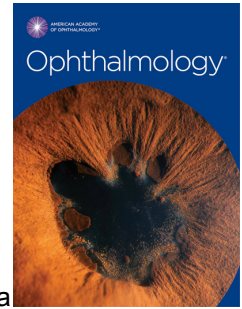


Journal Pre-proof



From data to deployment: the Collaborative Communities on Ophthalmic Imaging roadmap for artificial intelligence in age-related macular degeneration.

Eliot R. Dow, MD, PhD, Tiarnan D.L. Keenan, BM BCh, PhD, Eleonora M. Lad, MD, PhD, Aaron Y. Lee, MD, MSc, Cecilia S. Lee, MD, MS, Anat Lowenstein, MD, Malvina B. Eydelman, MD, Emily Y. Chew, MD, Pearse A. Keane, MD, FRCOphth., Jennifer I. Lim, MD, for the Collaborative Community for Ophthalmic Imaging executive committee and the working group for artificial intelligence in age-related macular degeneration

PII: S0161-6420(22)00002-1

DOI: <https://doi.org/10.1016/j.ophtha.2022.01.002>

Reference: OPTHHA 11941

To appear in: *Ophthalmology*

Received Date: 17 September 2021

Revised Date: 16 December 2021

Accepted Date: 4 January 2022

Please cite this article as: Dow ER, Keenan TDL, Lad EM, Lee AY, Lee CS, Lowenstein A, Eydelman MB, Chew EY, Keane PA, Lim JI, for the Collaborative Community for Ophthalmic Imaging executive committee and the working group for artificial intelligence in age-related macular degeneration, From data to deployment: the Collaborative Communities on Ophthalmic Imaging roadmap for artificial intelligence in age-related macular degeneration., *Ophthalmology* (2022), doi: <https://doi.org/10.1016/j.ophtha.2022.01.002>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology

From data to deployment: the Collaborative Communities on Ophthalmic Imaging roadmap for artificial intelligence in age-related macular degeneration.

Eliot R. Dow MD, PhD¹, Tiarnan D. L. Keenan BM BCh, PhD², Eleonora M. Lad MD, PhD³, Aaron Y. Lee MD, MSc⁴, Cecilia S. Lee MD, MS⁴, Anat Lowenstein MD⁵, Malvina B. Eydelman, MD⁶, Emily Y. Chew MD^{2*}, Pearse A. Keane MD, FRCOphth.^{7*}, and Jennifer I. Lim MD^{8*} for the Collaborative Community for Ophthalmic Imaging executive committee and the working group for artificial intelligence in age-related macular degeneration.⁹

Author affiliations:

1 Byers Eye Institute of Stanford University, Palo Alto, California, USA

2 Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, Maryland, USA

3 Department of Ophthalmology, Duke University Medical Center, Durham, North Carolina, USA

4 Department of Ophthalmology, University of Washington, Seattle, Washington, USA

5 Division of Ophthalmology, Tel Aviv Medical Center, Tel Aviv, Israel

6 Office of Health Technology 1, Center of Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA

7 NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK

8 Department of Ophthalmology, University of Illinois at Chicago, Chicago, Illinois, USA

9 Members are listed in Appendix A

**Corresponding authors:*

Pearse A. Keane, MD, Medical Retina Service, Moorfields Eye Hospital NHS Foundation Trust, London, UK
pearse.keane1@nhs.net

Jennifer I. Lim, MD, Department of Ophthalmology, U. of Illinois College of Medicine, 1855 W Taylor St # 210, Chicago, IL 60612, jennylim@uic.edu

Emily Y. Chew, MD, NIH, Building 10, CRC, Room 3-2531, 10 Center Drive, MSC 1204, Bethesda, MD 20892-1204
echew@nei.nih.gov

Conflicts of interest:

A.L. - Consultant to Allergan, Bayer, Beyeonics, Forsightlabs, Notal Vision, Novartis, Roche

C.S.L. - no financial disclosures

A.Y.L. - Consultant to Genentech/Roche, Johnson and Johnson and Gyroscope. Speaking fees from Topcon, employed by the US FDA, Grants from Regeneron, Carl Zeiss Meditec, Microsoft

E.Y.C.- No financial disclosures. Co-inventor on a patent application (Methods and Systems for Predicting Rates of Progression of Age-related Macular Degeneration)

E.R.D. - no financial disclosures

M.B.E.- no financial disclosures

P.A.K. - Consultant to DeepMind, Roche, Novartis, Apellis, and BitFount. Equity owner in Big Picture Medical. Speaking fees from Heidelberg Engineering, Topcon, Allergan, and Bayer.

T.D.L.K. - no financial disclosures; co-inventor on a patent application (Methods and Systems for Predicting Rates of Progression of Age-related Macular Degeneration)

E.M.L.- Consultant to Roche, Novartis, Apellis, Allegro, Annexon Bio, Retrotape, Gemini Therapeutics, IvericBio. Equity owner in Retrotape. Funding through Duke University from Apellis, LumiThera, Novartis, Roche, NGM Bio. Co-inventor on a patent application (A System and Method to Predict Progression of Age-Related Macular Degeneration)

J.I.L. - Consultant to Aura, Cognition, Eyenuk, Genentech/Roche, Iveric, Luxa, Novartis, Opthea, Regeneron, Santen, Quark; Clinical Studies Grants to University of Illinois from Aldeyra, Chengdu, Genentech/Roche, Greybug, NGM, Regeneron, Stealth

FDA participates in the AMD Ocular Imaging Working Group as a member of the Collaborative Community on Ophthalmic Imaging. This manuscript reflects the views of the authors and should not be construed to represent FDA's views or policies

Financial support:

C.S.L - NIH/NIA R01AG060942

A.Y.L - NIH/NEI K23EY029246

E.R.D. - None

P.A.K. - Moorfields Eye Charity Career Development Award (R190028A), UK Research & Innovation Future Leaders Fellowship (MR/T019050/1).

M.B.E.- None

E.Y.C.-National Eye Institute, National Institutes of Health Intramural program funding (employment)

T.D.L.K. - National Eye Institute, National Institutes of Health (employment)

E.M.L.- Veterans Affairs Merit Award I01 CX002116

J.I.L. - University of Illinois Core Grant Ey01792 and an unrestricted grant from Research to Prevent Blindness.

Abbreviations:

AI - artificial intelligence

AMD - age-related macular degeneration

AREDS - age-related eye disease study

AUROC - area under the receiver-operating characteristic curve

CCOI - Collaborative Community on Ophthalmic Imaging

CFP - color fundus photography
CNV - choroidal neovascularization
DL - deep learning
FA - fluorescein angiography
FDA - United States Food and Drug Administration
GA - geographic atrophy
ML - machine learning
nAMD - neovascular (or wet) age-related macular degeneration
OCT - optical coherence tomography
ROC - receiver-operating characteristic curve
SaMD - software as a medical device
UK NHS - United Kingdom National Health Service
VA - Veterans Affairs

Running head: Artificial intelligence in age-related macular degeneration

Keywords:

artificial intelligence, machine learning, deep learning, age-related macular degeneration, optical coherence tomography, color fundus photography

Word Count: Abstract: 337 words, Text: 10,258 words

Abstract

IMPORTANCE: Healthcare systems worldwide are challenged to provide adequate care for the 200 million individuals with age-related macular degeneration (AMD). Artificial intelligence (AI) has the potential to make a significant positive impact on the diagnosis and management of patients with AMD. However, the development of effective AI devices for clinical care faces numerous considerations and challenges, a fact evidenced by a current absence of FDA-approved AI devices for AMD.

OBJECTIVES: To delineate the state of AI for AMD including current data, standards, achievements, and challenges.

EVIDENCE Members of the Collaborative Community on Ophthalmic Imaging working group for AI in AMD attended an inaugural meeting on September 7, 2020 to discuss the topic. Subsequently, they undertook a comprehensive review of the medical literature relevant to the topic. Members engaged in meetings and discussion through December 2021 to synthesize the information and arrive at consensus.

FINDINGS: Existing infrastructure for robust AI development for AMD includes several large, labeled datasets of color fundus photography (CFP) and optical coherence tomography (OCT) images. However, image data often does not contain meta-data necessary for the development of reliable, valid, and generalizable models. Data sharing for AMD model development is made difficult by restrictions on data privacy and security, although potential solutions are under investigation. Computing resources may be adequate for current applications, but knowledge of machine learning (ML) development may be scarce in many clinical ophthalmology settings. Despite these challenges, researchers have produced promising AI models for AMD for screening, diagnosis, prediction, and monitoring. Future goals include defining benchmarks to facilitate regulatory authorization and subsequent real-world generalization.

CONCLUSIONS AND RELEVANCE: Delivering an FDA-authorized, AI-based device for clinical care in AMD involves numerous considerations including the identification of an appropriate clinical application, acquisition and curation of a large, high-quality data set, development of the AI architecture, training and validation of the model, and functional interactions between the model output and clinical end-user. The research efforts undertaken to date represent starting points for the medical devices that will eventually benefit providers, healthcare systems, and patients.

Age-related macular degeneration (AMD) is the leading cause of legal blindness in developed countries. As the average age of the global population rises, AMD is an increasingly urgent matter of public health. However, the specialized training necessary for the diagnosis and management of AMD challenges healthcare systems to deliver appropriate care to the 200 million affected individuals worldwide ¹.

In the past decade, artificial intelligence (AI), particularly the subset of machine learning (ML) known as deep learning (DL), has made significant advances that have the potential to transform clinical care for AMD. Nevertheless, no AI-based medical device has yet been authorized for market distribution in the United States for clinical management of AMD. Delivering a production-level AI-based software as a medical device (SaMD) for clinical care involves numerous considerations including the identification of an appropriate clinical application, acquisition and curation of a large, high-quality data set, development of the AI architecture, training and validation of the model, and useful interactions between the model output and clinical end-user.

Recognizing the dearth of clinically impactful AI for AMD, the Collaborative Community on Ophthalmic Imaging (CCOI) convened a working group to address the challenges facing the field. The CCOI is comprised of diverse stakeholders including clinicians, patients, researchers, medical device manufacturers, not-for-profit organizations, and government entities such as the US Food and Drug Administration (FDA) and the National Eye Institute. Through ongoing dialog, the CCOI seeks to identify obstacles, best practices, strategies, and standards to accelerate innovation in ophthalmic imaging that improves patient outcomes. This manuscript presents the perspective of the CCOI working group on AI in AMD on how to navigate the path from patient data to an authorized, AI-based SaMD.

Infrastructure

Effective AI is built on a foundation of robust, reliable data arising from appropriate infrastructure. Large ophthalmic image data sets exist for the development of AI, but each has shortcomings. An ideal data set would be large, diverse, representative, and well labeled. Efforts to achieve this through data-sharing are challenged by issues of data privacy, but several solutions may mitigate these risks.

Data

Current Practices

To date, the majority of AI research in AMD has employed supervised learning, thus requiring large sets of labeled data. Because AMD is a chorioretinal disease diagnosed by visible signs of pathology, research has predominantly involved the two most prevalent modalities of posterior segment imaging, color fundus photography (CFP) and optical coherence tomography (OCT) (Figure 1). CFP involves the acquisition of a two-dimensional image of the fundus through the pupil by analog or digital photography. The longstanding use of CFP and its ability to capture numerous anatomical structures within the posterior segment in a representation similar to a clinical fundoscopic exam has resulted in its basis for many clinical standards in ophthalmology. Consequently, CFP is widely used within eye care clinics, as well as some general medical practices and emergency rooms. The large number of CFP images and devices make it an attractive modality for ML-based screening.

For example, one widely used data set in ML is derived from the Age-Related Eye Disease Study (AREDS), which includes more than 120,000 cataract and fundus images from 4,757 participants from 11 clinical centers collected

annually over 10 years from 1992 to 2005². Although the AREDS image data represents a large, well-labeled training set, it poses several issues in producing models that will generalize well to real-world clinical settings. First, the data were collected 10-30 years ago using 30-degree fundus cameras with Kodachrome or similar film, rather than the digital fundus photography equipment currently in use. Second, the data were acquired in a clinical trial setting by highly trained personnel working under a uniform protocol, which may differ from the conditions in which data are acquired during routine eye care in the community. Additionally, anti-VEGF therapy was not available at the time of the study, resulting in differences in disease phenotypes. Finally, the ethnic and socioeconomic makeup of the patient population enrolled in the trial does not fully represent the demographic makeup of the United States or global populations.

The images from the subsequent AREDS2 are currently submitted to dbGaP (database for Genotype and Phenotype) for public access³. AREDS2 enrolled 4203 participants who were followed from 2006 to 2012 for a randomized trial of oral supplements. These participants had annual digitized color fundus photographs while a subset had ocular images at year 10. A subset of AREDS2 participants also had fundus autofluorescence images and ultra-widefield fundus photographs as well as spectral domain OCT in a longitudinal fashion during the clinical trial. The AREDS2 images carry the same limitations as those of the AREDS images except that they are all original digital images taken during the period in which anti-VEGF therapies were approved as the standard of care.

At more than 30 million imaging procedures per year in the US alone, OCT has also found widespread use in ophthalmology offices and rapid access clinics⁴. This non-invasive imaging method uses near-infrared light to acquire a series of two-dimensional cross-sectional images across several millimeters of the posterior pole. The micrometer-resolution of the modality captures the major pathological characteristics of AMD including the retinal fluid that is characteristic of neovascular AMD (nAMD). Consequently, OCT has also been instrumental in several major clinical trials investigating anti-VEGF agents including HARBOR, which collected more than 50,000 OCTs from its participants⁵. Notably, Kermany and colleagues have assembled a database of 109,000 OCT scans from eye clinics in the US and China that are available for public use⁶.

Obstacles

A recent effort by Khan and colleagues to exhaustively identify publicly available ophthalmic imaging data sets cataloged 18 data sets involving AMD patients including 11 CFP data sets with a median of 601 images and 7 OCT data sets with a median of 3,231 images⁷. However, among these data sets, as well as 79 more representing ophthalmic diseases other than AMD, there was a concerning dearth of metadata. Although every data set reported the imaging modality, total number of images, and image format, and a majority reported country of origin and acquisition device, fewer than 20% reported on demographic data including age, sex, or ethnicity. Likewise, only 15% described the inclusion and exclusion criteria for the data. The absence of these metadata increases the risk of poor model generalizability due to spectrum bias, differences such as disease prevalence between the population in which a model was developed and the one to which it is applied.

Finally, although nearly two thirds of the data sets included diagnostic or feature labels, few specified the origin of the labels. Among those that did report the source of the labels, the labelers included ophthalmologists, researchers, optometrists, medical students, general medical doctors, and reading-center technicians. The process for consensus and adjudication of labeling decisions was rarely disclosed. Since ground truth is of great importance to the accurate training and testing of AI models, details about labeling are essential. Image data is often insufficient for ML applications in the absence of image metadata, which as yet, most public data sets lack.

Strategies

The Digital Imaging and Communications in Medicine (DICOM) standard is a set of globally adopted specifications for how image data and metadata are formatted and exchanged. In radiology, radiation oncology, and cardiology imaging device manufacturers have a high degree of compliance with the DICOM standard. In addition to improving patient access to data and interoperability for efficient continuity of care, the uniform data structure has supported the rapid development of ML applications in these fields. DICOM has established specifications for ophthalmic imaging devices, however, manufacturer compliance with the standards is low in ophthalmology. Thus, different imaging devices will produce different digital representations of the same fundus⁸. Recently, the American Academy of Ophthalmology along with the National Eye Institute called for ophthalmic device makers to standardize their data formatting⁹. The adoption of these standards would ensure that pixel data, acquisition information, patient information, demographics, and other essential data would be uniformly formatted for the benefit of ML and other data analysis applications.

In addition to data standardization, it is also essential to carefully consider exclusion criteria of training and testing data. AI models must be robust to the degree of quality and variation that they will encounter after real-world deployment. Research endeavors that exclude a significant proportion of images due to quality or employ a data set of only a few pathological classes without comorbidities are unlikely to perform adequately in a real-world setting. Authors of the CONSORT-AI extension guidelines recommend that clinical trial investigators offer a rigorous explanation for the exclusion of participants and data^{10, 11}.

Data sharing and data security

Current Practices

AI development efforts for AMD have employed data sets ranging from hundreds to hundreds of thousands of images. These figures are modest in comparison to the data used for general image analysis applications, which often extend into the tens or hundreds of millions of images. Although few ophthalmic researchers are able to obtain such numbers of images within their individual institutions, collaboration across healthcare systems has the potential to achieve data on this scale. Each year, tens of millions of OCT and CFP images are acquired and labeled yet remain siloed in individual healthcare systems. Often, the data are made nearly inaccessible because of administrative and paperwork barriers, even to researchers affiliated with the healthcare system, let alone to those outside it.

Some efforts at data sharing are under way. The UK National Health Service, which captures 25 million ophthalmic images each year, has created the INSIGHT Health Data Research Hub for Eye Health to facilitate data-sharing with research partners¹². A key aspect of the organization is the involvement of the public, patients, clinicians, and other stakeholders in a data trust advisory board to determine by whom and how the data may be used. The board determines the purpose, value, and public benefit of data sharing and considers the 'five safes framework' of data storage and access. The advisory board will publish as much information as possible including the intended aims of the partners, except where there are competitive or other sensitive interests, for public accountability.

When multiple healthcare systems are each contributing data, the partners must verify that the clinical and data-collection protocols are uniform. This includes having agreed clinical definitions for diagnoses, disease stages, and pathologic features, standardized clinical metadata, and consistent procedures for labeling. Image reading centers

or a centralized coordinating center similar to those found in multi-center clinical trials may be useful to maintaining uniform collection and treatment of the data.

Obstacles

Nevertheless, data across institutions may widely vary. One major challenge to uniform data is the pervasive unreliability of electronic health record documentation. A recent study by Henriksen found that 45-75% of all text in a patient note is copied into the document¹³. Documentation styles and patterns may differ between individual providers and across institutions. This may not only make extraction of clinical labels and other metadata challenging by natural language processing of electronic health record, but may even mean that no valid metadata exists for extraction even if reviewed by a human expert. Since the issue arises from the idiosyncratic documentation habits of individuals, a solution would likely require a major behavioral change across providers. Finally, as mentioned previously, in the absence of DICOM compliance different institutions or even different clinics within an institution may not share image standards.

The sharing of data across institutions also brings up issues of data security and ownership. This is made additionally challenging due to evolving definitions of biometric data as protected health information under HIPAA guidelines, recommendations by the National Institute of Standards and Technology, and institutional policies within individual healthcare systems. Moreover, an additional challenge in the US is a state-by-state treatment of health data ownership, including many states without any applicable law.

Strategies

Several possible technological approaches to minimizing data-sharing risk have been proposed. First, institutions could contribute to a 'data mart', a secure, cloud-based server that can be accessed by partners through a virtual private cloud. The user's code base would be transferred to the server and the result of the analysis transferred back, without any distribution of raw data from the data mart. Verana Health has established a similar computing backend for partners accessing deidentified tabular clinical data derived from the IRIS Registry¹⁴.

Two additional options include transfer learning and federated learning. In transfer learning, a ML model is trained with a large data set to establish general network parameters, before the model is transferred to a local machine where the training is refined with data more representative of the anticipated distribution. Similarly, in a federated learning paradigm, an AI model is sequentially trained on several local databases, transferring the model but not the data. Mehta and colleagues, for example, trained a modified U-Net autoencoder to perform semantic segmentation of intraretinal fluid using 1289 OCT scans with sparse manual segmentation. Although the resulting model performed the task with high accuracy on a test set of images acquired at the same institution with the same type of device, it performed poorly on an independent data set from a second university medical center acquired on a different device. However, after the model underwent a brief period of retraining using 400 annotated images from the second clinical site using the compute of a single desktop processor, the accuracy of the model's segmentation output showed no statistically significant difference to the performance of retinal specialists¹⁵.

Finally, a third approach to addressing data security is to share derived data rather than raw data. As one example, a DL network from DeFauw and colleagues generated exhaustive semantic segmentation maps from OCT images that were, in turn, used to train a second DL model to triage AMD. The group found that the performance of the triage model did not improve when trained with the raw image data, compared to the segmentation maps. This

data representation has the added benefit of enhancing a clinician's ability to interpret the model's output¹⁶. Synthetic data produced by generative algorithms may also offer valid, deidentified training data, although its use in regulated devices currently lacks FDA guidance^{17, 18}.

Although health data privacy and security are areas of active discussion, one guiding principle is transparency. Patients should be informed of how their data is being used and by whom. Although obtaining verbal or written consent from every patient in a retrospective database may be infeasible, patients seen prospectively should have a mechanism for opting out of non-clinical use of their data. The UK NHS already makes such a service available to all of its patients¹⁹. Ultimately, though no system can be completely secure, the risk of a data breach must be balanced against the public health cost of failing to utilize available data. Therefore, the paperwork and committee approval required of researchers to access data should be reduced to the minimum needed to provide rapid, ethical decision-making.

Compute

Current Practices

With a data set in hand, one must next consider the compute and human resources needed to turn it into a clinically meaningful model. Although state-of-the-art models outside of healthcare run by commercial enterprises have trained models with up to 1.6 trillion parameters, ML models in AMD have shown sufficient accuracy with a fraction of the complexity²⁰. Many efforts have been undertaken with a scientific computing workstation or even consumer-grade desktop computers. With the availability of cloud-computing services from Amazon, Google, and Microsoft, among others, applying significant compute to ML problems may be within the budget of well-resourced academic departments and research grants.

Obstacles

Although some ophthalmologists have the technical knowledge to develop AI models, computer science and machine learning expertise remains, in general, a scarce resource within ophthalmology. Researchers within academic ophthalmology departments who possess domain knowledge but lack the technical knowledge to develop ML algorithms for scientific research may look to their local computer science department for masters and doctoral students who are interested in joining a project.

Strategies

AutoML tools available from several major technology companies may offer one solution to accelerate the development of clinical AI models. AutoML allows a clinician with a data set to train a DL model without any coding knowledge. Although there is no regulatory precedent for FDA clearance of a device based upon an AutoML-generated model, it nevertheless offers a way for domain experts to rapidly experiment with DL before additional resources are put toward development. Several studies have demonstrated the ability of AutoML platforms to produce accurate models for ophthalmic data^{21 22 23}. There have also been recent calls to provide more physicians with computer programming skills, although, if action is taken, a labor pool will likely not be manifest for several years²⁴.

Best Practices for Data Collection

ImageNet, a set of 14 million photographs labeled across 20,000 nested categories, was critical to the development of commercial applications in computer vision for both training and benchmarking. No similar data set yet exists in ophthalmology, let alone for AMD. What would a data set that catalyzed the development of AI for AMD look like?

The ideal data set for AMD would be large, on the order of ten million images from a million or more patients, and across a year or more of clinical encounters. Although low-shot learning aspires to train ML models with a fraction of the training examples, as yet it has not been demonstrated to yield the accuracy needed for the healthcare setting. Additionally, although few models have incorporated time series data, it is expected that it would improve model accuracy compared to a prediction from a single time point. The data would contain paired right and left eyes, since findings in one eye are often informative of the state of the fellow eye. The data would be acquired across multiple geographic regions with demographically diverse patient populations (age, sex, race, ethnicity, socioeconomic status, etc.) and in multiple practice settings including community hospitals, academic medical centers, private ophthalmology and optometry practices, primary care clinics, and emergency rooms. The data would be well-labeled by a validated process using standard disease definitions. Additional metadata would be uniformly attached to each image including demographic information, clinical data, and device information. In order to analyze response to treatment and other longitudinal trends, the metadata should contain the date of the encounter, treatment information, and a unique patient identifier to track patients across time and to prevent train-test leakage. The data set would be freely accessible, and transparent with clearly enumerated inclusion and exclusion criteria. The size and diverse sources of the data set would minimize the potential for patient reidentification through metadata. Patients would be aware of their data's inclusion in the database and would be able to opt out of usage.

The data would ideally come from OCT, CFP, or both, since the technologies are sufficiently mature as well as prevalent (Figure 1). The OCT data would preferably be three-dimensional, in order to preserve information about neighboring sections and acquired by a standard imaging protocol (6 mm x 6 mm 25-line raster scan). Since the predominant findings of AMD are concentrated in the posterior pole, standard 50-degree CFP would likely be sufficient and thus a data set could be comprised of images from traditional fundus cameras as well as cropped and scaled widefield devices.

Other less common imaging modalities are likely to be useful for certain applications in AMD or to establish a reliable ground truth. Fundus autofluorescence was traditionally used to track the progression of geographic atrophy (GA), though increasingly OCT offers higher resolution of the retinal pigment epithelium (RPE). Fluorescein angiography, ICG angiography, OCT angiography, and near-infrared reflectance all produce complex data that many comprehensive ophthalmologists and optometrists are not well trained to interpret and therefore could be democratized by assistive or autonomous AI. If home-use OCT is authorized by the FDA and becomes widely adopted for AMD monitoring, AI would be invaluable for analyzing the large data volume that it produces [Keenan et al, in review].

Finally, although the foregoing discussion of data describes the ideal data set for model training and testing, further validation would ideally arise from prospective collection and analysis of patient data rather than only retrospective analysis. Like other clinical trials, such studies would benefit from masked randomization and patient enrollment through consortia of multiple hospitals.

Clinical applications

Several key events in the clinical course of AMD would benefit from the application of ML and AI methods, namely screening, diagnosis, prediction, and monitoring (Figure 2). To date, researchers have applied ML to each of these key clinical stages.

Screening

Fundus photography

The first step in the initiation of eye care for AMD is the identification of individuals with the condition. A shortage of ophthalmologists has resulted in many adults receiving routine eye care by non-ophthalmologists, who lack advanced training in the diagnosis and management of AMD. Even among primary eye care providers like optometrists and comprehensive ophthalmologists, up to 25% of eyes with AMD fail to be diagnosed²⁵. AI devices may enhance the capability of primary eye care providers to address this gap in screening.

As described above, the prevalence of CFP, as well as its ability to capture much of the posterior segment anatomy, makes it an excellent modality for community-level screening. In one example of a ML analysis of CFP, Burlina and Bressler used 130,000 CFP from 4613 patients enrolled in AREDS to train a DL algorithm to perform two-class classification of images into no AMD/early AMD, which does not require treatment, and intermediate/late AMD, which should be referred to a specialist for treatment. The model, based on the AlexNet convolutional neural network (CNN), performed with 88.4-91.6% accuracy in the referral decision, when compared with the labels applied by AREDS reading center graders²⁶.

Since AMD has a worldwide prevalence of 2-4% and is often not visually significant until late in the disease course, most patients undergoing AMD screening in the primary eye care setting will be asymptomatic and negative. Thus, an efficient and economically viable screening product should ideally have the capability to detect at least several of the most common disorders of the retina and optic nerve. As one example of achieving this through multiclass output, Ting and colleagues trained a DL model to detect diabetic retinopathy and AMD as well as possible glaucoma from CFP using 494,661 images from the Singapore National Diabetic Retinopathy Screening Program. The model achieved AUROC (area under the receiver-operator characteristic curve) for AMD of 0.931, for DR of 0.936, and for glaucoma risk of 0.942, when compared with experienced reading center graders without medical training²⁷.

Like other medical applications of AI, these multiclass classifiers are best evaluated by a confusion matrix weighted by the cost of a clinical error, which in the case of AMD differs by stage. A false negative result in the detection of early AMD has a low cost to the patient since treatment is not recommended at this stage and the patient would likely be reevaluated before the disease progresses. A missed diagnosis of intermediate AMD would have greater consequence since the initiation of AREDS supplements decreases the likelihood of disease progression. A failure to identify nAMD would be associated with the greatest clinical cost, since the prompt initiation of intravitreal anti-VEGF agents affects visual prognosis very substantially. GA, although often vision threatening, would be associated with a lower cost, since there is no currently approved therapy for the condition. Hence the best general screening programs are those with higher sensitivity and lower specificity as a trade-off to avoid missing the more severe forms of AMD.

In the European Union, RetCAD (Thirona, Nijmegen, Netherlands) is a CE-certified software product for the automated detection of AMD from CFP. The model enlists an ensemble of three CNNs to analyze RGB channels of the photograph, as well as an additional three CNNs to analyze contrast-enhanced images. The device outputs a score from 0-100 that represents the likelihood of referable AMD, with each quartile score corresponding approximately to the 4-step Beckman AMD Severity Scale²⁸ [Thirona whitepaper]. Application of the RetCAD model to two publicly available data sets, DR-AMD derived from multiple European university medical centers and the AREDS data set, yielded an AUROC of 94.9% and 92.7% for the detection of referable AMD²⁹. There is currently no FDA-authorized AI-based SaMD for the screening of AMD.

Optical Coherence Tomography (OCT)

Although OCT of the macula has more limited availability, is more expensive, and captures a smaller area of the posterior segment compared to fundus photography, it depicts the main pathologic lesions of AMD in high-resolution and in three-dimensions. In particular, OCT can identify trace amounts of subretinal and intraretinal fluid indicative of nAMD, which would necessitate more urgent evaluation and treatment by a retinal specialist. Considering that just 12% of US counties have a retinal specialist, a second screening step by optometrists and comprehensive ophthalmologists using OCT-based AI may be needed in order to reduce the number of false positives cases and further stratify the true positive cases generated by fundus photograph screening³⁰.

DeFauw and Ronneberger demonstrated a proof-of-concept OCT-based AI screening system for several macular conditions including AMD, using 14,884 OCT scans collected from 7,621 patients at multiple clinical sites associated with Moorfields Eye Hospital. Using 877 images with sparse manual segmentation of normal and pathologic features, the group trained a U-Net architecture to perform semantic segmentation on 6 normal and 9 pathological OCT features. The model yielded nearly 15,000 segmented OCT studies that then served as inputs to a second DL model, which labeled each segmented OCT as requiring urgent, semi-urgent, or routine referral, or observation. Ground truth labels for the DL model were assigned by eye care providers with retinal training who reviewed patients' clinical data and images. The model achieved a 5.5% error rate with 96-99% AUROC for each pathology, which matched or outperformed each of the 4 retinal specialists and 4 optometrists with medical retina training against whom the model was benchmarked. Moreover, when the error rate was weighted by the clinical costs of an incorrect referral decision, the model outperformed the human evaluators, and it made no severe errors (i.e. recommending observation for a patient requiring urgent referral)¹⁶.

For any AI SaMD, particularly in the primary eye care setting, it is essential to consider the action-outcome pairing. Specifically, a product should be designed with a clear understanding for each of the SaMD's outputs of what action is consequently performed, by whom, and when in order to maximize its utility for the care provider who receives it. Moreover, the significance of information provided by SaMD to the healthcare decisionmaker in combination with the state of healthcare situation or condition determines risk categorization of SaMD³¹. For instance, many providers at the screening level may not know the next appropriate step in clinical management if a model outputs the segmentation of pathologic features in a retinal image or even produces a diagnostic label. By definition, an autonomous diagnostic AI system may evaluate for the presence of a disease or condition and notify the user whether the disease or condition is present without showing how the AI system arrived at the decision. Since community providers may lack the knowledge and training to make an accurate disease assessment, autonomous AI is the preferred system for screening in a non-specialist use-case setting³².

Diagnosis

For patients who have been elevated to the specialist, the diagnosis of AMD should be ascertained, in particular by excluding a variety of clinical entities that mimic AMD or are AMD subtypes that deviate from the typical natural history of the disease. Whereas during screening, community providers may best interface with a fully autonomous decision tool, the specialist may desire a model that is assistive. Utilizing an assistive diagnostic AI model, a clinician receives specific aspects of the inputs that indicate disease specific abnormalities or their absence, and the clinician determines the final diagnosis by their own clinical assessment³².

AI models, including assistive models, that identify known pathologic features of AMD within routine imaging studies could improve the efficiency of image interpretation and aid clinicians in the extraction of insights from the data. To date, numerous studies have been published that apply ML methods to CFP, OCT, OCT-A, indocyanine green (ICG), or FA, to segment pathologic features including drusen, subretinal and intraretinal fluid, reticular pseudodrusen, GA, hyperreflective foci, subretinal hyperreflective material, and pigment epithelial detachment^{33 34 35 36 37, 38}. As a representative example, Moraes and colleagues extended the three-dimensional U-Net architecture for OCT scans created by DeFauw and Ronneberger as described previously to perform voxel-level multiclass semantic segmentation of seven pathologies characteristic of AMD including fibrovascular PED, serous PED, drusen, SRF, IRF, SHRM, HRF, and central subfield thickness^{39 16}.

A clinically useful segmentation model would have the ability to identify at a minimum the lesions necessary for application of the Beckman classification and the AREDS simplified severity scale, both of which are in widespread clinical use to quantify patients' risk of progression to late AMD². AMD staging is also important for clinical trial enrollment since therapeutic approaches differ by stage. Peng and colleagues used 58,402 CFP to train DeepSeeNet, a combination of two CNN sub-networks for the classification of soft drusen (none/small, medium, large) and pigmentary abnormalities, the lesions used for the AREDS simplified scale, as well as a third sub-network to identify late AMD⁴⁰. The model exceeded the accuracy of retinal specialists for classification on the AREDS simplified severity scale and had comparable performance to retinal specialists for the detection of late AMD. Increasingly, additional lesions like reticular pseudodrusen or subretinal drusenoid deposits are also being recognized for their prognostic value including for development of late AMD, particularly GA⁴¹.

AMD Mimics and Subtypes

A variety of clinical entities mimic AMD or represent subtypes of the disease that differ in their natural history and response to treatment. These include polypoidal choroidal vasculopathy (PCV), retinal angiomatous proliferation, adult-onset vitelliform macular dystrophy, central serous chorioretinopathy, pathologic myopia, macular telangiectasia type 2, angiod streaks, pentosan, hydroxychloroquine maculopathy, posterior uveitis, and ABCA4-related retinopathy, pattern dystrophies, and other inherited retinal diseases. Yang and Yu used 475 images from indocyanine green angiography (ICG), the gold-standard test for the diagnosis of PCV, to train a two-step classifier created by Google Cloud AutoML. The model performed with accuracy of 0.83 along with sensitivity of 0.87 and specificity of 0.80⁴². Xu and Chen similarly curated a data set of OCT and CFP images depicting either dry AMD, nAMD, PCV, or a normal fundus. A CNN trained on these four classes had an 87.4% accuracy in labeling subsequent images, a performance that exceeded that of three retinal experts performing the same task⁴³. It is important to note that it is unlikely that all diagnoses on the differential can be made solely by unimodal DL analysis of image data: some AMD mimics are, to current clinical knowledge, phenotypically indistinguishable, and the diagnosis is determined by the clinician with knowledge of the patient's history and observed across time.

In addition to mimics of AMD, unsupervised ML may be useful to further subdivide AMD into novel disease entities. Hosoda and Tsujikawa used unsupervised deep learning to identify patients with a pachychoroid neovasculopathy variant of AMD⁴⁴. Using k-means clustering with 61 diverse feature inputs including multimodal imaging, clinical and demographic data, past medical history, biometry, and environmental factors, the cohort of 537 Japanese patients were assigned to one of two clusters, pachychoroid-variant AMD or non-pachychoroid AMD. The model achieved 89% agreement with the assessment of human experts. Their work additionally suggested that the patients with pachychoroid-variant disease may differ in their early response to anti-VEGF therapy⁴⁵. Similarly, a study by Treder and Eter used a deep CNN to perform two-step binary classification of fundus autofluorescence (FAF) images to identify the rapidly progressive diffuse-trickling subtype of GA. Accuracy of the model was 77% when tested on an independent validation set of 20 images⁴⁶.

Finally, although there has been no published research to date investigating the use of ML-based anomaly detection to flag possible misdiagnosed cases or AMD mimics, such an approach may have the potential to identify atypical individuals who could benefit from greater clinical scrutiny. For instance, an anomaly detection model trained on gold-standard examples of AMD may be able to identify patients with outlying phenotypes. Such an application could inform the clinician about the need to order additional testing like ICG or FA, or could draw attention to salient features of the image that deviate from typical features.

Prediction

An important event in the natural history of AMD is the transition to late disease, which includes the two disease phenotypes of GA and nAMD. To predict the likelihood of progression to late disease, AREDS CFP data was used to produce a 9-stage manual classification of AMD based upon six grades of drusen size and five grades of abnormal pigmentation. Longitudinal outcomes of AREDS patients established a risk of progression to late AMD that ranged from less than 1% in stage 1 to 50% in stage 9⁴⁷. Although the system provided prognostic insight, it was difficult to perform particularly for non-specialists. AI could assist clinicians to carry out this classification system or give rise to a novel prediction model either of which could inform the interval of the patient's follow up examination, the use of ancillary testing, or decisions about treatment.

Burlina and Bressler, for instance, used more than 65,000 CFP from the AREDS trial to train ResNet50 CNNs to label images with either an AREDS 4-step or 9-step severity scale score. For the 4-step classification, the model had a mean estimation error for 5-year risk of 3.5% to 5.3% and a classification accuracy of 75.7%, which exceeded the performance of human graders⁴⁸. AREDS severity scale has clinical acceptance as well as biological plausibility allowing the model to avoid providers' discomfort about working with 'black-box' predictions.

Peng and colleagues built a two-step prediction model to perform survival analysis for time to development of GA or to nAMD annually from one to ten years from the time of analysis. The first step used the previously described DeepSeeNet CNN to classify pathologic features in CFP from AREDS and AREDS2, and then the 16 highest-weight features from DeepSeeNet, along with age and smoking status (with or without genotype information) were used to train a Cox proportional hazards model. The model produced a C-statistic for 5-year prediction of late AMD of 86.4, a figure that exceeded that of retinal specialists performing predictions from commonly used progression-risk calculators. Importantly, the model was trained using time to disease progression on the level of individual patients rather than from 4- or 9-stage patient classes. Additionally, the model showed its robustness by maintaining its predictive accuracy when trained on data from AREDS and then tested on AREDS2 data⁴⁹.

In another example, Banerjee and Rubin trained a recurrent neural network (RNN), a DL architecture suitable for sequential data, with best-corrected visual acuity (BCVA), demographic data, and 21 engineered radiomic features derived from the size, shape, and position of drusen in OCT data from the HARBOR trial. The network predicted conversion to nAMD at 3- and 21-months after an encounter. When applied to an external data set from a university medical center, the model performed with AUROC of 0.82 at 3 months and 0.68 at 21 months⁵⁰.

It is as yet unclear whether the current known single nucleotide polymorphisms (SNPs) associated with AMD make a significant improvement to model prediction beyond what is possible with image data. In the report from Peng and colleagues described above, neither the variants at CFH/ARMS2 nor a 52-SNP-based AMD genetic risk score substantially enhanced the prediction of nAMD or GA at 5 years⁴⁹. Likewise, Yan and colleagues found that using the output from the Inception v3 CNN trained with CFP from the AREDS study along with the 52 risk variants as inputs to a second CNN only modestly improved the model's ability to predict AMD progression over 2 to 7 years from 0.81 to 0.84 on the AUROC⁵¹. Genome-wide ML analyses have not yet been reported in the literature. It is likely that this additional data may benefit model accuracy to the extent that imaging data is deficient in quality or label accuracy. Interestingly, since genotype data is unavailable for most patients, a non-reliance of genomic data on DL models for AMD may actually improve adoption and accessibility.

A subset of AMD patients may also deserve special attention: individuals with nAMD in one eye, have an 18% risk of development of nAMD in the fellow eye within one year and a 45% risk of conversion within five years, a clinical event that is made even more undesirable since it affects the patient's remaining "good" eye⁵² [Lechanteur and Klaver, ARVO 2021]. To address this occurrence, Yim and colleagues identified 2795 patients from Moorfields Eye Hospital who had a first eye with nAMD and a fellow eye without nAMD. They trained a CNN with OCT images from the fellow eye that were obtained every 1-12 months to output a score representing the likelihood of conversion to nAMD within six months. The authors identified two operating points on the receiver-operator characteristic curve that may be employed in different clinical scenarios: a liberal operating point with 80% sensitivity and 55% specificity, and a conservative operating point with 34% sensitivity and 90% specificity. This performance was sufficient to exceed 5 out of 6 retinal specialists and was non-inferior to the 6th provider. The work also identified a potentially clinically meaningful gap between the date of likely conversion and the date of the patient's first intravitreal anti-VEGF injection underscoring the importance of automated, longitudinal surveillance⁵³.

Monitoring

Patients with nAMD require regular monitoring to determine the need for retreatment with anti-VEGF medication, a regimen that may last for many years. Studies have shown that real-world outcomes with anti-VEGF therapies frequently fall short of those reported in clinical trials, in part, because of the difficulty of adherence to the appropriate treatment schedule by both patients and physicians⁵⁴. AI may be a useful tool in the longitudinal surveillance of these patients.

Several research efforts have produced ML models for the segmentation of retinal fluid, a principal pathologic feature that guides clinicians' decision to administer anti-VEGF therapy. As one example, Keenan and colleagues applied a ML model to the detection of intraretinal fluid and subretinal fluid from OCTs generated from AREDS2 10-year visit. A ground truth label for the presence or absence of each fluid type had been assigned to each OCT by reading center graders. The algorithm achieved a sensitivity of 0.76 and specificity of 0.92 for the detection of IRF (compared to 0.40 and 0.97 by retinal specialists), and 0.94 and 0.85 for SRF (compared to 0.58 and 0.97 for

reading center graders). The algorithm's superior ability to detect retinal fluid compared to retinal specialists was in large part due to relatively low sensitivity among retinal specialists⁵⁵.

One potential commercial application for retinal fluid detection may be home-use OCT in which nAMD patients perform regular imaging on a consumer grade OCT device that is monitored with AI. However, this technology has not received FDA authorization⁵⁶. With increasing interest in teleophthalmology and fewer barriers to practice, home-monitoring is likely to grow in demand in the future.

Monitoring may also be aided by a prediction of the future need for retreatment. One research example involved the classification of patients into a high (greater than 16) or low (less than 5) treatment requirement for ranibizumab therapy during the two years following three initial loading doses. The study used demographic features, BCVA, and 525 automatically acquired features from OCT data collected during the HARBOR trial to train a random forest classifier. Prediction of high- and low-treatment patients had an AUROC of 0.77 and 0.70, and better sensitivity but lower specificity compared to a human grader^{57 58}.

Clinical trials

Apart from routine clinical care, AI can also make an important impact in clinical trials for AMD through advances in recruitment, data collection, and cohort identification. As one example, GA is an advanced form of AMD characterized by loss of photoreceptors, RPE, and choriocapillaris. Despite the visual devastation rendered by these changes, there is an incomplete understanding of its pathogenesis as well as challenges of patient selection and primary outcomes for clinical trials. Currently, there are no FDA-authorized disease-modifying treatments for GA.

In two-thirds of cases, GA initially affects the perifovea and later affects central vision⁵⁹. As such, microperimetry, a testing modality that can be used to measure central retinal sensitivity, has served as a secondary outcome measure in the phase-3 studies of Lampalizumab for GA⁶⁰. However, the testing regimen is time consuming and may be difficult for patients. A ML approach sought to predict visual function, as represented in mesopic, dark-adapted cyan, and dark-adapted red microperimetry, from retinal anatomy alone, specifically using macular OCT features, infrared reflection intensity images, and fundus autofluorescence images. On a cross-validation design, the model achieved mean absolute error (MAE) of 3.94 dB for mesopic, 4.93 dB for dark-adapted cyan, and 4.02 dB for dark-adapted red testing. When the patient underwent an abbreviated 5-minute microperimetry examination the predicted pointwise MAE significantly improved⁶¹.

In addition to functional assessment, automated anatomical analyses may also benefit investigators seeking to establish secondary endpoints, identify patient subpopulations, or make early predictions of treatment success or failure. Zhang and colleagues used a U-Net architecture trained on 5049 manually segmented OCT B-scans from patients enrolled in a clinical trial for the treatment of GA. Three independent models recognized RPE loss, hypertransmission, or photoreceptor degeneration with median Dice similarity coefficient greater than 0.95, which exceeded human assessment⁶².

Finally, AI may accelerate GA research by identifying for inclusion in clinical trials patients who are at highest risk of converting to GA. This could involve large-scale screening of databases for patients with intermediate AMD for rapid enrollment of a large sample of patients. Additionally, it is possible that AI could identify AMD patients who are at high risk of conversion to GA so that therapeutics could be tested for efficacy in a shorter timeframe.

Clinical validation of AI models

Benchmarking

Validation of an AI model is possible to the extent that a ground truth can be assigned. For some tasks like feature segmentation, the ground truth is well-founded on the representation of anatomy in image data. On the other hand, a diagnosis or disease stage is subject to greater perceptual variability and occasional change in how the entity is categorized. Thus, a valid algorithm will approximate the output of a human expert, or in some cases, the average or majority consensus of multiple experts.

Studies to date have employed humans with a variety of experience levels to provide data labeling including fellowship-trained vitreoretinal specialists, comprehensive ophthalmologists, optometrists, reading center technicians, and medical students. They have also used different amounts of grading redundancy and strategies for arriving at consensus. One study by Maloca and Scholl compared the labeling of the internal limiting membrane, choriocapillaris boundary, and choroid-sclera interface in OCT images across three groups of graders, expert ophthalmologists, reading center graders, and lay persons with basic instructions. There was no statistically significant difference in the accuracy of the graders, when judged against the ground truth of three expert ophthalmologists. However, the lay annotators had greater inter-rater variability, particularly for labeling the choroid-scleral interface, which was considered less distinct than the other boundaries. This suggests that, for some labeling tasks, multiple non-specialists may provide adequate ground truth labeling but that, for more heuristic tasks like labeling the presence of RPE changes, multiple specialists or subspecialists will likely be required to perform redundant labeling with an adjudication strategy to reach consensus⁶³.

Interesting but also as yet unstudied is benchmarking the performance of a combined human and AI model workflow. There is a wide range of outputs that clinicians can receive from an AI model, including feature segmentation, differential diagnoses with confidence levels, saliency maps, and more. The extent to which the performance of a clinician using an AI model may be augmented or worsened has not yet been assessed in AMD but will be important, since implementations of AI in the clinical setting are likely to interface with human providers.

In terms of validating SaMD algorithms, although metrics like the AUROC can be useful for guiding development, ultimately the judgment of a success is the extent to which it makes a positive impact on patients' lives. If a device is screening for critical illness, a favorable high-sensitivity operating point may be selected over an alternative algorithm with a higher AUROC. With this in mind, the true characteristic of a clinically meaningful SaMD is whether it improves clinical outcomes like visual acuity when it is applied to patients in a particular healthcare setting. This result would be best demonstrated in a prospective or randomized clinical trial.

Equity

The accurate diagnosis of AMD with ML may be challenged by racial and ethnic differences in prevalence and phenotype of the disease. GA is much more frequent in individuals of European ancestry than those of African or Asian ancestry despite similar levels of soft drusen in both populations, and PCV may be three to five times more likely in African and Asian populations compared to those of European ancestry^{1, 64}. These differences in prevalence exist against the backdrop of documented anatomical differences in retinal characteristics, including

retinal thickness and retinal pigmentation, among individuals with different ethnic backgrounds⁶⁵. Even beyond biases in data sets due to geographic, economic, and sociological factors, these differences in prevalence may be a hindrance to classification parity of fairness and other strategies that seek to achieve similar predictive performance across demographic groups.

Some researchers have utilized synthetic data from generative adversarial networks (GANs) to address a dearth of data for demographic minority groups. A recent study found that a progressive growing GAN modeled on the ResNet50 CNN architecture and trained on CFP images from AREDS could produce synthetic fundus images that were nearly indistinguishable from real images by retinal specialists. Moreover, a CNN trained to classify images as having referable versus non-referable AMD performed with an AUROC of 0.9706 when trained with real images and 0.923 when trained with synthetic images¹⁷. A subsequent study by the same group focusing on diabetic retinopathy rather than AMD found that using a GAN to expand EyePACS with synthetic data from the latent space of racial and ethnic minority groups improved the disparity of model performance trained on the new data set from 12.5% to 0.5%¹⁸. These results suggest that accruing images of rare AMD mimics and variants as well as patients from underrepresented demographic groups could be used to generate a much larger set of images that may help in the development and training of more accurate ML algorithms.

In the recently published Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan⁶⁶, the FDA recognized the crucial importance for medical devices to be well suited for a racially and ethnically diverse intended patient population and the need for improved methodologies for the identification and development of ML robust algorithms. The FDA committed to support regulatory science efforts to develop methodology for the evaluation and improvement of ML algorithms, including for the identification and elimination of bias, and for the evaluation and promotion of algorithm robustness.

Additionally, in 2017, the FDA issued an Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies Guidance⁶⁷. In this Guidance, the Agency outlined expectations and provided recommendations for the evaluation and reporting of age-, race-, and ethnicity-specific data in medical device clinical studies. The guidance encourages the collection and consideration during the study design stage of relevant age, race, ethnicity, and associated covariates for devices for which safety, effectiveness or benefit-risk profile is expected to vary across these groups. This guidance is intended for all devices that include clinical information in support of a marketing submission. Thus, clinical validation trials of SaMD for AMD should include diverse populations that reflect the intended use population.

Regulation

AI-based SaMD for AMD is currently uncharted territory and the regulatory treatment by the FDA would, naturally, depend on many factors including the indication for use, operator and clinical setting, and other sources of risk and benefit. Although not a predicate, one touchstone is the IDx-DR, an AI-based device for the autonomous screening of diabetic retinopathy. The IDx-DR SaMD was authorized for marketing through the De Novo process, which also resulted in creating a new regulation (21 CFR 886.1100) that applies to any device that is intended for use as a prescription software device that incorporates an adaptive algorithm to evaluate ophthalmic images for diagnostic screening to identify retinal diseases or condition. Later, a similar device, EyeArt, was approved through the 510(k) pathway as a class II device using IDx-DR as a predicate device. Therefore, while there are currently no FDA-authorized AI-based SaMD indicated to screen for AMD, any future SaMD for this indication would likely fall within the scope of this new regulation and require marketing clearance through the premarket notification (510(k))

pathway prior to commercial distribution in the United States⁶⁸. However, non-screening applications or those involving novel technology are unlikely to be able to consider IDx-DR as a predicate device.

Also of note, AI-based devices must be approved for a specific imaging hardware make and model. For instance, IDx-DR was approved for use with the TopCon NW400 fundus camera. In January 2021, the FDA published the AI/ML SaMD Action Plan where the agency committed to issuance of Draft Guidance on the Predetermined Change Control Plan to update the previously proposed framework for AI/ML-based SaMD. This will help to delineate a path for extending an approved SaMD's use with other imaging devices^{66,69, 70}.

The same FDA guidance also addresses a pathway to continuous learning. ML models can be locked, such that after regulatory clearance, their parameters are no longer updated, or they can undergo continuous learning, in which the model is periodically retrained with new data. Continuous learning models have the potential to improve in accuracy and may even be necessary to prevent degradation of performance over time due to changes in workflow or patient population. Although continuous ML exists in many commercial AI applications from self-driving cars to movie recommendations, all FDA-approved SaMD to date are locked. One concern with continuous learning is catastrophic forgetting, a phenomenon in which new data interferes with previous learning and the model performance decreases as it is updated. For supervised learning, continuous learning would also require that a qualified human label new training data in a way that is consistent with prior data labeling.

AI used for scientific research may meet the definition of a medical device, but because the product is not offered for sale (i.e., commercial distribution), it may not need FDA marketing authorization. Marketing authorization notwithstanding, there may still be a need for FDA oversight through the investigational device exemption (IDE) pathway. If, however, a health care provider self-develops and uses an AI in their practice without offering the AI for sale, these 'homebrew' AI systems may be used by the clinicians who developed it under the umbrella of "practice of medicine." Even greater technical validation efforts are required to attain marketing authorization for AI systems that meet the definition of a regulated medical device. Those that do meet the definition of a medical device and that are actively regulated by FDA must meet applicable pre- and post-market requirements for safety, effectiveness, and performance monitoring by the overseeing regulatory agencies in the intended marketing countries.

Data to Deployment

Characterizing a clinical problem and its AI-based solution

In summary of the preceding discussion, what does the roadmap look like for someone who is interested in developing AI-based SaMD for AMD (Figure 3)? The first and most fundamental step in the development of an AI-based SaMD is to identify a genuine clinical need that is best solved by a feasible AI-based solution. As stated by Paul Yock, founder of the Stanford Byers Center for Biodesign, "A well-characterized need is the DNA of a great invention." The clinical need should delineate a problem in a specific patient population and a desired outcome after intervention. Additionally, the AI solution should have a defined user in a specific care setting in the patient's specific phase of care (screening, diagnosis, prediction, monitoring), and the device output should allow that user to address the clinical need (action-outcome pairing). Will the device segment SRF to assist an ophthalmologist in treatment decisions, flag likely AMD mimics for further work-up, or predict conversion to nAMD from home OCT data? A initial need supported by the authors is for the autonomous screening of vision-threatening AMD requiring further evaluation by an ophthalmologist, similar to the path traversed by AI-based screening for more-than-mild

(referable) diabetic retinopathy. Starting with a useful and tractable end in mind is essential for structuring the AI development process.

Once a clinical need and potential solution are delineated, one should evaluate whether data will support the proposed application. This includes having data modalities that represent a strong ground truth for the clinical condition. For instance, how can the investigator be sure that macular conditions that mimic AMD, such as inherited retinal diseases with macular atrophy, are excluded from a database of AMD patients? The careful selection of validated databases or a process for *de novo* validation is thus crucial for the ground truth. Additionally, one must carefully determine which other disease states to include in the data set in order to train the AI algorithm to recognize what is not the disease state, both normal eyes and disease states with similar phenotypes.

Finally, the investigator should decide on the broad training approach needed. Supervised and semi-supervised learning rely on varied levels of human-labeling of the data, whereas unsupervised learning does not. A final broad category of AI is reinforcement learning, which models sequential actions by an agent to maximize a reward determined by the system.

Data-set curation

Next, a data set must be assembled for training and testing the AI model. In general, retrospective analysis of existing data is the faster and cheaper than prospective collection. Many data sets can be downloaded from public websites, and additional data can be acquired from health systems and community practices through research collaborations or financial arrangements. The post-market performance of a production-level AI model is related to the extent to which the development data set represents the patient population for which the device is used. To cover as much of the potential clinical population as possible, the investigator should likewise seek a large and diverse development data set from multiple clinical sites and settings through data sharing agreements or federated learning. In light of the decrease in performance across different makes and models of imaging hardware, and since during the regulatory process, an image-based AI model must be paired with a single piece of imaging hardware, the data set should reflect the final intended image acquisition use case. As data is obtained from multiple sources, the investigator should preserve data provenance by retaining metadata like unique patient ID, demographic information, date of service, and acquisition hardware.

Some applications may require prospective data collection either because the needed data does not exist, may not be reliable, or may not be accessible. Although greater time and resources are required to collect prospective data sets, there can be significant benefit in the performance of AI models trained with data collected with an exact use in mind.

Data cleaning and manipulation

Having assembled a data set, the investigator next investigates the data for quality. Although cases of AMD with comorbid retinal disease can confound the model, it is recommended not to exclude such cases since they will inevitably be encountered in the real world. AI-based SaMD can contain a module that initially evaluates new patient data for quality, so excluding noisy or erroneous data from a training data set may be appropriate; nevertheless, the data should reflect the anticipated level of data quality that the model will encounter in the real world.

Additional preprocessing and data curation steps will be specific to the AMD application. For instance, OCT data could be analyzed as individual B-scans or as three-dimensional image data; further, the three-dimensional data can be greyscale data or rendered as a semantic segmentation map; and further still, the segmentation map may have variable types of normal and abnormal feature classes trained by distinct approaches to human-generated labeling. Class imbalance should be corrected by putting each label category into approximate parity. Data augmentation can make the model more attuned to biological features than idiosyncratic features of image acquisition.

At this stage, exploratory data analysis may help the investigator further refine the approach before committing additional time and resources to model development. AutoML platforms may be a convenient way for investigators to get preliminary feedback on the feasibility of their approach before ML programming is undertaken.

Model training and testing

The ultimate goal of AI model training is to arrive at a function, as represented by the ML parameters, that minimizes error in its approximation of real-world clinical data. For model training, the data set is generally split into either a training, validation, and testing samples (70%, 15%, and 15% are accepted divisions) or a k-fold cross validation strategy is used. All data arising from a single patient, including time series data, should fall into one split rather than being divided to prevent data leakage.

A description of AI architectures is beyond the scope of the manuscript and given the pace of the field is subject to continued change in the near future. Broadly, CNN, RNN, support vector machines, random forest, decision trees, K-means, Bayesian deep learning, graphical models, and transformers are all examples of relevant ML methods that AI-based SaMD may contain. Based on the application, the model may address a regression task that outputs a continuous variable or a classification task that outputs a categorical label; among classification tasks there may be binary classification or multi-class classification. The end-to-end algorithm may include an ensemble of models, especially if more than one mode of data contributes to the output. As previously mentioned, architectures may be developed for training on two-dimensional or three-dimensional image data. Finally, sets of labeled general images and ophthalmology-specific images are publicly available for transfer learning, and a variety of publicly available DL models come already pretrained.

During training, incomplete convergence of model performance over time may indicate underfitting and prompt the investigator to collect more data. After convergence to a reliable set of hyperparameters, model performance is evaluated on a test set (or derived from multiple test sets on k-fold cross validation) by plotting a receiver-operating characteristic (ROC) curve. Markedly poorer performance on the test set compared to the training set is characteristic of overfitting and may be addressed by reducing model complexity and imposing regularization methods.

From the ROC, an operating point should be calculated that has a clinically meaningful sensitivity and specificity. For the desired application, what is an acceptable false positive or false negative rate? What are the clinical consequences of each, and outside of the AI device, are safeguards in place that will mitigate these errors? These choices should reflect a clinical reference standard based on published literature or consensus in the field.

Regulation

Any medical device marketed in the United States must receive clearance from the U.S. FDA. As discussed previously, the regulation established by De Novo Pathway clearance of the IDx-DR (and that also includes EyeArt through 510(k) clearance) for a prescription software device that incorporates an adaptive algorithm to evaluate ophthalmic images for diagnostic screening to identify retinal diseases or condition (21 CFR 886.1100) may pertain to similar diagnostic screening devices for AMD. However, there is no exact regulatory precedent for AI-based SaMD for AMD. In addition to special regulatory considerations, AI-based devices face the same general regulations that pertain to all other SaMD. Finally, a device maker may consider submitting a pre-determined change control plan to try to create a flywheel effect: large amounts of data are obtained from real-world use of the device that, in turn, improves model performance through retraining, which further increases clinical utilization of the device, which generates even more training data. Finally, the FDA mandates that companies conduct post-market surveillance of SaMD and may require training and ongoing customer support.

Of interest, there are other reporting standards for AI studies published recently, including Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) and AI extension to the Standards for Reporting of Diagnostic Accuracy Studies that are currently under development^{11, 71}. While potentially beneficial, it is not yet proven that such standards may provide sufficient information for regulatory oversight as the FDA and other regulatory bodies have not yet evaluated them. However, when possible, it would be beneficial to align with these other standards such as the Clinical Evaluation of SaMD, STARD, CONSORT-AI, the American Telemedicine Association Telehealth Practice Guidelines for Diabetic Retinopathy, Digital Communications in Medicine (DICOM) and FDA's "Software as A Medical Device: Clinical Evidence Guidelines"^{11, 71-74}.

Implementation

The real-world financial viability of an AI model will likely depend on whether the device is covered under a reimbursement code. Two reimbursement codes may be relevant to the screening and monitoring of AMD, 92227 and 92228. FDA-approved devices for diabetic retinopathy screening are covered under the unique reimbursement code 92229, and future AMD devices may prompt the creation of unique codes as well.

Conclusion

Although no artificial intelligence device for AMD has been approved for clinical use in the United States, the research efforts undertaken to date represent starting points for the devices that will eventually benefit AMD patients. These benefits may include: scaling access to diagnosis and monitoring through the empowerment of non-specialists; increasing specialists' insights into clinical and imaging data for greater efficiency of monitoring or efficacy of treatment regimen, potentially leading to better patient outcomes at a lower cost; improved reliability and validity across patterns of management by aggregating data from large patient populations to aid in clinical decision-making; and the automation of rote tasks to decrease clinician fatigue and improving the doctor-patient relationship. Although numerous data types may support the development of models for AMD, CFP and OCT image data will likely be foundational for many applications given their widespread clinical use and high-quality depiction of AMD pathology. Realizing the potential of AI for AMD will require robust infrastructure and continued efforts within the AI and ophthalmology communities.

References

1. Wong, W.L., Su, X., Li, X., Cheung, C.M.G., et al., *Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis*. The Lancet Global Health, 2014. **2**(2): p. e106-e116.
2. Ferris, F.L., Davis, M.D., Clemons, T.E., Lee, L.-Y., et al., *A simplified severity scale for age-related macular degeneration: AREDS Report No. 18*. Arch Ophthalmol, 2005. **123**(11): p. 1570-4.
3. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., et al., *The NCBI dbGaP database of genotypes and phenotypes*. Nat Genet, 2007. **39**(10): p. 1181-6.
4. Windsor, M.A., Sun, S.J.J., Frick, K.D., Swanson, E.A., et al., *Estimating Public and Patient Savings From Basic Research-A Study of Optical Coherence Tomography in Managing Antiangiogenic Therapy*. Am J Ophthalmol, 2018. **185**: p. 115-122.
5. Busbee, B.G., Ho, A.C., Brown, D.M., Heier, J.S., et al., *Twelve-month efficacy and safety of 0.5 mg or 2.0 mg ranibizumab in patients with subfoveal neovascular age-related macular degeneration*. Ophthalmology, 2013. **120**(5): p. 1046-56.
6. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., et al., *Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning*. Cell, 2018. **172**(5): p. 1122-1131 e9.
7. Khan, S.M., Liu, X., Nath, S., Korot, E., et al., *A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability*. The Lancet Digital Health, 2021. **3**(1): p. e51-e66.
8. Yanagihara, R.T., Lee, C.S., Ting, D.S.W. and Lee, A.Y., *Methodological Challenges of Deep Learning in Optical Coherence Tomography for Retinal Diseases: A Review*. Transl Vis Sci Technol, 2020. **9**(2): p. 11.
9. Lee, A.Y., Campbell, J.P., Hwang, T.S., Lum, F., et al., *Recommendations for Standardization of Images in Ophthalmology*. Ophthalmology, 2021. **128**(7): p. 969-970.
10. Cruz Rivera, S., Liu, X., Chan, A.W., Denniston, A.K., et al., *Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension*. Nat Med, 2020. **26**(9): p. 1351-1363.
11. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J., et al., *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension*. Nat Med, 2020. **26**(9): p. 1364-1374.
12. Service, U.N.H. *INSIGHT Health Data Research Hub for Eye Health* April 1, 2021]; Available from: <https://theodi.org/project/insight/>. .
13. Henriksen, B.S., Goldstein, I.H., Rule, A., Huang, A.E., et al., *Electronic Health Records in Ophthalmology: Source and Method of Documentation*. Am J Ophthalmol, 2020. **211**: p. 191-199.
14. VeranaHealth. *IRIS Registry EHR Integration and MIPS Reporting FAQs*. 2021 September 7, 2021]; Available from: <https://www.veranahealth.com/aao-partnership-expansion-faqs/>.
15. Mehta, N., Lee, C.S., Mendonca, L.S.M., Raza, K., et al., *Model-to-Data Approach for Deep Learning in Optical Coherence Tomography Intraretinal Fluid Segmentation*. JAMA Ophthalmol, 2020. **138**(10): p. 1017-1024.
16. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., et al., *Clinically applicable deep learning for diagnosis and referral in retinal disease*. Nat Med, 2018. **24**(9): p. 1342-1350.
17. Burlina, P.M., Joshi, N., Pacheco, K.D., Liu, T.Y.A., et al., *Assessment of Deep Generative Models for High-Resolution Synthetic Retinal Image Generation of Age-Related Macular Degeneration*. JAMA Ophthalmol, 2019. **137**(3): p. 258-264.

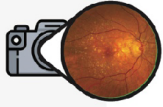
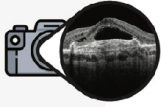


18. Burlina, P., Joshi, N., Paul, W., Pacheco, K.D., et al., *Addressing Artificial Intelligence Bias in Retinal Diagnostics*. *Transl Vis Sci Technol*, 2021. **10**(2): p. 13.
19. Service, U.N.H. *National data opt out*. April 15, 2021]; Available from: <https://digital.nhs.uk/services/national-data-opt-out>.
20. Fedus, W., Zoph, B. and Shazeer, N., *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*. Arxiv, 2021.
21. Korot, E., Pontikos, N., Liu, X., Wagner, S.K., et al., *Predicting sex from retinal fundus photographs using automated deep learning*. *Sci Rep*, 2021. **11**(1): p. 10286.
22. Faes, L., Wagner, S.K., Fu, D.J., Liu, X., et al., *Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study*. *The Lancet Digital Health*, 2019. **1**(5): p. e232-e242.
23. Korot, E., Guan, Z., Ferraz, D., Wagner, S.K., et al., *Code-free deep learning for multi-modality medical image classification*. *Nature Machine Intelligence*, 2021. **3**(4): p. 288-298.
24. Keane, P. and Topol, E., *AI-facilitated health care requires education of clinicians*. *Lancet*, 2021. **397**(10281).
25. Neely, D.C., Bray, K.J., Huisinigh, C.E., Clark, M.E., et al., *Prevalence of Undiagnosed Age-Related Macular Degeneration in Primary Eye Care*. *JAMA Ophthalmol*, 2017. **135**(6).
26. Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., et al., *Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks*. *JAMA Ophthalmol*, 2017. **135**(11): p. 1170-1176.
27. Ting, D.S.W., Cheung, C.Y., Lim, G., Tan, G.S.W., 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): p. 2211-2223.
28. Ferris, F.L., Wilkinson, C.P., Bird, A., Chakravarthy, U., et al., *Clinical classification of age-related macular degeneration*. *Ophthalmology*, 2013. **120**(4).
29. Gonzalez-Gonzalo, C., Sanchez-Gutierrez, V., Hernandez-Martinez, P., Contreras, I., et al., *Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration*. *Acta Ophthalmol*, 2020. **98**(4): p. 368-377.
30. Pandit, R.R., Wibbelsman, T.D., Considine, S.P., Jenkins, T.L., et al., *Distribution and Practice Patterns of Retina Providers in the United States*. *Ophthalmology*, 2020. **127**(11).
31. Group, I.M.D.R.F.S.a.a.M.D.S.W. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. 2014 September 7, 2021]; Available from: <http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf>.
32. Abramoff, M.D., Cunningham, B., Patel, B., Eydelman, M.B., et al., *Foundational Considerations for Artificial Intelligence Using Ophthalmic Images*. *Ophthalmology*, 2021: p. 1-19.
33. Mishra, Z., Ganegoda, A., Selicha, J., Wang, Z., et al., *Automated Retinal Layer Segmentation Using Graph-based Algorithm Incorporating Deep-learning-derived Information*. *Sci Rep*, 2020. **10**(1): p. 9541.
34. Saha, S., Nassisi, M., Wang, M., Lindenberg, S., et al., *Automated detection and classification of early AMD biomarkers using deep learning*. *Sci Rep*, 2019. **9**(1): p. 10990.
35. Waldstein, S.M., Vogl, W.D., Bogunovic, H., Sadeghipour, A., et al., *Characterization of Drusen and Hyperreflective Foci as Biomarkers for Disease Progression in Age-Related*







- Macular Degeneration Using Artificial Intelligence in Optical Coherence Tomography*. JAMA Ophthalmol, 2020. **138**(7): p. 740-747.
36. Schmidt-Erfurth, U., Vogl, W.D., Jampol, L.M. and Bogunovic, H., *Application of Automated Quantification of Fluid Volumes to Anti-VEGF Therapy of Neovascular Age-Related Macular Degeneration*. Ophthalmology, 2020. **127**(9): p. 1211-1219.
 37. Schmitz-Valckenberg, S., Gobel, A.P., Saur, S.C., Steinberg, J.S., et al., *Automated Retinal Image Analysis for Evaluation of Focal Hyperpigmentary Changes in Intermediate Age-Related Macular Degeneration*. Transl Vis Sci Technol, 2016. **5**(2): p. 3.
 38. Wang, J., Hormel, T.T., Gao, L., Zang, P., et al., *Automated diagnosis and segmentation of choroidal neovascularization in OCT angiography using deep learning*. Biomed Opt Express, 2020. **11**(2): p. 927-944.
 39. Moraes, G., Fu, D.J., Wilson, M., Khalid, H., et al., *Quantitative Analysis of OCT for Neovascular Age-Related Macular Degeneration Using Deep Learning*. Ophthalmology, 2021. **128**(5): p. 693-705.
 40. Peng, Y., Dharssi, S., Chen, Q., Keenan, T.D., et al., *DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs*. Ophthalmology, 2019. **126**(4): p. 565-575.
 41. Domalpally, A., Agron, E., Pak, J.W., Keenan, T.D., et al., *Prevalence, Risk, and Genetic Association of Reticular Pseudodrusen in Age-related Macular Degeneration: Age-Related Eye Disease Study 2 Report 21*. Ophthalmology, 2019. **126**(12): p. 1659-1666.
 42. Yang, J., Zhang, C., Wang, E., Chen, Y., et al., *Utility of a public-available artificial intelligence in diagnosis of polypoidal choroidal vasculopathy*. Graefes Arch Clin Exp Ophthalmol, 2020. **258**(1): p. 17-21.
 43. Xu, Z., Wang, W., Yang, J., Zhao, J., et al., *Automated diagnoses of age-related macular degeneration and polypoidal choroidal vasculopathy using bi-modal deep convolutional neural networks*. Br J Ophthalmol, 2021. **105**(4): p. 561-566.
 44. Pang, C.E. and Freund, K.B., *Pachychoroid neovascularopathy*. Retina, 2015. **35**(1).
 45. Hosoda, Y., Miyake, M., Yamashiro, K., Ooto, S., et al., *Deep phenotype unsupervised machine learning revealed the significance of pachychoroid features in etiology and visual prognosis of age-related macular degeneration*. Sci Rep, 2020. **10**(1): p. 18423.
 46. Treder, M., Lauermann, J.L. and Eter, N., *Deep learning-based detection and classification of geographic atrophy using a deep convolutional neural network classifier*. Graefes Arch Clin Exp Ophthalmol, 2018. **256**(11): p. 2053-2060.
 47. Davis, M.D., Gangnon, R.E., Lee, L.-Y., Hubbard, L.D., et al., *The Age-Related Eye Disease Study severity scale for age-related macular degeneration: AREDS Report No. 17*. Arch Ophthalmol, 2005. **123**(11).
 48. Burlina, P.M., Joshi, N., Pacheco, K.D., Freund, D.E., et al., *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**(12): p. 1359-1366.
 49. Peng, Y., Keenan, T.D., Chen, Q., Agron, E., et al., *Predicting risk of late age-related macular degeneration using deep learning*. NPJ Digit Med, 2020. **3**: p. 111.
 50. Banerjee, I., de Sisternes, L., Hallak, J.A., Leng, T., et al., *Prediction of age-related macular degeneration disease using a sequential deep learning approach on longitudinal SD-OCT imaging biomarkers*. Sci Rep, 2020. **10**(1): p. 15434.
 51. Yan, Q., Weeks, D.E., Xin, H., Swaroop, A., et al., *Deep-learning-based Prediction of Late Age-Related Macular Degeneration Progression*. Nature Machine Intelligence, 2020. **2**(2).
 52. Bek, T. and Klug, S.E., *Incidence and risk factors for neovascular age-related macular degeneration in the fellow eye*. Graefes Arch Clin Exp Ophthalmol, 2018. **256**(11): p. 2061-2068.

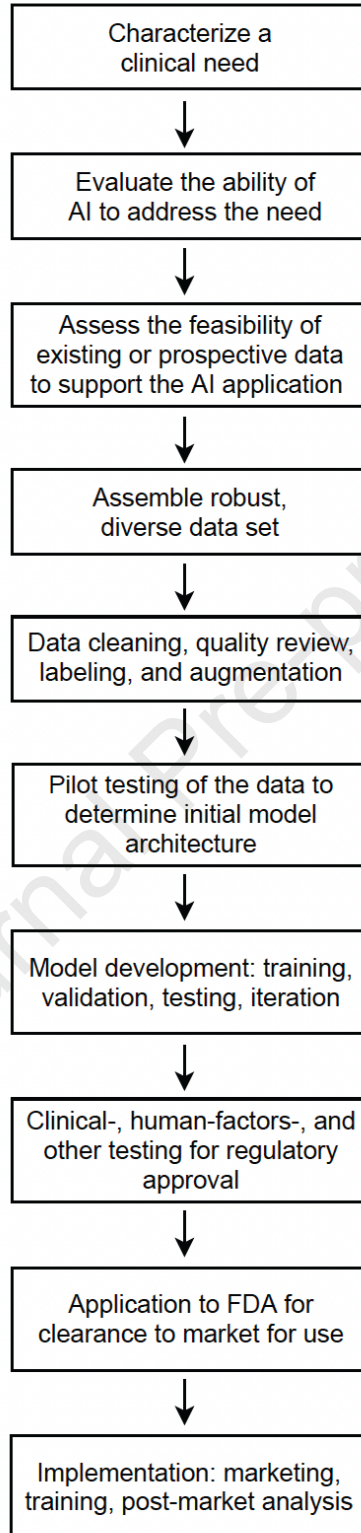
53. Yim, J., Chopra, R., Spitz, T., Winkens, J., et al., *Predicting conversion to wet age-related macular degeneration using deep learning*. Nat Med, 2020. **26**(6): p. 892-899.
54. Mehta, H., Tufail, A., Daien, V., Lee, A.Y., et al., *Real-world outcomes in patients with neovascular age-related macular degeneration treated with intravitreal vascular endothelial growth factor inhibitors*. Prog Retin Eye Res, 2018. **65**: p. 127-146.
55. Keenan, T.D., Dharssi, S., Peng, Y., Chen, Q., et al., *A Deep Learning Approach for Automated Detection of Geographic Atrophy from Color Fundus Photographs*. Ophthalmology, 2019. **126**(11): p. 1533-1540.
56. Ho, A.C., Heier, J.S., Holekamp, N.M., Garfinkel, R.A., et al., *Real-World Performance of a Self-Operated Home Monitoring System for Early Detection of Neovascular Age-Related Macular Degeneration*. J Clin Med, 2021. **10**(7).
57. Pfau, M., Sahu, S., Rupnow, R., Romond, K., et al., *Probabilistic Forecasting of the Anti-VEGF Treatment Frequency in Neovascular Age-Related Macular Degeneration*. Transl Vis Sci Technol, 2021. **in press**.
58. Bogunovic, H., Montuoro, A., Baratsits, M., Karantonis, M.G., et al., *Machine Learning of the Progression of Intermediate Age-Related Macular Degeneration Based on OCT Imaging*. Invest Ophthalmol Vis Sci, 2017. **58**(6): p. BIO141-BIO150.
59. Keenan, T.D., Agron, E., Domalpally, A., Clemons, T.E., et al., *Progression of Geographic Atrophy in Age-related Macular Degeneration: AREDS2 Report Number 16*. Ophthalmology, 2018. **125**(12): p. 1913-1928.
60. Holz, F.G., Sadda, S.R., Busbee, B., Chew, E.Y., et al., *Efficacy and Safety of Lampalizumab for Geographic Atrophy Due to Age-Related Macular Degeneration: Chroma and Spectri Phase 3 Randomized Clinical Trials*. JAMA Ophthalmol, 2018. **136**(6).
61. von der Emde, L., Pfau, M., Dysli, C., Thiele, S., et al., *Artificial intelligence for morphology-based function prediction in neovascular age-related macular degeneration*. Sci Rep, 2019. **9**(1): p. 11132.
62. Zhang, G., Dun, J.F., Liefers, B., Livia, F., et al., *Clinically relevant deep learning for detection and quantification of geographic atrophy from optical coherence tomography: a model development and external validation study*. The Lancet Digital Health, 2021. **3**(10): p. E665-E675.
63. Maloca, P.M., Lee, A.Y., de Carvalho, E.R., Okada, M., et al., *Validation of automated artificial intelligence segmentation of optical coherence tomography images*. PLoS One, 2019. **14**(8): p. e0220063.
64. Klein, R., Peto, T., Bird, A. and Vannewkirk, M.R., *The epidemiology of age-related macular degeneration*. Am J Ophthalmol, 2004. **137**(3).
65. Kashani, A.H., Zimmer-Galler, I.E., Shah, S.M., Dustin, L., et al., *Retinal thickness analysis by race, gender, and age using Stratus OCT*. Am J Ophthalmol, 2010. **149**(3): p. 496-502 e1.
66. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*, U.F.a.D.A.C.f.D.a.R. Health, Editor. January 2021. p. 1-8.
67. Administration, U.F.a.D., *Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies: Guidance for Industry and Food and Drug Administration Staff*. 2017. p. 1-36.
68. Administration, U.F.a.D. *Premarket Notification 510(k)*. 2021 [September 5, 2021]; Available from: <https://www.fda.gov/medical-devices/premarket-submissions/premarket-notification-510k>.
69. Vokinger, K.N., Feuerriegel, S. and Kesselheim, A.S., *Continual learning in medical devices: FDA's action plan and beyond*. The Lancet Digital Health, 2021. **3**(6): p. e337-e338.

70. Lee, C.S. and Lee, A.Y., *Clinical applications of continual learning machine learning*. The Lancet Digital Health, 2020. **2**(6): p. e279-e281.
71. JF, C., DA, K., DG, A. and al, e., *STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration*. BMJ Open, 2016. **6**(11): p. e012799.
72. J, C., MG, L., I, Z.-G. and al, e., *Telehealth practice recommendations for diabetic retinopathy*. Telemed J E Health, 2004. **10**(4): p. 469-482.
73. U.S. Food & Drug Administration (FDA), I.M.D.R.F., *SOFTWARE AS A MEDICAL DEVICE (SAMD): CLINICAL EVALUATION*. 2016.
74. MD, A., T, L., DSW, T. and al, e., *Automated and Computer-Assisted Detection, Classification, and Diagnosis of Diabetic Retinopathy*. Telemed J E Health, 2020. **26**(4): p. 544-550.

Journal Pre-proof

IMAGING DATA		CLINICAL DATA	GENETIC DATA
			
COLOR FUNDUS PHOTOGRAPHY	OPTICAL COHERENCE TOMOGRAPHY		
<p>Description: Two-dimensional image of the fundus through the pupil by analog or digital photography.</p> <p>Benefits:</p> <ul style="list-style-type: none"> • Non-invasive. • Simple operation. • Widespread clinical use. • Captures numerous structures of the posterior pole. • Large, labeled data sets from clinical trials (e.g. AREDS) 	<p>Description: A series of two-dimensional cross-sectional images spanning several millimeters of the posterior pole using near-infrared light.</p> <p>Benefits:</p> <ul style="list-style-type: none"> • Non-invasive. • High-resolution imaging AMD lesions. • Widespread clinical use. • Large, labeled data sets from clinical trials (e.g. HARBOR) 	<p>TYPES</p> <ul style="list-style-type: none"> • Age • Visual acuity • Co-morbidities • Gender • Smoking • Ethnicity <p>Challenges:</p> <ul style="list-style-type: none"> • Documentation across and within institutions may be heterogenous and unstructured. 	<p>TYPES</p> <ul style="list-style-type: none"> • CFH • ARMS • 52-SNP variants from International AMD Genomics Consortium • Genome-wide association studies <p>Challenges:</p> <ul style="list-style-type: none"> • Unclear benefit beyond image data. • Few patients have genotype data.

<p style="text-align: center;">SCREENING CFP</p>  <p>Setting: Primary care clinic, Optometry office Modality: Standard or widefield CFP Output: Refer for further testing/observe. Likely fully autonomous clinical decision. High sensitivity operating point. Opportunity: Expand access.</p>	<p style="text-align: center;">SCREENING OCT</p>  <p>Setting: Optometry office, Comprehensive ophthalmology clinic Modality: Macula OCT Output: Urgent referral/routine referral/observe for nAMD detection. High specificity operating point. Opportunity: Expand access. Improve outcomes.</p>	<p style="text-align: center;">DIAGNOSIS</p>  <p>Setting: Comprehensive ophthalmology clinic, Retina specialist office Modality: CFP, OCT, OCT-A, FA, ICG Output: AMD stage flagging mimics or subtypes. Likely assistive rather than autonomous. Opportunity: Increase efficiency. Greater diagnostic precision.</p>
<p style="text-align: center;">PREDICTION</p>  <p>Setting: Retina specialist office Modality: OCT, CFP Output: Risk of first or fellow eye conversion to late AMD. Frequency of reexamination or ancillary testing. May be assistive or autonomous. Opportunity: Improve patient outcomes.</p>	<p style="text-align: center;">MONITORING</p>  <p>Setting: Comprehensive ophthalmology office, Retina specialist office, Home use Modality: OCT Output: Anti-VEGF retreatment interval or likelihood of benefit. Autonomous or assistive. Opportunity: Expand access. Increase efficiency. Improve outcomes</p>	<p style="text-align: center;">CLINICAL TRIAL</p>  <p>Setting: Retina specialist office Modality: Multiple modalities Output: Identify stage- or phenotype-specific patients for trial recruitment. Comprehensive data collection. Opportunity: Greater insight. Faster recruitment. Precision medicine.</p>



Précis:

The Collaborative Community on Ophthalmic Imaging working group for artificial intelligence (AI) in age-related macular degeneration discusses the state of the current data, standards, achievements, and the challenges to the development of effective AI.

;

Journal Pre-proof

Journal Pre-proof