Artificial Intelligence Algorithms to Diagnose Glaucoma and Detect Glaucoma Progression: Translation to Clinical Practice

Anna S. Mursch-Edlmayr1, Wai Siene Ng2, Alberto Diniz-Filho3, David C. Sousa4, Louis Arnould5, Matthew B. Schlenker6, Karla Duenas-Angeles7, Pearse A. Keane8, Jonathan G. Crowston9,10, and Hari Jayaram8

1 Department of Ophthalmology Johannes Kepler University, Linz, Austria
2 Cardiff Eye Unit, University Hospital of Wales, Cardiff, UK
3 Department of Ophthalmology and Otorhinolaryngology, Federal University of Minas Gerais, Belo Horizonte, Brazil
4 Department of Ophthalmology, Hospital de Santa Maria, Lisbon, Portugal
5 Department of Ophthalmology, University Hospital, Dijon, France
6 Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Canada
7 Department of Ophthalmology, Universidad Nacional Autónoma de Mexico, Mexico City, Mexico
8 NIHR Biomedical Research Centre for Ophthalmology, UCL Institute of Ophthalmology & Moorfields Eye Hospital, London, UK
9 Centre for Vision Research, Duke-NUS Medical School, Singapore
10 Singapore Eye Research Institute, Singapore National Eye Centre, Singapore

Correspondence: Hari Jayaram, NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK. e-mail: h.jayaram@ucl.ac.uk

Received: September 30, 2019
Accepted: September 18, 2020
Published: October 15, 2020

Keywords: glaucoma; machine learning; artificial intelligence

Purpose: This concise review aims to explore the potential for the clinical implementation of artificial intelligence (AI) strategies for detecting glaucoma and monitoring glaucoma progression.

Methods: Nonsystematic literature review using the search combinations “Artificial Intelligence,” “Deep Learning,” “Machine Learning,” “Neural Networks,” “Bayesian Networks,” “Glaucoma Diagnosis,” and “Glaucoma Progression.” Information on sensitivity and specificity regarding glaucoma diagnosis and progression analysis as well as methodological details were extracted.

Results: Numerous AI strategies provide promising levels of specificity and sensitivity for structural (e.g. optical coherence tomography [OCT] imaging, fundus photography) and functional (visual field [VF] testing) test modalities used for the detection of glaucoma. Area under receiver operating curve (AROC) values of > 0.90 were achieved with every modality. Combining structural and functional inputs has been shown to even more improve the diagnostic ability. Regarding glaucoma progression, AI strategies can detect progression earlier than conventional methods or potentially from one single VF test.

Conclusions: AI algorithms applied to fundus photographs for screening purposes may provide good results using a simple and widely accessible test. However, for patients who are likely to have glaucoma more sophisticated methods should be used including data from OCT and perimetry. Outputs may serve as an adjunct to assist clinical decision making, whereas also enhancing the efficiency, productivity, and quality of the delivery of glaucoma care. Patients with diagnosed glaucoma may benefit from future algorithms to evaluate their risk of progression. Challenges are yet to be overcome, including the external validity of AI strategies, a move from a “black box” toward “explainable AI,” and likely regulatory hurdles. However, it is clear that AI can enhance the role of specialist clinicians and will inevitably shape the future of the delivery of glaucoma care to the next generation.
Translational Relevance: The promising levels of diagnostic accuracy reported by AI strategies across the modalities used in clinical practice for glaucoma detection can pave the way for the development of reliable models appropriate for their translation into clinical practice. Future incorporation of AI into healthcare models may help address the current limitations of access and timely management of patients with glaucoma across the world.

Introduction

Glaucoma is a leading cause of irreversible worldwide blindness. It is estimated to be the cause of visual impairment for almost six million people and blindness for three million people across the world. It is responsible for approximately 10% of those registered as blind within the United States, and for almost one third of sight loss certifications in England. However, the disease is asymptomatic unless at an advanced stage, and therefore an unacceptable number of affected patients continue to remain undiagnosed. Early diagnosis is a crucial factor that has significant impact upon prognosis. As the economic and personal burden associated with glaucoma escalates with the extent of disease progression, the ability to provide an early diagnosis and initiation of appropriate treatment becomes of crucial importance.

The prevalence of glaucoma is projected to increase by almost 50% over the next 20 years as the global population lives longer. The burden of glaucoma care will therefore continue to grow, without a corresponding growth in the number of ophthalmologists or available resources. Although the number of ophthalmologists is increasing, the population aged over 60 years of age is growing at almost twice the rate. As a result, the demand for the provision of glaucoma care will likely exceed capacity and ophthalmologists may face the insurmountable task of attempting to prioritize care for those patients at highest risk of visual loss, or developing novel models of delivering care. Delays to timely glaucoma care have already resulted in significant harm to patients in the United Kingdom. Thus, current models of glaucoma care are unsustainable and there is a need to look toward innovation in order to address the mismatch between capacity and demand.

Artificial intelligence (AI) strategies may provide a potential solution to the growing demand and have demonstrated the potential to redefine how clinicians can deliver health care to the next generation. Significant developments within the field of retinal disease have placed ophthalmology at the forefront of this area of innovation. Implementation of AI strategies offers an opportunity to address the global challenge to meet the increasing need for glaucoma care and significant research has been performed to explore this field. The increasing availability in primary ophthalmic care settings of advanced imaging and perimetry technologies, digital data acquisition, and the development of large clinically phenotyped datasets from routine clinical glaucoma care will continue to facilitate translational research in this area. AI strategies therefore may have the potential to develop novel methodologies to develop effective glaucoma population-based screening to identify undiagnosed glaucoma and also detect clinically relevant glaucoma progression in existing patients.

This review summarizes contemporary developments of AI strategies using fundus photography, optical coherence tomography (OCT) imaging, and perimetry in glaucoma diagnosis and the detection of glaucoma progression, and will contextualize its potential in helping shape the future of glaucoma service delivery.

Artificial Intelligence

The concept of AI is widely considered to have emerged at the Dartmouth Summer Research Project in 1956. It is a branch of computer science that aims to mimic intelligent human behavior. The term is sometimes used interchangeably with machine learning (ML) and deep learning (DL), however, in reality, AI is an umbrella term that includes ML, which itself encompasses DL.

ML is an extension of statistical modeling, whereas in artificial neural networks (ANNs), data analysis is through interconnected nodes with modifiable weights. With ML and ANN, the machine is able to identify the best parameters for a given algorithm to perform a particular task. For example, the task may involve the separation of “glaucoma” from “not glaucoma” from a cohort of optic disc photographs. It is able to detect relationships between multiple input parameters and a definition or diagnosis, although not necessarily providing insight into how these classifications are derived.
In supervised learning, a “training dataset” is required, which, for example, can be a large number of optic disc photographs. Experts need to go through the dataset and label each one with a correct diagnosis known as the “ground truth.” This information is given to the machine, which then uses a learning structure (e.g., random forest, support vector machine, and Gaussian mixture model) to identify the correct diagnosis. It adjusts itself by retesting multiple times until the desired output is achieved. Learning can also be unsupervised or semisupervised and is often relevant if the data has no labels. This approach has the ability to not only model the distribution of the data but also to classify data into groups, including groups that were not initially intended. DL is a modern extension of the classical neural network technique using deep neural networks (DNNs). A DNN is an ANN with multiple intermediate “hidden layers” where each level can transform its input signal into a gradually more abstract feature representation. This is achieved by successively combining outputs from the preceding layer, therefore utilizing fewer artificial neurons than a comparable shallow ANN. The advantage of DL is that more complex inputs, such as an entire image, can be used, however, this requires a much larger training dataset.

Glaucoma Diagnosis

The use of OCT imaging, visual field (VF) assessment using standard automated perimetry (SAP), and clinical examination of the optic disc underpin the diagnosis of glaucomatous optic nerve injury in a clinical setting. In order to accurately diagnose glaucoma, we require tests with both high sensitivity and specificity. Fundus photographs may be a suitable candidate for population-based glaucoma screening for diagnosing glaucoma as it is the simplest and most widely established modality of optic disc assessment. It represents a simple, relatively inexpensive approach and has shown promise for case detection among defined populations. However, the sheer workload generated through the need to manually grade images, the associated inter and intra-observer discrepancies, and confounding factors, such as extremes of refractive error, may challenge diagnostic accuracy. Therefore, automated systems for image grading and AI-based algorithms for improving the diagnostic efficiency of automated glaucoma diagnosis from large image sets is an attractive solution.

In 1999, Sinthanayothin et al. first described a process for the automated detection of the optic disc, fovea, and blood vessels from color fundus images. Since then, successful automated segmentation has been reproduced by several groups and has been considered a prerequisite for algorithm-based diagnosis of glaucoma from fundus photographs.

Subsequently, fundus photographs have been widely used as an input dataset for evaluating glaucoma diagnosis using AI strategies, and are summarized in Table 1. Segmentation and structured learning appear to be the most robust methods through which the analysis of fundus photographs can be utilized to detect glaucoma, with reported accuracy of over 95% in making a positive diagnosis.

In addition, artificial neural networks based on features, such as the cup-to-disc (C/D) ratio, achieved an area under the receiver operating curve (AROC) of up to 0.90 for discriminating healthy from glaucomatous eyes. Using DL algorithms, several groups have reported AROC values for glaucoma detection between 0.84 and 0.99. More recently, a remarkable sensitivity and specificity of 98% for glaucoma diagnosis was achieved through training a neural network with 1426 fundus images. A comprehensive DL algorithm to quantify glaucomatous optic nerve injury from fundus photographs has also been described. This involves features from spectral-domain OCT images being used to train a DL algorithm in order to predict neuroretinal damage from optic disc photographs and shows great promise. Such DL algorithms have even been shown to perform better than human grading of fundus photographs in discriminating between eyes with normal and abnormal VF tests.

Several authors have evaluated AI-based fundus photograph analysis for its utility for detecting glaucoma (see Table 1). In 2013, Cheng et al. reported an AROC of 0.82 in a population-based dataset in Singapore.

With recent advances in OCT imaging, spectral-domain OCT has evolved to improve the resolution, repeatability, and speed of image acquisition and will further increase with advent of swept-source technology. The retinal nerve fiber layer (RNFL) thickness remains the most common parameter utilized for glaucoma diagnosis and is a major focus in ML approaches using OCT imaging data. Starting in 2005, several studies have reported promising results with various ML algorithms analyzing OCT imaging data from peripapillary RNFL thickness maps and the macular ganglion cell complex for discriminating between glaucomatous and normal eyes, with AROC values ranging from 0.69 to 0.99. A recent report proposed a DL network able to classify eyes as normal or glaucomatous based upon unsegmented OCT volumes of the optic nerve head. This achieved...
Table 1. Summary of Studies Using AI to Detect Optic Nerve Head Abnormalities and/or Glaucoma From Fundus Photographs

<table>
<thead>
<tr>
<th>Study</th>
<th>Aim of Study</th>
<th>No. of Eyes/Images</th>
<th>ML Classifiers</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheng et al. (2013)</td>
<td>Glaucoma detection</td>
<td>2,326 images from 2,326 subjects</td>
<td>SVM</td>
<td>AROC of 0.8</td>
</tr>
<tr>
<td>Raja et al. (2015)</td>
<td></td>
<td>1,586 images, 74 glaucomatous eyes, 84 normal eyes</td>
<td>SVM</td>
<td>Maximum accuracy 98.2%</td>
</tr>
<tr>
<td>Ting et al. (2017)</td>
<td></td>
<td>494,661 retinal images; possible glaucoma: 125,189 images</td>
<td>Deep learning</td>
<td>AROC 0.942, sensitivity 96.4%, specificity 87.2%</td>
</tr>
<tr>
<td>Carmona et al. (2008)</td>
<td>Automated location and segmentation of the optic nerve head</td>
<td>10 eyes; 25 with glaucoma, 85 with ocular hypertension</td>
<td>Genetic algorithms</td>
<td>Generalization capability: 96%</td>
</tr>
<tr>
<td>Mookiah et al. (2013)</td>
<td>Glaucoma detection</td>
<td>100 images; 30 normal eyes, 39 glaucomatous eyes, 31 eyes with diabetic retinopathy</td>
<td>SVM</td>
<td>Maximum accuracy 98.2%</td>
</tr>
<tr>
<td>Raja et al. (2015)</td>
<td></td>
<td>40 images, 158 images, 74 glaucomatous eyes, 84 normal eyes</td>
<td>SVM</td>
<td>Maximum accuracy 98.2%</td>
</tr>
<tr>
<td>Ting et al. (2017)</td>
<td></td>
<td>494,661 retinal images; possible glaucoma: 125,189 images</td>
<td>Deep learning</td>
<td>AROC 0.942, sensitivity 96.4%, specificity 87.2%</td>
</tr>
<tr>
<td>Carmona et al. (2008)</td>
<td>Automated location and segmentation of the optic nerve head</td>
<td>10 eyes; 25 with glaucoma, 85 with ocular hypertension</td>
<td>Genetic algorithms</td>
<td>Generalization capability: 96%</td>
</tr>
<tr>
<td>Mookiah et al. (2013)</td>
<td>Glaucoma detection</td>
<td>100 images; 30 normal eyes, 39 glaucomatous eyes, 31 eyes with diabetic retinopathy</td>
<td>SVM</td>
<td>Maximum accuracy 98.2%</td>
</tr>
<tr>
<td>Raja et al. (2015)</td>
<td></td>
<td>40 images, 158 images, 74 glaucomatous eyes, 84 normal eyes</td>
<td>SVM</td>
<td>Maximum accuracy 98.2%</td>
</tr>
<tr>
<td>Ting et al. (2017)</td>
<td></td>
<td>494,661 retinal images; possible glaucoma: 125,189 images</td>
<td>Deep learning</td>
<td>AROC 0.942, sensitivity 96.4%, specificity 87.2%</td>
</tr>
<tr>
<td>Carmona et al. (2008)</td>
<td>Automated location and segmentation of the optic nerve head</td>
<td>10 eyes; 25 with glaucoma, 85 with ocular hypertension</td>
<td>Genetic algorithms</td>
<td>Generalization capability: 96%</td>
</tr>
</tbody>
</table>

ANN, artificial neural network; SVM, support vector machine; RNFL, retinal nerve fiber layer; AROC, area under the receiver operating characteristic curve.
an AROC of 0.94 and also showed that the neuroretinal rim, the optic disc area, and the lamina cribrosa and its surrounding regions were significantly associated with classification as glaucomatous. These studies are summarized in Table 2.

Although the studies described above report the general success of AI systems in identifying glaucomatous eyes, the majority of studies were unable to demonstrate superiority in diagnostic accuracy in comparison to using the best single conventional OCT parameter (e.g. rim area and average RNFL thickness). More complex transformations of the OCT data, including super-pixel segmentation in supervised ML, a hybrid DL approach, and use of the Mahalanobis distance, were, however, able to demonstrate superiority compared to using conventional OCT parameters achieving AROC values between 0.86 and 0.99.

The algorithms reported to date have been trained and validated on specific patient cohorts or on collections of disc photographs. A potential limitation of this approach is whether the quoted sensitivities and specificities will be generalizable to real-world patient populations where prevalent comorbidities, such as cataract and ocular surface disease, exist, negatively impacting the quality of images used as input data.

AI strategies to diagnose glaucoma using datasets derived from VF testing have been studied since 1994 and are summarized in Table 3. Using standard automated perimetry (SAP) perimetry data, AI can classify the severity of field loss from early to advanced damage from a single field. In 1994, the use of a back-propagation strategy (i.e. with no clinical diagnostic parameters incorporated) with an ANN showed that neural networks can be as proficient as a trained specialist in distinguishing normal from glaucomatous VFs, with agreement seen in 74% of cases. In the same year, an unsupervised ML classifier was shown to be capable of the identification of typical patterns of VF loss seen with clinical experience. Without being guided by a prior diagnosis, this approach was able to place 98% of normal visual fields within the same cluster and successfully classify 71% of glaucomatous fields across 4 other disease-specific clusters, showing good agreement with glaucoma specialists and pattern standard deviation. Andersson et al. were the first to report the potential outperformance of clinicians by a trained ANN in making a diagnosis of glaucoma based upon visual field test data. The ANN performed comparably to clinicians with specificities of 90% and 91%, respectively, however, with significantly improved sensitivity (91% vs. 83%).

Other studies have demonstrated that evaluation of VF tests with ML classifiers and trained ANNs perform as well as, or if not better, at identifying glaucomatous VFs than conventional parameters, such as the Glaucoma Hemifield Test, Mean Deviation, and Pattern Standard Deviation.

In order to identify early glaucomatous injury, ML has also been applied to frequency doubling perimetry data with promising results. Bowd et al. used an unsupervised ML classifier to differentiate normal VFs from glaucomatous visual fields with 93.1% specificity and 82.8% sensitivity. Recently, two papers were published reporting results of DL algorithms to diagnose glaucoma from VF data. The algorithm of Li et al. which involved a DNN, outperformed the diagnostic accuracy of glaucoma experts as well as traditional indices in differentiating normal from glaucomatous VFs, with a specificity of 83% and sensitivity of 93%. Kucur et al. have also developed an algorithm using a convolutional neural network capable of discriminating between normal and early glaucomatous VFs with an average precision score of 87%. However, in general, neural network performance is affected by training sets, which need to be large in size and well balanced in phenotype with respect to the normal and glaucomatous datasets, as well as in defect severity and defect location.

Misclassification may still be an issue in more challenging diagnostic scenarios, including patients with tilted and myopic optic discs.

Combining structural and functional inputs, such as standard automated perimetry and OCT parameters, does improve the ability for AI strategies to diagnose glaucoma with an AROC of up to 0.98 using an ANN. These approaches are summarized in Table 4. Incorporating other clinical parameters, including advancing age, intraocular pressure (IOP), and corneal thickness per se appear to contribute little to improving the diagnostic accuracy of the algorithms. This may not be surprising as the fundus appearance and retinal ganglion cell (RGC) functions are a manifestation of the disease itself rather than being directly determined risk factors for the disease.

### Glaucoma Progression

Detection of glaucoma progression is a key component of the clinical management of patients with glaucoma, in order to identify those individuals at risk of developing glaucoma-related visual impairment. Identifying progression over shorter time intervals is often challenging and requires the identification...
Table 2. Area Under the Receiver Operating Characteristic Curve (AROC) Values of Different Machine Learning (ML) Classifiers Using OCT Imaging for Glaucoma Diagnosis

<table>
<thead>
<tr>
<th>Study</th>
<th>Input Data</th>
<th>No. of Eyes/Images</th>
<th>ML Classifiers</th>
<th>AROC</th>
<th>OCT Parameter with Best Diagnostic Accuracy</th>
<th>AROC</th>
<th>Significance Level (best ML Approach Versus Conventional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burgansky-Eliash Z. et al. (2005)</td>
<td>38 conventional OCT parameters (macular and ONH)</td>
<td>27 early glaucoma, 20 advanced glaucoma, 42 healthy eyes</td>
<td>LDA</td>
<td>0.979</td>
<td>Rim area</td>
<td>0.969</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVML</td>
<td>0.981</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RPART</td>
<td>0.885</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GLM</td>
<td>0.975</td>
<td>Mean RNFL</td>
<td>0.938</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GAM</td>
<td>0.854</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang et al. (2005)</td>
<td>56 OCT parameters</td>
<td>89 glaucoma, 100 healthy</td>
<td>LDA</td>
<td>0.824</td>
<td>Inferior quadrant thickness</td>
<td>0.832</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MD</td>
<td>0.849</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>0.821</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naithani et al. (2007)</td>
<td>Peripapillary RNFL and ONH parameters (HRT) 19 parameters</td>
<td>30 early glaucoma 30 moderate glaucoma 60 healthy</td>
<td>LDA</td>
<td>0.982</td>
<td>Average RNFL thickness</td>
<td>0.953</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ANN</td>
<td>0.938</td>
<td></td>
<td>0.977</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTREE</td>
<td>0.979</td>
<td>Global transformed A-scan data</td>
<td>0.977</td>
<td>n/a</td>
</tr>
<tr>
<td>Bizios et al. (2010)</td>
<td>28 RNFL parameters</td>
<td>62 glaucoma, 90 healthy</td>
<td>ANN</td>
<td>0.958 to 0.995</td>
<td>Cup/disc area ratio</td>
<td>0.846</td>
<td>0.542</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVML</td>
<td>0.690</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barella et al. (2013)</td>
<td>23 parameters (RNFL thickness and ONH topography)</td>
<td>57 glaucoma, 46 healthy</td>
<td>ANN</td>
<td>0.958 to 0.995</td>
<td>Cup/disc area ratio</td>
<td>0.846</td>
<td>0.542</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BAG</td>
<td>0.804</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NB</td>
<td>0.818</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVMG</td>
<td>0.753</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MLP</td>
<td>0.768</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RBF</td>
<td>0.839</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RAN</td>
<td>0.877</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ENS</td>
<td>0.793</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CTREE</td>
<td>0.687</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ADA</td>
<td>0.839</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xu J et al. (2013)</td>
<td>OCT with super pixel analysis</td>
<td>59 glaucoma suspects 84 glaucoma 44 healthy</td>
<td>LDA</td>
<td>0.903</td>
<td>Average RNFL thickness</td>
<td>0.707</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Log</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamosa et al. (2015)</td>
<td>RNFL thickness: 2 semi-circles, 4 quadrants, and 6, 8, 12, 16, 24, 32, 64, and 768 sectors</td>
<td>117 glaucoma</td>
<td>ANN</td>
<td>0.770 to 0.845</td>
<td>12 peripapillary RNFL thickness sectors</td>
<td>0.845</td>
<td>0.0001</td>
</tr>
<tr>
<td>Muhammad et al. (2017)</td>
<td>RNFL thicknesses and retinal ganglion cell plus inner plexiform layer</td>
<td>123 healthy</td>
<td>RAN</td>
<td>0.77 to 0.97</td>
<td>Average RNFL thickness</td>
<td>0.973</td>
<td>n/a</td>
</tr>
<tr>
<td>Maetschke et al. (2019)</td>
<td>RNFL thicknesses, rim area, disc area, cup-to-disc ratio, vertical cup-to-disc ratio, cup volume</td>
<td>263 healthy, 847 glaucoma</td>
<td>DL</td>
<td>0.94</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

KNN, k-nearest neighbor; LDA, linear discriminant analysis; SVML, support vector machine linear; RPART, recursive partitioning and regression tree; GLM, generalized linear model; GAM, generalized additive model; MD, Mahalanobis distance; ANN, artificial neural network; Log, LogitBoost adaptive boosting; BAG, bagging; NB, naive-bayes; SVMG, support vector machine Gaussian; MLP, multi-layer perception; RBF, radial basis function; RAN, random forest; ENS, ensemble selections; CTREE, classification tree; ADA, AdaBoost M1; SAP, standard automatic perimetry.
Table 3. Summary of Studies Using Machine Learning (ML) Classifiers to Detect Glaucoma From Perimetric Datasets

<table>
<thead>
<tr>
<th>Study</th>
<th>Input Data</th>
<th>No. of Eyes/Images</th>
<th>ML Classifiers</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goldbaum et al. (1994)</td>
<td>Central 24° of standard automated perimetry with Humphrey Visual Field 24-2 or 30-2 SITA Standard visual field test</td>
<td>120 eyes, 60 normal 60 glaucomatous</td>
<td>Trained two layer ANN</td>
<td>Experts versus two-layered neural network. Sensitivity: 59% vs. 65% Specificity: 74% vs 71% Agreement 74%</td>
</tr>
<tr>
<td>Goldbaum et al. (2002)</td>
<td>SAP Humphrey visual field 24-2 or 30-2</td>
<td>189 normal eyes and 156 glaucomatous eyes</td>
<td>MLP, SVM, MoG, MGG</td>
<td>AROC 0.922, sensitivity 79%, specificity 90%</td>
</tr>
<tr>
<td>Chan et al. (2002)</td>
<td>SAP</td>
<td>189 normal eyes and 156 glaucomatous eyes</td>
<td>MLP, SVM, LDA, QDA, Parzen window, MoG, MGG</td>
<td>AROC 0.88-0.92 sensitivity 58.3–78.2% specificity 90%</td>
</tr>
<tr>
<td>Sample et al. (2004)</td>
<td>Standard automated perimetry with Humphrey visual field 24-2 or 30-2 SITA standard visual field test</td>
<td>345 eyes, 189 normal</td>
<td>vbMFA (unsupervised)</td>
<td>Comparing clusters versus GHT = 0.913–0.875 versus PSD = 0.905–0.863 versus expert = 0.873–0.829</td>
</tr>
<tr>
<td>Bizios et al. (2007)</td>
<td>Standard automated perimetry with Humphrey visual field 30-2</td>
<td>100 glaucoma eyes, 116 normal eyes</td>
<td>Trained artificial neural network compared to PSD</td>
<td>ANN: AROC 0.984, sensitivity 93%, specificity 94%</td>
</tr>
<tr>
<td>Andersson et al. (2013)</td>
<td>Standard automated perimetry with Humphrey visual field 30-2 SITA standard visual field test</td>
<td>99 glaucoma patients, 66 healthy subjects</td>
<td>Trained artificial neural network</td>
<td>30 physicians (varying experience) versus trained artificial neural network Sensitivity: 83% vs. 93% Specificity: 90% vs. 91%</td>
</tr>
<tr>
<td>Bowd et al. (2014)</td>
<td>FDT perimetry with Humphrey matrix (24-2 test pattern)</td>
<td>1976 eyes</td>
<td>Variational Bayesian independent component analysis-mixture model</td>
<td>compared to FDT sensitivity: 82.8% specificity: 93.1%</td>
</tr>
<tr>
<td>Asaoka et al. (2016)</td>
<td>Standard automated perimetry with Humphrey visual field 30-2 SITA standard visual field test</td>
<td>108 healthy eyes, 171 pre-perimetric glaucoma eyes</td>
<td>Deep FNN RF NN</td>
<td>AROC: Deep FNN 92.6% RF 77.6% NN 66.7%</td>
</tr>
<tr>
<td>Cai et al. (2017)</td>
<td>Standard automated perimetry with Humphrey visual field 24-2 SITA standard visual field test</td>
<td>243 eyes mean MD = –11.0 ± 8.7dB and PSD 9.5 ± 4.1dB</td>
<td>Archetypal analysis (unsupervised)</td>
<td>AT2 (superior defect) and ptosis P &lt; 0.001 AT12 cluster and stroke presence (temporal hemianopia) P = 0.02 AT1 (no focal defect) and GHT within normal limits P &lt; 0.001</td>
</tr>
<tr>
<td>Li et al. (2018)</td>
<td>Standard automated perimetry with Humphrey visual field 24-2 and 30-2 SITA standard visual field test</td>
<td>1623 normal eyes and 87 glaucomatous eyes (early stage)</td>
<td>DL</td>
<td>Sensitivity 93.2%, specificity 82.6%</td>
</tr>
<tr>
<td>Kucur et al. (2018)</td>
<td>OCTOPUS 101G1 program and the Humphrey Field Analyzer 24-2</td>
<td>158 normal eyes and 307 glaucomatous eyes</td>
<td>DL</td>
<td>Average precision 87.40%</td>
</tr>
</tbody>
</table>

ANN, artificial neural network; MD, mean deviation; GHT, Glaucoma Hemifield test; PSD, pattern standard deviation; FDT, frequency doubling technology; AROC, area under the receiver operating characteristic curve; vbMFA, variational Bayesian mixture of factor analysis; FNN, feed-forward neural network; RF, random forests; NN, neural network; MLP, multilayer perception; SVM, support vector machines; MoG, mixture of Gaussian; MGG, mixture of generalized Gaussian classifiers; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; SAP, standard automated perimetry; DL, deep learning.

Note: The table entries are based on the extracted text from the image.
Table 4. Area Under the Receiver Operating Characteristic Curve (AROC), Sensitivity and Specificity Values of Different Machine Learning Classifiers Using Optical Coherence Tomography (OCT) or Standard Automated Perimetry (SAP) Alone, or in Combination for Glaucoma Diagnosis

<table>
<thead>
<tr>
<th>Study</th>
<th>Input Data</th>
<th>No. of Eyes/Images</th>
<th>ML Classifiers</th>
<th>AROC/Sensitivity and Specificity OCT Parameter Alone</th>
<th>AROC/Sensitivity and Specificity SAP Parameters Alone</th>
<th>AROC/Sensitivity and Specificity for Combined Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brigatti et al. (1996)</td>
<td>SAP indices (mean defect, corrected loss variance, and short-term fluctuation) and structural data (cup/disk ratio, rim area, cup volume, and nerve fiber layer height)</td>
<td>185 glaucoma, 54 healthy</td>
<td>NN</td>
<td>87% sensitivity and 56% specificity</td>
<td>84% sensitivity and 86% specificity</td>
<td>90% sensitivity and 84% specificity</td>
</tr>
<tr>
<td>Bowd et al. (2008)</td>
<td>RNFL thickness + SAP</td>
<td>69 glaucoma, 156 healthy</td>
<td>RVM</td>
<td>0.809</td>
<td>0.815</td>
<td>0.845</td>
</tr>
<tr>
<td>Grewal et al. (2008)</td>
<td>Age, sex, myopia, intraocular pressure (IOP), optic nerve head, and retinal nerve fiber layer (RNFL), SAP and GDx parameters</td>
<td>35 glaucoma, 30 glaucoma suspects, 35 healthy</td>
<td>SSMoG</td>
<td>0.817</td>
<td>0.841</td>
<td>0.896</td>
</tr>
<tr>
<td>Bizios et al. (2011)</td>
<td>SAP and OCT</td>
<td>135 glaucoma, 125 healthy</td>
<td>ANN</td>
<td>0.970</td>
<td>0.945</td>
<td>0.978</td>
</tr>
<tr>
<td>Sugimoto et al. (2013)</td>
<td>VF damage, age, gender, right or left eye, axial length, 237 different OCT measurements</td>
<td>224 glaucoma, 69 healthy</td>
<td>RAN</td>
<td>m-RNFL (0.86), cp-RNFL (0.77), GCL + IPL (0.80), rim area (0.78)</td>
<td>0.9 (all parameters)</td>
<td></td>
</tr>
<tr>
<td>Silva et al. (2013)</td>
<td>SD-OCT parameters and global indices of SAP</td>
<td>62 glaucoma, 48 healthy</td>
<td>Conventional</td>
<td>0.574—0.813</td>
<td>0.828—0.915</td>
<td></td>
</tr>
<tr>
<td>Kim et al. (2017)</td>
<td>Age, IOP, corneal thickness, RNFL, GHT, MD, PSD</td>
<td>178 glaucoma, 164 healthy</td>
<td>C5.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANN, artificial neural network; MLC, machine learning classifier; RVM, relevance vector machine; BAG, bagging; NB, naive Bayes; NN, neural network; MLP, multilayer perception; RBF, radial basis function; RAN, random forest; ENS, ensemble selection; CTREE, classification tree; ADA, AdaBoost M1; SVML, support vector machine linear; SVMG, support vector machine Gaussian; SSMoG, subspace mixture of Gaussians; KNN, k-nearest neighbor; SAP, standard automatic perimetry; GHT, Glaucoma Hemifield test.
of structural or functional change at the earliest possible time point. Because AI algorithms have the potential to incorporate structural or functional changes over time, they have the potential to provide more accurate and timely identification of likely glaucoma progression.

**Structural Aspects: Imaging Techniques**

With the widespread availability of ever increasingly sophisticated imaging technology, there will be further opportunities to develop longitudinal analytical approaches to detect glaucoma progression. Although AI technologies have been developed for glaucoma screening using fundus photographs, this approach has not been evaluated to detect progression.

Multiclass support vector machines (SVMs), a form of supervised ML, have been used to simultaneously discriminate between normal, nonprogressing, and progressing eyes through the analysis of confocal scanning laser ophthalmoscopy (CSLO) images with a correct classification rate of 88%.

The incorporation of pixel-wise rates of change from CSLO image analysis has been shown to reduce the overall false-positive rate in detecting glaucoma progression. This strategy demonstrated a sensitivity of 86% in progressing eyes, compared to 39% using conventional approaches. Higher sensitivity for progression with similar specificity was shown compared to statistical image mapping, suggesting an improved ability to detect glaucomatous progression. The sensitivity and specificity of a unified framework for detection of glaucomatous progression using CSLO images was reported as 86% and 88%, respectively.

A hierarchical framework for detecting glaucoma progression using spectral-domain OCT images encompassing the whole three-dimensional volume of the optic nerve head has also been tested. The control dataset for training of the algorithm included both healthy normal and stable nonprogressive glaucoma eyes, which resulted in a very robust algorithm. This technique was able to demonstrate high diagnostic accuracy with 78% sensitivity for detecting glaucoma progression, compared to 69% using an ANN evaluating RNFL thickness alone. More recently, the application of computational techniques to a large set of swept-source OCT images to identify structural features associated with glaucoma progression has been described. These features outperformed glaucoma detection using conventional measures (e.g. SAP, peripapillary OCT, and RNFL scans) with an AROC of 0.95, compared to 0.90 for average global peripapillary RNFL thickness and 0.86 for SAP mean deviation.

**Functional Aspects: Visual Field Analysis**

In clinical practice, glaucoma progression is often identified through the analysis of serial VF tests using SAP and is considered to be the gold standard, despite its test-retest variability. Early work in 1997 by Brigatti et al. demonstrated glaucoma progression through analyzing serial fields with a neural network. They reported a sensitivity and specificity of 73% and 88% with good concordance of neural network observers. To date, the networks have used supervised learning techniques, but, in 2005, Sample et al. used unsupervised ML to identify areas of progression in glaucomatous VF tests comparable or even better than clinical criteria. In tandem with this study, a sister paper was published by Goldbaum et al. detailing the application of ML in identifying and validating patterns of glaucomatous VF defects, reporting an impressive 98.4% specificity.

Various ML approaches have been tested for their clinical effectiveness for detecting VF progression of which the strongest performed strongest with an AROC of 0.86, 89.9% sensitivity, and 93.8% specificity. Detailed summaries are presented in Table 5.

Of particular clinical significance, AIs have been shown to be able to detect progressing eyes 20 months earlier than using conventional approaches, such as global, region-wise, and point-wise indices. This was without the need for a further visit for confirmation and showed particular strength in detecting slowly progressing eyes. More recently, Wen et al. used an unfiltered real-world dataset of over 30 thousand VF tests and 1.7 million perimetry points to train a DL ANN that was able to predict future VF test performance over a 5-year period given only a single input field test. Further validation of this approach from other groups may enable future incorporation of this strategy into clinical risk stratification models.

**Combining Structure and Function**

As observed with glaucoma diagnosis, the detection of progression was improved using a combination of different modalities with ML generating good AROCs, but not remarkably higher than with single modality inputs as would be intuitively expected.

The combination of structural and functional parameters using different ML classifiers generated AROC curves for progression detection from 0.83 to 0.88. A Bayesian joint longitudinal model to integrate structural and functional information from longitudinal measures has also been evaluated. Information derived from one test influenced the inferences obtained from the other test. Therefore, a SAP change
Table 5. Summary of Studies Using Artificial Intelligence to Detect Progression in Glaucomatous Eyes

<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Eyes/Images</th>
<th>Follow-up, y</th>
<th>Instrument</th>
<th>Approach</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brigatti et al. (1997)⁹⁰</td>
<td>233 n/a</td>
<td>SAP</td>
<td>Supervised ML</td>
<td>Sensitivity 73%; specificity 88%; AROC 0.88</td>
<td></td>
</tr>
<tr>
<td>Lin et al. (2003)¹³⁶</td>
<td>80 7.2</td>
<td>SAP</td>
<td>Supervised ML</td>
<td>Sensitivity 86%; specificity 88%; AROC 0.92</td>
<td></td>
</tr>
<tr>
<td>Sample et al. (2005)¹¹¹</td>
<td>191 6.2</td>
<td>SAP</td>
<td>Unsupervised ML</td>
<td>Sensitivity, Specificity, AROC n/a Comment: The classifier separated the data based on the patterns of visual field loss, placing 98.4% of the healthy eyes within the same cluster and spreading 70.5% of the eyes with glaucoma across the other clusters, in good agreement with conventional methods</td>
<td></td>
</tr>
<tr>
<td>Goldbaum et al. (2012)⁹²</td>
<td>478 suspects, 150 glaucoma, and 55 stable glaucoma</td>
<td>4.0 SAP</td>
<td>Unsupervised ML</td>
<td>Sensitivity 98.4%; Specificity, AROC n/a Comment: Use of variational Bayesian independent component analysis mixture model in identifying patterns of glaucomatous visual field defects and its validation</td>
<td></td>
</tr>
<tr>
<td>Medeiros et al. (2012)¹²⁷</td>
<td>380 suspects, 331 glaucoma, and 50 stable glaucoma</td>
<td>5.0 SAP</td>
<td>Bayesian hierarchical model</td>
<td>Presented a method of integrating event- and trend-based analyses of visual field progression that performed better than either isolated analyses alone</td>
<td></td>
</tr>
<tr>
<td>Murata et al. (2014)¹²⁸</td>
<td>5049 (training data) and 911 (test data)</td>
<td>4.4 SAP</td>
<td>Unsupervised ML</td>
<td>Sensitivity, Specificity, AROC n/a Comment: Variational Bayes model predicts more accurately future SAP progression in glaucoma patients compared to conventional methods, especially in short series</td>
<td></td>
</tr>
<tr>
<td>Yousefi et al. (2016)¹⁴⁶</td>
<td>859 abnormal SAP and 1117 normal SAP</td>
<td>9.1 SAP</td>
<td>Unsupervised ML</td>
<td>AROC 0.82 for VIM-POP, 0.86 for GEM-POP, 0.81 for permutation of point-wise linear regression, 0.69 for linear regression of MD, and 0.76 for linear regression of VFI</td>
<td></td>
</tr>
<tr>
<td>Yousefi et al. (2018)⁹⁵</td>
<td>939 abnormal SAP and 1146 normal SAP in the cross-sectional and 270 glaucoma in the longitudinal dataset</td>
<td>9.0 SAP</td>
<td>Unsupervised ML</td>
<td>Sensitivity 34.5-63.4% at specificity 87% Comment: It took 3.5 years for the ML analysis to detect progression while it took over 3.9 years for other methods to detect progression in 25% of the eyes</td>
<td></td>
</tr>
<tr>
<td>Wang et al. (2019)¹²⁹</td>
<td>11817 (method developing cohort) and 397 (clinical validation cohort)</td>
<td>7.6 and 6.3</td>
<td>SAP</td>
<td>Unsupervised ML</td>
<td>AROC of the archetype method 0.77</td>
</tr>
</tbody>
</table>

Downloaded from tvst.arvojournals.org on 11/22/2021
### Table 5. Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>No. of Eyes/Images</th>
<th>Follow-up, y</th>
<th>Instrument</th>
<th>Approach</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. (2013)</td>
<td>96</td>
<td>3.3</td>
<td>SLP</td>
<td>Supervised ML</td>
<td>AROC 0.82</td>
</tr>
</tbody>
</table>
| Balasubramanian et al. (2014) | 36 progressing, 210 non-progressing and 21 healthy controls | 4.1, 3.6 and 0.5 | CSLO        | Supervised ML | Sensitivity 39-86%
| Comment: Progression detected by pixelwise rates of retinal height changes in non-progressing eyes was associated with early signs of SAP change |
| Belghith et al. (2014) | 36 progressing, 210 non-progressing and 21 healthy controls | 4.1, 3.6 and 0.5 | CSLO        | Reinforcement ML | Sensitivity 86%; specificity 88% |
| Belghith et al. (2015) | 27 progressing, 26 stable glaucoma and 40 healthy controls | 2.4, 0.1 and 2.0 | SD-OCT | Supervised ML | Sensitivity 78%; specificity in normal eyes 93%, 94% in non-progressing eyes |
| Christopher et al. (2018) | 179 glaucoma and 56 healthy controls | 2.1 and 1.8 | SS-OCT | Unsupervised ML | AROC 0.95 for RNFL principal component analysis |
| Medeiros et al. (2011) | 434 glaucoma and suspects | 4.2 | Combined (SAP and SLP) | Bayesian hierarchical model | Bayesian method: Sensitivity 74%, Specificity 100%, AROC 0.9-0.94 |
| OLS method: sensitivity 37%, specificity 100%, AROC 0.77-0.79 |
| Bowd et al. (2012) | 264 suspects (47 progressing and 217 stable) | 5.4 and 5.1 | Combined (SAP and CSLO) | Supervised ML | AROC between 0.640 and 0.805, sensitivity 21–72% at 75% specificity |
| Medeiros et al. (2012) | 242 glaucoma | 6.4 | Combined (SAP and CSLO) | Bayesian hierarchical model | Sensitivity, specificity, AROC n/a |
| Comment: Bayesian joint regression model combining structure and function resulted in more accurate and precise estimates of slopes of change compared to the conventional method of ordinary least squares linear regression |
| Medeiros et al. (2012) | 352 glaucoma | 8.1 | Combined (SAP and information on risk factors and structural damage) | Bayesian hierarchical model | Sensitivity, specificity, AROC n/a |
| Comment: incorporating structural and risk factor information resulted in more precise estimation of glaucomatous visual field progression |
| Yousef et al. (2014) | 107 progressing and 73 stable glaucoma | 2.2 and 0.1 | Combined (SAP and SD-OCT) | Unsupervised ML | AROC from 0.83 to 0.88 |

SAP, standard automated perimetry; SLP, scanning laser polarimetry; CSLO, confocal scanning laser ophthalmoscopy; SD-OCT, spectral domain optical coherence tomography; SS-OCT, swept source optical coherence tomography; ML, machine learning; AROC, area under the receiver operating characteristic curve.
that would otherwise be declared not statistically significant by analysis of SAP data alone could become significant after taking into consideration structural changes occurring in the same eye.98,99 This approach resulted in more accurate and precise estimates of rates of change compared to the conventional method. Glaucoma progression has been successfully predicted from baseline CSLO and SAP through relevance vector machine (RVM) classifiers.100 Incorporation of known risk factors and information from additional tests into the assessment of change resulted in a better accuracy of the risk detection for development of functional impairment in individual patients101 (detail on individual studies is summarized in Table 5).

Discussion

This review summarizes the current status of AI strategies with regard to glaucoma detection and diagnosis and in assessing progression of the disease, and highlights the potential future role of this sphere of innovation in shaping how glaucoma care may be delivered to the next generation.

ML algorithms developed using almost 50,000 fundus images have been shown to identify referable glaucomatous optic neuropathy with an AROC of 0.90.38 Further DL algorithms trained on matched fundus and OCT images of over 30,000 eyes are able to discriminate between glaucomatous and healthy eyes with an AROC of 0.98,47 and may even be superior to human grading.46 Algorithms incorporating further clinical parameters and information from VF testing and OCT imaging were able to identify patients with glaucoma with an AROC of 0.98, even when only using under 200 subjects.75,81

Despite the progress that has been made in developing AI strategies for glaucoma diagnosis, several significant hurdles still need to be overcome before these advances can be translated to clinical practice. Establishing a ground truth for glaucoma diagnosis can be contentious even among experts in the field.102 This becomes evident as studies have variable levels of agreement between glaucoma specialists in differentiating patients with glaucoma from subjects without the disease.103–105 Ultimately, any supervised ML or DL approach is dependent on the “ground truth” as its reference standard, which in the case of glaucoma diagnosis can prove to be challenging. Establishing a ground truth for glaucoma progression is equally contentious. However, a potential solution to this is to utilize datasets from patients with long-term follow-up. As glaucoma is a progressive disease, absolute confirmation of diagnosis may only be possible in some cases through the evaluation of extended longitudinal data.

Although impressive AROC values have been demonstrated by many study algorithms, it is difficult to compare the clinical applicability between different studies with differing methodologies. Algorithms may vary between clinic settings, the diversity of inputs from commercially available devices, and also due to the subjective and variable nature of patient-reported data.106 Furthermore, current published research has not been designed to account for the natural variability that exists within populations, including the impact of ethnicity, extremes of refractive error, and age. AI strategies show great promise in their ability to discriminate between glaucomatous and healthy subjects. However, further large-scale population-based algorithm validation is essential in order to confidently implement these advances toward assisting glaucoma diagnosis in the general population. In addition, AI strategies need to be transferable in order to accommodate input data from different machines using standardized methodological approaches.

Even though the results on AI strategies using VF inputs for progression analysis shows considerable promise, even AI cannot overcome one of the major challenges in glaucoma care, which is how we define progression using a test that is prone to significant test–retest variability. The studies mentioned used a variety of methods to define glaucoma progression. These include event-based and trend-based approaches to detect visual field progression107 (see Table 5). Event-based analysis compares the sensitivities of the current VF to established thresholds from baseline examinations. In trend-based analyses, VF sensitivities of all tests during the follow-up period are analyzed to identify any statistically significant change over time. This is usually done by using a linear regression approach. In addition, even with an AI approach to VF analysis, the algorithms are still dependent on patient factors like fixation losses.

As highlighted before, establishing confirmed perimetric progression to define the “ground truth” is not without challenge. Even the “expert opinion” of glaucoma specialists of detecting glaucoma progression from assessment of the optic disc alone cannot be regarded as a “gold standard.”108 Numerous objective protocols have been therefore developed to identify VF progression and are frequently used in routine clinical care. However, considerable inter-protocol variability exists. This is an ongoing challenge even in current clinical practice and will further impact the generalizability of innovations derived through AI to individual clinical settings.109–111 In addition,
considerable ocular variability occurs within patient populations depending upon factors, such as age, gender, refractive error, medical comorbidities, and ethnicity. Before ML strategies can be translated to everyday clinical practice, further validation across diverse global patient populations is necessary. There is ongoing debate about the relationship between structure-function correlations and glaucoma progression, and how any mismatch between these should be addressed. AI studies have the potential to integrate all available data and provide a more reliable and objective conclusion.

**Future Prospects**

Although superiority of AI technologies to humans is frequently reported in the media and even supported by some emerging studies, this area of innovation should be regarded as a tool to supplement the skills of clinicians who face the challenge of delivering high quality glaucoma care to an aging population with an increasing life expectancy. In the future, AI may become an essential adjunct to glaucoma diagnosis, which will not replace the clinical skills but facilitate decision making. AI strategies have the potential to transform how clinical glaucoma care may be delivered in future years. This transformation will undoubtedly be facilitated by the digital era of constantly improving technologies, connectivity, imaging, and electronic medical records, and will enable the improved efficiency and workflow of a glaucoma service, as visualized in the Figure.

AI algorithms may be developed to serve as a glaucoma referral refinement scheme to manage referrals from community based screening programs and optometrists, as has been suggested for diabetic retinopathy screening. Through DL approaches, analysis of fundus photographs reports a diagnostic accuracy of more than 99% in detecting glaucoma, a clinically acceptable performance level for translation to patient care. This is not an unrealistic goal given that for diabetic retinopathy screening, a DL system developed by Abramoff et al. has obtained a US Food and Drug Administration approval with a sensitivity of 87.2% and specificity of 90.7%. Comparable and even superior levels of performance have been demonstrated by several groups using AI algorithms to diagnose glaucoma, however, not in a population-based study (see Tables 1–3 for summaries).

We propose that in cases of an established diagnosis of glaucoma, AI strategies may have the potential to function as an additional adjunct to the glaucoma assessment in making a clinical diagnosis in more challenging cases by helping to support the diagnosis or reject it. The detection of glaucoma progression at earlier stage using DL algorithms compared with conventional approaches may enable earlier intervention and therefore further reduce the risk of patients developing glaucoma-related visual impairment in their lifetimes. Future studies and training datasets of anterior segment OCT images may facilitate throughput in remote monitoring clinics by providing direction on safe pupil dilation to nonmedical staff through the identification of occludable drainage angles.

A diagnosis of glaucoma is based upon expert evaluation, which can be challenging to replicate in a population-based screening program. A major barrier to the widespread implementation of glaucoma screening at a global level is related to the lack of a simple and reliable screening test. More sensitive tests will detect real cases of glaucoma better, whereas more specific tests will better detect healthy cases. Together, high specificity and sensitivity will prevent unnecessary clinic reviews of individuals who do not have glaucoma, which in turn enables more efficient use of the available healthcare resources. Health
economic analyses suggest that although whole-population screening may not be cost-effective, programs focusing on higher risk groups may be worthwhile. A 5-yearly glaucoma screening program for older patients would require a test specificity greater than 96% in order to be cost-effective.

Systematic reviews have provided no evidence in support of an individual test or group of tests that show superiority for glaucoma screening, however, these analyses were performed prior to the advent of advances in OCT technology. Nevertheless, the best sensitivity/specificity balance with an acceptable cost-effectiveness may be achieved through the combination of parameters, including IOP measurement, SAP, and vertical C/D ratio. The incorporation of OCT-based parameters can only further improve this performance, and should be the focus of DL algorithms in the future.

There have been major advances in teleophthalmology in recent years, in both developed as well as developing countries. “Teleglaucoma” involves remote analysis of imaging data like stereoscopic disc photographs or results of functional testing. Remote review of fundus photographs has been shown to be clinically effective and more cost-effective than face to face consultations. This approach offers benefits to both patients and healthcare systems, including early diagnosis, reduced travel, increased targeted specialist referral rates, and cost-efficiency savings. Further research into how AI strategies may further refine and stratify telemedicine referrals within referral refinement schemes would certainly improve the efficiency of healthcare systems and improve the overall quality of glaucoma care.

We have already highlighted that the diagnosis of glaucoma (i.e. the gold standard or “ground truth” in terms of AI), shows considerable variability even between expert observers. This may prove to be a challenge in the provision of training datasets for supervised ML algorithms. However, in reality, the major practical advantage of an AI based screening protocol would be to discriminate between “likely glaucoma” and “not glaucoma.” (see the Figure) Sources of dispute between expert clinicians often arise in more complex cases, for example, with atypical optic disc appearances (e.g. myopic optic discs) and patterns of VF loss. Cases such as these can be challenging to classify, and would likely require a face to face consultation for definitive diagnosis and to ensure that potential confounding pathologies are not missed. This would also serve as a safety net mechanism to minimize the risk of misclassification and incorrect diagnosis. Conversely the use of DL algorithms may enable the identification of novel parameters associated with glaucoma - the so-called “unknown unknowns,” which can help support or reject a diagnosis. This may enable the discovery of biomarkers that may facilitate the identification and prediction of glaucomatous change at an earlier stage in the disease than is currently achievable. This may also expedite drug discovery pipelines for novel molecular and therapeutic approaches toward goals, such as neuroprotection and neuroregeneration, which may have only been aspirational prior to the advent of AI.

Fundus photographs are the simplest and most readily available input modality for ML algorithms. Clinical databases within healthcare systems contain vast numbers of archived fundus photographs often with corresponding OCT imaging and perimetry data that can be used for training datasets. VF testing, on the other hand, is reliant upon patient compliance, is more time-consuming, less widely available, and exhausting for the patient compared to a fundus photograph. For this reason, fundus photographs were the first dataset type to be tested using AI approaches both in retinal disease and glaucoma. Currently, ML approaches using photographs alone and augmented by training with OCT datasets can obtain diagnostic specificities in excess of 95%, which are at an acceptable level for direct translation to patient care. Accelerating the translation of AI interpretation of fundus photographs for glaucoma screening is a realistic and reasonable goal, considering that automated analysis of fundus photographs are already in place for diabetic retinopathy screening.

In order to truly maximize the potential power of AI, both in terms of diagnostic ability and to improve the efficiency of healthcare delivery systems, a longer-term aim should be to incorporate the latest advances in imaging technology and perimetry into future algorithms in the same manner as in routine clinical practice. However, caution in interpretation and validation of outputs should be taken to ensure that ML classifiers are based upon glaucoma-related parameters as opposed to other population-based features that may demonstrate a strong correlation with patients who have glaucoma. The implementation of ML approaches in discriminating between stable eyes and those with glaucoma progression with fewer tests and in a shorter timescale would have major impact upon glaucoma research. It is likely that these strategies would lead to the development of novel end points for future clinical trials of drug or surgical interventions, which would enable results to be more rapidly obtained, therefore accelerating the translation of innovation to patient care.

Despite this promise, there are still many hurdles that need to be overcome in order to implement AI strategies in a clinical setting. The advances discussed in this review have largely been performed on highly curated smaller training datasets from individual
institutions. In reality, training datasets may need to contain up to 100,000 images covering all stages of the disease spectrum, and the outcome of algorithm will be dependent on image quality, which may need to be standardized and with accurate phenotyping. There are also numerous sources of variability within the global population and therefore further validation studies and/or training datasets will need to be tested in a variety of populations in order to maximize the external validity of novel AI algorithms. For example, it is not currently known whether every ML approach is as effective in every ethnicity as in “one size fits all.”

A significant barrier to the acceptability of AI strategies within healthcare is the “black box” phenomenon. The ability of clinicians to accept and trust outputs of an algorithm, when the decision-making process is not apparent or comprehensible to them, may prove to be an obstacle to adoption. Ultimately, the responsibility for individual healthcare decisions lies with the responsible physician, who may fear liability from adverse outcomes arising from clinical decisions based upon AI tools of which they have an inherent suspicion. Medical training is based upon appraising available evidence to make a rational and considered clinical decision in the best interest of an individual patient. The computational and more abstract approach used by DL algorithms to make similar decisions can be unsettling for the intellectual mindset of clinicians. Acceptability to medical professionals and regulatory agencies may be increased if there is enhanced understanding as to how an algorithm arrives at its decision. This approach was adopted by the Moorfields/DeepMind collaboration by generating relevant tissue segmentations for clinicians to interpret as a device-independent representation of the algorithm. Moving forward, further research in to so called “Explainable AI” may provide the necessary transparency, trust and accountability desired by the healthcare profession. Elze et al. used an archetypal analysis to develop a framework more meaningful to clinicians to quantify the various subtypes of glaucomatous VF loss. This approach was developed further by Yousefi et al. to study glaucoma progression, by using a ML-driven approach to cluster longitudinal VF data of glaucoma patients to generate an “AI-enabled glaucoma dashboard,” which showed a specificity of 94% for “likely nonprogression.” This has the potential to provide a clinician-friendly tool to help determine the severity of glaucomatous VF deficit and a means for monitoring disease progression.

Regulatory permissions will need to be secured from regional authorities, such as the US Food and Drug Administration and the European Medicines Agency. The required performance standards required for glaucoma are yet to be discussed and will likely require further international discussion and consensus. The precise regulated role of where AI approaches may sit in the clinical care pathway will be challenging to define. Despite the promising performance statistics presented in published papers, the real-world impact of false-positive or in particular false-negative results derived from AI technologies remains unclear. Clinical decisions based upon AI may even confer increased medicolegal liability upon manufacturers in the case of missed diagnoses, which may ultimately influence the cost and rate of adoption of such innovation.

Ultimately, the uptake of AI technologies within clinical glaucoma practice will be dependent upon clinicians themselves. AI algorithms may help to augment referral refinement in order to efficiently triage those patients who need to be seen by a specialist, and those who do not. The integration of AI within new models of care delivery will be driven by the combined opportunity to optimize both resource utilization and the workload of clinicians, thus enabling the provision of high-quality glaucoma care to a population that continues to increase in both number and age.

**Acknowledgments**

A.S.M.-E., W.S.N., A.D.-F., D.C.S., L.A., M.B.S., K.D.-A., J.G.C., and H.J. are participants in an independent educational program (“International Glaucoma Panel”) supported by the Allergan Medical Institute. P.A.K. is supported by a Moorfields Eye Charity Career Development Award and a UK Research and Innovation Future Leaders Fellowship. H.J is supported by the Moorfields Eye Charity. P.A.K. and H.J. are grateful for the support of the National Institute for Health Research Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The views expressed in this paper are those of the authors and not necessarily those of any funding body or the UK Department of Health.

**Disclosure:** A.S. Mursch-Eldmayr, None; W.S. Ng, None; A. Diniz-Filho, None; D.C. Sousa, None; L. Arnould, None; M.B. Schlenker, None; K. Duenas-Angeles, None; P.A. Keane: Heidelberg Engineering (F), Topcon (F), Carl Zeiss Meditec (F), Haag-Streit (F), Allergan (F), Novartis (F, S), Bayer (F, S),
DeepMind (C), Optos (C); J. Crowston, Allergan (C); H. Jayaram, Allergan (C)

References


75. Bizios D, Heijl A, Bengtsson B. Integration and fusion of standard automated perimetry and optical coherence tomography data for improved...
Artificial Intelligence in Glaucoma


