin and Varadarajan *et al* (Poplin, Varadarajan *et al.* 2018) used DL to build a model that predicted cardiovascular risk factors using retinal fundus images from 48,101 patients from the UK Biobank study (2017) and 236,234 from the EyePACS population. (2017) 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. (Ting and Wong 2018) 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, 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, body mass index (BMI) was used as a proxy while calculating the SCORE risk. (Cooney, Dudina *et al.* 2009, Dudina, Cooney *et al.* 2011, 2017)

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 (Simonyan K 2017) for a few predictions is shown in **Figure 8**.

Despite these promising results, efforts to improve the performance and interpretability of these DL models seems indicated, especially for MACE. In this

study, Poplin *et al* study did not include blood tests such as lipid panels in the analysis because it was not available. (Poplin, Varadarajan et al. 2018) 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. Lastly, It is also important to explore how this DL algorithm can be incorporated into the current cardiovascular risk calculators to improve the predictive power for 5-year MACE risks.

#### D. AI for refractive error

In the previous examples with CVD 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 *et al* (Varadarajan AV 2017) showed that DL 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. 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 9**). 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. The accuracy of the refractive error prediction seemed to decrease with a smaller field of view, poorer image quality, and possibly macular lesions.

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.

#### **4. Potential challenges for AI implementation within clinical practice**

First, 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.(Gomes 2014) 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.

Second, 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 an 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. Nonetheless, all the AI research groups worldwide should continue to collaborate to rectify this barrier, aiming to harness the power of big data and DL to advance the discovery of scientific knowledge.

Third, 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 intellectual property 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 lose interest in the topic.

Fourth, in the training set, a large number of images is required that need to be well phenotyped for different diseases (e.g. DR, glaucoma and AMD). 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.

Fifth, while 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 (Fiorini 2016, Menti, Bonaldi et al. 2016). 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.

Sixth, 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. 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 OCT-angiography or Doppler OCT (Dobhoff-Dier, Schmetterer et al. 2014, Leitgeb, Werkmeister et al. 2014), this may have considerable potential for diagnosis, classification and progression analysis, this is an important challenge for the future.

Seventh, 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 effort 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.

Eighth, there is lack of ethical and legal regulations for DL algorithms. These concerns can occur during the data sourcing, product development and clinical deployment stage. (Char, Shah et al. 2018, Vayena, Blasimme et al. 2018) Char et al stated that the intent behind the design of DL algorithms also needs to be considered. (Char, Shah et al. 2018) One needs to be careful about building racial biases into the healthcare algorithms, especially when the healthcare deliveries already varies by race. Moreover, given the growing importance of quality indicators for public evaluations and reimbursement rates, there may be a tendency to design the DL algorithms that would result in better performance metrics, but not necessarily better clinical care for the patients. Traditionally, a physician could withhold the patients' information from the medical record in order to keep it confidential. In the era of digital health record integrated with the deep-learning-based decision support,

it would be hard to withhold patients' clinical data from the electronic system. Hence, the medical ethics surrounding these issues may need to evolve over time.

## 5. Conclusions

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, and it 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, Ms Valentina Bellemo, Ms Yuchen Xie and Ms Michelle Yip (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

- "World Health Organization (WHO). Global Health and Ageing. 2018. URL: [http://www.who.int/ageing/publications/global\\_health.pdf](http://www.who.int/ageing/publications/global_health.pdf) [Accessed on 17th November, 2018]."
- (2017). "About UK Biobank." Retrieved 26 March, 2017, from <http://www.ukbiobank.ac.uk/about-biobank-uk/>.



(2017). "Cardiovascular Disease (10-year risk)." Retrieved 21 June 2017, from <https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/>.

(2017). "Welcome to EyePACS " Retrieved 31 July, 2017, from <http://www.eyepacs.org/>.

Abbey, A. M., C. G. Besirli, D. C. Musch, C. A. Andrews, A. Capone Jr, K. A. Drenser, D. K. Wallace, S. Ostmo, M. Chiang and P. P. Lee (2016). "Evaluation of screening for retinopathy of prematurity by ROPtool or a lay reader." *Ophthalmology* **123**(2): 385-390.

Abramoff, M. D., P. T. Lavin, M. Birch, N. Shah and J. C. Folk (2018). "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices." *NPJ Digital Medicine* **39**: 1-8.

Abramoff, M. D., Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk and M. Niemeijer (2016). "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning." *Invest Ophthalmol Vis Sci* **57**(13): 5200-5206.

Anwar, S. M., M. Majid, A. Qayyum, M. Awais, M. Alnowami and M. K. Khan (2018). "Medical Image Analysis using Convolutional Neural Networks: A Review." *J Med Syst* **42**(11): 226.

Asaoka, R., H. Murata, K. Hirasawa, Y. Fujino, M. Matsuura, A. Miki, T. Kanamoto, Y. Ikeda, K. Mori, A. Iwase, N. Shoji, K. Inoue, J. Yamagami and M. Araie (2018).

"Using Deep Learning and transform learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images." *Am J Ophthalmol*.

Ataer-Cansizoglu, E., V. Bolon-Canedo, J. P. Campbell, A. Bozkurt, D. Erdogmus, J. Kalpathy-Cramer, S. Patel, K. Jonas, R. P. Chan and S. Ostmo (2015). "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* **4**(6): 5-5.

Baeza, M., D. Orozco-Beltran, V. F. Gil-Guillen, V. Pedrera, M. C. Ribera, S. Pertusa and J. Merino (2009). "Screening for sight threatening diabetic retinopathy using non-mydratic retinal camera in a primary care setting: to dilate or not to dilate?" *Int J Clin Pract* **63**(3): 433-438.

Blencowe, H., T. Vos, A. C. Lee, R. Philips, R. Lozano, M. R. Alvarado, S. Cousens and J. E. Lawn (2013). "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 **74 Suppl 1**: 4-16.
- Bourne, R. R. A., S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keefe, J. H. Kempen, J. Leasher, H. Limburg, K. Naidoo, K. Pesudovs, S. Resnikoff, A. Silvester, G. A. Stevens, N. Tahhan, T. Y. Wong, H. R. Taylor and G. Vision Loss Expert (2017). "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis." Lancet Glob Health **5**(9): e888-e897.
- Brandl, C., M. E. Zimmermann, F. Gunther, T. Barth, M. Olden, S. C. Schelter and F. Kronenberg (2018). "On the impact of different approaches to classify age-related macular degeneration: Results from the German AugUR study." **8**(1): 8675.
- Bressler, N. M. (2004). "Age-related macular degeneration is the leading cause of blindness." JAMA **291**(15): 1900-1901.
- Bressler, N. M., Q. V. Doan, R. Varma, P. P. Lee, I. J. Suter, C. Dolan, M. D. Danese, E. Yu, I. Tran and S. Colman (2011). "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 **129**(6): 709-717.
- Brown, J. M., J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Ioannidis, J. Kalpathy-Cramer, M. F. Chiang, Imaging and C. Informatics in Retinopathy of Prematurity Research (2018). "Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks." JAMA Ophthalmol **136**(7): 803-810.
- Bunce, C. (2009). "Correlation, agreement, and Bland-Altman analysis: statistical analysis of method comparison studies." Am J Ophthalmol **148**(1): 4-6.
- Burlina, P., N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong and N. M. Bressler (2018). "Utility of Deep Learning Methods for Referability Classification of Age-Related Macular Degeneration." JAMA Ophthalmol.
- Burlina, P. M., N. Joshi, K. D. Pacheco, D. E. Freund, J. Kong and N. M. Bressler (2018). "Use of Deep Learning for Detailed Severity Characterization and Estimation of 5-Year Risk Among Patients With Age-Related Macular Degeneration." JAMA Ophthalmol.
- Burlina, P. M., N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund and N. M. Bressler (2017). "Automated Grading of Age-Related Macular Degeneration From Color

Fundus Images Using Deep Convolutional Neural Networks." JAMA Ophthalmol **135**(11): 1170-1176.

Cai, S., T. Elze, P. J. Bex, J. L. Wiggs, L. R. Pasquale and L. Q. Shen (2017).

"Clinical Correlates of Computationally Derived Visual Field Defect Archetypes in Patients from a Glaucoma Clinic." Curr Eye Res **42**(4): 568-574.

Campbell, J. P., E. Ataer-Cansizoglu, V. Bolon-Canedo, A. Bozkurt, D. Erdogmus, J. Kalpathy-Cramer, S. N. Patel, J. D. Reynolds, J. Horowitz and K. Hutcheson (2016).

"Expert diagnosis of plus disease in retinopathy of prematurity from computer-based image analysis." JAMA ophthalmology **134**(6): 651-657.

Campbell, J. P., J. Kalpathy-Cramer, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro and M. X. Repka (2016). "Plus disease in retinopathy of prematurity: a continuous spectrum of vascular abnormality as a basis of diagnostic variability." Ophthalmology **123**(11): 2338-2344.

Campbell, J. P., M. C. Ryan, E. Lore, P. Tian, S. Ostmo, K. Jonas, R. P. Chan and M. F. Chiang (2016). "Diagnostic discrepancies in retinopathy of prematurity classification." Ophthalmology **123**(8): 1795-1801.

Carin, L. and M. J. Pencina (2018). "On Deep Learning for Medical Image Analysis." JAMA **320**(11): 1192-1193.

Chakravarthy, U., S. P. Harding, C. A. Rogers, S. M. Downes, A. J. Lotery, L. A.

Culliford, B. C. Reeves and I. s. investigators (2013). "Alternative treatments to inhibit VEGF in age-related choroidal neovascularisation: 2-year findings of the IVAN randomised controlled trial." Lancet **382**(9900): 1258-1267.

Chan, R. P., S. L. Williams, Y. Yonekawa, D. J. Weissgold, T. C. Lee and M. F.

Chiang (2010). "Accuracy of retinopathy of prematurity diagnosis by retinal fellows." Retina (Philadelphia, Pa.) **30**(6): 958.

Chang, R. T. and K. Singh (2016). "Glaucoma Suspect: Diagnosis and Management." Asia Pac J Ophthalmol (Phila) **5**(1): 32-37.

Char, D. S., N. H. Shah and D. Magnus (2018). "Implementing Machine Learning in Health Care - Addressing Ethical Challenges." N Engl J Med **378**(11): 981-983.

Chauhan, B. C. and C. F. Burgoyne (2013). "From clinical examination of the optic disc to clinical assessment of the optic nerve head: a paradigm change." Am J Ophthalmol **156**(2): 218-227 e212.

Chauhan, B. C., V. M. Danthurebandara, G. P. Sharpe, S. Demirel, C. A. Girkin, C. Y. Mardin, A. F. Scheuerle and C. F. Burgoyne (2015). "Bruch's Membrane Opening

- Minimum Rim Width and Retinal Nerve Fiber Layer Thickness in a Normal White Population: A Multicenter Study." Ophthalmology **122**(9): 1786-1794.
- Cheung, C. Y., W. T. Tay, M. K. Ikram, Y. T. Ong, D. A. De Silva, K. Y. Chow and T. Y. Wong (2013). "Retinal microvascular changes and risk of stroke: the Singapore Malay Eye Study." Stroke **44**(9): 2402-2408.
- Chiang, M. F., L. Jiang, R. Gelman, Y. E. Du and J. T. Flynn (2007). "Interexpert agreement of plus disease diagnosis in retinopathy of prematurity." Archives of ophthalmology **125**(7): 875-880.
- Conroy, R. M., K. Pyorala, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, P. Ducimetiere, P. Jousilahti, U. Keil, I. Njolstad, R. G. Oganov, T. Thomsen, H. Tunstall-Pedoe, A. Tverdal, H. Wedel, P. Whincup, L. Wilhelmsen, I. M. Graham and S. p. group (2003). "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project." Eur Heart J **24**(11): 987-1003.
- Cooney, M. T., A. Dudina, D. De Bacquer, A. Fitzgerald, R. Conroy, S. Sans, A. Menotti, G. De Backer, P. Jousilahti, U. Keil, T. Thomsen, P. Whincup and I. Graham (2009). "How much does HDL cholesterol add to risk estimation? A report from the SCORE Investigators." Eur J Cardiovasc Prev Rehabil **16**(3): 304-314.
- Crowston, J. G., C. R. Hopley, P. R. Healey, A. Lee, P. Mitchell and S. Blue Mountains Eye (2004). "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 **88**(6): 766-770.
- D'Agostino, R. B., Sr., R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro and W. B. Kannel (2008). "General cardiovascular risk profile for use in primary care: the Framingham Heart Study." Circulation **117**(6): 743-753.
- Daniel, E., G. E. Quinn, P. L. Hildebrand, A. Ells, G. B. Hubbard, 3rd, A. Capone, Jr., E. R. Martin, C. P. Ostroff, E. Smith, M. Pistilli, G. S. Ying and R. O. P. C. G. e (2015). "Validated System for Centralized Grading of Retinopathy of Prematurity: Telemedicine Approaches to Evaluating Acute-Phase Retinopathy of Prematurity (e-ROP) Study." JAMA Ophthalmol **133**(6): 675-682.
- De Fauw, J., J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman,

- J. Cornebise, P. A. Keane and O. Ronneberger (2018). "Clinically applicable deep learning for diagnosis and referral in retinal disease." Nat Med.
- Divo, M. J., C. H. Martinez and D. M. Mannino (2014). "Ageing and the epidemiology of multimorbidity." Eur Respir J **44**(4): 1055-1068.
- Doblhoff-Dier, V., L. Schmetterer, W. Vilser, G. Garhofer, M. Groschl, R. A. Leitgeb and R. M. Werkmeister (2014). "Measurement of the total retinal blood flow using dual beam Fourier-domain Doppler optical coherence tomography with orthogonal detection planes." Biomed Opt Express **5**(2): 630-642.
- Dudina, A., M. T. Cooney, D. D. Bacquer, G. D. Backer, P. Ducimetiere, P. Jousilahti, U. Keil, A. Menotti, I. Njolstad, R. Oganov, S. Sans, T. Thomsen, A. Tverdal, H. Wedel, P. Whincup, L. Wilhelmsen, R. Conroy, A. Fitzgerald and I. Graham (2011). "Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators." Eur J Cardiovasc Prev Rehabil **18**(5): 731-742.
- Early Treatment for Retinopathy of Prematurity Cooperative, G., W. V. Good, R. J. Hardy, V. Dobson, E. A. Palmer, D. L. Phelps, B. Tung and M. Redford (2010). "Final visual acuity results in the early treatment for retinopathy of prematurity study." Arch Ophthalmol **128**(6): 663-671.
- Elze, T., L. R. Pasquale, L. Q. Shen, T. C. Chen, J. L. Wiggs and P. J. Bex (2015). "Patterns of functional vision loss in glaucoma determined with archetypal analysis." J R Soc Interface **12**(103).
- Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature **542**(7639): 115-118.
- Ferris, F. L., 3rd, C. P. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, S. R. Sadda and C. Beckman Initiative for Macular Research Classification (2013). "Clinical Classification of Age-related Macular Degeneration." Ophthalmology **120**(4): 844-851.
- Fiorini, S. B., L.; Trucco, E.; Ruggeri, A. (2016). "Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network." Procedia Computer Science **90**: 54-60.
- Flaxman, S. R., R. R. A. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen, J. Leasher, H. Limburg, K. Naidoo, K. Pesudovs, A. Silvester, G. A. Stevens, N. Tahhan, T. Y. Wong, H. R. Taylor and S. Vision Loss Expert Group of the Global Burden of Disease (2017).

"Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis." Lancet Glob Health **5**(12): e1221-e1234.

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

Fleck, B. W., C. Williams, E. Juszczak, K. Cocker, B. J. Stenson, B. A. Darlow, S. Dai, G. A. Gole, G. E. Quinn, D. K. Wallace, A. Ells, S. Carden, L. Butler, D. Clark, J. Elder, C. Wilson, S. Biswas, A. Shafiq, A. King, P. Brocklehurst, A. R. Fielder and B. I. R. I. D. A. Group (2018). "An international comparison of retinopathy of prematurity grading performance within the Benefits of Oxygen Saturation Targeting II trials." Eye (Lond) **32**(1): 74-80.

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

Foster, P. J., R. Buhrmann, H. A. Quigley and G. J. Johnson (2002). "The definition and classification of glaucoma in prevalence surveys." Br J Ophthalmol **86**(2): 238-242.

Fujimoto, J. and E. Swanson (2016). "The Development, Commercialization, and Impact of Optical Coherence Tomography." Invest Ophthalmol Vis Sci **57**(9): OCT1-OCT13.

Garcia, G. P., K. Nitta, M. S. Lavieri, C. Andrews, X. Liu, E. Lobaza, M. P. Van Oyen, K. Sugiyama and J. D. Stein (2018). "Using Kalman Filtering to Forecast Disease Trajectory for Patients with Normal Tension Glaucoma." Am J Ophthalmol.

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

Gelman, S. K., R. Gelman, A. B. Callahan, M. E. Martinez-Perez, D. S. Casper, J. T. Flynn and M. F. Chiang (2010). "Plus disease in retinopathy of prematurity: quantitative analysis of standard published photograph." Archives of Ophthalmology **128**(9): 1217-1220.

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

Goff, D. C., Jr., D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R.

Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, Jr., P. Sorlie, N. J. Stone, P. W. Wilson, H. S. Jordan, L. Nevo, J. Wnek, J. L. Anderson, J. L. Halperin, N. M. Albert, B. Bozkurt, R.

- G. Brindis, L. H. Curtis, D. DeMets, J. S. Hochman, R. J. Kovacs, E. M. Ohman, S. J. Pressler, F. W. Sellke, W. K. Shen and G. F. Tomaselli (2014). "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 **129**(25 Suppl 2): S49-73.
- Gole, G. A., A. L. Ells, X. Katz, G. Holmstrom, A. R. Fielder, A. Capone Jr, J. T. Flynn, W. G. Good, J. M. Holmes and J. McNamara (2005). "The international classification of retinopathy of prematurity revisited." JAMA Ophthalmology **123**(7): 991-999.
- Gomes, L. (2014). "Machine-learning maestro michael jordan on the delusions of big data and other huge engineering efforts." IEEE Spectrum, Oct **20**.
- Graham, I., D. Atar, K. Borch-Johnsen, G. Boysen, G. Burell, R. Cifkova, J. Dallongeville, G. De Backer, S. Ebrahim, B. Gjelsvik, C. Herrmann-Lingen, A. Hoes, S. Humphries, M. Knapton, J. Perk, S. G. Priori, K. Pyorala, Z. Reiner, L. Ruilope, S. Sans-Menendez, W. S. Op Reimer, P. Weissberg, D. Wood, J. Yarnell, J. L. Zamorano, E. Walma, T. Fitzgerald, M. T. Cooney, A. Dudina, A. Vahanian, J. Camm, R. De Caterina, V. Dean, K. Dickstein, C. Funck-Brentano, G. Filippatos, I. Hellemans, S. D. Kristensen, K. McGregor, U. Sechtem, S. Silber, M. Tendera, P. Widimsky, A. Altiner, E. Bonora, P. N. Durrington, R. Fagard, S. Giampaoli, H. Hemingway, J. Hakansson, S. E. Kjeldsen, L. Larsen m, G. Mancina, A. J. Manolis, K. Orth-Gomer, T. Pedersen, M. Rayner, L. Ryden, M. Sammut, N. Schneiderman, A. F. Stalenhoef, L. Tokgozoglul, O. Wiklund and A. Zampelas (2007). "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 **14 Suppl 2**: S1-113.
- Grassmann, F., J. Mengelkamp, C. Brandl, S. Harsch, M. E. Zimmermann, B. Linkohr, A. Peters, I. M. Heid, C. Palm and B. H. F. Weber (2018). "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular Degeneration from Color Fundus Photography." Ophthalmology **125**(9): 1410-1420.



- Group, C. R., D. F. Martin, M. G. Maguire, G. S. Ying, J. E. Grunwald, S. L. Fine and G. J. Jaffe (2011). "Ranibizumab and bevacizumab for neovascular age-related macular degeneration." N Engl J Med **364**(20): 1897-1908.
- Group, S.-R. M. S. (2000). "Supplemental therapeutic oxygen for prethreshold retinopathy of prematurity (STOP-ROP), a randomized, controlled trial. I: primary outcomes." Pediatrics **105**(2): 295-310.
- Gschließer, A., E. Stifter, T. Neumayer, E. Moser, A. Papp, N. Pircher, G. Dorner, S. Egger, N. Vukojevic and I. Oberacher-Velten (2015). "Inter-expert and intra-expert agreement on the diagnosis and treatment of retinopathy of prematurity." American journal of ophthalmology **160**(3): 553-560. e553.
- Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega and D. R. Webster (2016). "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." JAMA **316**(22): 2402-2410.
- Hewing, N. J., D. R. Kaufman, R. P. Chan and M. F. Chiang (2013). "Plus disease in retinopathy of prematurity: qualitative analysis of diagnostic process by experts." JAMA ophthalmology **131**(8): 1026-1032.
- Hira, R. S., K. Kennedy, V. Nambi, H. Jneid, M. Alam, S. S. Basra, P. M. Ho, A. Deswal, C. M. Ballantyne, L. A. Petersen and S. S. Virani (2015). "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 **65**(2): 111-121.
- Hong, S. W., H. Koenigsman, R. Ren, H. Yang, S. K. Gardiner, J. Reynaud, R. M. Kinast, S. L. Mansberger, B. Fortune, S. Demirel and C. F. Burgoyne (2018). "Glaucoma Specialist Optic Disc Margin, Rim Margin, and Rim Width Discordance in Glaucoma and Glaucoma Suspect Eyes." Am J Ophthalmol **192**: 65-76.
- Hubel, D. H. and T. N. Wiesel (1968). "Receptive fields and functional architecture of monkey striate cortex." J Physiol **195**(1): 215-243.
- Hwang, E. J., S. Park, K. N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J. J. Yim, C. M. Park, D. Development and G. Evaluation (2018). "Development and Validation of a Deep Learning-Based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs." Clin Infect Dis.



- Kalpathy-Cramer, J., J. P. Campbell, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J. D. Reynolds, K. Hutcheson, M. J. Shapiro and M. X. Repka (2016). "Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis." Ophthalmology **123**(11): 2345-2351.
- Keane, P. A. and S. R. Sadda (2014). "Retinal imaging in the twenty-first century: state of the art and future directions." Ophthalmology **121**(12): 2489-2500.
- Keck, K. M., J. Kalpathy-Cramer, E. Ataer-Cansizoglu, S. You, D. Erdogmus and M. F. Chiang (2013). "Plus disease diagnosis in retinopathy of prematurity: vascular tortuosity as a function of distance from optic disc." Retina (Philadelphia, Pa.) **33**(8): 1700.
- Kermany, D. S., M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia and K. Zhang (2018). "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning." Cell **172**(5): 1122-1131 e1129.
- Kim, S. J., J. P. Campbell, J. Kalpathy-Cramer, S. Ostmo, K. Jonas, R. P. Chan and M. F. Chiang (2018). "Plus disease in retinopathy of prematurity: should diagnosis be eye-based or quadrant-based?" Journal of American Association for Pediatric Ophthalmology and Strabismus {JAAPOS} **22**(4): e78.
- Klein, R., S. M. Meuer, C. E. Myers, G. H. Buitendijk, E. Rochtchina, F. Choudhury, P. T. de Jong, R. McKean-Cowdin, S. K. Iyengar, X. Gao, K. E. Lee, J. R. Vingerling, P. Mitchell, C. C. Klaver, J. J. Wang and B. E. Klein (2014). "Harmonizing the classification of age-related macular degeneration in the three-continent AMD consortium." Ophthalmic Epidemiol **21**(1): 14-23.
- Koreen, S., R. Gelman, M. E. Martinez-Perez, L. Jiang, A. M. Berrocal, D. J. Hess, J. T. Flynn and M. F. Chiang (2007). "Evaluation of a computer-based system for plus disease diagnosis in retinopathy of prematurity." Ophthalmology **114**(12): e59-e67.
- Krause, J., V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng and D. R. Webster (2018). "Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy." Ophthalmology **125**(8): 1264-1272.

- Krizhevsky, A., H. Sutskever and G. E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks." NIPS.
- Kwon, J., K. R. Sung and J. M. Park (2017). "Myopic glaucomatous eyes with or without optic disc shape alteration: a longitudinal study." Br J Ophthalmol **101**(12): 1618-1622.
- Lakhani, P. and B. Sundaram (2017). "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks." Radiology **284**(2): 574-582.
- LeCun, Y., Y. Bengio and G. Hinton (2015). "Deep learning." Nature **521**(7553): 436-444.
- Lee, C. S., D. M. Baughman and A. Y. Lee (2017). "Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images." Ophthalmol Retina **1**(4): 322-327.
- Lee, C. S., A. J. Tying, N. P. Deruyter, Y. Wu, A. Rokem and A. Y. Lee (2017). "Deep-learning based, automated segmentation of macular edema in optical coherence tomography." Biomed Opt Express **8**(7): 3440-3448.
- Lee, C. S., A. J. Tying, N. P. Deruyter, Y. Wu, A. Rokem and A. Y. Lee (2017). "Deep-learning based, automated segmentation of macular edema in optical coherence tomography." Biomed. Opt. Express **8**(7): 3440-3448.
- Leitgeb, R. A., R. M. Werkmeister, C. Blatter and L. Schmetterer (2014). "Doppler optical coherence tomography." Prog Retin Eye Res **41**: 26-43.
- Li, F., Z. Wang, G. Qu, D. Song, Y. Yuan, Y. Xu, K. Gao, G. Luo, Z. Xiao, D. S. C. Lam, H. Zhong, Y. Qiao and X. Zhang (2018). "Automatic differentiation of Glaucoma visual field from non-glaucoma visual field using deep convolutional neural network." BMC Med Imaging **18**(1): 35.
- Li, Z., Y. He, S. Keel, W. Meng, R. T. Chang and M. He (2018). "Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs." Ophthalmology **125**(8): 1199-1206.
- Li, Z., S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, P. Y. Lee, J. Shaw, D. Ting, T. Y. Wong, H. Taylor, R. Chang and M. He (2018). "An Automated Grading System for Detection of Vision-Threatening Referable Diabetic Retinopathy on the Basis of Color Fundus Photographs." Diabetes Care **41**(12): 2509-2516.

- Ludwig, C. A., T. A. Chen, T. Hernandez-Boussard, A. A. Moshfeghi and D. M. Moshfeghi (2017). "The Epidemiology of Retinopathy of Prematurity in the United States." Ophthalmic Surg Lasers Imaging Retina **48**(7): 553-562.
- Masumoto, H., H. Tabuchi, S. Nakakura, N. Ishitobi, M. Miki and H. Enno (2018). "Deep-learning Classifier With an Ultrawide-field Scanning Laser Ophthalmoscope Detects Glaucoma Visual Field Severity." J Glaucoma **27**(7): 647-652.
- McCarthy, J., M. L. Minsky, N. Rochester and C. E. Shannon (1955). "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." URL: <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf> [Accessed on December 18th, 2018].
- Menti, E., L. Bonaldi, L. Ballerini, A. Ruggeri and E. Trucco (2016). Automatic Generation of Synthetic Retinal Fundus Images: Vascular Network, Cham, Springer International Publishing.
- Mitchell, P., N. Bressler, Q. V. Doan, C. Dolan, A. Ferreira, A. Osborne, E. Rochtchina, M. Danese, S. Colman and T. Y. Wong (2014). "Estimated cases of blindness and visual impairment from neovascular age-related macular degeneration avoided in Australia by ranibizumab treatment." PLoS One **9**(6): e101072.
- Moleta, C., J. P. Campbell, J. Kalpathy-Cramer, R. P. Chan, S. Ostmo, K. Jonas, M. F. Chiang, Imaging and I. i. R. R. Consortium (2017). "Plus disease in retinopathy of prematurity: diagnostic trends in 2016 versus 2007." American journal of ophthalmology **176**: 70-76.
- Muhammad, H., T. J. Fuchs, N. De Cuir, C. G. De Moraes, D. M. Blumberg, J. M. Liebmann, R. Ritch and D. C. Hood (2017). "Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects." J Glaucoma **26**(12): 1086-1094.
- Myung, J. S., R. V. P. Chan, M. J. Espiritu, S. L. Williams, D. B. Granet, T. C. Lee, D. J. Weissgold and M. F. Chiang (2011). "Accuracy of retinopathy of prematurity image-based diagnosis by pediatric ophthalmology fellows: implications for training." Journal of American Association for Pediatric Ophthalmology and Strabismus **15**(6): 573-578.
- Nagiel, A., M. J. Espiritu, R. K. Wong, T. C. Lee, A. K. Lauer, M. F. Chiang and R. P. Chan (2012). "Retinopathy of prematurity residency training." Ophthalmology **119**(12): 2644-2645. e2642.

- National Cholesterol Education Program Expert Panel on Detection, E. and A. Treatment of High Blood Cholesterol in (2002). "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 **106**(25): 3143-3421.
- Peters, D., B. Bengtsson and A. Heijl (2014). "Factors associated with lifetime risk of open-angle glaucoma blindness." Acta Ophthalmol **92**(5): 421-425.
- Poplin, R., A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng and D. R. Webster (2018). "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning." Nature Biomedical Engineering. **2**: 158-164.
- Quelleg, G. and M. D. Abramoff (2014). "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**: 154-157.
- Quinn, G. E. (2016). "Retinopathy of prematurity blindness worldwide: phenotypes in the third epidemic." Eye Brain **8**: 31-36.
- Rao, R., N. J. Jonsson, C. Ventura, R. Gelman, M. A. Lindquist, D. S. Casper and M. F. Chiang (2012). "Plus disease in retinopathy of prematurity: diagnostic impact of field of view." Retina (Philadelphia, Pa.) **32**(6): 1148.
- Raumviboonsuk, P., J. Krause, P. Chotcomwongse, R. Sayres, R. Raman, K. Widner, B. J. L. Campana, S. Phene, K. Hemarat, M. Tadarati, S. Silpa-Archa, J. Limwattanayingyong, C. Rao, O. Kuruvilla, J. Jung, J. H. Tan, S. Orprayoon, C. Kangwanwongpaisan, R. Sukumalpaiboon, C. Luengchaichawang, J. Fuangkaew, P. Kongsap, L. Chualinpha, S. Saree, S. Kawinpanitan, K. Mitvongsa, S. Lawanasakol, C. Thepchatri, L. Wongpichedchai, G. S. Corrado, L. Peng and D. L. Webster (2019). "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program." NPJ Digital Medicine **2**(25).
- Reynolds, J. D., V. Dobson, G. E. Quinn, A. R. Fielder, E. A. Palmer, R. A. Saunders, R. J. Hardy, D. L. Phelps, J. D. Baker and M. T. Trese (2002). "Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies." Archives of ophthalmology **120**(11): 1470-1476.

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

Roth, G. A., C. Johnson, A. Abajobir, F. Abd-Allah, S. F. Abera, G. Abyu, M. Ahmed, B. Aksut, T. Alam, K. Alam, F. Alla, N. Alvis-Guzman, S. Amrock, H. Ansari, J. Arnlov, H. Asayesh, T. M. Atey, L. Avila-Burgos, A. Awasthi, A. Banerjee, A. Barac, T. Barnighausen, L. Barregard, N. Bedi, E. Belay Ketema, D. Bennett, G. Berhe, Z. Bhutta, S. Bitew, J. Carapetis, J. J. Carrero, D. C. Malta, C. A. Castaneda-Orjuela, J. Castillo-Rivas, F. Catala-Lopez, J. Y. Choi, H. Christensen, M. Cirillo, L. Cooper, Jr., M. Criqui, D. Cundiff, A. Damasceno, L. Dandona, R. Dandona, K. Davletov, S. Dharmaratne, P. Dorairaj, M. Dubey, R. Ehrenkranz, M. El Sayed Zaki, E. J. A. Faraon, A. Esteghamati, T. Farid, M. Farvid, V. Feigin, E. L. Ding, G. Fowkes, T. Gebrehiwot, R. Gillum, A. Gold, P. Gona, R. Gupta, T. D. Habtewold, N. Hafezi-Nejad, T. Hailu, G. B. Hailu, G. Hankey, H. Y. Hassen, K. H. Abate, R. Havmoeller, S. I. Hay, M. Horino, P. J. Hotez, K. Jacobsen, S. James, M. Javanbakht, P. Jeemon, D. John, J. Jonas, Y. Kalkonde, C. Karimkhani, A. Kasaeian, Y. Khader, A. Khan, Y. H. Khang, S. Khera, A. T. Khoja, J. Khubchandani, D. Kim, D. Kolte, S. Kosen, K. J. Krohn, G. A. Kumar, G. F. Kwan, D. K. Lal, A. Larsson, S. Linn, A. Lopez, P. A. Lotufo, H. M. A. El Razek, R. Malekzadeh, M. Mazidi, T. Meier, K. G. Meles, G. Mensah, A. Meretoja, H. Mezgebe, T. Miller, E. Mirrakhimov, S. Mohammed, A. E. Moran, K. I. Musa, J. Narula, B. Neal, F. Ngalesoni, G. Nguyen, C. M. Obermeyer, M. Owolabi, G. Patton, J. Pedro, D. Qato, M. Qorbani, K. Rahimi, R. K. Rai, S. Rawaf, A. Ribeiro, S. Safiri, J. A. Salomon, I. Santos, M. Santric Milicevic, B. Sartorius, A. Schutte, S. Sepanlou, M. A. Shaikh, M. J. Shin, M. Shishehbor, H. Shore, D. A. S. Silva, E. Sobngwi, S. Stranges, S. Swaminathan, R. Tabares-Seisdedos, N. Tadele Atnafu, F. Tesfay, J. S. Thakur, A. Thrift, R. Topor-Madry, T. Truelsen, S. Tyrovolas, K. N. Ukwaja, O. Uthman, T. Vasankari, V. Vlassov, S. E. Vollset, T. Wakayo, D. Watkins, R. Weintraub, A. Werdecker, R. Westerman, C. S. Wiysonge, C. Wolfe, A. Workicho, G. Xu, Y. Yano, P. Yip, N. Yonemoto, M. Younis, C. Yu, T. Vos, M. Naghavi and C. Murray (2017). "Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015." J Am Coll Cardiol **70**(1): 1-25.

Samuel, A. L. (1959). "Some Studies in Machine Learning Using the Game of Checkers." IBM Journal of Research and Development.

- Savini, G., M. Carbonelli and P. Barboni (2011). "Spectral-domain optical coherence tomography for the diagnosis and follow-up of glaucoma." Curr Opin Ophthalmol **22**(2): 115-123.
- Schell, G. J., M. S. Lavieri, J. D. Stein and D. C. Musch (2013). "Filtering data from the collaborative initial glaucoma treatment study for improved identification of glaucoma progression." BMC Med Inform Decis Mak **13**: 137.
- Schmidl, D., L. Schmetterer, G. Garhofer and A. Popa-Cherecheanu (2015). "Pharmacotherapy of glaucoma." J Ocul Pharmacol Ther **31**(2): 63-77.
- Schmidt-Erfurth, U., H. Bogunovic, A. Sadeghipour, T. Schlegl, G. Langs, B. S. Gerendas, A. Osborne and S. M. Waldstein (2018). "Machine Learning to Analyze the Prognostic Value of Current Imaging Biomarkers in Neovascular Age-Related Macular Degeneration." Ophthalmol Retina **2**(1): 24-30.
- Schmidt-Erfurth, U., A. Sadeghipour, B. S. Gerendas, S. M. Waldstein and H. Bogunovic (2018). "Artificial intelligence in retina." Prog Retin Eye Res **67**: 1-29.
- Schmidt-Erfurth, U., S. M. Waldstein, S. Klmscha, A. Sadeghipour, X. Hu, B. S. Gerendas, A. Osborne and H. Bogunovic (2018). "Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence." Invest Ophthalmol Vis Sci **59**(8): 3199-3208.
- Seidemann, S. B., B. Claggett, P. E. Bravo, A. Gupta, H. Farhad, B. E. Klein, R. Klein, M. Di Carli and S. D. Solomon (2016). "Retinal Vessel Calibers in Predicting Long-Term Cardiovascular Outcomes: The Atherosclerosis Risk in Communities Study." Circulation **134**(18): 1328-1338.
- Shibata, N., M. Tanito, K. Mitsuhashi, Y. Fujino, M. Matsuura, H. Murata and R. Asaoka (2018). "Development of a deep residual learning algorithm to screen for glaucoma from fundus photography." Sci Rep **8**(1): 14665.
- Simonyan K, V. A., Zisserman A (2017). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." Computer Vision and Pattern Recognition (cs.CV).
- Slidsborg, C., J. L. Forman, A. R. Fielder, S. Crafoord, K. Baggesen, R. Bangsgaard, H. C. Fledelius, G. Greisen and M. La Cour (2012). "Experts do not agree when to treat retinopathy of prematurity based on plus disease." British Journal of Ophthalmology **96**(4): 549-553.
- Stone, N. J., J. G. Robinson, A. H. Lichtenstein, C. N. Bairey Merz, C. B. Blum, R. H. Eckel, A. C. Goldberg, D. Gordon, D. Levy, D. M. Lloyd-Jones, P. McBride, J. S.



- Schwartz, S. T. Shero, S. C. Smith, Jr., K. Watson, P. W. Wilson and G. American College of Cardiology/American Heart Association Task Force on Practice (2014). "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 **63**(25 Pt B): 2889-2934.
- Tasman, W. (1988). "Multicenter trial of cryotherapy for retinopathy of prematurity." Archives of Ophthalmology **106**(4): 463-464.
- Tham, Y. C., X. Li, T. Y. Wong, H. A. Quigley, T. Aung and C. Y. Cheng (2014). "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis." Ophthalmology **121**(11): 2081-2090.
- Ting, D. S. and T. Y. Wong (2018). "Eyeing Cardiovascular Risk Factors." Nature Biomedical Engineering. **2**: 140-141.
- Ting, D. S. W., C. Y. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C. Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee and T. Y. Wong (2017). "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes." JAMA **318**(22): 2211-2223.
- Ting, D. S. W., C. Y. Cheung, N. D. Quang, C. Sabanayagam, G. Lim, Z. Lim, G. S. W. Tan, Y. Q. Soh, L. Schmetterer, Y. X. Wang, J. B. Jonas, R. Varma, M. L. Lee, W. Hsu, E. L. Lamoureux, C. Y. Cheng and T. Y. Wong (2019). "Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: a multi-ethnic study." NPJ Digital Medicine **2**(24).
- Ting, D. S. W., L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane and T. Y. Wong (2018). "Artificial intelligence and deep learning in ophthalmology." Br J Ophthalmol.
- Treder, M., J. L. Lauermann and N. Eter (2018). "Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning." Graefes Arch Clin Exp Ophthalmol **256**(2): 259-265.
- Varadarajan, A. V., R. Poplin, K. Blumer, C. Angermueller, J. Ledsam, R. Chopra, P. A. Keane, G. S. Corrado, L. Peng and D. R. Webster (2018). "Deep Learning for



Predicting Refractive Error From Retinal Fundus Images." Invest Ophthalmol Vis Sci **59**(7): 2861-2868.

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

Vayena, E., A. Blasimme and I. G. Cohen (2018). "Machine learning in medicine: Addressing ethical challenges." PLoS Med **15**(11): e1002689.

Wallace, D. K., S. F. Freedman, M. E. Hartnett and G. E. Quinn (2011). "Predictive value of pre-plus disease in retinopathy of prematurity." Archives of ophthalmology **129**(5): 591-596.

Wallace, D. K., J. A. Kylstra and D. A. Chesnutt (2000). "Prognostic significance of vascular dilation and tortuosity insufficient for plus disease in retinopathy of prematurity." Journal of American Association for Pediatric Ophthalmology and Strabismus **4**(4): 224-229.

Wallace, D. K., G. E. Quinn, S. F. Freedman and M. F. Chiang (2008). "Agreement among pediatric ophthalmologists in diagnosing plus and pre-plus disease in retinopathy of prematurity." Journal of American Association for Pediatric Ophthalmology and Strabismus **12**(4): 352-356.

Wang, J. J., G. Liew, T. Y. Wong, W. Smith, R. Klein, S. R. Leeder and P. Mitchell (2006). "Retinal vascular calibre and the risk of coronary heart disease-related death." Heart **92**(11): 1583-1587.

Wang, M., L. R. Pasquale, L. Q. Shen, M. V. Boland, S. R. Wellik, C. G. De Moraes, J. S. Myers, H. Wang, N. Baniyadi, D. Li, R. N. E. Silva, P. J. Bex and T. Elze (2018). "Reversal of Glaucoma Hemifield Test Results and Visual Field Features in Glaucoma." Ophthalmology **125**(3): 352-360.

Wheatley, C. M., J. L. Dickinson, D. A. Mackey, J. E. Craig and M. M. Sale (2002). "Retinopathy of prematurity: recent advances in our understanding." Br J Ophthalmol **86**(6): 696-700.

Wicherts, J. M., C. L. Veldkamp, H. E. Augusteijn, M. Bakker, R. C. van Aert and M. A. van Assen (2016). "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." Front Psychol **7**: 1832.

Wilson, C. M., K. Wong, J. Ng, K. D. Cocker, A. L. Ells and A. R. Fielder (2012). "Digital image analysis in retinopathy of prematurity: a comparison of vessel

- selection methods." Journal of American Association for Pediatric Ophthalmology and Strabismus **16**(3): 223-228.
- Wilson, P. W., R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz and W. B. Kannel (1998). "Prediction of coronary heart disease using risk factor categories." Circulation **97**(18): 1837-1847.
- Windsor, M. A., S. J. J. Sun, K. D. Frick, E. A. Swanson, P. J. Rosenfeld and D. Huang (2018). "Estimating Public and Patient Savings From Basic Research-A Study of Optical Coherence Tomography in Managing Antiangiogenic Therapy." Am J Ophthalmol **185**: 115-122.
- Witt, N., T. Y. Wong, A. D. Hughes, N. Chaturvedi, B. E. Klein, R. Evans, M. McNamara, S. A. Thom and R. Klein (2006). "Abnormalities of retinal microvascular structure and risk of mortality from ischemic heart disease and stroke." Hypertension **47**(5): 975-981.
- Wittenberg, L. A., N. J. Jonsson, R. P. Chan and M. F. Chiang (2012). "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity." Journal of pediatric ophthalmology and strabismus **49**(1): 11-19.
- Wong, R. K., C. V. Ventura, M. J. Espiritu, Y. Yonekawa, L. Henchoz, M. F. Chiang, T. C. Lee and R. V. P. Chan (2012). "Training fellows for retinopathy of prematurity care: a Web-based survey." Journal of American Association for Pediatric Ophthalmology and Strabismus **16**(2): 177-181.
- Wong, T. Y. and N. M. Bressler (2016). "Artificial Intelligence With Deep Learning Technology Looks Into Diabetic Retinopathy Screening." JAMA **316**(22): 2366-2367.
- Wong, T. Y., A. Kamineni, R. Klein, A. R. Sharrett, B. E. Klein, D. S. Siscovick, M. Cushman and B. B. Duncan (2006). "Quantitative retinal venular caliber and risk of cardiovascular disease in older persons: the cardiovascular health study." Arch Intern Med **166**(21): 2388-2394.
- Wong, T. Y., S. C. Loon and S. M. Saw (2006). "The epidemiology of age related eye diseases in Asia." Br J Ophthalmol **90**(4): 506-511.
- Wong, W. L., X. Su, X. Li, C. M. Cheung, R. Klein, C. Y. Cheng and T. Y. Wong (2014). "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis." Lancet Glob Health **2**(2): e106-116.

- Worrall, D. E., C. M. Wilson and G. J. Brostow (2016). Automated retinopathy of prematurity case detection with convolutional neural networks. Deep Learning and Data Labeling for Medical Applications, Springer: 68-76.
- Xu, Y., K. Yan, J. Kim, X. Wang, C. Li, L. Su, S. Yu, X. Xu and D. D. Feng (2017). "Dual-stage deep learning framework for pigment epithelium detachment segmentation in polypoidal choroidal vasculopathy." Biomed Opt Express **8**(9): 4061-4076.
- Yeboah, J., R. L. McClelland, T. S. Polonsky, G. L. Burke, C. T. Sibley, D. O'Leary, J. J. Carr, D. C. Goff, P. Greenland and D. M. Herrington (2012). "Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals." JAMA **308**(8): 788-795.
- Yip, W., P. G. Ong, B. W. Teo, C. Y. Cheung, E. S. Tai, C. Y. Cheng, E. Lamoureux, T. Y. Wong and C. Sabanayagam (2017). "Retinal Vascular Imaging Markers and Incident Chronic Kidney Disease: A Prospective Cohort Study." Sci Rep **7**(1): 9374.
- Yousefi, S., M. H. Goldbaum, M. Balasubramanian, F. A. Medeiros, L. M. Zangwill, J. M. Liebmann, C. A. Girkin, R. N. Weinreb and C. Bowd (2014). "Learning from data: recognizing glaucomatous defect patterns and detecting progression from visual field measurements." IEEE Trans Biomed Eng **61**(7): 2112-2124.
- Zayit-Soudry, S., I. Moroz and A. Loewenstein (2007). "Retinal pigment epithelial detachment." Surv Ophthalmol **52**(3): 227-243.

## 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:** 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 (Abramoff, Lou et al. 2016)	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 specialists 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 (Gulshan, Peng et al. 2016)	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	97.50%	93.40%
				Messidor-2	Mydriatic	Image-level		1748	0.17%	(0.974)* 0.94	(96.7%)* 96.10%	(84%)* 93.90%
Gargeya and Leng (Gargeya and Leng 2017)	2017	75,137 images from Kaggle competition graded by "a panel of retinal specialists" (with no additional 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 (Ting, Cheung et al. 2017)	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		Guangdong	Non-mydratic	Image-level	retinal specialist generated final grade 2 graders; arbitration by 1 retinal specialist	15,798	1.40%	0.949*	98.70%*	81.60%*
				SIMES	Mydratic	Image-level	1 grader; 1 retinal specialist	3052	1.80%	0.889*	97.10%*	82%*
				SINDI	Mydratic	Image-level	1 grader; 1 retinal specialist	4512	2.10%	0.917*	99.30%*	73.30%*
				SCES	Mydratic	Image-level	1 grader; 1 retinal specialist	1936	1.00%	0.919*	100%*	76.30%*
				BES	Mydratic	Image-level	2 ophthalmologists	1052	0.40%	0.929*	94.40%*	88.50%*
				AFEDS	Mydratic	Image-level	2 retinal specialists	1968	4.20%	0.98*	98.80%*	86.50%*
				RVEEH	Mydratic	Image-level	2 graders	2302	10.90%	0.983*	98.90%*	92.20%*
				Mexican	Mydratic	Image-level	2 retinal specialists	1172	0.50%	0.95*	91.80%*	84.80%*
				CUHK	Mydratic	Image-level	2 retinal specialists	1254	0.00%	0.948*	99.30%*	83.10%*
				HKU	Mydratic	Image-level	2 optometrists	7706	0.00%	0.964*	100%*	81.30%*
Krause et al(Krause, Gulshan et al. 2018)	2018	1.67M images with clinical grades for train set 3,737 fully adjudicated images for tune set	Inception-V3	EyePACS-2*	Mostly Non-Mydratic	Image-level	Adjudication by 3 retinal specialists until full consensus for all cases using a single 45 degree FOV image	--	0%	0.986	97.1%	92.3%
Abramoff et al(Abramoff,	2018	10,000 to 1,250,000	Customized CNN	FDA Pivotal Trial	23.6% Mydratic	Patient-level	Reading center grading of	892	8.20%	-	87.2%	90.7%

Lavin et al. 2018)		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(Li, Keel et al. 2018)	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 = $\geq$ 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 3:** A summary of artificial intelligence system using deep learning for detection of glaucoma suspect and glaucoma

Author	Year	Disease definition	Development Dataset	CNN	Clinical Validation	Mydriatic or non-myd	Granularity	Ground Truth	Imaging Modality	Number of images	AUC	Sensitivity	Specificity
Li et al.(Li, He et al. 2018)	2018	CDR $\geq$ 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.(Ting, Cheung et al. 2017)	2017	CDR $\geq$ 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.(Shibata, Tanito et al. 2018)	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. <sup>75</sup>	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. <sup>76</sup>	2018	Glaucoma	1,117 images (Tsukazaki Hospital)	Customized CNN	282 images (Tsukazaki Hospital)	Non-mydratic	Image-level	2 glaucoma specialists	Optos wide-field fundus photos	1,399	0.872*	81.3%*	80.2%*
Li et al. <sup>80</sup>	2018	Glaucoma	3,712 images (3 ophthalmic centers in China)	VGG-15	300 images	--	Image-level	9 ophthalmologists (3 glaucoma experts, 3 attending ophthalmologists, 3 resident ophthalmologists)	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 4:** A summary of artificial intelligence system using deep learning for detection of age-related macular degeneration (AMD)

Author	Year	Disease	Development Dataset	CNN	Clinical Validation	Mydriatic or non-mydriatic	Granularity	Ground Truth	Number of retinal images	AUC	Sensitivity	Specificity	Remark
Burlina et al(Burlina, Joshi et al. 2017)	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(Burlina, Joshi et al. 2018)	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(Ting, Cheung et al. 2017)	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(Grassmann, Mengelkamp et al. 2018)	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)	

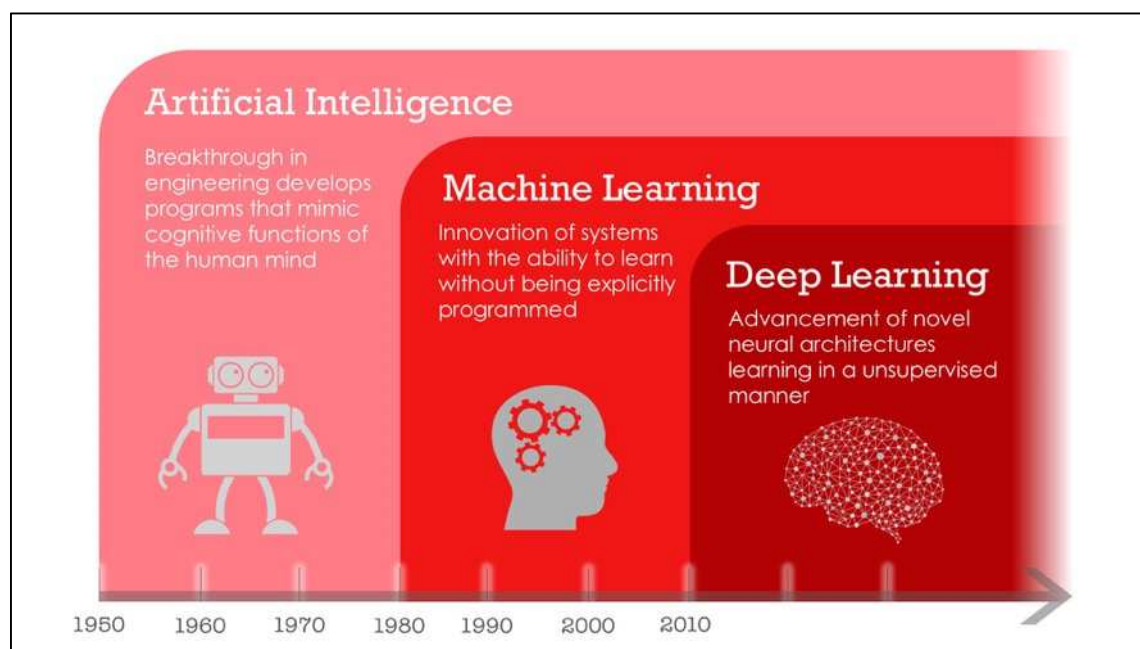
**Table 5:** A summary of artificial intelligence system using deep learning for optical coherence tomography for age-related macular degenerations (AMD)

DL systems	Year	Disease	OCT machines	Development Dataset	CNN	Test Images	AUC	Accuracy	Sensitivity	Specificity
<b>Disease Detection</b> Lee et al(Lee, Baughman et al. 2017)	2017	Exudative AMD	Spectralis	80,839 images	VGG-16	20613 images	0.928	87.60%	84.60%	91.50%
Treder et al(Treder, Lauermann et al. 2018)	2018	Exudative AMD	Spectralis	1,012 images (University of Muenster Medical Center)	Inception-V3	100 images	NR	100%	92%	96%
Kermany et al(Kermany, Goldbaum et al. 2018)	2018	CNV DME Drusen 1. Multi-class comparison 2. Limited model 3. Binary model CNV vs normal DME vs normal Drusen vs normal	Spectralis	108,312 images	Inception-V3	1,000 images	0.999	96.60%	97.80%	97.40%
							0.988	93.40%	96.60%	94.00%
							1	100%	100%	100%
							0.999	98.20%	96.80%	99.60%
							0.999	99.00%	98.00%	99.20%
De Fauw et al(De Fauw, Ledsam et al. 2018)	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					
<b>Disease Prediction</b> Ursula Schmidh(Schmidt-Erfurth, Waldstein et al. 2018)	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.



**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

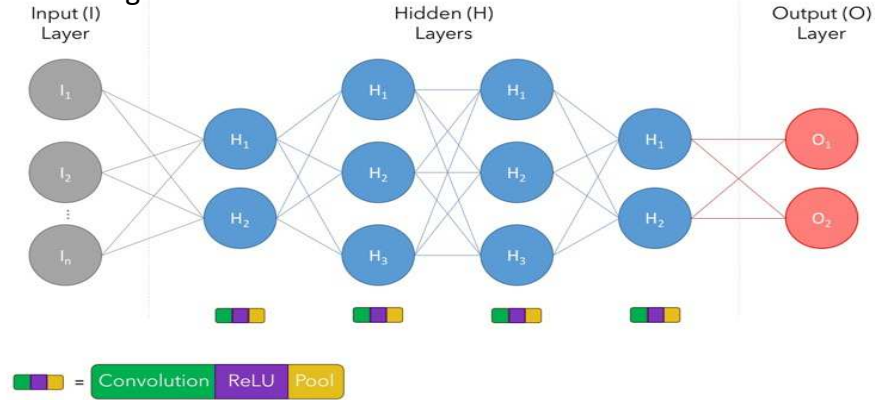
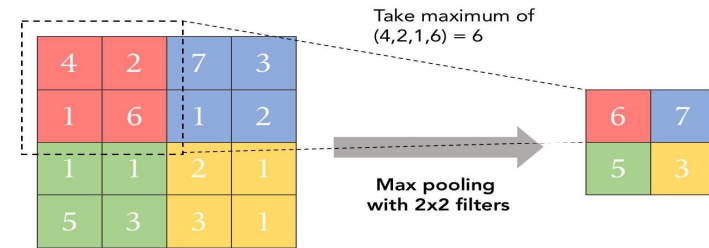
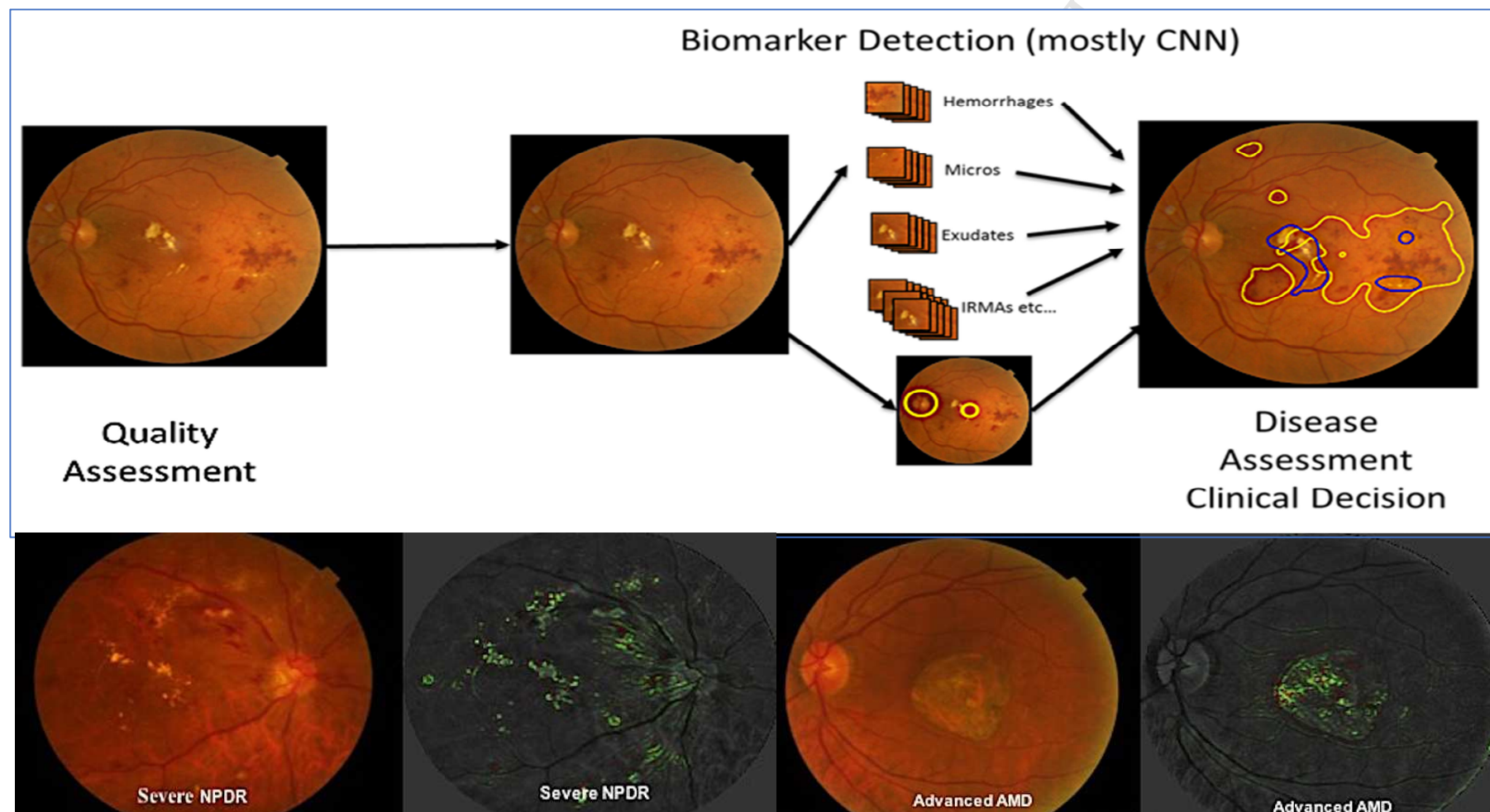


Figure 2B: Max pooling





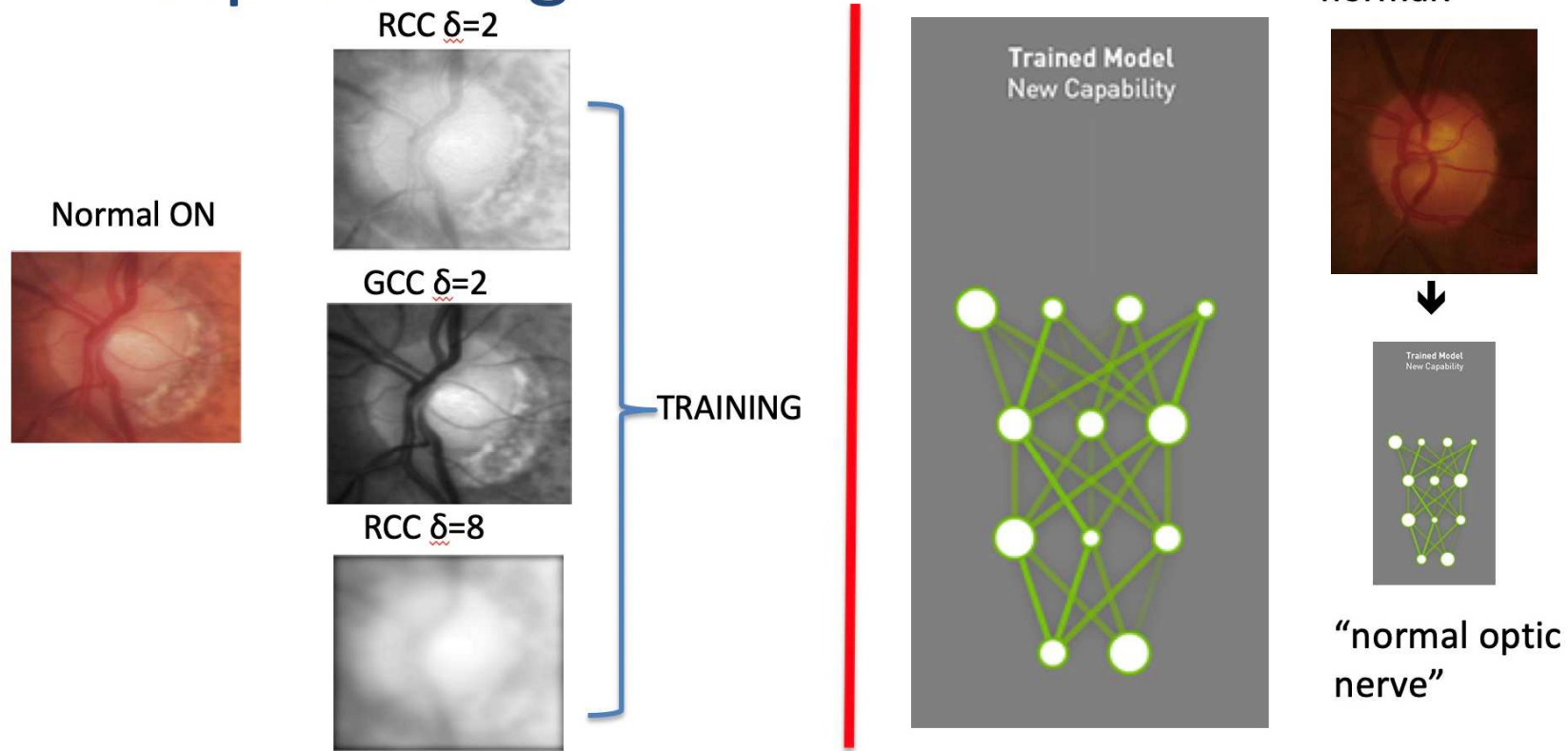
**Figure 3:** The workflow of a deep learning system in detecting referable diabetic retinopathy and age-related macular degeneration, further demonstrated by the heat map



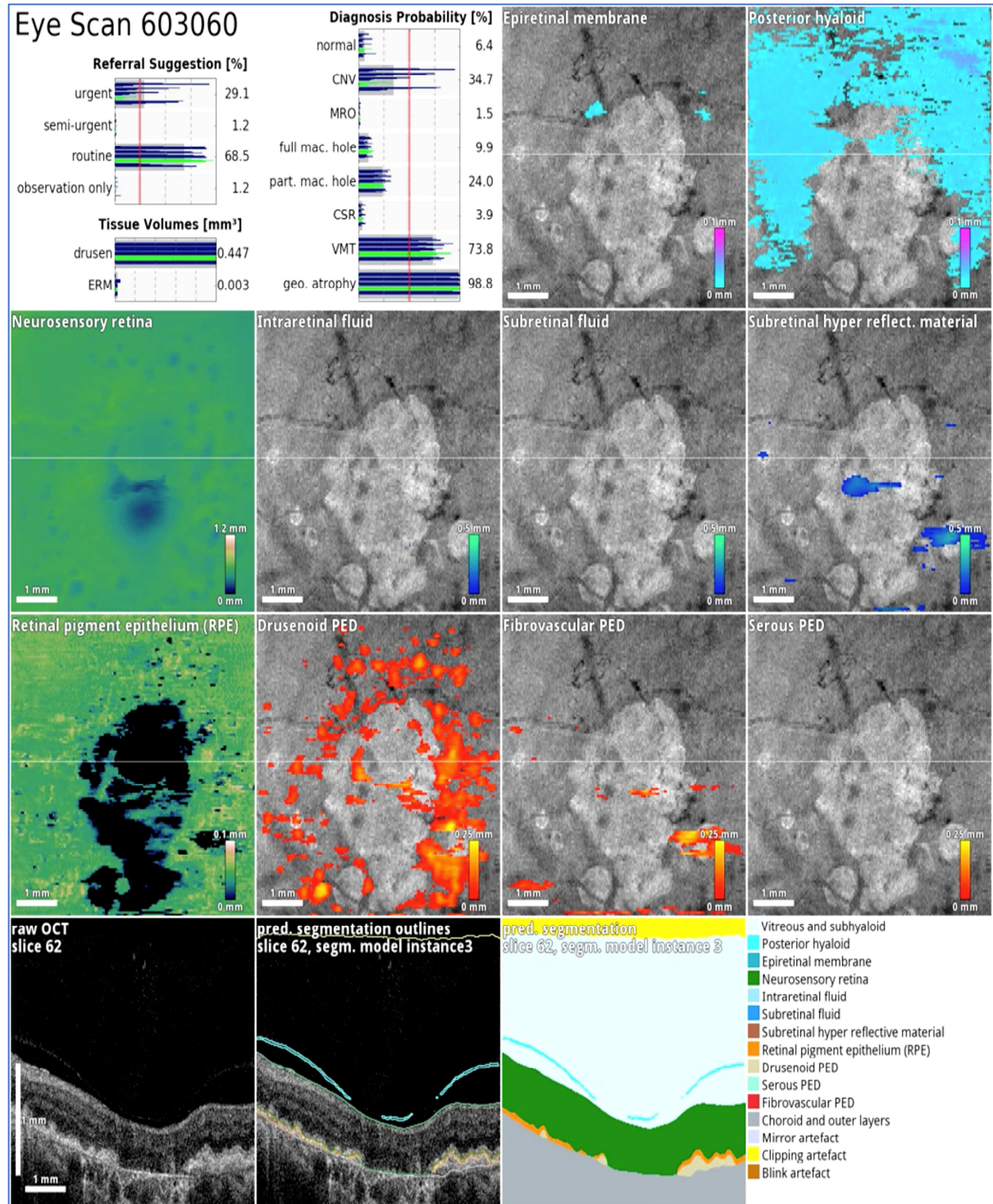
**Figure 4:** Deep learning system for detection of glaucomatous optic disc using optic disc imaging

RCC= Red channel convolution; GCC = Green channel convolution

## Deep learning

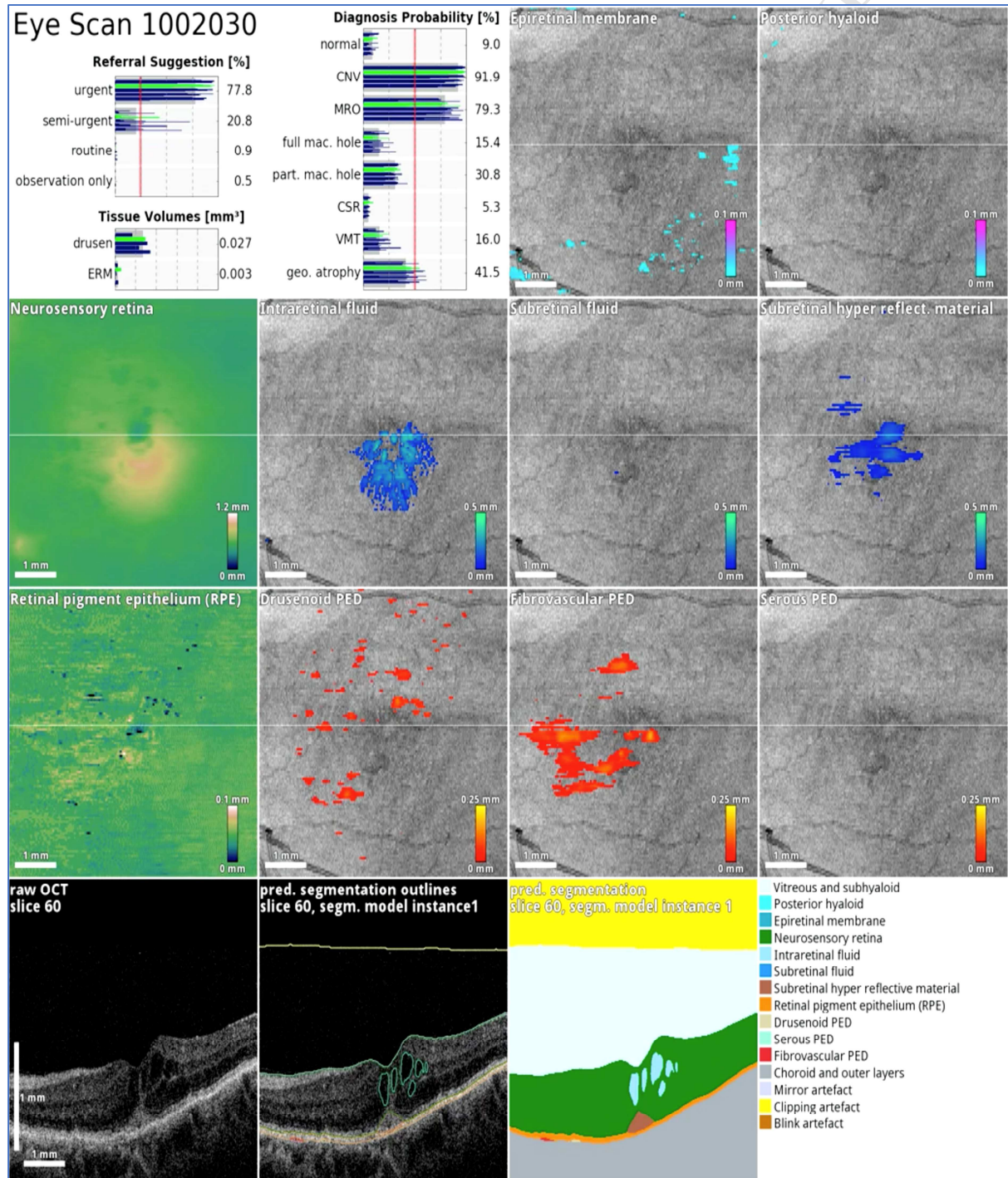


**Figure 5:** The application of deep learning to the segmentation of retinal optical coherence tomography (OCT) images – the prototype OCT viewer for the Moorfields-DeepMind deep learning system. In this case, the system correctly segments loss of the retinal pigment epithelium (RPE) highlighting an 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.





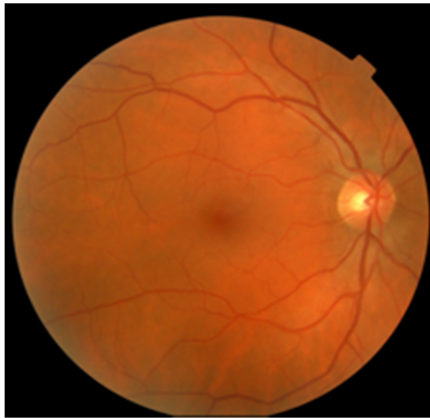
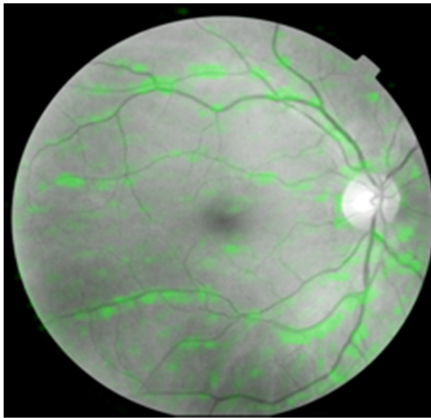
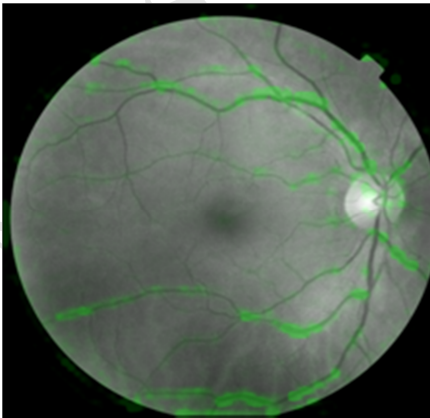
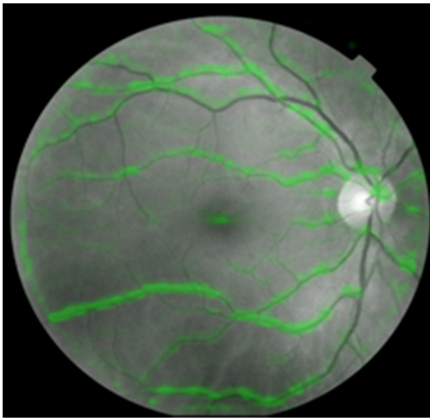
**Figure 6:** The application of deep learning to the segmentation of retinal optical coherence tomography (OCT) images – the prototype OCT viewer for the Moorfields-DeepMind deep learning system. In this challenging case of retinal angiomatous proliferation (RAP), the system correctly segments an area 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.



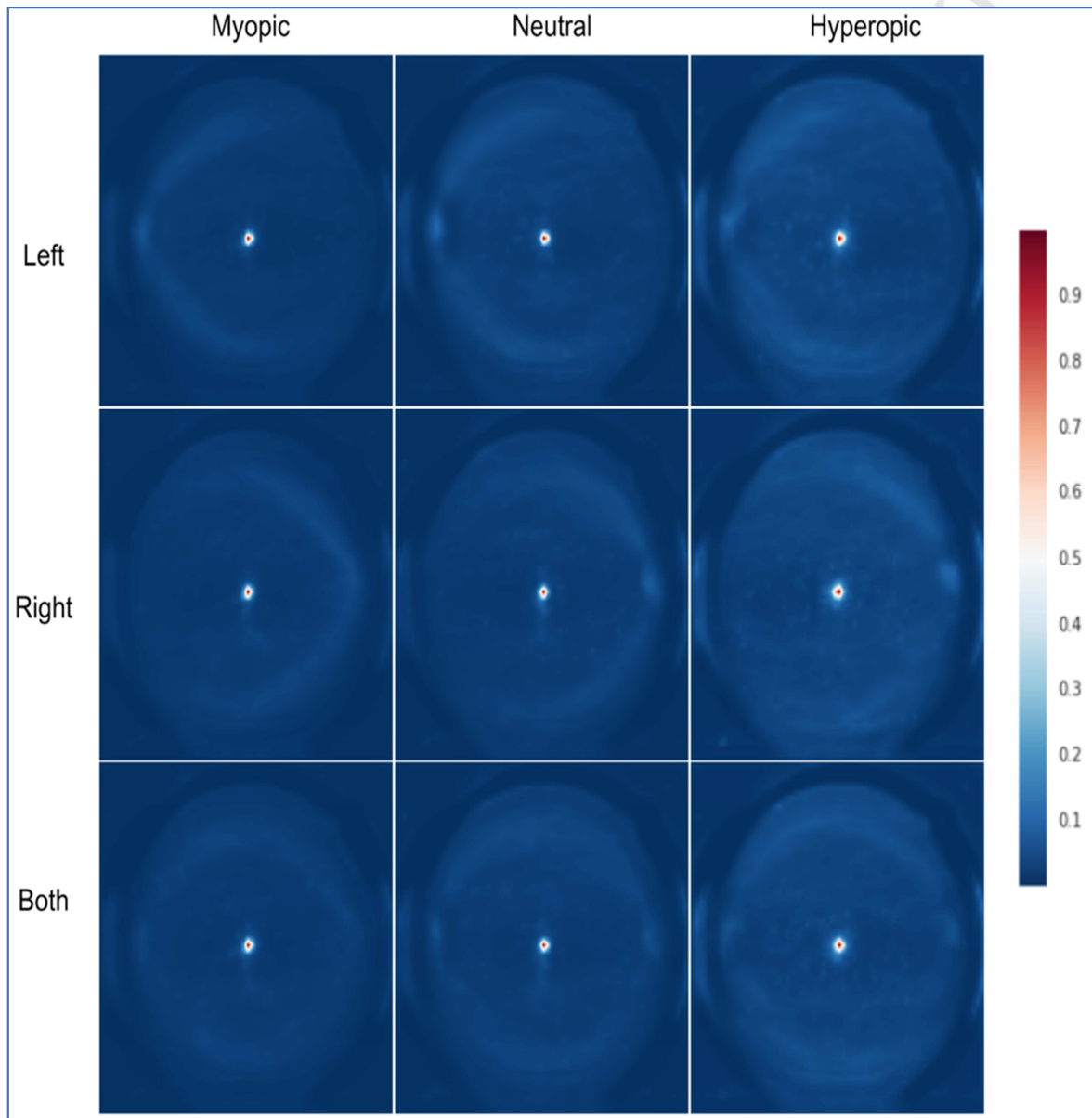
**Figure 7:** Continuous spectrum of retinal vascular findings in retinopathy of prematurity (ROP). (A) shows normal posterior retinal vessels. (B) shows pre-plus disease with mild retinal vascular dilation and tortuosity. (C) shows plus disease with significant retinal vascular dilation and tortuosity.



**Figure 8: 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 9.** 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 be one.





ACCEPTED MANUSCRIPT

**Authors Statement**

DT, LP, AV, PK, PB, MC, LS, LP, NB, DW, MA and TW have all contributed to the initial drafting, critical review and final approval of the manuscript.