SriniVas R. Sadda, MD
Section Editor

Sharon Fekrat, MD, FACS
Section Editor

Exploring the Role of Reading Centers in the Era of Artificial Intelligence

In recent years, artificial intelligence (AI) has begun to revolutionize medicine and medical research—and in some cases, available algorithms can outperform clinicians. The debate is everywhere: Will machines someday monopolize or even take over clinical care? Here, we specifically address the involvement of AI in reading centers—organizations that perform image analysis in clinical trials and that happen to be the perfect playground for AI.

Panelists

Barbara A. Blodi, MD
Fundus Photograph Reading Center
University of Wisconsin-Madison
Madison, Wisconsin

Amitha Domalpally, MD
Fundus Photograph Reading Center
University of Wisconsin-Madison
Madison, Wisconsin

Muneeswar Gupta Nittala, MS
Doheny Image Reading Center
University of California-Los Angeles
Los Angeles, California

Michael S. Ip, MD
Doheny Image Reading Center
University of California-Los Angeles
Los Angeles, California

'The debate is everywhere: Will machines someday monopolize or even take over clinical care?'

Terminology and technology

Artificial intelligence (AI) is an umbrella term for any machine capable of imitating human behavior, eg, an algorithm that can distinguish classical music from jazz. **Machine learning** is a subfield of AI that uses statistical techniques to enable computer systems to learn without being explicitly programmed; learning without programming is the key concept in AI.

All machine learning is AI, but not all AI is machine learning. **Deep learning** is an approach to machine learning that uses layers of interconnected units (termed *neurons*) that are able to learn features and patterns. The term *neural network* is inspired by the structure of the human brain's synaptic nodes, particularly the visual cortex pathway.

The neural network most popular in identifying imaging patterns is called a *convolutional neural network (CNN)*. In a simplified model, the CNN consists of an input layer, multiple hidden layers, and an output layer (Figure 1). The input layer receives images and the hidden layer parses the data to identify patterns, producing the final result through the output layer. Deep learning using CNN is particularly useful in identifying patterns in complex data and has yielded major advances in the field of medical image processing.

From a reading center perspective, human interpretation of ocular images follows a pattern-recognition process involving image scanning for pertinent pathological lesions to interpret the disease stage (Figure 1). Traditional machine-learning models are similar to human interpretation and use a process called *feature selection* to identify patterns from retinal images, eg, microaneurysms vs hemorrhages vs artifacts such as cotton-wool spots, to detect diabetic retinopathy (Figure 2).

'All machine learning is AI, but not all AI is machine learning.'

Deep-learning algorithms can bypass the feature selection process and interpret disease stage directly. With deep learning, training images tagged with the presence or absence of diabetic retinopathy are fed into the algorithm, which eventually "learns" to detect the disease. It is unknown whether the hidden CNN layers use lesions such as microaneurysms to detect diabetic retinopathy or have developed their own unique pattern recognition model. The unknown nature of the decision-making process has created both a fascination and distrust for AI and has led to the term *black box*.

AI in ophthalmology

AI has become an ophthalmology buzzword in recent years. With the first AI screening methods for diabetic retinopathy approved by the US Food and Drug Administration (FDA), we know that deep learning is going to change the future of ophthalmology. In 2015, Kaggle, a platform for predictive modeling and analytics, held a competition for deep-learning algorithms for diabetic retinopathy classification.[1] With more than 7000 entries, the competition brought to light the potential application of deep-learning algorithms in retinal imaging.

‘The unknown nature of the decision-making process has created both a fascination and distrust for AI and has led to the term *black box*.’

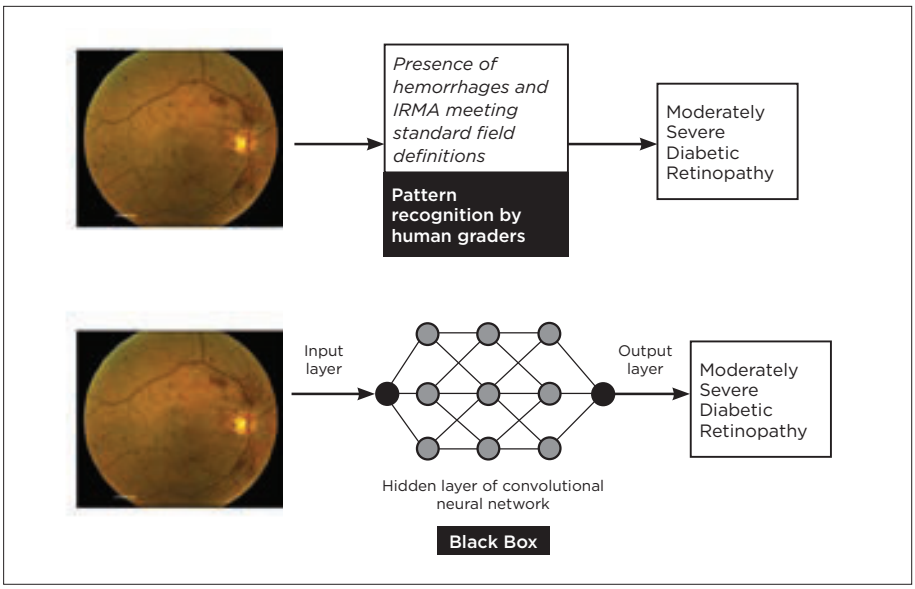

Figure 1. A simplified model showing the approach of human graders in identifying features corresponding to diabetic retinopathy classification (top). A deep-learning neural network arrives at the same classification using an unknown process (bottom).

Around the same time, Google DeepMind published one of the first validated algorithms toward diabetic retinopathy screening using fundus photographs.[2,3] More recently, IDx, LLC has developed the first FDA-approved algorithm to screen for diabetic retinopathy and diabetic macular edema.[4] In addition to diabetic retinopathy screening, deep-learning algorithms have been created to use fundus photographs to identify and screen for glaucoma and age-related macular degeneration.[5,6]

Algorithms are being designed to identify biomarkers on optical coherence tomography (OCT), such as the presence of fluid, which may predict disease progression.[7,8] Retinal imaging is now a hub for various AI companies such as Visulytix, Google DeepMind, and Eyenuk, Inc. We are not too far from the day when smartphone retinal imaging and AI diagnosis will be possible.

To date, deep-learning algorithm development has focused on screening programs. These algorithms cannot be grandfathered for use in clinical trials. Use of AI in clinical trials requires specific validation involving the current gold standard for image interpretation in clinical trials, ie, masked independent graders.

The imaging protocols used in clinical trials are complex, such as 7-field color fundus photography and montaged ultra-widefield imaging. The grading scales are elaborate and require detailed lesion evaluation. Most deep-learning algorithms currently available aim at disease classification. In contrast, reading center evaluation utilizes detailed feature detection in addition to disease subtyping.

Artificial intelligence and reading centers

The role of reading centers in clinical trials

Clinical trial standardization and rigor have come a long way since the seminal trials in ophthalmology more than 4 decades ago with the Diabetic Retinopathy Study.[9] Since then, ophthalmology has witnessed a number of landmark clinical trials supported by the National Eye Institute and the pharmaceutical industry. Endpoints for ophthalmic clinical trials continue to be developed and include visual as well as anatomic outcomes.[10,11]

‘We are not too far from the day when smartphone retinal imaging and AI diagnosis will be possible.’

Clinical trials for retinal diseases have a wide selection of imaging modalities for evaluating anatomical endpoints, which are typically assessed at central reading centers. The rise in digital imaging techniques such as ultra-widefield fundus photography, OCT, fluorescein angiography (FA), and fundus autofluorescence, among many others, has increased the reliance on reading centers. Centralized image interpretation by uniformly trained masked readers or graders, with an

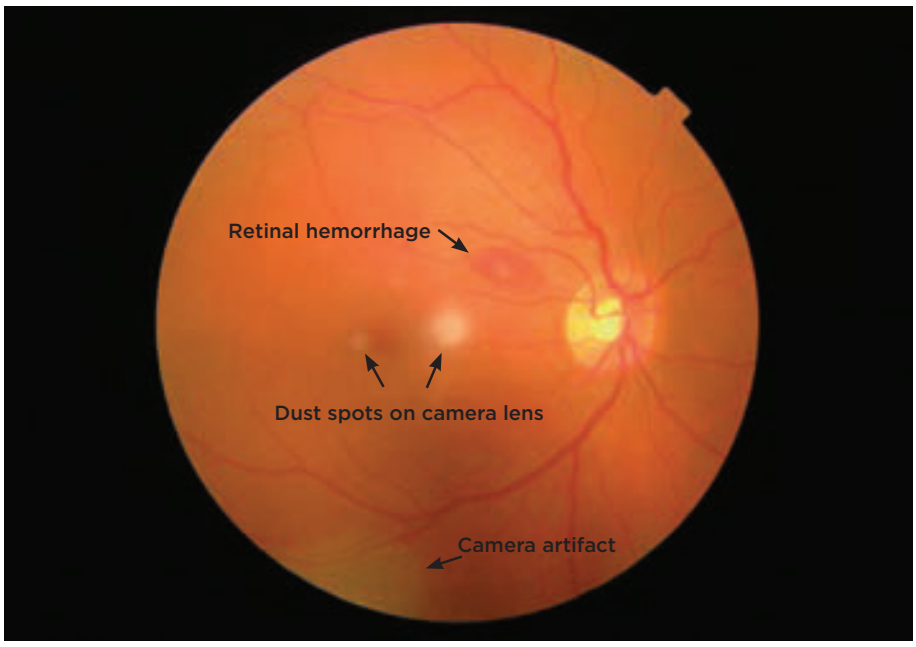

Figure 2. Annotated fundus photograph to train machine learning algorithms to distinguish artifacts from pathology.

emphasis on image quality, ensures consistency in clinical trial data.[12]

While image interpretation is the key function performed at reading centers, multiple associated tasks are incorporated into the workflow that facilitate the success of clinical trials. Reading center participation in clinical trials starts with certification of imaging technicians and equipment to standardize imaging protocols. All study image submissions are monitored for image quality and feedback is provided to sites. Other methods of quality control monitoring include grader reproducibility and data queries to review study outcomes. The typical workflow of a reading center is presented in Figure 3.

Neural networks as graders

Unlike radiology, where images are interpreted by diagnostic radiologists, most ophthalmology reading centers employ non-ophthalmologist graders to evaluate images. The grader certification program is much like that of the development of deep-learning algorithms, with training, tuning, and testing components, although thousands of images are not required.

> ‘Centralized image interpretation by uniformly trained masked readers or graders, with an emphasis on image quality, ensures consistency in clinical trial data.’

The grader training program involves basics of ocular anatomy and diseases. Graders are typically trained in one imaging modality and a particular disease type, such as color fundus photography for diabetic retinopathy or OCT for macular edema. Certification tests are required before graders evaluate clinical trial images. Over time, an individual grader’s training extends across multiple imaging types and ocular pathologies. Performance is monitored by rigorous quality control with periodic assessments of inter- and intra-grader agreement.

Reading centers, as the hub of image analysis, are the perfect arena for deploying AI. There are many advantages to incorporating AI into reading center workflow, with speed topping the list. Programs such as AlphaGo have demonstrated the extraordinary speed of AI.

Human resources are expensive and come with the risk of staff turnover. Quantification of lesions will be more accurate and reproducible with automated algorithms. The work environment for human graders can be tedious and intense, involving hours at a computer screen with frequent mouse clicking for deploying software tools.

The FDA and reading centers recognize grader fatigue and its impact on potential errors in image interpretation. On the other hand, grading with deep-learning algorithms is fast, less expensive, and more reproducible.

Why do we not yet have a reading center model with automated image evaluation?

1. **Training datasets are needed.** Developing reproducible deep-learning algorithms is not easy. Training deep-learning algorithms requires many thousands of images captured with a variety of cameras in a diverse population with widely distributed pathology.[13] The first step in creating a deep-learning algorithm involves annotation or labeling of images by multiple graders to train the algorithm, eg, presence vs absence of diabetic retinopathy.

   High-quality ground truth—reference standards generated by humans—is required to train the algorithm to avoid poor classification performance. Google DeepMind used 1,662,646 images in its training dataset, with 54 US-licensed ophthalmologists or trainees providing the diabetic retinopathy classification.[2]

> ‘Reading centers, as the hub of image analysis, are the perfect arena for deploying AI.’

   Training is followed by fine tuning of the algorithm, where errors are identified and corrected. Finally, the algorithm is validated using a masked, labeled image set, and accuracy is measured. During this step, unlabeled images are presented to the

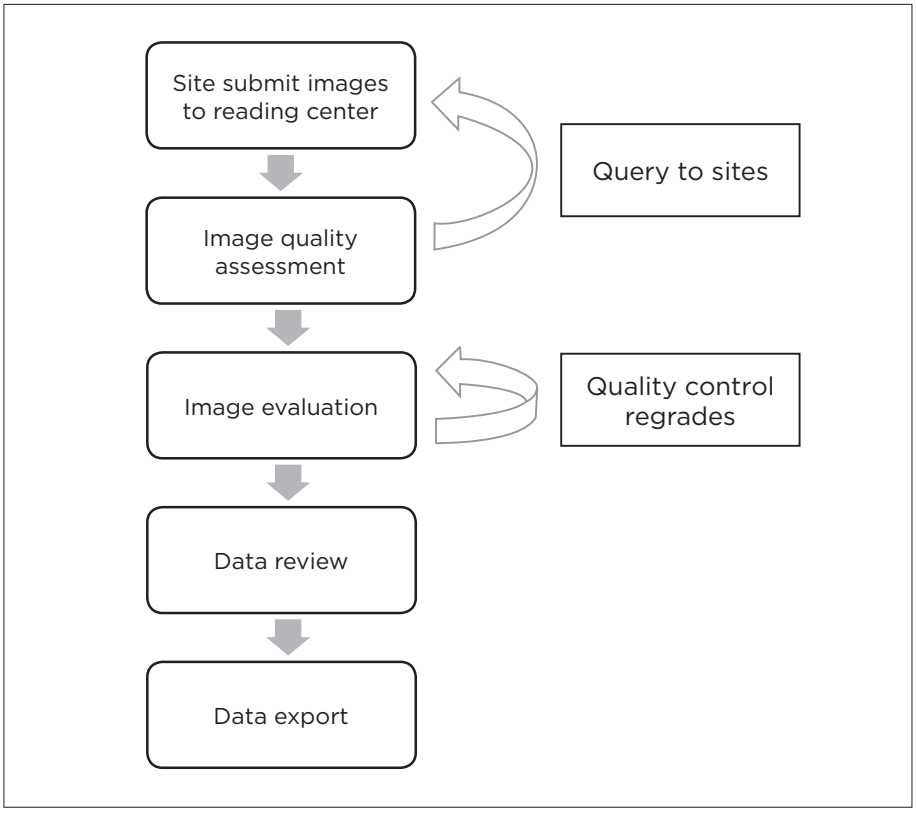

Figure 3. Flowchart representing traditional reading center workflow.

algorithm, and the AI-generated response is compared with the human's reference standard—a man vs machine comparison.

It is important to have independent training, tuning, and testing-image sets for an accurately functioning algorithm. Developing neural networks that can mimic the training of graders will require thousands of annotated images with detailed lesion identification.

2. **Also needed: AI quality-control guidelines.** Continuous quality control will be required in deep-learning environments. These guidelines, similar to calibration of equipment in any lab, must be established for reproducibility testing of deep-learning diagnostics. Repeated reading of images on an ongoing basis will be needed to test contemporaneous reproducibility, both within the algorithm and compared against humans.

   Even experienced human graders have been known to drift from their training as their understanding of the disease and patterns increases. Self-learning algorithms could experience the same phenomenon.

   Scheduled checks against a set reference standard will be required to keep tabs on the development of temporal drift. While algorithms can be monitored similarly to grader reproducibility, corrective actions may be more complicated. If a human grader drifts in his or her reproducibility, others can step in and take over the workload while retraining is performed. What if temporal drift occurs or reproducibility is affected in an automated algorithm?

3. **Rare diseases = too-small datasets.** We are in a great era of evolving therapies for inherited retinal degenerations. Unlike pathologies such as diabetic retinopathy and macular degeneration, disease quantification and imaging outcomes are less defined in rare diseases. Developing deep-learning algorithms requires big datasets and thus may not be a practical approach for evaluating outcomes in rare diseases.

4. **Bias and the "black box" pose risks.** The use of an unknown method for arriving at clinical trial outcomes in an FDA-regulated environment can be unsettling to many industry partners. It is known that training datasets can introduce inherent biases into deep-learning algorithms. The most commonly cited example is the gender bias in Google image searches; a search for a chief executive officer (CEO) shows mostly men.[14]

   Machines do not have biases, but inherit them from humans. However, an issue exclusive to machines is adversarial examples. Abramoff et al have shown adversarial examples in fundus photographs where a small amount of carefully constructed noise is added to an image, resulting in the CNN misclassifying the fundus image, despite the image looking the same to the human eye.[15]

   One developer in the Kaggle competition for diabetic retinopathy algorithms found that the AI performed more accurately when both eyes were presented, even though the diagnosis was required of one eye only. On further investigation, the CNN was found to be utilizing information from both eyes to identify camera artifacts, such as cotton-wool spots. The AI had learned to identify artifacts on its own.

'Developing neural networks that can mimic the training of graders will require thousands of annotated images with detailed lesion identification.'

'Machines do not have biases, but inherit them from humans.'

   Reading centers are typically masked to treatment to avoid bias in interpretation. What if the self-learning black box can unmask treatment randomizations based on images? While this may sound like an extreme example, the full potential of deep learning and the issues surrounding it will evolve as our understanding of neural networks, and the black box improves.

5. **Technological advances require testing and protocols.** There has been much progress in ocular imaging over the last few decades; advanced imaging techniques such as OCT angiography and ultra-widefield imaging are now the standard of care. With cellular-level imaging as the focus of next-generation imaging machines, we have a great future in retinal diagnostics.

   The initial decoding of the images and their translation for clinical use will require research and development. There will be a need for human graders to test analysis tools and develop evaluation protocols before deploying algorithms into unknown territory.

'What if the self-learning black box can unmask treatment randomizations based on images?'

6. **AI must be validated for clinical trials.** The explosion of deep learning in our everyday lives is partly due to the progress in smartphone operating systems. These smartphone-enabled neural engines make access to diagnostic apps available at our fingertips. While the future with AI seems exciting, caution is needed. Many low-quality screening apps with false scientific claims are already available. A recent example is the crackdown by the Federal Trade Commission on melanoma detection apps that claimed high screening rates without scientific evidence.[16]

   AI for medical imaging is considered "Software as a Medical Device" (SaMD) by the FDA and thus requires validation. Once approved, the deep-learning device can be used only under the specific label. AI developed for screening cannot be used for clinical trial outcome evaluation.

   Crowdsourcing of images has resulted in a flood of diabetic retinopathy screening algorithms in recent years, and the choices can be confusing. There are also challenges with regulating a self-improving algorithm; as the neural network improves itself, the version validated by the FDA may be different a year later. The FDA has started developing guidelines and is actively looking into validation of CNN under its Digital Health Program.

Integrated reading center model

The preceding scenarios provide sufficient evidence that reading centers, in the near

term, cannot be fully automated. Instead of computerizing image evaluation, an integrated approach where the advantages of neural networks and human skills can be combined may provide an ideal environment.

Rather than artificial intelligence, the term gaining popularity is *augmented intelligence*, in which machines increase the efficiency of humans. Generating heat maps that depict the change in images over time helps the grader focus on areas highlighted and increases the speed of grading.

AI can be effectively used in other areas of reading center workflow where rapid turnaround is required. Eligibility determination by reading centers is a requirement for multicenter clinical trials to standardize the interpretation of inclusion and exclusion features between various ophthalmologists in global trials.

Incorporating AI at clinics for real-time eligibility can be a very useful and patient-friendly application. Image quality determination to request retakes is another function that can be served well by neural networks, especially if incorporated in the camera. Figure 4 is an example of a reading center workflow with augmented grading where deep-learning algorithms assist humans in documenting clinical trial outcomes.

It is fortunate for our field that among all the parts of a human body, AI has set its sights on the eye. We have many challenges to overcome, but the future of ophthalmology and ophthalmic research is going to be fundamentally transformed by AI. Augmented grading with algorithms assisting humans can improve efficiency, accuracy, and reproducibility of clinical trial data, serving our patients in the best possible way. It is important that AI developers, reading centers, clinical trialists, and the FDA collaborate closely in developing and validating deep-learning technologies for use in clinical trials.

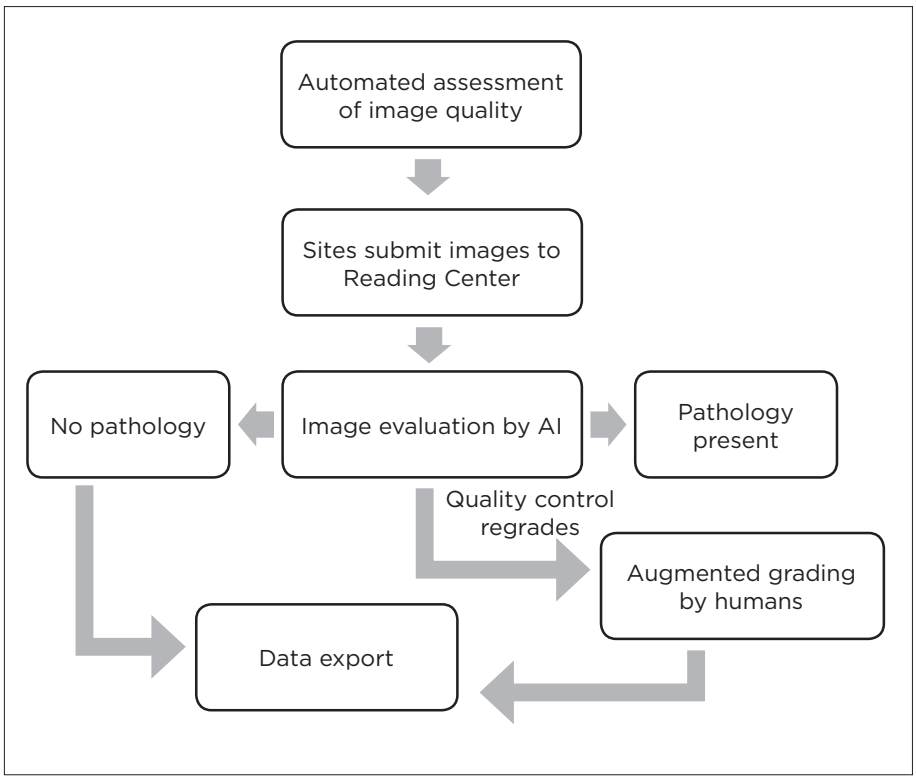

Figure 4. Possible modification to reading center workflow by integrating artificial intelligence.

References

1. Diabetic retinopathy detection: identify signs of diabetic retinopathy in eye images. Kaggle, Inc website. https://www.kaggle.com/c/diabetic-retinopathy-detection. Accessed May 23, 2018.
2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216
3. Abramoff MD, Lou Y, Erginay A, et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci.* 2016;57(13):5200-5206. doi:10.1167/iovs.16-19964
4. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [news release]. Silver Spring, MD: US Food and Drug Administration; April 11, 2018. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm. Accessed May 23, 2018.
4. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318(22):2211-2223. doi:10.1001/jama.2017.18152
5. 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(11):1170-1176. doi:10.1001/jamaophthalmol.2017.3782
6. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning [published online December 8, 2017]. *Ophthalmol.* 2018;125(4):549-558. doi:10.1016/j.ophtha.2017.10.031
8. Bogunovic H, Montuoro A, Baratsits M, et al. Machine learning of the progression of intermediate age-related macular degeneration based on OCT imaging. *Invest Ophthalmol Vis Sci.* 2017;58(6):bio141-bio150. doi:10.1167/iovs.17-21789
9. Preliminary report on effects of photocoagulation therapy. The Diabetic Retinopathy Study Research Group. *Am J Ophthalmol.* 1976;81(4):383-396. doi:10.1016/0002-9394(76)90292-0
10. Csaky K, Ferris F 3rd, Chew EY, Nair P, Cheetham JK, Duncan JL. Report from the NEI/FDA endpoints workshop on age-related macular degeneration and inherited retinal diseases. *Invest Ophthalmol Vis Sci.* 2017;58(9):3456-3463. doi:10.1167/iovs.17-22339
11. Nair P, Aiello LP, Gardner TW, Jampol LM, Ferris FL III. Report from the NEI/FDA diabetic retinopathy clinical trial design and endpoints workshop. *Invest Ophthalmol Vis Sci.* 2016;57(13):5127-5142. doi:10.1167/iovs.16-20356
12. Danis RP. The clinical site-reading center partnership in clinical trials. *Am J Ophthalmol.* 2009;148(6):815-817. doi:10.1016/j.ajo.2009.07.017
13. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy [published online March 2, 2018]. *Ophthalmol.* doi:10.1016/j.ophtha.2018.01.034
14. Kay M, Matuszek C, Munson SA. Unequal representation and gender stereotypes in image search results for occupations. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. April 18-23, 2015; Seoul, Republic of Korea. doi:10.1145/2702123.2702520
15. Roach L. Artifical intelligence: the next step in diagnostics. *EyeNet.* 2017;21(11):76-83.
16. "Melanoma detection" app sellers barred from making deceptive health claims [news release]. Washington, DC: Federal Trade Commission; August 13, 2015. https://www.ftc.gov/news-events/press-releases/2015/08/melanoma-detection-app-sellers-barred-making-deceptive-health. Accessed May 23, 2018.

Financial Disclosures

Dr. Blodi – None.

Dr. Domalpally – None.

Dr. Fekrat – ALCON LABORATORIES, INC: Other, Royalty; REGENERON PHARMACEUTICALS, INC: Consultant, Honoraria.

Dr. Ip – ALLEGRO OPHTHALMICS, LLC: Consultant, Honoraria; ALLERGAN, INC: Consultant, Honoraria; BOEHRINGER INGELHEIM: Consultant, Honoraria; GENENTECH, INC: Consultant, Honoraria; QUARK PHARMACEUTICALS, INC: Consultant, Honoraria; OMEROS CORPORATION: Consultant, Honoraria; THROMBOGENICS, INC: Consultant, Honoraria.

Mr. Nittala – None.

Dr. Sadda – ALLERGAN, INC: Consultant, Investigator, Grants, Honoraria; CARL ZEISS MEDITEC: Consultant, Investigator, Equipment (Department or Practice), Grants; CENTERVUE INC: Consultant, Other, Equipment (Department or Practice), Honoraria; GENENTECH, INC: Consultant, Investigator, Grants, Honoraria; HEIDELBERG ENGINEERING: Consultant, Other, Equipment (Department or Practice), Honoraria; ICONIC THERAPEUTICS, INC: Advisory Board, Honoraria; NIDEK CO, LTD: Other, Equipment (Department or Practice); NOVARTIS PHARMACEUTICALS CORPORATION: Consultant, Honoraria; OPTOS PLC: Consultant, Investigator, Equipment (Department or Practice), Grants, Honoraria; THROMBOGENICS, INC: Consultant, Honoraria; TOPCON MEDICAL SYSTEMS: Other, Equipment (Department or Practice), Honoraria.