OPEN

Deployment of Artificial Intelligence in Real-World Practice: Opportunity and Challenge

Mingguang He, MD, PhD†, Zhixi Li, MD, PhD*, Chi Liu, MS*‡*
Danli Shi, MD, and Zachary Tan, MBBS§¶*

Abstract: Artificial intelligence has rapidly evolved from the experimental phase to the implementation phase in many image-driven clinical disciplines, including ophthalmology. A combination of the increasing availability of large datasets and computing power with revolutionary progress in deep learning has created unprecedented opportunities for major breakthrough improvements in the performance and accuracy of automated diagnoses that primarily focus on image recognition and feature detection. Such an automated disease classification would significantly improve the accessibility, efficiency, and cost-effectiveness of eye care systems where it is less dependent on human input, potentially enabling diagnosis to be cheaper, quicker, and more consistent. Although this technology will have a profound impact on clinical flow and practice patterns sooner or later, translating such a technology into clinical practice is challenging and requires similar levels of accountability and effectiveness as any new medication or medical device due to the potential problems of bias, and ethical, medical, and legal issues that might arise. The objective of this review is to summarize the opportunities and challenges of this transition and to facilitate the integration of artificial intelligence (AI) into routine clinical practice based on our best understanding and experience in this area.

Key Words: artificial intelligence, deployment, real-world
(*Asia Pac J Ophthalmol (Phila)* 2020;9:299–307)

Downloaded from https://journals.lww.com/apjoo by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 06/17/2021

ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND DEEP LEARNING

Artificial intelligence (AI) was first proposed in print in a Dartmouth Summer Research Project in 1955.[1] AI is a broad term referring to a branch of computer science that is hypothetically committed to developing computer algorithms for the tasks that have traditionally been accomplished by human intelligence, such as the ability to learn and solve problems. Machine learning (ML) is a division of AI that provides knowledge in the form of data to computers, along with observations to optimize the goodness of fit between input—including text, image, or video data—and output as a classification. A conceptual and engineering breakthrough by pioneers of the field, Yoshua Bengio, Geoffrey Hinton, and Yann LeCun enabled the development of artificial neural networks and deep learning (DL) to become a subfield of ML. This technology requires multiple processing layers to learn and detect features ranging from simple ones such as lines, edges, textures, and intensity, to complex features like shapes, lesions, and a whole image in a hierarchical structure.

Neural networks, inspired by simulating the neurons in the brain, include algorithms that are commonly used for image analysis today. These neural networks are composed of a number of layers of connected nodes, where each node receives information from other nodes and also sends a signal to other groups of nodes. The goal of the overall network is to find an answer that matches a defined ground-truth label by changing the pattern and weights of node connections via thousands of millions of attempts until the best match of the ground truth is achieved. Many types of neural structures have been proposed, representing various ways to cluster those nodes. The most common type of neural network used for image recognition is a convolutional neural network (CNN).[2]

The "training" of neural networks is conducted either by supervised learning, where a training set of data with annotations by humans to match the disease outcome are used, or unsupervised learning, in which the training data do not have annotations and where the algorithm strives to cluster or organize to "understand" the underlying patterns. The majority of ML systems to date in ophthalmology are developed using supervised learning,[2,3] wherein the CNN analyzes pixel data from a large number of manually labeled images to determine a specific classification of disease type and severity.

CURRENT STATUS OF AI DEVELOPMENT IN OPHTHALMOLOGY

Ophthalmology is a branch of medicine that deals with the diagnosis and treatment of eye diseases. A number of imaging modalities have been used for the diagnosis of eye diseases; however, the interpretation of these images is highly dependent on the skill and experience of physicians, and this process is often subjective, with substantial interobserver variation.[4–6] Evidence supporting this includes a study wherein even senior glaucoma specialists could only achieve a "substantial" level of agreement

From the *State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China; †Centre for Eye Research Australia, Royal Victorian Eye & Ear Hospital, Melbourne, Australia; ‡School of Computer Science, University of Technology Sydney, Ultimo NSW, Australia; §Faculty of Medicine, The University of Queensland, Brisbane, Australia; and ¶Schwarzman College, Tsinghua University, Beijing, China.
Submitted February 29, 2020; accepted May 18, 2020.
Financial Support: Supported in part by the National Key R&D Program of China (2018YFC0116500), the Fundamental Research Funds of the State Key Laboratory in Ophthalmology, National Natural Science Foundation of China (81420108008), Science and Technology Planning Project of Guangdong Province (2013B20400003).
The authors report no conflicts of interest.
M.H. holds a patent for using deep learning models to process color fundus images (patent application number: ZL201510758675.5, patent granted date: May 31,2017)
Address Correspondence and reprint requests to: Mingguang He, MD, PhD, Center for Eye Research Australia, Royal Victorian Eye & Ear Hospital, University of Melbourne, Melbourne 3003, Australia. E-mail: mingguang.he@unimelb.edu.au
Copyright © 2020 Asia-Pacific Academy of Ophthalmology. Published by Wolters Kluwer Health, Inc. on behalf of the Asia-Pacific Academy of Ophthalmology. This is an open access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.
ISSN: 2162-0989
DOI: 10.1097/APO.0000000000000301

($\kappa = 0.63$) for the classification of glaucoma based on optic disc photography. This agreement could be even poorer among general ophthalmologists ($\kappa = 0.51$) and trainees ($\kappa = 0.50$).[6] Similarly, optometrists achieved sensitivity of 67% and specificity of 84% with diagnosing diabetic retinopathy (DR), which is below the recommended screening standard.[7]

Abràmoff et al[8] and Gulshan et al[9] published the first 2 articles on DL technology in the field of ophthalmology, with a task to detect DR based on retinal fundus photographs in 2016. These 2 studies both reported an area under the curve (AUC) achieving 0.98 to 0.99, a level of accuracy that is far better than any previous reports based on traditional pattern recognition of lesions. The power of DL, for the first time, captured the imagination of the whole field of ophthalmology. Since then, a number of research groups developed their algorithm based on similar CNNs and published their results on multiple eye diseases, including DR,[10–14] glaucoma,[15–17] age-related macular degeneration (AMD),[18–21] retinopathy of prematurity,[22–25] based on a variety of imaging modalities, including fundus photography, optical coherence tomography (OCT),[26–31] visual field,[32–34] and many others.[35,36] Despite this variation, DL algorithm (DLA) applications primarily focus on 3 major tasks: classification, segmentation, and prediction using static images collected from medical devices.

The most common DLA application is to generate a global classification from a specific image, either with or without disease or on a specific disease severity scale. Some review articles have been published to summarize the performance of these DLAs with a variety of diseases,[37–42] but several key points must be highlighted.

First, almost all the studies reported robust accuracy, described as AUC, sensitivity, and specificity, which is far better than what has been reported in human image graders and machine learning tools based on traditional pattern recognition algorithms. For example, in the classification of referrable DR, DLA achieves an accuracy of AUC 0.98 to 0.99 with a sensitivity of 0.97 (ranging from 0.89 to 0.99) and specificity of 0.96 (ranging from 0.98 to 0.99). These are comparable, if not better than, trained human graders where unanimous consensus grading by specialist experts was set as the benchmark.[43]

Second, almost all the studies used CNNs that are publicly available. Google Inception V3 is the most commonly used of these, followed by VGG-net, AlexNet, ResNet, and so on.[9,10,13,27,44,45] The adoption of a neural network model often depends on its availability when the study is conducted. Several studies have used pretrained CNNs and transfer learning, achieving similar accuracy with relatively smaller sample sizes.[17,26,46–48] In 1 study, 6 CNN models were simultaneously assessed on AMD classification; AlexNet yielded a better performance but the difference in performance among the networks was minimal.[18]

Third, CNN image classification is not used only with 2-dimensional (2-D) images but also with 3-dimensional (3-D) images such as OCT. A few studies have reported on the performance of OCT B-scans where 2-D data were used.[27,28,31,49,50] Another recently published study developed and reported on a 3-D DLA of an OCT image to classify glaucomatous optic neuropathy when 3-D parapapillary retinal nerve fiber layer (RNFL) data were used.[51]

Fourth, DL is not used only for image classification but also for lesion detection and image segmentation. This task is more complex than image classification, consisting of creating a boundary definition around the objects in an image and classifying each of them. U-net is the most commonly used algorithm for this task, and it has proved to be accurate in the detection of exudates, hemorrhage, and optic discs in fundus photographs. It has also been accurate with segmenting OCT structures and detecting intraretinal fluid and subretinal fluid, and OCT pathologies such as neovascularization, macular edema, drusen, geographic atrophy, epiretinal membrane, vitreous traction, and macular holes.[52–56]

Fifth, in image classification tasks, DL makes classifications based on the global image instead of generating a classification after the detection of a specific lesion. This raises "black box" concerns where this classification is based on a one-size-fits-all neural network architecture that is not specific to disease. Heatmaps have become a popular method to highlight the pixel regions that contribute most to the DL classification. DL research has increasingly published their heatmap results where frequently, heatmap regions do not always necessarily match with the features that clinicians commonly used to differentiate disease, highlighting that the CNN may "see" things differently to humans.[21,27,51,57,58] In fact, some studies have synthesized images through very small perturbations of the pixels, which can easily fool the CNN to produce a completely inaccurate output.[59,60]

Sixth, DL may go beyond simple classification of an image; it can also predict the prognosis or outcomes of a treatment when progression data are used as ground truth for training the algorithm. DL has demonstrated that image data alone, without referring to other known risk factors, is able to achieve reasonably good performance in predicting the prognosis for DR, AMD progression, and structural and functional progression in glaucoma, although most of these DLAs have yet to be independently validated.[19,61–63]

Seventh, DL is able to "see" the features that are not differentiable by humans in certain classification tasks, such as cardiovascular disease risk factors and smoking status. Poplin et al from Google demonstrated that a DLA trained with UK Biobank and US EyePACS datasets was able to classify age, current smoking status, blood pressure, body mass index, and even 5-year myocardial infarction with reasonably good accuracy among independent datasets.[64] Their DLA was unexpectedly validated in a small group of 239 patients selected from a randomly selected Asian database during the publication peer review process.[42] This finding is intriguing because it was able to prove that the features of the retina, as a biologically relevant end-organ for the vascular and neural systems, could be used to classify cardiovascular disease risks when other known risk factors were not included.

Despite all the fascinating advancements in AI technology, developments in how to translate and deploy AI technology into real-world ophthalmology practice remains challenging.

Challenges of AI algorithms in ophthalmology

Challenge #1: Adequate Quantity of Training Data

A useful training dataset should exhaust all possible variations of disease phenotypes, including but not limited to the variations of disease severity, ethnicity, artifacts, types of fundus camera, and confounding of coexisting diseases. In this scenario, the clinical characteristics of the training set should be clearly delineated. For instance, an algorithm that was trained based on a dataset from a screening setting might not be appropriately used in hospital clinical settings, where disease severity is substantially different. An algorithm developed based on subjectively defined glaucoma without referring to the visual field or relevant clinical diagnostic data may not be appropriate for real-world deployment

© 2020 Asia-Pacific Academy of Ophthalmology.

as an end product for glaucoma diagnosis, as the benchmark used in the training process is different from the purpose of deployment.

This dependency on large amounts of data for accurate algorithm development has become an impediment to the adoption of AI in clinical practice. Hospitals may have a large amount of data but not have good access to computer science and AI experts. The data in hospitals are often not well organized for meaningful data mining, and other obstacles such as regulation considerations, privacy protection, ethical issues, and legal concerns may further hinder data sharing. Similarly, computer science companies have the computing power and AI expertise, but they do not have access to clinical datasets. Although the "Big Nine" (Google, Alibaba, Amazon, Tencent, Apple, Baidu, Facebook, IBM, and Microsoft) have invested tremendous amounts of resources on AI development, that does not mean they had sufficient access to clinical data. It would be tremendously helpful to create freely available disease or device-specific shared data resources for computer experts to use in testing different algorithm designs. The publicly available ImageNet dataset that has been used to generate many breakthroughs in image recognition; the public datasets for DR, organized by Kaggle, which has been used by many developers for DR algorithms; the UK Biobank's open-source data for eye disease classification and prediction model development are good examples of successful data sharing. Nevertheless, alternative learning methods have recently been proposed that can simulate how the human brain works and learn from fewer examples. Using generative adversarial networks, one group of researchers[65–68] created or synthesized a large number of diverse and random computed tomography and magnetic resonance imaging images from scratch and claimed that the set of images could be used as a training set for future CNN development. These efforts however remain yet to be proven and have not achieved great success to date.[69]

Challenge #2: Appropriate Definition of Ground Truth and Labeling

A majority of algorithms are developed based on images retrieved directly from medical devices, and then a number of experts are asked to subjectively grade (or label) these images. The ground truth is determined by the unanimity of the graders along a continuum, such as normal, probably normal, indeterminate, probably abnormal, or definitely abnormal; or by simply classifying the images into normal or abnormal. This approach is often subject to significant misclassification, errors, and insufficiency because expert classification is subjective, and there is significant interobserver variation.[70,71] The label from subjective classification does not necessarily contain clinically important information such as the likelihood of progression, potential treatments, responses to treatment, and so on. An ideal ground truth should be based on criterion standard definitions and retrieved from real-world clinical data such as pathology reports and electronic medical records, which often depend on multiple modalities of imaging devices and clinical procedure details that comply with diagnostic standards.

Challenge #3: Assessment of the Accuracy of AI Algorithms

Critical assessment of the accuracy of an AI algorithm should be based on widely recognized principles of evidence-based medicine, just like other new medicines or devices. An article in preprint that is released via online repositories, for instance, arXiv.org; peer-reviewed articles describing technical developments published in computer science journals; or even peer-reviewed articles reporting diagnostic performance in clinical journals do not necessarily fully establish the accuracy of technologies or justify the adoption of the technology in real-world clinical practice. In this context, an AI algorithm developed for diagnosis or classification should comply with the Standards for Reporting Diagnostic Accuracy (STARD) statement, and an AI algorithm developed for prediction should follow the Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement for transparent reporting of study settings, study population, definitions and measurements of outcomes, time, and interval of follow-up, and so on, as AI algorithms are sensitive to the target population, equipment used, imaging protocols, and referral standards.[72,73] A number of initiatives to develop specific guidelines for AI-based clinical trials are currently under deliberation.[74–76] These include the Consolidated Standards of Reporting Trial (CONSORT)-AI, Specific Protocol Items: Recommendations for Interventional Trials (SPIRIT)-AI and TRIPOD-ML statements, which are extensions of existing guidelines that provide clinical trial protocol and reporting standards.

Overfitting is the most common bias in AI algorithm development, where the algorithm is overfitted to the training set. This can be driven by a mistake in training set development, for instance, using one type of fundus camera to collect images for the disease and another type of fundus camera for the normal group, so that the algorithm essentially ends up classifying the type of fundus camera instead of disease and nondisease. Reliable external validation of the data collected in newly recruited patients or at different sites, or using different models of device as an independent study, is the best way to mitigate overfitting problems. There are, in general, 2 approaches to external validation of algorithms noted in published studies—either a publicly available image dataset or a prospective study. In ophthalmology, Abràmoff et al,[8] Gargeya and Leng,[58] and Gulshan et al[9] used the Messidor-2 dataset, E-Ophtha databases, or EyePACS-1 as a reference for external validation. However, an ideal validation should be based on images that have never been used in the training set, and a validation process conducted by researchers is often not able to ensure this is the case. Therefore, an ideal validation for an image dataset should be done on larger standardized datasets provided by an independent party using a setup similar to Kaggle's public image recognition challenges, where the ground truth has not been disclosed, so that a neutral objective benchmark can be established.

Similar to the assessment of other new medicines or devices, the assessment of AI algorithms should ideally be carried out in a prospective clinical trial. Recently, Lin et al[77] conducted a clinical trial to compare an AI-assisted cataract classification technique versus ophthalmologists in real-world settings, and found the AI technique was less accurate in diagnosing childhood cataracts than ophthalmologists but was more efficient. There have been 25 items registered under eye diseases on the clinical trials website (https://clinicaltrials.gov/) using terms such as "convolutional neural network," or "deep learning," or "machine learning," or "artificial intelligence," and "ophthalmology," with most focused on DR, glaucoma, cataracts, visual acuity assessment, and so on.

© 2020 Asia-Pacific Academy of Ophthalmology.

Challenge 4: Mode of Care of AI Adoption

Currently available AI algorithms in ophthalmology fundamentally enable tasks of classification, detection, or segmentation. Classification refers to assigning an entire image or lesion in a particular image to a category, for example, the DR severity scale. Detection is done to identify a specific abnormality within an image, for example, choroidal neovascular lesions in an OCT image. Segmentation is done to identify or isolate a specific structure of interest in an image, such as isolating the RNFL in an OCT image.

The outcomes of interest in AI classification can be multiple and varied. The most common outcome of interest is separating abnormal from normal patients. This perhaps appears to be a very simple day-to-day task for ophthalmologists, but it could be challenging for noneye professionals, such as asking an endocrinologist to classify an image as either with or without referable DR, despite this possibly being part of their professional training for the management of diabetic complications. The second common outcome of interest is to classify or assign the images into severity categories or grading schemes, such as classifying an image on the DR severity scale. This classification task is often more difficult and less accurate than a dichotomous classification because an effective differentiation between certain grades could be challenging when the difference is minimal, for instance, differentiating normal and mild DR where the only difference is the presence of a microaneurysm. The third outcome of interest is to predict the prognosis of a disease, such as to differentiate progressive glaucoma from stable ones.

Challenge #5: Integrate AI Into Clinical Pathways

Bossuyt et al[78] proposed 3 clinical pathways (triage, replacement, and add-on) to integrate new diagnostic tests into existing clinical pathways that would be appropriate for AI deployment.

In the triage model, AI algorithms can be used as a triage tool for opportunistic screening in noneye clinical settings or as a tool for assisting integration into pathways of grading in reading centers. In Australia, our research team has installed an AI system in endocrinology and primary care settings to enable opportunistic screening and targeted referral so that DR patients are referred to eye professionals (Fig. 1), whereas the normal ones stay for routine glucose management. In this research deployment model, the AI system was used to generate a report that would be reviewed by a qualified physician such as an endocrinologist who would sign off on the diagnosis. In this case, the AI is considered a decision support tool for diagnosis. In China, our research team works with a local software company to provide technical support for a nationwide DR screening program where the AI is used as the first triage tool for identifying "super-normal" cases, defined as classifications of normal and gradable in all 4 images collected for an individual (Fig. 2). In the case of super-normal, a report from AI is generated and sent to the patient at their point-of-care. A pilot study proved that this model could reduce the workload for telemedicine grading centers by >50%.

In a replacement model, the AI algorithm is used to replace the clinical diagnosis task of a clinician because the AI can be more accurate, rapid, and reproducible and less dependent on access to a clinician. This model of care is feasible only for tasks where AI is definitively superior to physicians (such as estimation of bone age in radiology) or for tasks that are simple enough to carry out, such as classifying the image as that of a right eye or left eye, but often require strict regulatory approval.

An add-on model of care refers to where AI is used as a procedure that is used in parallel with or after diagnosis from clinicians. This model is often used when the task involves time-consuming and repetitive work, such as counting lesions (eg, microaneurysms in DR grading) or automatically measuring RNFL thickness that involve lots of manual segmentation on a large number of OCT B-scans (Fig. 3).

Challenge #6: AI Clinical Adoption Is Beyond Clinical Consideration

Successful AI deployment in clinical practice requires the active involvement of all stakeholders, including patients, ophthalmologists, imaging technicians, hospital administrations, regulatory

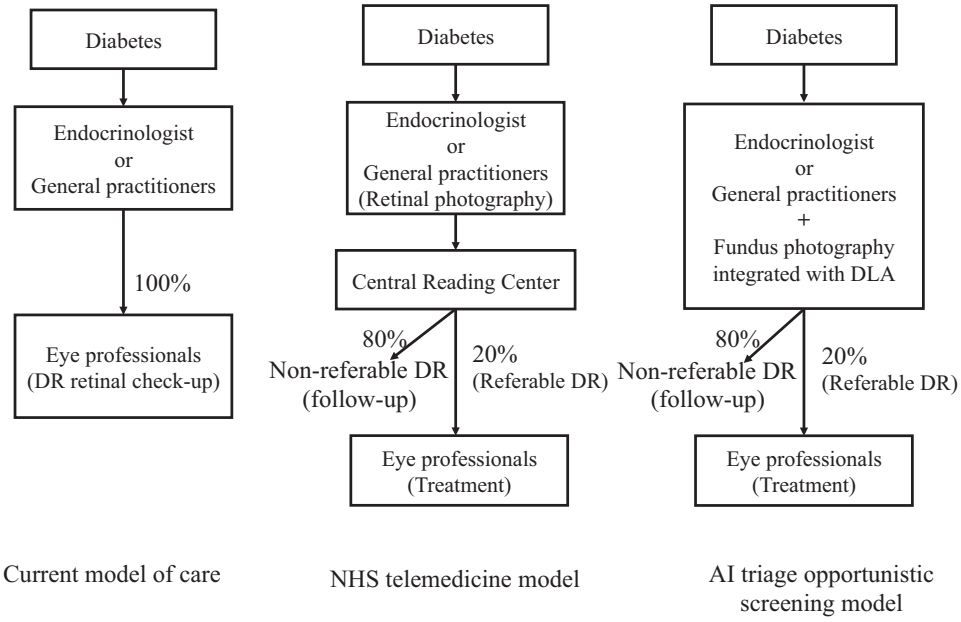

FIGURE 1. Artificial intelligence triage opportunistic screening model. DLA indicates deep learning algorithm; DR, diabetic retinopathy; AI, artificial intelligence.

© 2020 Asia-Pacific Academy of Ophthalmology.

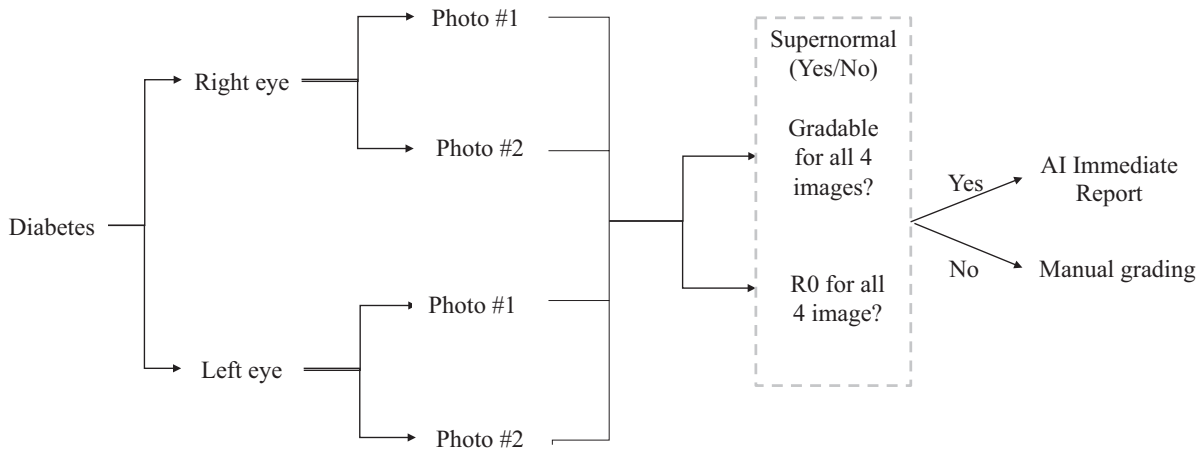

FIGURE 2. Artificial intelligence is used as an initial triage strategy to maximize efficiency of manual grading. AI indicates artificial intelligence.

bodies, and industry. It is important to ensure that all stakeholders will benefit from AI deployment and are willing to collaboratively facilitate the development of best practices by integrating ethics, patient consent, privacy protection, data ownership and sharing, integration with existing electronic medical system, and user-friendly software interface for targeted clinical settings.

In 2016, the 21st Century Cures Act was signed into law, aiming to accelerate the discovery, development, and delivery of new cures and treatments. The US Food and Drug Administration's (FDA's) current strategic policy emphasizes leveraging innovation, including digital health technologies, and is highlighted by new software as a medical device (SaMD) and digital health regulatory approval pathways for AI and computer vision algorithms.[79,80] To date, the FDA has approved several AI-based SaMDs, including the IDx ML-based software system for automated detection of DR and an AI product for use with computed tomography scans for indicators associated with stroke.[81,82] Most of these approvals are for algorithms that are locked in before going to market, but the FDA is currently also considering SaMD regulatory frameworks for continuous learning and adaptive algorithms that are potentially able to be adapted to improve performance in real time.

Challenge #7: Security Issues in AI Deployment

Recently, various attack methods have proved effective against existing DL models. For example, some studies found that by adding adversarial noises to a raw image, a well-trained model could be successfully fooled into making a wrong decision that is totally opposite to the ground truth.[83] The adversarial noise can be carefully designed to ensure the manipulated image is visually the same as the raw image such that it is almost impossible for humans to detect. Worse, this kind of adversarial attack can be conducted as a black box attack, meaning that no previous knowledge of the model details, such as information about the model's structure or parameters, is required by an attacker.

The vulnerability of DL models has spurred the community to seriously rethink the security and robustness of AI in real-world deployments, especially in the medical domain.[84] Scaling up AI systems for clinical use without any defendable countermeasures means that any falsified diagnosis could lead to considerable risks.

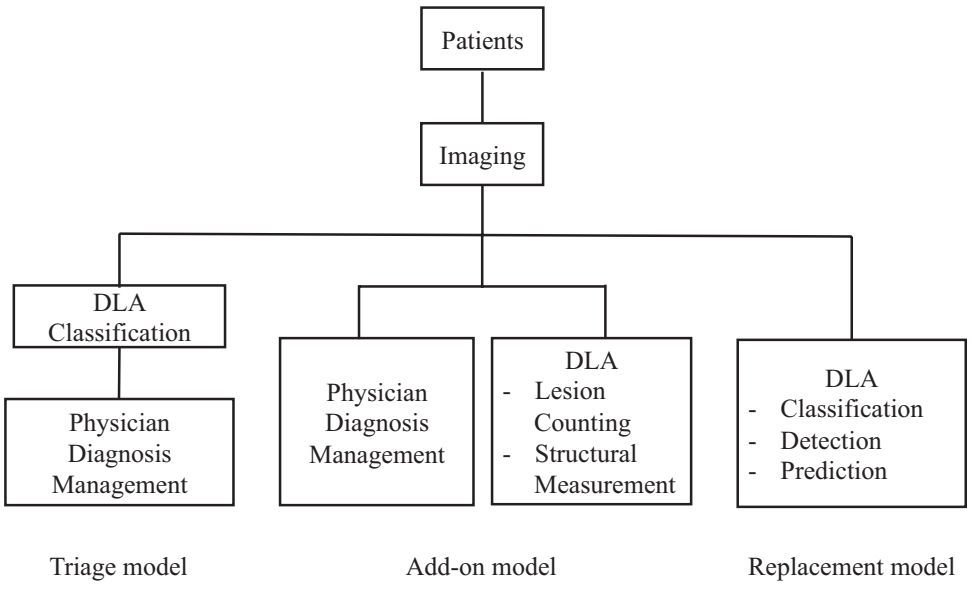

FIGURE 3. Three models of integration: triage, add-on, and replacement models. DLA indicates deep learning algorithm.

© 2020 Asia-Pacific Academy of Ophthalmology.

Case Study: IDx-DR and Real-World Challenges of DLA

The limitations of retrospective in silico validation of DLA are significant. Real-world prospective trials are now increasingly essential for clinical uptake and regulatory approval of DLA systems. In a landmark study by Abràmoff et al,[12] the efficacy of the "IDx-DR" system, a fully autonomous screening system for more-than-mild DR (mtmDR), was evaluated in a real-world prospective trial of 900 patients across 10 primary care sites in the United States.

This trial is notable as it both addresses and reflects many of the key challenges described in this article. The DLA was assessed in a prospective clinical trial, where inclusion criteria and the clinical setting were strictly defined. Eligible participants were asymptomatic patients, diagnosed with diabetes and not previously diagnosed with DR, in a primary care setting. All images were captured with 1 retinal camera model, the Topcon NW400 system. Captured images were evaluated by 2 IDx-DR image quality and diagnostic algorithms, which determined the presence of mtm DR. Results produced by the DLA were compared with high-quality ground-truth Wisconsin Fundus Photograph Reading Center widefield stereoscopic photography and OCT.

The system achieved a sensitivity and specificity of 87.2% and 90.7%, respectively, meeting endpoints predetermined by the US FDA. This allowed it to achieve the first regulatory approval ever for an autonomous AI-based diagnostic system.[85] Following this precedent, prospective real-world validation trials are now essential for future regulatory approval of DLA systems. Notably, however, regulatory approval for the IDx-DR system remains limited to the cohort that was defined in the system's validation trial. This includes the detection of mtmDR only in adults diagnosed with diabetes who have not been previously diagnosed with DR, and in nondilated images captured by the Topcon NW400 camera.[86]

The real-world validation of DLA systems in less-controlled real-world settings remains immature. In the first human-centered observational study of a DLA in clinical care published in April 2020, Google Health researchers working in 11 clinics across Thailand encountered a number of socioenvironmental factors that limited the accuracy and adoption of a DR screening DL system.[87] Challenges include clinic screening conditions, image gradeability affecting system performance, internet speed and connectivity, and the impact of referrals on patient time. For instance, several clinics reported issues with image gradeability as fundus images were captured in nondarkened clinics that resulted in insufficient pupil dilation and insufficient quality images. Alternative darkened clinic rooms could not be found. Images were often rejected for grading by the DLA, requiring multiple attempts that added frustration and work to an already busy clinic. Furthermore, the DLA system required images to be uploaded to the cloud for assessment. The study sites often experienced slower and less reliable connections that slowed down the overall screening workflow, and reduced the number of patients a clinic could screen daily. Lastly, a large number of patients were discouraged from participation in the study, after understanding during the consent process that a positive screening result would require further assessment in a hospital an hour drive away, opting out to avoid possible additional time burden.

Thus, as described in this article, successful AI deployment in clinical practice requires the involvement of all stakeholders, including patients, ophthalmologists, imaging technicians, hospital administrations, regulatory bodies, and industry. End-users and their environment determine implementation, which may be as important as the accuracy of the algorithm itself. Early and material consideration of these real-world factors will be essential to the successful future clinical deployment of DLA systems.

THE FUTURE

There have been arguments made and concerns expressed that AI will replace professionals in future practice. However, one should note that currently, supervised ML is typically trained to discriminate features based on a trusted training set for only a limited assigned task, whereas humans are able to transfer experience and expertise to a new task through reasoning. A DLA system may be able to classify the presence and severity of a limited number of predefined diseases, to segment image structures, or even predict disease prognosis more accurately and perhaps more efficiently than humans. However, it is unable to make a valid diagnosis of diseases that it has not been trained for, nor able to accurately perform clinical reasoning based on multimodality data and experience, interact with patients properly, and perform treatment procedures like human doctors. To paraphrase the prominent AI expert Ng, the measure of a good AI technology is that it does well what humans can do, but easier and quicker, in 1 second.[88] This will likely remain true until the development of the "singularity": a hypothetical future point in time when "general AI" becomes available such that machines can learn, reason, and create like humans, undoubtedly and unforeseeably changing human civilization.

REFERENCES

1. McCarthy J, Minsky M, Rochester N, Shannon C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*. 2006;27:12–14.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
3. Belthangady C, Royer LA. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat Methods*. 2019;16:1215–1225.
4. Breusegem C, Fieuws S, Stalmans I, Zeyen T. Agreement and accuracy of non-expert ophthalmologists in assessing glaucomatous changes in serial stereo optic disc photographs. *Ophthalmology*. 2011;118:742–746.
5. O'Neill EC, Gurria LU, Pandav SS, et al. Glaucomatous optic neuropathy evaluation project: factors associated with underestimation of glaucoma likelihood. *JAMA Ophthalmol*. 2014;132:560–566.
6. Kong YX, Coote MA, O'Neill EC, et al. Glaucomatous optic neuropathy evaluation project: a standardized internet system for assessing skills in optic disc examination. *Clin Exp Ophthalmol*. 2011;39:308–317.
7. Sundling V, Gulbrandsen P, Straand J. Sensitivity and specificity of Norwegian optometrists' evaluation of diabetic retinopathy in single-field retinal images—a cross-sectional experimental study. *BMC Health Serv Res*. 2013;13:17.
8. Abràmoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*. 2016;57:5200–5206.

© 2020 Asia-Pacific Academy of Ophthalmology.

9. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402–2410.
10. Ting D, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211–2223.
11. Tufail A, Rudisill C, Egan C, et al. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. *Ophthalmology*. 2017;124:343–351.
12. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
13. Li Z, Keel S, Liu C, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. *Diabetes Care*. 2018;41:2509–2516.
14. Verbraak FD, Abràmoff MD, Bausch GCF, et al. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. *Diabetes Care*. 2019;42:651–656.
15. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125:1199–1206.
16. Phene S, Dunn RC, Hammel N, et al. Deep learning and glaucoma specialists: the relative importance of optic disc features to predict glaucoma referral in fundus photographs. *Ophthalmology*. 2019;126:1627–1639.
17. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353–1360.
18. Grassmann F, Mengelkamp J, Brandl C, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*. 2018;125:1410–1420.
19. Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J, Bressler NM. Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. *JAMA Ophthalmol*. 2018;136:1359–1366.
20. Keenan TD, Dharssi S, Peng Y, et al. A deep learning approach for automated detection of geographic atrophy from color fundus photographs. *Ophthalmology*. 2019;126:1533–1540.
21. Keel S, Li Z, Scheetz J, et al. Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs. *Clin Exp Ophthalmol*. 2019;47:1009–1018.
22. Brown JM, Campbell JP, Beers A, et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol*. 2018;136:803–810.
23. Gupta K, Campbell JP, Taylor S, et al. A quantitative severity scale for retinopathy of prematurity using deep learning to monitor disease regression after treatment. *JAMA Ophthalmol*. 2019;137:1029–1036.
24. Taylor S, Brown JM, Gupta K, et al. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. *JAMA Ophthalmol*. 2019;137:1022–1028.
25. Tan Z, Simkin S, Lai C, Dai S. Deep learning algorithm for automated diagnosis of retinopathy of prematurity plus disease. *Transl Vis Sci Technol*. 2019;8:23.
26. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
27. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122–1131.
28. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. 2018;125:549–558.
29. Sun Z, Tang F, Wong R, et al. OCT angiography metrics predict progression of diabetic retinopathy and development of diabetic macular edema: a prospective study. *Ophthalmology*. 2019;126:1675–1684.
30. Medeiros FA, Jammal AA, Thompson AC. From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology*. 2019;126:513–521.
31. Christopher M, Bowd C, Belghith A, et al. Deep learning approaches predict glaucomatous visual field damage from OCT optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology*. 2020;127:346–356.
32. Sample PA, Chan K, Boden C, et al. Using unsupervised learning with variational bayesian mixture of factor analysis to identify patterns of glaucomatous visual field defects Glaucoma detection based on deep convolutional neural network. *Invest Ophthalmol Vis Sci*. 2004;45:2596–2605.
33. Asaoka R, Murata H, Iwase A, Araie M. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology*. 2016;123:1974–1980.
34. Wang M, Tichelaar J, Pasquale LR, et al. Characterization of central visual field loss in end-stage glaucoma by unsupervised artificial intelligence. *JAMA Ophthalmol*. 2020;138:190–198.
35. Son J, Shin JY, Kim HD, Jung KH, Park KH, Park SJ. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*. 2020;127:85–94.
36. Varadarajan AV, Bavishi P, Ruamviboonsuk P, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning identifying medical diagnoses and treatable diseases by image-based deep learning. *Nat Commun*. 2020;11:130.
37. Cheung CY, Tang F, Ting DSW, Tan GSW, Wong TY. Artificial intelligence in diabetic eye disease screening. *Asia Pac J Ophthalmol (Phila)*. 2019;8:158–164.
38. Kapoor R, Whigham BT, Al-Aswad LA. Artificial intelligence and optical coherence tomography imaging. *Asia Pac J Ophthalmol (Phila)*. 2019;8:187–194.
39. Li Z, Keel S, Liu C, He M. Can artificial intelligence make screening faster, more accurate, and more accessible? *Asia Pac J Ophthalmol (Phila)*. 2018;7:436–441.
40. Tan Z, Scheetz J, He M. Artificial intelligence in ophthalmology: accuracy, challenges, and clinical application. *Asia Pac J Ophthalmol (Phila)*. 2019;8:197–199.
41. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunovic H. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1–29.
42. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: the technical and clinical considerations Artificial

© 2020 Asia-Pacific Academy of Ophthalmology.

Intelligence in Diabetic Eye Disease Screening. *Prog Retin Eye Res*. 2019;72:100759.

43. Raumviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:25.
44. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. 2019;126:552–564.
45. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digit Health*. 2019;1:e35–e44.
46. Quellec G, Lamard M, Conze PH, Massin P, Cochener B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Med Image Anal*. 2020;61:101660.
47. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135:1170–1176.
48. Burlina P, Pacheco KD, Joshi N, et al. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated amd analysis automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *Comput Biol Med*. 2017;82:80–86.
49. Samagaio G, Estevez A, Moura J, Novo J, Fernandez MI, Ortega M. Automatic macular edema identification and characterization using OCT images. *Comput Methods Programs Biomed*. 2018;163:47–63.
50. Chakravarthy U, Goldenberg D, Young G, et al. Automated identification of lesion activity in neovascular age-related macular degeneration. *Ophthalmology*. 2016;123:1731–1736.
51. Ran AR, Cheung CY, Wang X, et al. Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis. *Lancet Digit Health*. 2019;1:e172–e182.
52. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. New York, NY: Springer; 2015, 234–241.
53. Zheng R, Liu L, Zhang S, et al. Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network. *Biomed Opt Express*. 2018;9:4863–4878.
54. Zhao H, Sun N. U-net model for nerve segmentation. International Conference on Image and Graphics. New York, NY: Springer; 2017, 496–504.
55. Burewar S, Gonde AB, Vipparthi SK. Diabetic retinopathy detection by retinal segmentation with region merging using CNN. 2018 IEEE 13th International Conference on Industrial and Information Systems; 2018. New York, NY: IEEE; 2018, 136–142.
56. Tennakoon R, Gostar AK, Hoseinnezhad R, Bab-Hadiashar A. Retinal fluid segmentation in Oct images using adversarial loss based convolutional neural networks. 2018 IEEE 15th International Symposium on Biomedical Imaging. New York, NY: IEEE; 2018, 136–142.
57. Keel S, Wu J, Lee PY, Scheetz J, He M. Visualizing deep learning models for the detection of referable diabetic retinopathy and glaucoma. *JAMA Ophthalmol*. 2019;137:288–292.
58. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. 2017;124:962–969.
59. Moosavi-Dezfooli SM, Fawzi A, and Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. arXiv preprint arXiv:1511.04599 (2015). 2574–2582.
60. Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*. 2019;23:828–841.
61. Rohm M, Tresp V, Muller M, et al. Predicting visual acuity by using machine learning in patients treated for neovascular age-related macular degeneration. *Ophthalmology*. 2018;125:1028–1036.
62. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med*. 2019;2:1–9.
63. Christopher M, Belghith A, Weinreb RN, et al. Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression. *Invest Ophthalmol Vis Sci*. 2018;59:2748–2756.
64. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2:158–164.
65. Maspero M, Savenije MH, Dinkla AM, et al. Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy. *Phys Med Biol*. 2018;63:185001.
66. Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van der Berg CA, Išgum I. Deep MR to CT synthesis using unpaired data. International workshop on simulation and synthesis in medical imaging; 2017. New York, NY: Springer; 2017, 14–23.
67. Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *Trans Med Imaging*. 2018;37:1348–1357.
68. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal*. 2019;58:101552.
69. George D, Lehrach W, Kansky K, et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science*. 2017;358:eaag2612.
70. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med*. 1994;331:1493–1499.
71. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*. 2015;313:1122–1132.
72. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem*. 2015;61:1446–1452.
73. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;102:148–158.
74. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25:1467–1468.
75. Liu X, Faes L, Calvert M, Denniston AK. CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. *Lancet*. 2019;394:1225.
76. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577–1579.
77. Lin H, Li R, Liu Z, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine*. 2019;9:52–59.

© 2020 Asia-Pacific Academy of Ophthalmology.

78. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332:1089–1092.
79. FDA. Software as a Medical Device. 2018. Available at: https://www.fda.gov/medical-devices/digital-health/software-medical-device-samd (accessed 2020308).
80. FDA. Digital Health Innovation Action Plan. Available at: https://www.fda.gov/media/106331/download (accessed 20200308).
81. FDA. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. 2018. Available at: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye(accessed 20200513).
82. FDA. FDA permits marketing of clinical decision support software for alerting providers of a potential stroke in patients. 2018. Available at: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye. (accessed 20200513).
83. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*. 2018;6:14410–14430.
84. Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial attacks against medical deep learning systems. *arXiv preprint arXi*. 2018;1804.05296.
85. FDA administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems: U.S. Food and Drug Administration. 2018. Available at: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye. (accessed 20200513).
86. De Novo Classification Request for IDX-DR. 2018. Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf (accessed 20200515).
87. Beede E, Baylor E, Hersch F, et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020, Available at: https://dl.acm.org/doi/pdf/10.1145/3313831.3376718.
88. Ng A. What artificial intelligence can and can't do right now. *Harvard Business Review*. 2016;9.

© 2020 Asia-Pacific Academy of Ophthalmology.