[IN PRESS OPHTHALMOLOGY]

Citation: Abramoff MD, Cunningham B, Patel B, et al. Foundational Principles of Artificial Intelligence. Ophthalmology [in press]. 2021;

Foundational Considerations for Artificial Intelligence Utilizing Ophthalmic Images

Running Title: "Foundational Considerations for Artificial Intelligence"

Authors: Michael D. Abramoff,[1] Brad Cunningham, [2] Bakul Patel,[3] Malvina B. Eydelman, [2] Theodore Leng, [4] Taiji Sakamoto, [5,6] Barbara Blodi, [7] S. Marlene Grenon, [8] Risa M. Wolf, [9] Arjun K. Manrai, [10,11] Justin M. Ko, [12] Michael F. Chiang, [13] Danton Char, [14,15] on behalf of the *Collaborative Community on Ophthalmic Imaging Executive Committee* and *Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group**

1) University of Iowa, Iowa City, IA.
2) Center for Devices and Radiological Health, Office of Health Technology-1, US Food and Drug Administration, Silver Springs, MD.
3) Center for Devices and Radiological Health, Digital Health Center of Excellence, US Food and Drug Administration, Silver Springs, MD.
4) Byers Eye Institute at Stanford, Stanford University School of Medicine, Palo Alto, CA
5) Kagoshima University, Kyushu Pref, Japan
6) Japanese Vitreous Retina Society, Japan
7) Department of Ophthalmology, University of Wisconsin, Madison, WI
8) Innovation Ventures, University of California San Francisco, San Francisco, CA
9) Department of Pediatric Endocrinology, Johns Hopkins University School of Medicine, Baltimore, MD
10) Computational Health Informatics Program, Boston Children's Hospital, Boston, MA
11) Department of Biomedical Informatics, Harvard Medical School, Boston, MA
12) Department of Dermatology, Stanford University School of Medicine, Stanford, CA
13) National Eye Institute, Bethesda, MD
14) Department of Anesthesiology, Stanford University School of Medicine, Division of Pediatric Cardiac Anesthesia, San Francisco, CA
15) Center for Biomedical Ethics, Stanford University School of Medicine, San Francisco, CA

*) member list in appendix A

Conflicts of interest and disclaimers:

MDA: (F,C,I,P) Digital Diagnostics

FDA participates in the *Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation* 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.

Abstract

IMPORTANCE The development of Artificial Intelligence (AI) and other machine diagnostic systems, also known as Software as a Medical Device (SaMD), and its recent introduction into clinical practice, requires a deeply-rooted foundation in bioethics, for consideration by regulatory agencies and other stakeholders around the globe.

OBJECTIVES Initiate a dialogue on the issues to consider when developing a bioethically sound foundation for AI in medicine, based on images of eye structures, for discussion with all stakeholders.

EVIDENCE REVIEW The scope of the issues and summaries of the discussions under consideration by the Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group, as first presented during the Collaborative Community on Ophthalmic Imaging inaugural meeting on September 7, 2020, and afterwards in the working group.

FINDINGS AI has the potential to fundamentally improve healthcare access and patient outcome, while decreasing disparities, lowering cost, and enhancing the care team. Nevertheless, substantial concerns exist. Bioethicists, AI algorithm experts, as well as the Food and Drug Administration (FDA) and other regulatory agencies, industry, patient advocacy groups, clinicians and their professional societies, other provider groups, payors, ("stakeholders"), working together in collaborative communities to resolve the fundamental ethical issues of non-maleficence, autonomy and equity, is essential to attain this potential. Resolution impacts all levels of the design, validation and implementation of AI in medicine. Design, validation and implementation of AI warrant meticulous attention.

CONCLUSIONS AND RELEVANCE The development of a bioethically sound foundation may be possible if it is based in the fundamental ethical principles non-maleficence, autonomy and equity, for considerations for the design, validation and implementation for AI systems. Achieving such a foundation will be helpful for continuing successful introduction into medicine, before consideration by regulatory agencies. Important improvements in accessibility and quality of healthcare, decrease in health disparities, and lower cost can thereby be achieved. These considerations should be discussed with all stakeholders and expanded upon as a useful initiation of this dialogue.

1 Introduction

The Collaborative Community on Ophthalmic Imaging (CCOI) formed in 2019 to advance innovation of ophthalmic imaging with a focus on Medical Devices utilizing Artificial Intelligence. [1 2]

The CCOI's *Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation* Working Group (FPOAI), was established in March 2020 to generate consensus on a bioethical foundation for Artificial Intelligence (AI) of Ophthalmic Imaging, for consideration by all stakeholders in the healthcare system, including but not limited to the US Food and Drug Administration (FDA) and other regulatory agencies. Its processes draw on the expertise of bioethicists, [3, 4] AI algorithm experts, FDA and other regulatory agencies, as well as industry, patients and patient advocacy groups, clinicians and their professional societies, and payors,[5] to identify best practices for addressing novel issues emerging with AI conception, evaluation, and implementation, including validation, reference standards, performance metrics, accountability for output, bias, and impacts on workflow.

The terms Artificial Intelligence * and Augmented Intelligence (AI) are used interchangeably for systems that perform tasks that mimic human cognitive capabilities.[1] Such anthropomorphic AI systems, which are becoming more common, are not explicitly programmed, and instead learn from data that reflect highly cognitive tasks, typically performed by trained healthcare professionals. In some cases, these AI systems are used to aid healthcare professionals.[6] The introduction of AI in medicine has the potential to improve quality, reduce costs, diminish health disparities and increase accessibility, as well as enhance the care team, at both the individual and population levels.[7, 8] Its introduction thus aligns with the American Medical Association's principle of *quadruple aim* of improved outcomes, lower cost, improved patient experience, and improved clinician experience.[9] After the first FDA *De Novo* clearance for an autonomous AI,[10] in other words, an AI that makes a clinical decision without human oversight,[10] AI has entered mainstream healthcare, including standards of care.[11] The use of AI in the ophthalmic setting has been studied for many applications [12] including in diseases such as diabetic retinopathy, [13] retinopathy of prematurity,[14] and macular degeneration,, [15] glaucoma,[16] and cancer,[17] as well as many other ocular conditions, such as those of the cornea[18] and other parts of the anterior segment.[19]

To maximize AI's benefits, many ethical, economic, and scientific issues, including algorithmic bias, safety, efficacy and equity - terms that will be explained below - need to be addressed in a transparent fashion, for acceptance by all stakeholders. So far, studies to establish scientific evidence for the safety, and other criteria of AI in general are quite limited, with few exceptions.[20] In a meta-analysis of 81 AI clinical trials, only nine were prospective and just six were tested in a real-world clinical setting.[21] The relationship of the AI's diagnostic accuracy to clinical outcomes in this widely cited study was not even mentioned, and more generally, in an analysis of 126 published diagnostic accuracy studies, only 12% reported any statistical test of a hypothesis related to the study objectives. [22]

Recently, reporting standards for AI studies have been published, such as CONSORT-AI[23] , as well as an AI extension to Standards for Reporting of Diagnostic Accuracy Studies (STARD) [24] currently under development. While potentially beneficial, such standards may not provide sufficient information to help inform regulatory evaluation and have not been recognized by FDA. See also FDA's Recognized Consensus Standards. [25] While reporting standards may have benefits in improving consistency, there may be additional considerations beyond these recommendations that are needed for regulatory evaluation, many of which are the subject of these "Considerations."

This first 'Considerations' article to come from our FPOAI working group, presents the scope of the issues and concepts, and briefly summarizes the discussion on diagnostic AI and other Software as a Medical Device (SaMD) systems that use images of the eye, as first presented during the Collaborative Community on Ophthalmic Imaging inaugural meeting on September 7,2020, and later discussed within the FPOAI working group.[26] Specifically, it describes both clinical constraints for AI systems, as well as bioethically founded constraints, derived from the three major bioethical principles non-maleficence, equity and autonomy. While, as FPOAI stakeholders, we realize the tremendous potential advantages of AI systems, we also realize that substantial concerns also exist from the scientific and clinical communities, as well as society at large. Therefore, involvement of [3-5]all stakeholders to resolve ethical issues including non-maleficence, autonomy and equity, is key.

*) The term artificial intelligence refers to the concept of programming computer systems to perform tasks to mimic human cognitive capabilities- such as understanding language, recognizing objects and sounds, learning, and problem solving – by using logic, decision trees, machine learning, or deep learning.

Design, validation and implementation of diagnostic AI systems warrant meticulous attention. We limit the scope of these considerations, for the time being, to AI intended for diagnosis. While therapeutic AI, including autonomous AI for prescribing as well as autonomous AI for surgery are on the horizon, we decided that these are currently beyond the scope of these "Considerations" given the multiple ethical and even theoretical problems that need to be resolved. Furthermore, there is no regulatory guidance for AI systems using images of the eye.

Obviously, the considerations will be commensurate with the risk of harm to the patient, with different indications for use, conditions diagnosed, autonomy of the AI, consequences of a missed diagnosis, the population at risk and ~~many~~ other factors. Thus, the right balance needs to be considered between resource requirements and burden on AI creators[27] [28] to align with proposed ethical principles on the one hand, and risk of patient harm from lack of access to AI systems on the other hand, in order for patients, patient populations, and the wider healthcare system, to benefit .

In addition, while some AI systems are "marketed" medical devices, and under regulatory oversight, other AI systems are never marketed. Such 'homebrew' AI is used, by the clinicians who developed it, or others, in patient care, but is never marketed, and their safety and equity can be of concern.[29]

There are many useful resources, such as the reporting guidelines mentioned (e.g., Clinical Evaluation of SaMD,[26] STARD, [24] , CONSORT-AI,[23]), clinical practice guidelines (e.g., the American Telemedicine Association Telehealth Practice Guidelines for Diabetic Retinopathy[30, 31]), standards (e.g.,Digital Communications in Medicine (DICOM), [32]), and FDA guidance (e.g., "Software as A Medical Device: Clinical Evidence guidelines"[26]) that can be referenced to help mitigate aforementioned concerns. Ultimately, we incorporated these useful resources as initial steps in developing "best practices," and AI tailored regulatory frameworks, including Good Machine Learning Practice (GMLP), and other equivalents to the more familiar good manufacturing practices (GMP), as has been called for by the US General Accounting Office, in its report GAO-21-7SP: *Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care*, [33] as well as by regulatory agencies, such as FDA in its Digital Health Center of Excellence's recent *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.*[1]

2. Clinical Considerations for AI systems

These Considerations divide the requirements for AI systems into two categories: the clinical requirements, covered in this section, and the ethical requirements, derived from a bioethical foundation, which will be covered in Section 3. Therefore, this current Section 2 discusses the various clinical aspects of AI systems that use images of the eye in some form, conform the scope of the Collaborative Community on Ophthalmic Imaging.[2] We define *images of the eye* as topologically ordered sets of intensities, which represent physical and pathophysiological processes occurring in the eye, and that may reflect conditions of the eye and other conditions of parts of the patient's body. Specifically, we cover intended use, impact, inputs and outputs, and human factor design aspects of the AI system.

2.1 Intended use of the diagnostic AI system

The rationale for designing, developing, validating and deploying AI systems includes improving individual patient care, population health, and scientific research. Specifically for individual patients - improving their quality of care, lowering cost, increase access, decrease health disparities and improving

efficiency. For scientific research - discovering new disease mechanisms, and gaining a better understanding of a disease.

2.2 Impact of the diagnostic AI system

Once the use is identified, the impact of the AI systems can be assessed. AI systems span a wide range of impact, from having no direct impact on an individual patient or group of patients (e.g., *inform* a provider), to having an important, decision-making impact on an individual patient (i.e., *drive* or *treat*). [34] From a regulatory perspective, many AI systems are considered medical devices –SaMD - whereas other systems may not meet the definition of a medical device, as definitions differ across regulatory agencies.[35, 36] We refer to the US FDA's more narrow definition of medical devices under section 201h, [35] as modified under section 3060 of the 21st Century CURES act,[37] as well as the broader definition used by the International Medical Device Regulators Forum (IMDRF).[34] Based on those definitions, AI systems can be subdivided by impact as follows

Use case	**Description**	**Examples**	**FDA oversight**
Population care	prioritization and triage with potential impact on groups of patients and individual patients	Care pathway assignment	Likely[35]
Individual patient care			
Assistive AI	assists a clinician who determines the patient's management	Provides a probability or likelihood of a disease or condition, or may highlight potential lesions that should be reviewed by a specialist.	Likely[35]
Autonomous AI	makes a medical decision without input from a clinician.	For example, an autonomous AI system may evaluate for the presence of a disease, such as diabetic retinopathy and macular edema, or condition and notify the user whether the disease/condition is present	Likely[35]
Scientific research	not used for individual patient or population care, though the results of the research may impact populations or patients downstream.	Healthcare Analytics	Unlikely
Operations and data management	where this does not impact individual patient or population care. These often exist within the realm of Health Information Technology systems as they relate to administrative purposes.	VIM Referral Guidance, a triage system from EHRs	Unlikely
Clinical Decision Support (CDS)	Informs the clinician by aggregating, reformatting, or visualizing data, without providing analytical insights of the data, in a manner that allows the clinician to independently review the basis of the information provided by the software	AI system that suggests a G6PD test before prescribing an antimalarial therapy.[38]	Depends.[†]

[†] See 'Clinical Decision Support Software Guidance'. This explains when a software function qualifies as non-device CDS as well as device CDS, and which of these are actively regulated or for which compliance with applicable regulation would not be enforced, here: https://www.fda.gov/media/109618/download

General wellness –	collects physiologic information from devices, sensors including wearables.	Smart Watch that captures heart rate	Depends.[‡]

Table 1. AI System Impact

An important aspect of these AI systems is their theoretically unlimited scalability. ~~If the AI system is not locked after validation, there is also potentially unlimited configurability~~. Once designed and validated, the algorithms of a single AI system can be used on hundreds of millions of patients. While the number of patients a human clinician may encounter varies greatly based on the health setting and geography (e.g., 800-1000 unique patients per year, or during their entire career, no more than approximately 30,000-40,000 unique patients[8]), the scale is significantly different than for an AI system. Thus, the impact of any benefits or risks stemming from the use of the AI system is massively scaled – and in just one year of implementation, possibly a thousand-fold or more than the impact any individual clinician can have in their lifetime.

The training and practice of an individual clinician may be optimal for a specific (sub-)population, based on demographics, geographic proximity, and other facts[39] – we define this as *vernacular medicine*. Such vernacular medicine may be less generalizable than is often acknowledged. For an AI system at scale, such optimality may not necessarily be present, depending on training data as well as other factors. While this may increase its value for multiple, but geographically or demographically different groups, it may be less optimized for specific groups, and thus this needs to be considered carefully – we cover this in more detail in the Ethical Considerations section. Privacy, confidentiality and other clinical data security aspects may differ across regions as well. Recently, a concept of federated machine learning has been introduced that allows for an aggregated, scalable AI system to fine-tune from independent training datasets.[40] A more recent concept of federated ML enables remote devices (e.g., mobile phones) to collaboratively engage in model learning and improvement that can take place at a more local level. Such an approach decouples the machine learning from any global training data that would ordinarily be derived from a single discrete storage system. Rather, model training obtains multiple, different, localized, and vernacular, datasets. For deployment, the trained AI model ~~the~~ contains no reference to the local training data that were used to refine/tune the model. This technique, similar to edge-computing, may appear to have benefits. However, there may also be novel risk considerations, relating to algorithm or model iteration that would need to be captured for accurate documentation. These include training data characterization, GMLP, model version and updates, as well as assumption that multiple vernacular datasets are normally distributed can be reduced to a simple distribution function. Probable risks of patient harm and benefits of such a federated approach have not been sufficiently studied.

2.3 AI system outputs

The intended use and impact of an AI system constrains its outputs. According to the IMDRF's definitions of the type of the output (inform, drive, diagnose and treat), as well as the significance of the condition (non-serious, serious and critical), outputs can be categorized as follows[26]:

[‡] See 'Clinical Decision Support Software Guidance'. This explains when a software function qualifies as non-device CDS as well as device CDS, and which of these are actively regulated or for which compliance with applicable regulation would not be enforced, here: https://www.fda.gov/media/109618/download

Type of output	Significance of the condition	Category	Clinical context
Inform	non-serious, serious or critical	Risk prediction	Suggest specific test-types that may be implemented as part of a diagnostic workup of a patient based on clinician suspicion
Drive	non-serious, serious or critical	Likelihood, probability, or prediction of disease	Used by clinician who understands how to interpret the input image (e.g., ophthalmic clinician).
		Saliency, such as highlighting regions of interest or specific lesions in an image.	Used by clinician who understands how to interpret the input image (e.g., ophthalmic clinician).
Diagnose or Treat	non-serious, serious or critical	Disease staging.	Assistive use case: clinician receives specific aspects of the inputs that indicate the disease stage and decides the stage.
		Disease staging.	Autonomous use case: the user receives the disease stage.
		Screening.	Assistive use case: Clinician receives specific aspects of the inputs that indicate abnormalities and decides whether the disease may be present.
		Screening.	Autonomous use case, the user receives output on whether the disease may be present.
		Diagnosis	Assistive use case, a clinician receives specific aspects of the inputs that indicate disease specific abnormalities and the absence of disease specific abnormalities and decides the diagnosis by excluding other disease.
		Diagnosis	Autonomous use case~~, if a specific disease is present the system excludes other diseases and~~ the user receives a diagnosis. An autonomous AI system may evaluate for the presence of a disease or condition and notify the user whether the disease/condition is present *without* showing how the AI system arrived at the decision.

Table 2. AI System outputs

AI system outputs may be aligned with preferred practice patterns or other standards of care, in order to maximize the potential of the AI system to positively impact clinical outcome. This is discussed in more detail in Section 3.2, *Non-maleficence*.

The term *assistive* is usually used for those systems where the clinician makes the ultimate medical decision, and carries liability for the AI performance, while *autonomous* is reserved for those systems where the AI system makes the ultimate medical decision, and the AI creator carries the liability for the AI performance.[6] This distinction, assistive versus autonomous, coupled with intended use, including the significance of the condition, have important bearings on the interpretation of risk as well as other regulatory implications (e.g., clinical study design). The interaction between physician and AI – who risk becoming, as it were, *physicians of the magenta,*[41] and too dependent on monitoring diagnostic AI

devices - is of crucial importance here. Potentially, assistive systems may need subdivision into additional categories that more specifically delineate the roles of human vs AI.

2.4 AI system use environment

The AI outputs, including for whom it is meant, the information provided, etc., may de facto dictate the use environment, including the operator, for the AI system:

Use case setting	**Description**
Home	AI system is used by the patient, and patient images him/herself without clinician or other operator assistance, or imaging by the general home healthcare provider. The output may be provided to the user (patient or home healthcare provider) or may be provided to a remote clinician.
Non specialist (primary care or other non-ophthalmologist)	AI system is used by clinicians and operators, who have minimal experience with imaging the eye or the evaluations of ocular images or other input. The specific interpretation of the image may be important for that clinician to manage the patient in the context of a disease - evaluation of fundus photos for presence of diabetic retinopathy while managing diabetes - or to determine the presence or severity of a systemic disease or disease in another organ system than that being managed, such as determining neurological disease from retinal images.
Specialist (Ophthalmologist or other eye care provider)	AI system is used by clinicians and operators that have experience with ocular imaging and with evaluation of ocular images, but not necessarily with the specific AI output. An example is an AI system for retinal vessel analysis that outputs vascular beading or caliber metrics.

Table 3. AI system use environment

2.5 AI system human factor considerations

Considering the use environment leads to consideration of human factors and impact and outputs of the AI system:

Operator expertise level	patient operated
	untrained operator
	ophthalmic photographer
	Certified ophhalmic photographer
Operator AI assistance level	Differering levels of assistance during the imaging process and protocol. These may include evaluation of image quality evaluation, field, and sequence order.

Table 4. AI System human factors

2.6 AI system inputs

An AI system achieves its intended use through sampling inputs that are analyzed via the algorithm. One goal of an AI system is to obtain a reliable, consistent output while minimizing the number of inputs (samples and types) to help improve robustness of an algorithm to changes in input signal quality and environment, among other factors. For ophthalmic images, inputs can range from image sets from an entire population with multiple images for each member of that population (for population risk assessment) to multiple images from a single patient (for diagnosis). The number and extent of these images are typically dictated by the intended use of the algorithm and the use environment. A non-exhaustive list of input types (image and non-image input types) follows:

Input	**Characteristic**	**Examples**
Image based	Image modality	fundus imaging
		slit lamp photography,
		Optical Coherence Tomography
		ultrasound

		scanning laser ophthalmoscope, topography
		aberrometry
		perimetry (functional)
		multifocal electroretinagram (functional)
		computed tomography, including orbital CT
		magnetic resonance imaging, including orbital MRI
	Image characteristics: while there is currently no required standard, standardization of image metadata such as defined by the DICOM standard 91 [32, 42] will benefit these considerations	Sample area
		x, y or *en face* resolution,
		X, y or depth/axial resolution
		Field of view / area of retina covered
		Number of fields
		Stereo images vs mono
		depth penetration limit
		center wavelength(s)
		momentary pupil diameter
		compression characteristics
		Ambient light level and other environmental conditions
Non-image	Input from modalities that do not meet the definition of a medical device (i.e., that are not FDA-regulated as a medical device), including	patient history
		medication history,
		systemic comorbidities
	Input from modalities that do meet the definition of a medical device (i.e., that are FDA-regulated)	axial eye length
		IOP
		pachymetry
		keratometry
		visual acuity
		heart rate
		blood pressure
		hemoglobin A1C (HbA1c

Table 5. AI system inputs

3 Ethical consideration for AI systems

3.1 Bioethical foundation

In addition to their clinical requirements, such as intended use, human factors, and input and output requirements, as set forward in section 2, AI systems will have to meet ethical requirements to function. This has both practical and philosophical importance: AI systems should follow ethical standards because the field of medicine has defined these standards as guiding principles for the appropriate delivery of healthcare; if AI systems are perceived as unethical or not bound by ethical constraints, stakeholders will not trust these systems, may refuse to engage with them, and this promising technology will fail to reach the populations it is designed to impact.

Consequently, this section introduces the relevant bioethical foundation, [43] and then derives operational ethical dimensions or principles that can be used to create ethical requirements for AI systems.

All healthcare stakeholders, as well as society as large, are already concerned with the use of AI in healthcare, even when they understand the potential efficiency gains. They are concerned with AI systems':

- safety[44]
- actual patient outcome benefit [45]
- mitigation of healthcare disparities rather than worsening them
- potential for racial, ethnic or other inappropriate biases[46]
- usage of patient data, including Personal Health Information, during training and implementation[46]
- misuse, including off label use[47]
- liability, in other words, who can be held accountable or liable for any patient harm [6]

To address these concerns, an ethical framework to identify ethical concerns before they become consequential is considered essential. Several such ethical frameworks for AI[4] and autonomous AI[3] have been proposed and discussed. We focus on the primary bioethical principles of *non-maleficence*, *autonomy* and *justice*, per Beauchamp and Childress. [48] Instead of the term *justice*, which is widely used in the ethics literature, but which may have legal connotations and thus lead to confusion, here we use the more familiar term *equity* instead to describe freedom from bias or favoritism. *Accountability*, while strictly speaking not an ethical concern, leads to related requirements primarily related to *autonomy,* and will be discussed as well.

Such an ethical framework, as developed along the cited publications [3, 4], leads to the following:

a) ethical *requirements* to be created, metrics derived from each of the ethical principles, such as *population achieved sensitivity* which is derived from *equity* below;
b) the insight that it is non-orthogonal, as most ethical metrics are not independent axes, but instead partially overlap. If they were forming independent metrics / axes, this would allow an orthogonal framework;
c) the requirement for a balance to be found or defined between those three ethical principles we focus upon (beneficence, equity and autonomy). Thus, a so-called Pareto optimum needs to be defined, as it is impossible to perfectly meet all 3 ethical principles.

In effect, we use the three bioethical principles as (non-orthogonal) axes, along which to analyze and constrain AI systems, and define their *ethical requirements*. We emphasize that they exist in tension to each other, such that increasing one of them for a particular AI system may decrease another one. For example, for an autonomous AI for the diabetic eye exam, an acceptable balance needs to be found between a) improving access to a disadvantaged population (equity), while b) ensuring that increasing diabetic eye exam compliance leads to overall net-benefit in improvement in care, rather than just increasing diagnoses without access to treatment (non-maleficence), and c) while also maintaining sufficient transparency about the use of AI, training data limitations and data usage, so that patients can decide about their own participation, even if opting-out means losing access to AI benefits (autonomy). [49] Theoretically, health disparities can be mitigated through adjusting the output of the AI system for those patients that are considered advantaged according to some metric. While potentially increasing equity of the AI system, such an approach will likely conflict with non-maleficence and autonomy.

In addition, complicating ethical analyses, is that AI output will itself impact clinical workflows and clinical decisions, both of which may increase tensions between these bioethical axes. Much as

intention-to-treat is standard for Randomized Clinical Trial (RCT) evaluation, the downstream consequences from AI output, will need to be part of any ethical evaluation of a medical AI application.

Put differently, bioethical analysis *per se,* along these dimensions, cannot prescribe the right balance. Rather, it offers a framework to guide and evaluate such decisions. The – "Pareto optimal" - balance between non-maleficence, autonomy and equity, has to be determined by all stakeholders.

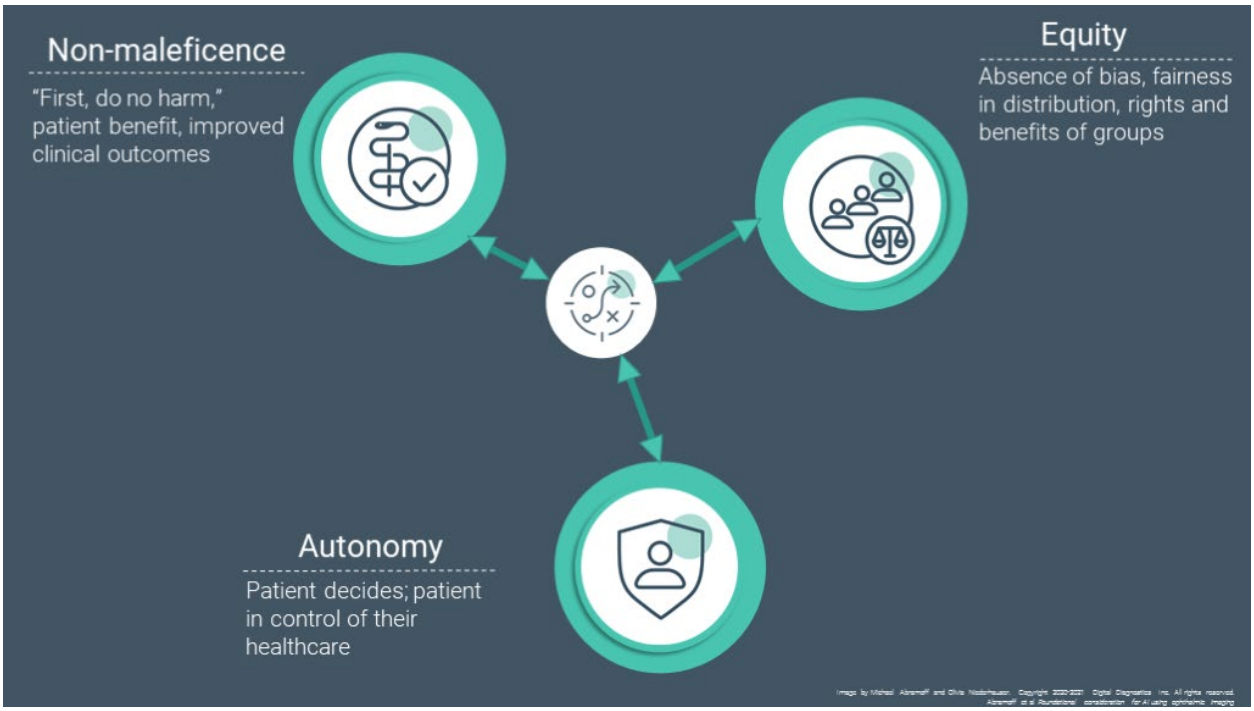

Figure 1. Balance and tension between the three bioethical principles non-maleficence, autonomy, and equity (justice).

Once determined, such a balance results in (ethical) constraints on the design, validation and implementation of AI systems. Thus, we next go through the different bioethical principles, and show how these principles affect AI system requirements. Ultimately, the goal of such ethical requirements is to address and answer the valid existing concerns about AI systems in healthcare that were introduced above.

3.2 Non-maleficence

This principle, "first, do no harm," is often interpreted as *safety* for the individual patient. It affects all aspects of autonomous AI systems including design, validation and implementation. An AI system's risk of harm is affected by intended use, impact, inputs and outputs, and use context, as explained above. However, there are additional considerations unique to AI systems that affect probable risk of harm and are specific to AI and machine learning: design and development, validation, and post-market validation and monitoring. We will explain how these considerations are related to non-maleficence, and how these may lead to more detailed ethical requirements. There are many considerations not unique to AI systems and instead common to all systems involving software that are not discussed here.

3.2.1 Design of the AI system and non-maleficence

In general, AI system design and development share many characteristics with non-AI software systems, and these requirements are laid down in standards like ISO 90003, [50] and for medical devices in ISO 13485.[51] In addition, there exist AI specific design considerations, related to *insight* into the AI, that derive from non-maleficence, and that can affect the risk of more or less harm to the patient. We differentiate three forms of insight that can be assessed for the design: *explainability*, the amount of insight the user (typically the physician) has into the clinical logic that determined the AI output for a specific patient; *transparency*, the amount of insight the user has into the clinical utility of the AI system for all patients; and *validability*, the amount of insight that exists into the non-clinical validity (analytical validity) of the AI system and which can be determined without clinical validation studies. The following are examples of relevant aspects that were discussed by FPOAI, that needs further consideration[52]:

- Transparency – the degree to which the user / clinician of the AI system has insight into the requirements and limitations for the AI system inputs, its training data characteristics, and how the AI outputs are derived from the inputs for the intended use (i.e., for the specific disease or condition).[1, 23] Transparency may also include how the AI system creator uses patient-derived data outside the use for this AI system's intended use. For example, whether patient-derived data can be monetized after the AI system output has been derived. This aspect of transparency primarily serves autonomy, see below.
- Explainability, while fundamentally related to transparency, refers more to how the output is related to clinical practice and scientific literature. For example, is the output clinically meaningful (e.g., diagnosis of a known condition, presence of a particular lesion), rather than something not well understood (e.g., disease severity on a scale that has not been clinically validated or widely recognized)? There are other aspects of transparency, beyond algorithmic functionality in the clinic, such as aspects relating to validation efforts (including analytical validation) that should be transparent to the user to help replicate the measured performance in real-world use. In fact, per the main principles of EQUATOR (which includes the CONSORT-AI extension),[23, 24] complete, accurate, and transparent reporting is an integral part of responsible research conduct. Thus, trial reporting should include a thorough description of the input-data handling, including image acquisition, selection and any pre-processing before feeding into an AI system for analysis. This transparency is integral to the replicability of the intervention beyond the clinical trial in real-world use.
- Validability – the degree to which the validity of the AI system can be assessed without clinical validation studies. That is, to what extent is it possible to self-validate an AI system without going through formal bench or clinical performance validation? This would include aspects like bugs, unresolved anomalies, open loops, etc. that would be found upon inspecting algorithm coding. For cases of black-box systems, there is not as much that can be inspected, which would decrease the overall validability of the system. Thus, validability qualifies our understanding of the analytical performance of the AI system and the impact of other systems on its performance. Examples of this may include the following:
  - AI algorithms structure and infrastructure, including unit level and code analysis, hardware, firmware, operating system
  - Use of federated hardware - dynamically allocated hardware – such as 'the cloud'. As more and more AI algorithms move to cloud-based environments, this may remove execution of the code from the original 'computational infrastructure' (hardware/firmware/software) where it was validated. Federated, or cloud-based, execution environments, for example, using Amazon Web Services (AWS), or Microsoft Azure, on one hand make it easier to have

only a single version of the codebase, rather than multiple different versions, thus enhancing deterministic-ness. On the other hand, the same code may now be executed on a diversity of computational infrastructures. Executing a code fragment may have differing floating point and other operations results, lowering deterministic-ness of the code fragment. Mitigation may require pre-validation of the computational infrastructure for a specific code fragment, or instead, may require constraining the range of computational infrastructure on which such a code fragment may be executed. Pre-validation may maximize a computational infrastructure agnostic approach and thus allow a single codebase globally, as well as providing easier maintenance costs and higher redundancy.

- Inspection of intellectual property that includes source code, patented and copyrighted components. Determining who has authority and expertise to evaluate validability, as well as what can be shared at which level, has large implications for AI creators. Such inspection may include algorithmic correctness verification.[53]
- AI system's use of priors. This may include analysis of whether the AI is designed as a black box (no validability) gray box (limited validability), or detector based (enumerated validability).[3] Here, validability is primarily concerned whether analysis of catastrophic and graceful failures of the AI system shows unanticipated risks - which has been shown to occur more often in black box than detector-based AI systems [54, 55]
- Full characterization of the training datasets at the patient level, which may include partial or full traceability to individual patients, as well as patient demographics and other patient specific characteristics. Compare the amount of information needed for validability to that needed for transparency, which could only require aggregate characteristics to be identified, for validability there could be more strict requirements.

As shown, both explainability and transparency primarily involve the clinically oriented AI system user, while validability primarily involves AI creators, regulators, and non-clinical (technical) AI system users.

3.2.2 Validation of the AI system and non-maleficence

The ethical principle of non-maleficence also leads to requirements for non-clinical and clinical validation or testing of the AI system. Non-clinical testing may include: input data compatibility, discussed below; software verification, including software/firmware description, hazard analysis, software requirements specifications, architecture design specifications, and code traceability.[51]

For clinical validation, common reporting standards, [24] CONSORT-AI,[23] preregistration of study and analysis protocols, [56 10, 57] and validated relationship to patient management[3] are important factors to enhance reproducibility. Given the many concerns about replicability, preregistration of the study protocol, in- and exclusion criteria, and statistical analysis, according to Good Clinical Practice (GCP), [58] or other standards, should be considered. While potentially beneficial, such standards may not provide sufficient information to help inform regulatory evaluation and have not been recognized by FDA. See also FDA's Recognized Consensus Standards. [25] An important decision is whether the AI system is locked before validation, as this affects the external validity and power of any validation study.

The requirements for clinical validation should be commensurate with risk of harm to the patient. Determining the right balance between resource requirements and burden on AI creators for validation on the one hand, and risk of patient harm from AI system usage on the other hand, is essential, in order for patients, patient populations, and the wider healthcare system, to benefit from healthcare AI done the right way.

3.2.2.1 Validation study design

As far as AI validation study design is concerned, prospective longitudinal or cross-sectional designs may be most appropriate for diagnostic AI validation studies. Incorporating as much of the real-world workflow as possible should be considered. Consider the importance of incorporating the actual workflow[59] into AI system validation, and the risk of leaving workflow out, in a purely observational validation study, as first shown by Fenton and colleagues.[3] In this pivotal retrospective cohort study, the outcomes of women undergoing breast cancer screening by a radiologist assisted by a previously FDA cleared (based on a study showing high accuracy of the AI compared to radiologists) assistive AI system, were compared to women who underwent breast cancer screening by a radiologist *without* an assistive AI. [60] When this assistive AI system was evaluated in the setting of actual workflow – where it assists a radiologist who makes the final clinical decision – outcomes were found to be worse for the women who underwent breast cancer screening with AI assistance. This finding and its implications highlight the importance of evaluating such technologies within the intended workflow. This applies to both the validation clinical trial design, as well as through continuing evaluation after actual deployment, as discussed below in 3.2.3. This also aligns well with the FDA's, and other regulatory agencies, trend towards use of real world data and increasing emphasis on continuous efficacy assessment in the post-market phase.

As far as study design is considered, for diagnostic AI prospective longitudinal or cross-sectional designs may be appropriate. Such study designs allow hypothesis testing of the effect of the AI diagnostic on patient outcome, or where diagnosis has already been linked to (untreated) clinical outcome, of the diagnostic accuracy of the AI. For example, diagnostic accuracy hypothesis testing may allow a prospective cohort study design, while outcome hypothesis testing design will likely require a randomized clinical trial design. While a null hypothesis of "no effect" works well in interventional validation studies, a null hypothesis of "not informative" in an RCT may be less desirable for validation of diagnostic AI systems, and especially for validation of autonomous AI systems.[61] [62] Consider that such an RCT needs an arm where the patient management is based on the autonomous AI output, including the need for intervention. To emphasize, in this arm the patient management can only be based on the diagnostic output of the autonomous AI, without the possibility of overruling by a clinician. (If clinician overruling is not ruled out, the effect measured would be that of the clinician and the AI combined rather than of the autonomous AI only). The autonomous AI may incorrectly output a diagnosis leading to not treatment, leaving a subject untreated where treatment would have been beneficial. Whether or not the AI made the incorrect call can only be known when the study is complete. [63, 64] As mentioned, where diagnosis can be linked to outcome, such a design is not necessary and cohort design may be sufficient.

3.2.2.2 Validation study reference standards

Consideration should be given to how AI outputs are validated, in other words, what these outputs are compared against. For a diagnostic AI system, such a comparison will typically be made against an appropriate reference standard, based on its diagnostic indication - informing a healthcare provider or patient, to driving treatment decisions and making a definitive diagnosis. Such reference standards can be categorical or continuous.[65]

From a non-maleficence principle, the effects of the AI system on *clinical outcomes* are most relevant, which may be indirect, as clinical outcomes may likely depend on medical decisions that are neither visible to nor affected by the AI system. Such clinical outcomes include events of which the patient is

aware and wants to avoid, including death, loss of vision, visual field loss, and other events causing a reduction in the patient's quality of life. [66] The resources required to objectively quantify such clinical outcomes can be immense, particularly for chronic disease. In contrast, for acute diseases or interventions, clinical outcomes can be immediate and therefore relatively easier to obtain, such as visual acuity improvement in response to an AI that assists in refraction. For the many chronic diseases to which an AI may be applied, such as diabetic retinopathy, glaucoma or macular degeneration, clinical outcomes may take years to manifest. There has been great interest in the development of alternative outcomes, or *surrogate endpoints*,[67] in the evaluation of investigational medical products to reduce the cost and shorten the duration of trials.

For diagnostic AI, interest in surrogate endpoints has focused on *prognostic standards*, where a patient's disease state has been related to a future clinical outcome. Obviously, these should be validated and directly correlated to clinical outcome.[68] The advantage of a prognostic standard over a surrogate outcome as endpoint is that it is not dependent on clinical decisions outside the intended use of the AI system – in other words its output. For example, within ophthalmology, a prognostic standard is available for diabetic retinopathy, and can be determined by an autonomous diagnostic AI system. However, an expert will make clinical decisions after the diagnosis was determined, such as whether or not to perform laser or deliver anti-vascular endothelial growth factor (anti-VEGF) treatment. Such clinical decisions impact the ultimate clinical outcome, but were not made or influenced directly by the AI system. Thus, there is an advantage of using a prognostic standard, rather than outcome, to evaluate an AI system to not inadvertently diminish or underestimate the benefits of the AI for some decisions outside its control in the context of the clinical outcome.

The Early Treatment of Diabetic Retinopathy Study (ETDRS) severity scale and the Diabetic Retinopathy Clinical Research network (DRCR.net) macular edema scale, as well as the AREDS macular degeneration scale, are representative of such prognostic standards.[69, 70] Ideally, the strength of a prognostic standard is determined by the evidence available to support its capacity to predict progression, or manifestation of a condition or disease, or the benefit of a treatment or management. Its strength is also determined by any evidence that shows that treatments based on the prognostic standard correspond to effects on clinical outcome.[71, 72] As the ETDRS, Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study (DCCT/EDIC),), and DRCR studies have established such evidence extensively, this applies to these prognostic standards.

While requiring less time and fewer resources than developing and validating clinical outcomes, quantifying prognostic standards may still require considerable effort. While dependent on the intended use, for autonomous diagnostic AI studies, this is likely an important reason why clinician-derived reference standards, instead of prognostic reference standards, are widely used in AI validation.[73]A widely cited meta-analysis of the quality of evidence of AI accuracy, while mentioning the potential of AI to improve outcome - takes as a given comparison to clinician-derived ground truth – and the relationship to prognostic standards or clinical outcome is not considered. [22] And indeed, it is a major strength of the Collaborative Community CCOI and its disease-specific subgroups, that it has started discussing the development of such prognostic standards for disease areas of interest.

Other factors that should be considered when evaluating potential reference standards - in addition to their validity or lack of that against outcome -:

- *reproducibility* of the reference standard – many studies have shown that multiple clinicians evaluate the same patient differently in 30-50% of cases[74-76]
- *repeatability* – many studies have shown that the same clinician evaluates the same patient differently in 20-30% of cases; [74-76]
- *diagnostic drift* – studies have shown that clinicians from different regions, countries or continents evaluate the same patient differently up to 50%, leading to *vernacular medicine* as explained above [39]
- *temporal diagnostic drift* – studies have shown that clinicians systematically evaluating the same hypothetical patient differently over generations of clinicians.[77]

As the evidence for a given treatment based on a given evaluation may have been derived decades ago, temporal drift may be hard to determine and difficult to correct for. We want to clarify that while temporal *drift* will typically be pernicious and undesirable, temporal diagnostic *shift*, where new, and better treatments led to a new prognostic standard, is often desirable. An example of temporal shift is the shift from the prognostic standard "clinical significant macular edema as defined by ETDRS" to the new prognostic standard "center-involved macular edema" which is derived from OCT, not fundus photographs, and was developed in conjunction with evaluating with novel anti-VEGF treatments for macular edema.[70, 78] Optimally, correction for reproducibility and repeatability with strict evaluation protocols and independent verification where possible is indicated. [66]

Given these considerations, and depending on an AI system's role, output type, SAMD risk categorization and risk of harm to the patient, certain types of reference standards may be differentiated based on the rigor or validity of the reference standard (Table I). While such a hierarchy, as shown here, may be useful for consideration of reference standard differences, there is no required level for a specific intended use. Generally, these Levels I-IV can be related to their rigor, with Level I as having the most rigor. Typically, an AI system that carries more risk of harm, such as personalized treatment (e.g., Artificial Pancreas), as a stand-alone diagnosis, or determination of disease level used in treatment decisions, would be compared to a more rigorous standard. It, therefore, remains up to regulatory agencies around the world to balance the intended use and risk category of the AI system, and potentially include the reference standard Level in this balance.

- **Level I:** A reference standard that is either a prognostic standard, clinical outcome, or a biomarker standard. If a prognostic standard, it is determined by an independent reading center. If either a prognostic standard or a biomarker, it is validated against clinical outcome, and temporal drift, reproducibility, and repeatability metrics are published.

- **Level II:** A reference standard established by an independent reading center. Temporal drift, reproducibility, and repeatability metrics are published. A level II reference standard has not been validated against clinical outcome or a prognostic standard.

- **Level III:** A reference standard created from the same modality as used by the AI, by adjudicating or voting of multiple independent expert readers. The readers are documented to be masked, and reproducibility and repeatability metrics are published. A level III reference standard has not been validated against clinical outcome or a prognostic standard, and does not have known temporal drift.

- **Level IV:** All other reference standards, created by single readers or non-expert readers, and may be without an established protocol. A level IV reference standard has not been validated

against clinical outcome or a prognostic standard, does not have known temporal drift, reproducibility or repeatability metrics, and the readers may not have been masked.

Table 6: reference standard levels

For level I and II reference standards, there is no reference to modality, as the modalities are entirely determined by the requirements for outcome, prognostic standard or reading center.

While a higher-level reference standard may at first glance seem more desirable, in many cases, this may not be the preferred choice. Such a higher level may not be available or even unachievable, and the requirement for a higher level needs to be balanced with the burden to obtain it. An example is Retinopathy of Prematurity (ROP), where the only prognostic standards are derived from expert clinicians – i.e., level II, are available. At this point in time, it is ethically impossible to determine a level I standard for ROP, and in fact Level II is the accepted reference standard in the clinical community. Collecting a Level I standard would require a study that may leave some treatable patients untreated, depending on how accurate the AI under study actually is, and thus requiring a level I for an AI creator would be an undue burden, and frankly an impossible hurdle to overcome.

It is worth reemphasizing that a) the level of reference standard is entirely independent from the AI system or its intended use; b) different intended use cases may require different levels of reference standard; and c) the level of reference standard is evaluated entirely separate from the minimally acceptable criteria for performance of the AI. The minimally acceptable criteria can only be understood for a given reference standard level.

3.2.2.3 Minimal acceptable criteria for validation

The minimal acceptable criteria for the AI system are the decision cutoffs for determining the safety, and efficacy of the AI, in hypothesis testing clinical trials, to estimate non-maleficence. Such minimal acceptable criteria include combinations of sensitivity, specificity, area under the Receiver Operator Characteristics (ROC) curve. While the concept of decision cutoffs for safety and efficacy of an AI system may be broadly accepted, it is also a major factor in the review processes by regulatory agencies.

As an example, for the first autonomous De Novo AI authorized by FDA, two hypotheses had to be confirmed in a preregistered clinical trial, with sensitivity, specificity characteristics all exceeding 80% at the population level. This corresponded to study-based endpoints of 85%for sensitivity andand 82.5% for specificity.[13]

Theoretically, such minimal acceptable criteria can be derived analytically, with the goal to minimize subjectivity and maximize external validity. Thus, approaches have been developed to come up with analytical solutions for diagnostic algorithm endpoints, including Pareto optimization, Youden and Euclidean indices for sensitivity – specificity combinations[79-82], quantitative cost-benefit derivative analysis, as well as (modified) Angoff approaches.[83, 84] Specifically, the (modified) Angoff approach has been validated for setting testing thresholds in educational settings. These analytical approaches are helpful in informing the choices to be made, by improving understanding of risks and benefits of any choices made.

Alternatively, minimal acceptable criteria for a diagnostic AI can be set to conform to existing diagnostic procedures. For example, when excluding pulmonary embolism, test negatives should have a 3-month thrombo-embolic risk of less than 3%, which is derived from the equivalent risk after a negative pulmonary angiography, the gold standard.[85] Understanding of the accuracy of comparable diagnostic processes performed by human clinicians and other human experts should be a requirement (note that the current diagnostic standard-of-care may not necessarily involve a clinician in the future as AI systems may, at some point, be considered as standard-of-care).

In contrast, the existing literature does not offer guidance on these minimal acceptable criteria for an autonomous AI performing the diabetic eye exam, as the standard of care by ophthalmologists only reaches 33% or 34% sensitivity.[75, 76]

Given such widespread lack of scientific evidence for specific minimal acceptable criteria, deciding these minimal acceptable criteria involves ethical, cost-effectiveness and other risk-benefit trade-offs by patients, clinicians, and payors. Such decisions will typically require the involvement of domain experts. As examples, minimally acceptable criteria for screening mammography were previously determined by a set of domain experts using a modified Angoff approach;[84] and a sampled survey of pediatricians was used to estimate the minimally acceptable sensitivity threshold for a 'streptococcal pharyngitis test' in children. [86] For such approaches to work, it is important that the experts involved fully grasp the spectrum of risks and benefits for patients of each alternative set of criteria. This may not always be the case: in the latter study, 80% of pediatricians proposed a sensitivity of at least 95%, which was not achievable by any feasible test under consideration.[86] The structured collection of patient preferences, , also known as Patient Preference Information (PPI) could also be included in shaping these decisions.[87] Thus, the following stages can be considered in isolation or in the aggregate for setting minimal acceptable criteria:

- Literature or meta-analysis review of existing minimal acceptable criteria, and assignment of weights to the consequences of test misclassifications, according to one or more metrics such as cost or quality-adjusted life years (QALY)). As an example, estimate whether the consequences of missing a case, such as increased morbidity or cost at a later stage when the disease manifests more clearly, outweigh the consequences of misclassifying a non-case as a case, such as unnecessary radical diagnostic or treatment decisions with major side-effects. Scientific evidence of comparable diagnostic processes, performed by human clinicians and other human experts, should be included if available, or may need to be collected, if not available.
- Analysis of a representative spectrum of sensitivity and specificity combinations, and determination of the downstream cumulative weight of consequences for patients[88, 89] and other stakeholders in the healthcare system, including PPI.
- A process of domain experts (e.g., Network of Experts)[90] can potentially generate consensus on minimal acceptable criteria. For example, using vignettes that condense analytical evidence, to ensure minimal bias among domain experts.

3.2.3 Post-market monitoring of AI systems and non-maleficence

Monitoring of the safety and efficacy of an AI system is important because it affects non-maleficence. Real-world performance monitoring after implementation can be achieved by putting a prospective monitoring protocol in place. Such a prospective monitoring protocol may be agreed upon by a regulatory agency, for example, implemented as part of a comprehensive Quality Management System

following 21 CFR 820, and accommodate user feedback, complaints, and reportable events. In addition, other AI system characteristics that are within creators' control such as usability, user experience, product performance, and necessary safety controls, including a comprehensive framework for cybersecurity, data protection, and data privacy, may also be monitored.

To ensure continued acceptable performance of an adaptive AI system, for example, a prospective monitoring protocol may require the AI system output to be compared to the same reference standard that was used in (pre-market) validation to be able to determine whether or not it still meets safety and efficacy standards in the post implementation real world. As discussed above, more rigorous, higher level, reference standards often require substantial resources, for patients and creators. Real world monitoring may require the collection of this reference standard for each monitored patient – which may diminish the reasons why the AI system was implemented in the first place, such as improved access, lower cost, and patient friendliness. Thus, prospective monitoring protocols will have to find a balance between burden on AI creators, as well as patients, on the one hand, and non-maleficence on the other.

3.2.3.1 Changing an AI system after validation

As an AI system is used on patients and continuous efficacy monitoring is in effect, there will be opportunities to improve the AI system technical specifications in terms of safety, efficacy, and/or equity (see Equity section). AI systems – in other words, SaMD that use AI or machine learning - have the unique capacity to be updated after implementation. In fact, if an AI system is not locked after validation, there is also potentially unlimited configurability.

It is important to determine that changes to the technical specifications, while intended to improve the AI system, do not negatively affect the ethical principle of non-maleficence. Traditionally, from a regulatory perspective, almost all technical specification changes to a SaMD that affect safety or effectiveness may require a new validation; cybersecurity changes may be the only ones currently possible without such full validation, depending on how one interprets current [91]FDA guidance.[91]

Thus, safely updating the AI system requires that appropriate controls and validation methodologies are in place. These controls and methodologies will be dependent on both the *type* of change, as well as on the *risk of patient harm*, and we differentiate the following types of changes:

- *Changes to AI system computations*
  These include
  - Changes to pre- and postprocessing algorithms
  - Changes in algorithmic infrastructure, including hardware and software.
  - Changes to AI algorithm architecture, including to improve performance
    - Types of classifiers
    - Hyperparameter and parameter (including model weights) values
    - Training data
- *Changes to AI system Input*
  While keeping the output type constant, these are some examples of such changes
  - Change in imaging system, such as optical, sensor, image compression, and imaging protocol;
  - Adding other information about the patient to the inputs, such as pulse, VA, IOP, that are used *along with* the original image by the AI to make a final determination.
- *Changes to the AI system Output e*

While keeping the input types unchanged, examples inclue
    - marking regions of interest when previously only a normal/abnormal output was validated
- *Changes to AI system indications and intended use*
  An example of *change to intended use* is accumulating scientific evidence that an AI system that was validated as a "referral tool" and authorized by a regulatory agency as such, can actually be used as a "diagnostic tool": as it becomes more accepted in the clinical community and its performance thresholds are adjusted to support such use. Other examples include
    - Inclusion or exclusion criteria - such as expansion to people with different risk of having the disease, age groups, ancestry, race or ethnicity that were not accounted for in the design or validation of the AI system to be improved.
    - Disease level or threshold
    - Disease type, for example macular degeneration, when previously validated for diabetic retinopathy.

An important component of AI system changes is the method of change validation that is used to establish safety, efficacy and equity of the changed AI system. AI systems may differ in the data that was collected for their validation. At one end of the spectrum of validations, a recent autonomous AI system required a full preregistered clinical trial – a pivotal trial - comparing against a prognostic standard.[13] Depending on the patient risk of harm, and the type of change, as set forth above, the following categories of such methods can be discerned. As an aside, many of these methods require the pivotal trial data, of the *index AI system*, to have been escrowed under a so-called *algorithmic integrity protocol*[13]:

- *Regression identity testing*, to establish, non-probabilistically, that for any input data, changes to the AI system do not result in any change whatsoever in diagnostic output.
- *Bench validation*, to formally test the statistical hypothesis that a change that can impact the AI algorithm, for example a change in GPU, has no impact on the diagnostic output for any input from a given group of subjects.
- *Recursive validation,* to formally test the statistical hypothesis that for example, a change in Input Type, such as a change in imaging system, has no impact on the diagnostic output, compared to the index AI system output. Recursive validation uses the index AI system output as the reference standard. [92]It is similar to a reproducibility study,[92] where the output of the index AI system is compared to a modified AI system with the inputs slightly perturbed.
- *Performance (safety, effectiveness, and equity) bracketing.* Analytically, the maximum change in performance metrics caused by a specific change in the algorithm can be calculated and bounded quantitatively; can be used to ensure maximum change continues to be within expectations and also exceed the minimally acceptable criteria that were determined for the pivotal trial.
- *Escrowed validation study iteration,* to statistically test the hypothesis, that an *AI system* is not inferior, or possibly superior, to the index AI system. This can be achieved by reusing the inputs of the index AI system validation dataset that were previously *escrowed*, and comparing the outputs of the changed system to that escrowed, established reference standard. There are limits on the number of iterations that can be achieved, as explored by Ioannidis et al, [93] as each dataset reuse increases the potential for overfitting to the escrowed validation data.[94] The degree to which escrowed dataset reuse leads to false positive claims and overfitting can be quantified through systematic frameworks, including the dataset positive predictive value (DPPV) framework. The success of this approach depends on parameters including the number

of available escrowed validation subjects, type 1 and type 2 error rates, and degree of dependence between outputs of the index AI system and the modified AI system. The validation study needs to have been *escrowed* as part of the preregistration algorithm integrity process for this to be a valid methodology.[13, 95]

- *Escrowed Validation study expansion,* to statistically test the hypothesis that the AI system is not inferior or possibly superior, due to a change in target patient population. Escrowed validation study expansion reuses the inputs of the index AI system validation dataset that has been *escrowed*, expands this dataset with subjects from the new target patient population, and then compares the outputs of the changed system to the reference standard. Either the identical workflow can be used, or a secondary analysis on the effect, if any, of a change in workflow is required. As new subjects are added to the original study for this expansion, information is gained and this may compensate for the information loss and risk of overfitting from dataset reuse.[95] As with Escrowed validation study iteration, it is critical to monitor the overall degree of dataset reuse.

Where *"index AI system"* refers to the AI system that was validated in a pivotal trial. The term *"escrowed under an algorithm integrity protocol"* implies that the human subject input data (including the corresponding reference standard), collected in this pivotal trial, is kept inaccessible by a third independent party. Thus, there will be a complete, arms length, documented record of any access or use of this data by the index AI system developer, for example for retraining a modified AI system, somewhat analogous to the concept of *clinical trial preregistration*.

The above studies can, ofcourse, be performed both by the AI creator, or by independent research groups.

3.3 Autonomy

Analysis of autonomy of the patient with respect to AI leads to at least two important considerations. First, considerations of the use of patient-derived data, which applies to both training data for the AI system algorithms, as well as to implementation, where the AI system collects this data to determine its outputs.

Transparency may include how the AI system creator uses patient-derived data outside the use for this AI system's intended use. An example is insight into whether patient-derived data is monetized for other purposes than the diagnosis by the AI. Autonomy is greater when the collection of patient-derived data is lawful and in compliance with laws and regulations and best practices. This may include compliance with the Health Insurance Portability and Accountability Act (HIPAA), the Health Information Technology for Economic and Clinical Health ((HITECH) act, and other data security aspects of 21 CFR 50, the Declaration of Helsinki, as well other statutory and regulatory rules in place, in a manner that is transparent about the purpose and scope for which the data will be used.[96] Ideally, patient-derived data used by AI creators is traceable to patient authorization to use that data. Those involved in the design of AI systems should have accountability with respect to protecting patient rights as stewards of patient-derived data. Auditable processes and security controls aid in ensuring that patient data is being used in accordance with the scope for which it was authorized, and to protect the data from unauthorized use or access.

A current controversy is the reward or recognition of clinicians contributing a reference standard to patient-derived data incorporated in the intellectual property of an AI system. Such contributions may

include their diagnostic work recorded in medical records, subsequently used to train or evaluate an AI system. [97] Such ownership collides with rising public desire for increased control over, and privacy with respect to, electronic data and emerging regulations to address these (General Data Protection Regulation (EU) 2016/679 (GDPR), and the California Consumer Privacy Act, Cal. Civ. Code § 1798.100 et seq.), as well as increasing patient activism for recognition for contributions to scientific advances.

Secondly, liability for the AI system malfunction is related to autonomy. Abramoff, *et al,* previously proposed that creators of *autonomous* AI systems assume liability for harm caused by the diagnostic output of the device when used properly and on-label.[3] In their paper, they state that this is essential for adoption – it may be inappropriate for clinicians using an autonomous AI, to make a clinical decision they are not comfortable making themselves, to still have full medical liability for harm caused by that autonomous AI. This view was recently endorsed by the AMA in its 2019 AI Policy.[6] Such a paradigm for responsibility is more complex for assistive AI, where medical liability may fall only on the provider using it – as they are ultimately responsible for the medical decision, or on a combination of both, where even the relative balance sheets of the AI user and the AI creator come into play.

Meanwhile, as Abramoff, *et al,* proposed elsewhere,[3] medical decisions by autonomous AI on *individual patients* typically cannot be unequivocally labeled as correct or incorrect, especially in chronic diseases where outcomes may emerge years later. On *populations of patients* however, the medical decisions can be compared statistically to the desired decisions, for example to claimed correctness, and it is thus there the liability will be focused. Another issue is that, while autonomous AI is preferably compared to patient outcome, or prognosis, these comparisons require enormous resources that will be not available for the individual patient where liability is at stake. Instead, the autonomous AI decision may be compared to an individual physician or group of physicians, lacking validation and thus, with unknown correspondence to outcome or surrogate outcome. As an aside, this can be an issue also for so-called continuous learning AI systems.

These distinctions will need to be resolved as various AI applications move forward. The legal responsibility for an AI system built in partnership with a large healthcare system and intended to be used on its patient population is, by definition, more diffuse and likely to vest in the sponsoring healthcare system or with some comparative or contributory analysis of fault. A privately designed system, sold as a finished product, may need to bear its own responsibility for autonomous output, absent superseding or intervening causation.

Responsibility for proper use and maintenance of the AI system, consistent with terms of service and FDA or other regulatory agency labeling, remains with the providers – the practice of medicine.

Finally, the output of the autonomous AI system, while valid as a diagnostic record from a regulatory perspective, is not currently defined as a medical record, when it is not signed off on by a physician. What is and is not, and who can and cannot, create a medical record is determined, in the US, primarily by the State Medical Boards or their equivalent. At present, such Boards do not consider an autonomous AI output to have the same medicolegal status as physician documentation, and the legal status of reports generated by AI has been brought to the attention of the US Federation of State Medical Boards.

3.4 Equity

The third bioethical principle is Equity – we mentioned above that we use this term rather than the traditional bioethical term of "Justice" for this concept. Equity primarily concerns itself at the impact on the patient at the population level, beyond the impact on an individual patient.

In the context of AI, this translates to estimating its differential impact of safety, or any other characteristics of the AI system, on members of a group with respect to members of other groups, called "health disparities".For example, inappropriate bias of the AI system may result in the AI system being less effective for some group, based on race,[98, 99] ethnicity, sex,[100] age, income, and other categories, than another, even though on average it was found to be safe. Any medical process has the potential to either increase or decrease health disparities, depending on how it is used. Because of the scale at which AI systems operate, their potential to increase or decrease disparities is also tremendously magnified.

Inappropriate bias, increase in health disparities, and thus decreased Equity can exist across the entire AI pipeline, as Char, *et al*, outlined,[4] including in the choice of intended use of the AI, its design, its validability, its validation and the choice of reference standards, as well as how, and where, it is implemented. For example, as far as design of the AI is considered, lower validability of a black box algorithmic approach, makes bias harder to anticipate, detect and mitigate, when it replaces explicit priors with properties that cannot be analyzed and evaluated. Another example with respect to design, incomplete or unrepresentative training data, or relying on complete and representative data that reflects and reproduces (at scale) pre-existing healthcare bias, increases the risk worsening health disparities. As far as validation is concerned, selection of study sites, biased inclusion and exclusion criteria, can all decrease validity for certain subgroups, and thereby exacerbate health disparities. Finally, as far as implementation is concerned, implementation of the AI system preferably in some locations may affect access to disadvantaged groups, again increasing health disparities.

When analyzing validation, this can be used to estimate equity by determining or testing for the presence or absence of an effect of predefined characteristics of subgroups on the characteristics of the AI system, such as sensitivity or specificity. Such characteristics will typically include race, ethnicity, age and gender on sensitivity and specificity.[101] In addition, differential usage in subgroups will affect equity, and such effects can be compared using metrics like population achieved sensitivity (see below).[98]

As mentioned, when analyzing the Equity dimension of an AI system, particularly in the context of health disparities, it is useful to consider the implementation context. Different diagnostic processes, including AI systems, may differ in patient friendliness, availability, access and direct and indirect cost, even with equal sensitivity and specificity (i.e. equally high non-maleficence).

With respect to intended use and implementation in the context of Equity, the goal of the diagnostic process at the population level, is to identify the maximum number of true disease cases identified in that population. A given diagnostic process, like a high performing AI system, may have a high sensitivity – in other words, non-maleficence is maximal for those patients that have access. However, if for example this AI system is only available in one place, the number of cases identified will not be maximal, as many in the population will simply never undergo its diagnostic process, and Equity is much lower.

Population achieved sensitivity, or *Access corrected sensitivity,* is used to analyze such effects on Equity. In other words, while an AI system – and any diagnostic process - with very high sensitivity is attractive

from an individual perspective, if only few people have access to the diagnostic AI, the population achieved sensitivity PAS, or effective sensitivity at the population level, will be much lower, and concomitantly its equity:

$$PAS = \frac{s_c c p_c}{c p_c + (1-c)\widehat{p_{nc}}} \cong s_c c$$

Where
$s_c = sensitivity\ (as\ determined\ in\ adherent\ population)$
$c = compliance$
$p_c = measured\ prevalence\ \ in\ the\ adherent\ population$
$\widehat{p_{nc}} = estimated\ prevalence, in\ the\ nonadherent\ \ population$

When we assume $p_c \cong \widehat{p_{nc}}$, i.e. the prevalence of the disease is the same in the non-adherent as in the adherent part of the population, we can use the simplified estimate $s_c c$. [102]For example, if compliance *c*, with the diabetic eye exam, is 15%,[102], and the minimal acceptable sensitivity is 85%,[13] the population achieved sensitivity (PAS) $= 0.13$. In other words, only 13% of all cases in the population will be identified correctly with this diagnostic system. In many cases, the prevalence in the part of the population that does not undergo the AI system is actually *higher* than in the adherent population, so that this estimate of PAS forms an upper bound. It is useful to consider PAS in determining minimal acceptable sensitivity. A more accessible AI system may have lower s_c but still result in higher PAS as adherence can expected to be higher.

4. Conclusions

The considerations in this article are a useful first step in the development of a bioethically sound foundation, based in non-maleficence, autonomy and equity, of considerations for the design, validation and implementation for AI systems CCOI's FPOAI exceptional and diverse experience means it is well placed to develop and evaluate such a foundation. Considerations of FPOAI's future consensus statements and cooperation among AI creators, industry, ethicists, [3, 4] clinicians, patients, and regulatory agencies, is key to facilitating rapid innovation of AI technologies and their successful implementation in clinical medicine. Such global collaboration will adhere to bioethical principles, guide development and use of clinical AI, helping to make fundamental improvements in accessibility and quality of health care, decreasing disparities and lowering the overall cost of health care.

Appendix A

FPOAI members as of writing

Michael Abramoff, MD, PhD (Chair, University of Iowa)

Malvina B. Eydelman, MD, (CDRH, OHT-1, US Food and Drug Administration)

Brad Cunningham, MSE (CDRH, OHT-1, US Food and Drug Administration)

Bakul Patel, MBA (CDRH, DHCoE, US Food and Drug Administration)

Karen A. Goldman, PhD, JD (OPP, US Federal Trade Commission)

Danton Char, MD MS (Stanford University, CA)

Taiji Sakamoto, MD (Kagoshima University, Japanese Ophthalmological Society)

Barbara Blodi, MD (Department of Ophthalmology, University of Wisconsin)

Risa Wolf, MD (Department of Pediatrics, Johns Hopkins University)

Jean-Louis Gassee (Apple)

Theodore Leng, MD, MS (Department of Ophthalmology, Stanford University School of Medicine)

Dan Roman (Director Diabetes Measures, National Committee of Quality Assurance)

Sally Satel (Yale, AEI, Data usage ethics)

Donald Fong (Kaiser Permanente)

David Rhew (Chief Medical Officer, Microsoft)

Henry Wei (Google Health)

Michael Willingham (Google Health)

Michael Chiang, MD, PhD (Director, National Eye Institute)

Mark Blumenkranz, MD (Facilitator, Stanford University)

While the members' main affiliations are stated, they do not in every case represent their institution/company.

CCOI Executive Committee

Michael Abramoff, MD, PhD

Mark Blumenkranz, MD

Emily Chew, MD

Michael Chiang, MD

Malvina Eydelman, MD

David Myung, MD, PhD

Joel S. Schuman, MD

Carol Shields, MD

References

1. U.S. Food & Drug Administration (FDA) Digital Health Center of Excellence C, ,. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. https://www.fda.gov/media/145022/download
2. Stanford University. Collaborative Community on Ophthalmic Imaging (CCOI). 2020:https://www.cc-oi.org/.
3. Abramoff MD, Tobey D, Char DS. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *Am J Ophthalmol*. 2020;214(1):134-142. doi:10.1016/j.ajo.2020.02.022
4. Char DS, Abràmoff MD, Feudtner C. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *The American Journal of Bioethics*. 2020/11/01 2020;20(11):7-17. doi:10.1080/15265161.2020.1819469
5. Abramoff MD. The autonomous point of care diabetic retinopathy examination. In: Klonoff DC, Kerr D, Mulvaney SA, eds. *Diabetes Digital Health*. Elsevier; 2020:chap 12.
6. American Medical Association (AMA) Board of Trustees Policy Summary. Augmented intelligence in healthcare. 2019. Updated 4 Nov. 2019. https://www.ama-assn.org/system/files/2019-08/ai-2018-board-policy-summary.pdf
7. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA*. Jun 18 2019;321(23):2281-2282. doi:10.1001/jama.2019.4914
8. Autonomous AI in Action. https://www.forbes.com/sites/oraclecloud/2020/01/16/autonomous-in-action-self-driving-cars-get-all-the-publicity-but-other-industries-are-already-getting-exceptional-value-from-ai-based-systems/#1ecc65d86e94

9. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. *Ann Fam Med*. Nov-Dec 2014;12(6):573-6. doi:10.1370/afm.1713
10. U.S. Food & Drug Administration (FDA). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. 2018. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm
11. American Diabetes A. 11. Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes-2020. *Diabetes Care*. Jan 2020;43(Suppl 1):S135-S151. doi:10.2337/dc20-S011
12. Ting DSW, Peng L, Varadarajan AV, et al. Deep learning in ophthalmology: The technical and clinical considerations. *Prog Retin Eye Res*. Sep 2019;72:100759. doi:10.1016/j.preteyeres.2019.04.003
13. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Nature Digital Medicine*. 2018/08/28 2018;1(1):39. doi:10.1038/s41746-018-0040-6
14. Gensure RH, Chiang MF, Campbell JP. Artificial intelligence for retinopathy of prematurity. *Curr Opin Ophthalmol*. Sep 2020;31(5):312-317. doi:10.1097/ICU.0000000000000680
15. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs. *Ophthalmology*. Apr 2019;126(4):565-575. doi:10.1016/j.ophtha.2018.11.015
16. Christopher M, Belghith A, Weinreb RN, et al. Retinal Nerve Fiber Layer Features Identified by Unsupervised Machine Learning on Optical Coherence Tomography Scans Predict Glaucoma Progression. *Invest Ophthalmol Vis Sci*. Jun 1 2018;59(7):2748-2756. doi:10.1167/iovs.17-23387
17. Kaiserman I, Rosner M, Pe'er J. Forecasting the prognosis of choroidal melanoma with an artificial neural network. *Ophthalmology*. Sep 2005;112(9):1608. doi:10.1016/j.ophtha.2005.04.008
18. Siddiqui AA, Ladas JG, Lee JK. Artificial intelligence in cornea, refractive, and cataract surgery. *Curr Opin Ophthalmol*. Jul 2020;31(4):253-260. doi:10.1097/ICU.0000000000000673

19. Yu F, Silva Croso G, Kim TS, et al. Assessment of Automated Identification of Phases in Videos of Cataract Surgery Using Machine Learning and Deep Learning Techniques. *JAMA Netw Open*. Apr 5 2019;2(4):e191860. doi:10.1001/jamanetworkopen.2019.1860
20. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. Oct 2019;1(6):e271-e297. doi:10.1016/S2589-7500(19)30123-2
21. Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MM. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology*. May 2013;267(2):581-8. doi:10.1148/radiol.12120527
22. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. Mar 25 2020;368:m689. doi:10.1136/bmj.m689
23. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*. 2020/09/01 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x
24. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. Nov 14 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799
25. U.S. Food & Drug Administration (FDA) CDRH. Recognized Consensus Standards. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfstandards/search.cfm
26. U.S. Food & Drug Administration (FDA); International Medical Device Regulators Forum. SOFTWARE AS A MEDICAL DEVICE (SAMD): CLINICAL EVALUATION. 2016.
27. Botkin JR, Goldenberg AJ, Rothwell E, Anderson RA, Lewis MH. Retention and research use of residual newborn screening bloodspots. *Pediatrics*. Jan 2013;131(1):120-7. doi:10.1542/peds.2012-0852
28. All of Us Research program NIoH, ,. "Ethical Considerations in the All of Us Research Program." 2018. 2021. https://www.bioethics.nih.gov/courses/pdf/2018/session5_blizinsky.pdf
29. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Oct 25 2019;366(6464):447-453. doi:10.1126/science.aax2342
30. Cavallerano J, Lawrence MG, Zimmer-Galler I, et al. Telehealth practice recommendations for diabetic retinopathy. *Telemed J E Health*. 2004;10(4):469-482.
31. Abramoff MD, Leng T, Ting DSW, et al. Automated and Computer-Assisted Detection, Classification, and Diagnosis of Diabetic Retinopathy. *Telemed J E Health*. Apr 2020;26(4):544-550. doi:10.1089/tmj.2020.0008
32. Digital Imaging and Communications in Medicine (DICOM) Standard. Supplement 91: Ophthalmic Photography Image SOP Classes. Rosslyn, VA, USA National Electrical Manufacturers Association (NEMA).
33. Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care. GAO-21-7SP. (US General Accounting Office,) (2020).
34. International Medical Device Regulators Forum - Software as a Medical Device (SaMD) Working Group. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. 2014. http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf
35. U.S. Food & Drug Administration (FDA). How to Determine if Your Product is a Medical Device. 2018. https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device
36. Digital Therapeutics Alliance (DTA). DIGITAL HEALTH INDUSTRY CATEGORIZATION. 2019.

37. U.S. Food & Drug Administration (FDA) CfDaRHaCfBEaR, ,. Changes to Existing Medical Software Policies Resulting from Section 3060 of the 21st Century Cures Act. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/changes-existing-medical-software-policies-resulting-section-3060-21st-century-cures-act
38. U.S. Food & Drug Administration (FDA) CDRH. Clinical Decision Support Software Draft Guidance for Industry and Food and Drug Administration Staff. 2019. https://www.fda.gov/media/109618/download
39. van Dijk HW, Verbraak FD, Kok PH, et al. Variability in photocoagulation treatment of diabetic macular oedema. *Acta Ophthalmol*. Dec 2013;91(8):722-7. doi:10.1111/j.1755-3768.2012.02524.x
40. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J Biomed Inform*. Nov 2019;99:103291. doi:10.1016/j.jbi.2019.103291
41. Geer D. Children of the Magenta. *IEEE Security & Privacy*. 2015;13(05):104-104. doi:http://doi.ieeecomputersociety.org/10.1109/MSP.2015.91
42. Lee A, Campbell J, Hwang T, Lum F, Chew E. Recommendations for Standardization of Images in Ophthalmology. *Ophthalmology [in press]*. 2021;
43. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med*. Mar 15 2018;378(11):981-983. doi:10.1056/NEJMp1714229
44. Kent J. Artificial Intelligence Falls Short in Detecting Diabetic Eye Disease. *Health IT Analytics*. 2021;
45. Artificial Intelligence (AI) Health Outcomes Challenge (2019).
46. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. Mar 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370
47. U.S. Food & Drug Administration (FDA) CDRH. Understanding Unapproved Use of Approved Drugs "Off Label". 2018. https://www.fda.gov/patients/learn-about-expanded-access-and-other-treatment-options/understanding-unapproved-use-approved-drugs-label
48. Beauchamp TL, Childress JF. *Principles of biomedical ethics*. Eighth edition. ed. Oxford University Press; 2019:pages cm.
49. Gayle HD, Childress JF. Race, Racism, and Structural Injustice: Equitable Allocation and Distribution of Vaccines for the COVID-19. *Am J Bioeth*. Mar 2021;21(3):4-7. doi:10.1080/15265161.2021.1877011
50. International Organization for Standardization (ISO). ISO/IEC/IEEE 90003:2018 Software engineering — Guidelines for the application of ISO 9001:2015 to computer software. 2018.
51. International Organization for Standardization (ISO). ISO 13485:2016 Medical devices — Quality management systems — Requirements for regulatory purposes. 2016.
52. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. Sep 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
53. Yang Y, Rinard M. Correctness Verification of Neural Networks. 2019:arXiv:1906.01030. Accessed June 01, 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190601030Y
54. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. Mar 22 2019;363(6433):1287-1289. doi:10.1126/science.aaw4399
55. Shah A, Lynch S, Niemeijer M, et al. Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms. 2018:1454-1457.
56. Kaplan RM, Irvin VL. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLoS One*. 2015;10(8):e0132382. doi:10.1371/journal.pone.0132382
57. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proceedings of the National Academy of Sciences*. 2018;115(11):2600-2606. doi:10.1073/pnas.1708274114

58. U.S. Food & Drug Administration (FDA). E6(R2) Good Clinical Practice: Integrated Addendum to ICH E6(R1) 2018. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e6r2-good-clinical-practice-integrated-addendum-ich-e6r1
59. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*. 2021/02/19 2021;4(1):31. doi:10.1038/s41746-021-00385-9
60. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007;356(14):1399-409.
61. Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. *Stat Med*. Apr 30 2013;32(9):1451-66. doi:10.1002/sim.5655
62. Pearl J, Mackenzie D. *The book of why : the new science of cause and effect*. Basic Books,; 2018:1 online resource.
63. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet*. Nov 25 2000;356(9244):1844-7. doi:10.1016/S0140-6736(00)03246-3
64. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagn Progn Res*. 2019;3:22. doi:10.1186/s41512-019-0069-2
65. Cash BD, Schoenfeld P, Rex D. An evidence-based medicine approach to studies of diagnostic tests: assessing the validity of virtual colonoscopy. *Clin Gastroenterol Hepatol*. Mar 2003;1(2):136-44. doi:10.1053/cgh.2003.50021
66. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. Oct 1 1996;125(7):605-13. doi:10.7326/0003-4819-125-7-199610010-00011
67. Temple R. A regulatory authority's opinion about surrogate endpoints. . In: Nimmo W, Tucker G, eds. *Clinical Measurement in Drug Evaluation* J Wiley; 1995.
68. U.S. Food & Drug Administration (FDA) C, ,. Design Considerations for Pivotal Clinical Investigations for Medical Devices. Guidance for Industry,Clinical Investigators, Institutional Review Boards and Food and Drug Administration Staff. 2013. https://www.fda.gov/media/87363/download
69. Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. *Ophthalmology*. 1991;98(5 Suppl):823-833.
70. Browning DJ, Glassman AR, Aiello LP, et al. Optical coherence tomography measurements and analysis methods in optical coherence tomography studies of diabetic macular edema. *Ophthalmology*. Aug 2008;115(8):1366-71, 1371 e1. doi:10.1016/j.ophtha.2007.12.004
71. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. Apr 1989;8(4):431-40. doi:10.1002/sim.4780080407
72. U.S. Food & Drug Administration (FDA). International conference on harmonisation; guidance on statistical principles for clinical trials; availability--FDA. . Fed Regist. 1998/08/20 ed1998. p. 49583-98.
73. Lee AY, Yanagihara RT, Lee CS, et al. Multicenter, Head-to-Head, Real-World Validation Study of Seven Automated Artificial Intelligence Diabetic Retinopathy Screening Systems. *Diabetes Care*. Jan 5 2021;doi:10.2337/dc20-1877
74. Lin AP, Katz LJ, Spaeth GL, et al. Agreement of visual field interpretation among glaucoma specialists and comprehensive ophthalmologists: comparison of time and methods. *Br J Ophthalmol*. Jun 2011;95(6):828-31. doi:10.1136/bjo.2010.186569
75. Lin DY, Blumenkranz MS, Brothers RJ, Grosvenor DM. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *AmJOphthalmol*. 2002;134(2):204-213.
76. Pugh JA, Jacobson JM, Van Heuven WA, et al. Screening for diabetic retinopathy. The wide-angle retinal camera. *Diabetes Care*. Jun 1993;16(6):889-95.

77. Abramoff MD, Lou Y, Erginay A, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest Ophthalmol Vis Sci*. Oct 01 2016;57(13):5200-5206. doi:10.1167/iovs.16-19964
78. Glassman AR, Beck RW, Browning DJ, Danis RP, Kollman C, Diabetic Retinopathy Clinical Research Network Study G. Comparison of optical coherence tomography in diabetic macular edema, with and without reading center manual grading from a clinical trials perspective. *Invest Ophthalmol Vis Sci*. Feb 2009;50(2):560-6. doi:10.1167/iovs.08-1881
79. Hajian-Tilaki K. The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation. *Statistical Methods in Medical Research*. 2018/08/01 2017;27(8):2374-2383. doi:10.1177/0962280216680383
80. van Stralen KJ, Stel VS, Reitsma JB, Dekker FW, Zoccali C, Jager KJ. Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. *Kidney Int*. Jun 2009;75(12):1257-1263. doi:10.1038/ki.2009.92
81. Sánchez MS, Ortiz MC, Sarabia LA, Lletí R. On Pareto-optimal fronts for deciding about sensitivity and specificity in class-modelling problems. *Analytica Chimica Acta*. 2005/07/15/ 2005;544(1):236-245. doi:https://doi.org/10.1016/j.aca.2004.12.084
82. Kupinski MA, Anastasio MA. Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves. *IEEE Transactions on Medical Imaging*. 1999;18(8):675-685. doi:10.1109/42.796281
83. Pepe MS, Janes H, Li CI, Bossuyt PM, Feng Z, Hilden J. Early-Phase Studies of Biomarkers: What Target Sensitivity and Specificity Values Might Confer Clinical Utility? *Clin Chem*. May 2016;62(5):737-42. doi:10.1373/clinchem.2015.252163
84. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. *Radiology*. May 2010;255(2):354-61. doi:10.1148/radiol.10091636
85. Righini M, Van Es J, Den Exter PL, et al. Age-adjusted D-dimer cutoff levels to rule out pulmonary embolism: the ADJUST-PE study. *JAMA*. Mar 19 2014;311(11):1117-24. doi:10.1001/jama.2014.2135
86. Gieseker KE, Roe MH, MacKenzie T, Todd JK. Evaluating the American Academy of Pediatrics diagnostic standard for Streptococcus pyogenes pharyngitis: backup culture versus repeat rapid antigen testing. *Pediatrics*. Jun 2003;111(6 Pt 1):e666-70. doi:10.1542/peds.111.6.e666
87. U.S. Food & Drug Administration (FDA). Patient Preference Information (PPI) in Medical Device Decision-Making. 2020. https://www.fda.gov/about-fda/cdrh-patient-science-and-engagement-program/patient-preference-information-ppi-medical-device-decision-making
88. U.S. Food & Drug Administration (FDA) CDRH. Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims 2009. https://www.fda.gov/media/77832/download
89. U.S. Food & Drug Administration (FDA) CDRH. Patient Preference Information – Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling: Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. 2016. https://www.fda.gov/media/92593/download
90. U.S. Food & Drug Administration (FDA). Network of Experts Program: Connecting the FDA with External Expertise. 2020. https://www.fda.gov/about-fda/center-devices-and-radiological-health/network-experts-program-connecting-fda-external-expertise#:~:text=The%20Network%20of%20Experts%20is,CDRH)%20and%20the%20Center%20for
91. U.S. Food & Drug Administration (FDA). Deciding When to Submit a 510(k) for a Software Change to an Existing Device. 2017. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-software-change-existing-device

92. Expert Working Group (Quality). Guideline for Industry: Text on Validation of Analytical Procedures. 1995. https://www.fda.gov/media/71724/download
93. Ioannidis JP. Why most published research findings are false. *PLoS Med*. Aug 2005;2(8):e124. doi:10.1371/journal.pmed.0020124
94. Shannon CE, Weaver W. *The mathematical theory of communication*. University of Illinois Press; 1949:v (i.e. vii), 117 p.
95. Xu A, Raginsky M. Information-theoretic analysis of generalization capability of learning algorithms. 2017:arXiv:1705.07809. Accessed May 01, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170507809X
96. Blumenthal D. Launching HITECH. *N Engl J Med*. Feb 4 2010;362(5):382-5. doi:10.1056/NEJMp0912825
97. Sloan Kettering Controversies: Trust is the Public Foundation of Medical Research. Bioethics.net.
98. U.S. Food & Drug Administration (FDA). Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies. 2017. https://www.fda.gov/media/98686/download
99. U.S. Food & Drug Administration (FDA) CDRH. Collection of Race and Ethnicity Data in Clinical Trials, Guidance for Industry and Food and Drug Administration Staff. 2016. https://www.fda.gov/media/75453/download
100. U.S. Food & Drug Administration (FDA) CDRH. Evaluation of Sex-Specific Data in Medical Device Clinical Studies Guidance for Industry and Food and Drug Administration Staff. 2014. https://www.fda.gov/media/82005/download
101. Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. 2018:arXiv:1810.03993. Accessed October 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv181003993M
102. Benoit SR, Swenor B, Geiss LS, Gregg EW, Saaddine JB. Eye Care Utilization Among Insured People With Diabetes in the U.S., 2010-2014. *Diabetes Care*. Mar 2019;42(3):427-433. doi:10.2337/dc18-0828