Progress in Retinal and Eye Research 90 (2022) 101034

Contents lists available at ScienceDirect

Progress in Retinal and Eye Research

journal homepage: www.elsevier.com/locate/preteyeres

Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice

Cristina González-Gonzalo [a,b,*], Eric F. Thee [c,d], Caroline C.W. Klaver [c,d,e,f], Aaron Y. Lee [g], Reinier O. Schlingemann [h,i], Adnan Tufail [j,k], Frank Verbraak [h], Clara I. Sánchez [a,l]

[a] Eye Lab, qurAI Group, Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands
[b] Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands
[c] Department of Ophthalmology, Erasmus Medical Center, Rotterdam, the Netherlands
[d] Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
[e] Department of Ophthalmology, Radboud University Medical Center, Nijmegen, the Netherlands
[f] Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland
[g] Department of Ophthalmology, School of Medicine, University of Washington, Seattle, WA, USA
[h] Department of Ophthalmology, Amsterdam University Medical Center, Amsterdam, the Netherlands
[i] Department of Ophthalmology, University of Lausanne, Jules Gonin Eye Hospital, Fondation Asile des Aveugles, Lausanne, Switzerland
[j] Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom
[k] Institute of Ophthalmology, University College London, London, United Kingdom
[l] Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, Amsterdam, the Netherlands

A R T I C L E I N F O

Keywords:
Artificial intelligence
Deep learning
Machine learning
Trustworthiness
Integration
Ophthalmic care

A B S T R A C T

An increasing number of artificial intelligence (AI) systems are being proposed in ophthalmology, motivated by the variety and amount of clinical and imaging data, as well as their potential benefits at the different stages of patient care. Despite achieving close or even superior performance to that of experts, there is a critical gap between development and integration of AI systems in ophthalmic practice. This work focuses on the importance of trustworthy AI to close that gap. We identify the main aspects or challenges that need to be considered along the AI design pipeline so as to generate systems that meet the requirements to be deemed trustworthy, including those concerning accuracy, resiliency, reliability, safety, and accountability. We elaborate on mechanisms and considerations to address those aspects or challenges, and define the roles and responsibilities of the different stakeholders involved in AI for ophthalmic care, i.e., AI developers, reading centers, healthcare providers, healthcare institutions, ophthalmological societies and working groups or committees, patients, regulatory bodies, and payers. Generating trustworthy AI is not a responsibility of a sole stakeholder. There is an impending necessity for a collaborative approach where the different stakeholders are represented along the AI design pipeline, from the definition of the intended use to post-market surveillance after regulatory approval. This work contributes to establish such multi-stakeholder interaction and the main action points to be taken so that the potential benefits of AI reach real-world ophthalmic settings.

1. Introduction

The potential of artificial intelligence (AI) in healthcare has become apparent in recent years with an increasing number of publications using deep learning (DL) and machine learning (ML) techniques for the automated analysis of clinical data (Litjens et al., 2017). AI systems have been shown to achieve close or even superior performance to that of clinical experts in different medical specialties (Liu et al., 2019b) and to provide valuable support tools for clinical decisions (Bulten et al., 2021). An increasing number of AI applications are being proposed in the field of ophthalmology (Ting et al., 2019b; Schmidt-Erfurth et al., 2018; Lee et al., 2017a), motivated by the variety and amount of clinical and imaging data, and the potential benefits of AI solutions in the different stages of patient care (González-Gonzalo et al., 2020b; De Fauw et al., 2018; Schmidt-Erfurth et al., 2021). Nevertheless, few prospective studies have been performed to validate proposed AI solutions in real-world settings (Heydon et al., 2020; Gulshan et al., 2019; Abràmoff

* Corresponding author. Eye Lab, qurAI Group, Informatics Institute, University of Amsterdam, P.O. Box 94323, 1090 GH, Amsterdam, the Netherlands.
E-mail address: c.gonzalezgonzalo@uva.nl (C. González-Gonzalo).

https://doi.org/10.1016/j.preteyeres.2021.101034
Received 9 September 2021; Received in revised form 3 December 2021; Accepted 6 December 2021
Available online 10 December 2021

Abbreviations

AAO	American Academy of Ophthalmology	FDA	United States Food and Drug Administration
AI	Artificial intelligence	GAN	Generative adversarial network
AMD	Age-related macular degeneration	GDPR	General Data Protection Regulation
CFP	Color fundus photography	HIPAA	Health Insurance Portability and Accountability Act
CNN	Convolutional neural network	HTA	Health Technology Assessment
CPT	Current Procedural Terminology	IRB	Institutional Review Board
DICOM	Digital Imaging and Communications in Medicine	ML	Machine learning
DL	Deep learning	MRI	Magnetic resonance imaging
DME	Diabetic macular edema	NIR	Near-infrared imaging
DNN	Deep neural network	NHS	National Health Service
DR	Diabetic retinopathy	OCT	Optical coherence tomography
EU	European Union	OCTA	Optical coherence tomography angiography
FAF	Fundus autofluorescence	PACS	Picture archiving and communication system
		UK	United Kingdom
		US	United States

et al., 2018), with very few AI systems obtaining regulatory approval for clinical use (Abràmoff et al., 2018; Heydon et al., 2020; González-Gonzalo et al., 2020b).

We and others observe a critical gap between development and deployment of AI systems in ophthalmic practice. In contrast to the 510 ophthalmic AI systems developed between 2010 and 2020 (Meskó and Görög, 2020), only 12 AI-based medical devices are approved for clinical use in Europe and 2 are approved in the US (Muehlematter et al., 2021). To address this gap, guidelines are being proposed that prepare the ground for standardization and facilitation of AI integration in healthcare (Rivera et al., 2020; Liu et al., 2020; Sounderajah et al., 2020; Collins and Moons, 2019). Other studies identify important considerations and current challenges (Abràmoff et al., 2021; Singh et al., 2020; Ting et al., 2019a; Kelly et al., 2019). These studies all agree on the need of *trustworthy* AI systems to facilitate the uptake of AI in healthcare.

The European *Ethics Guidelines for Trustworthy Artificial Intelligence* identify a set of requirements that AI systems should meet in order to be deemed trustworthy, including properties such as accuracy, resiliency, reliability, safety or accountability (European Commission, 2019). These guidelines, as well as others, articulated and debated concerns and principles to guide trustworthy AI development for the global AI community (European Commission, 2019) or considering specific fields in society such as healthcare (World Health Organization, 2021). Although the undertaken effort is necessary, it is deemed crucial to move beyond high-level principles to a focus on mechanisms for ensuring and measuring trustworthy behavior of AI systems. As a well-calibrated evidence of a trustworthy AI system, the definition of verifiable claims based on trustworthy properties, as well as mechanisms to support these

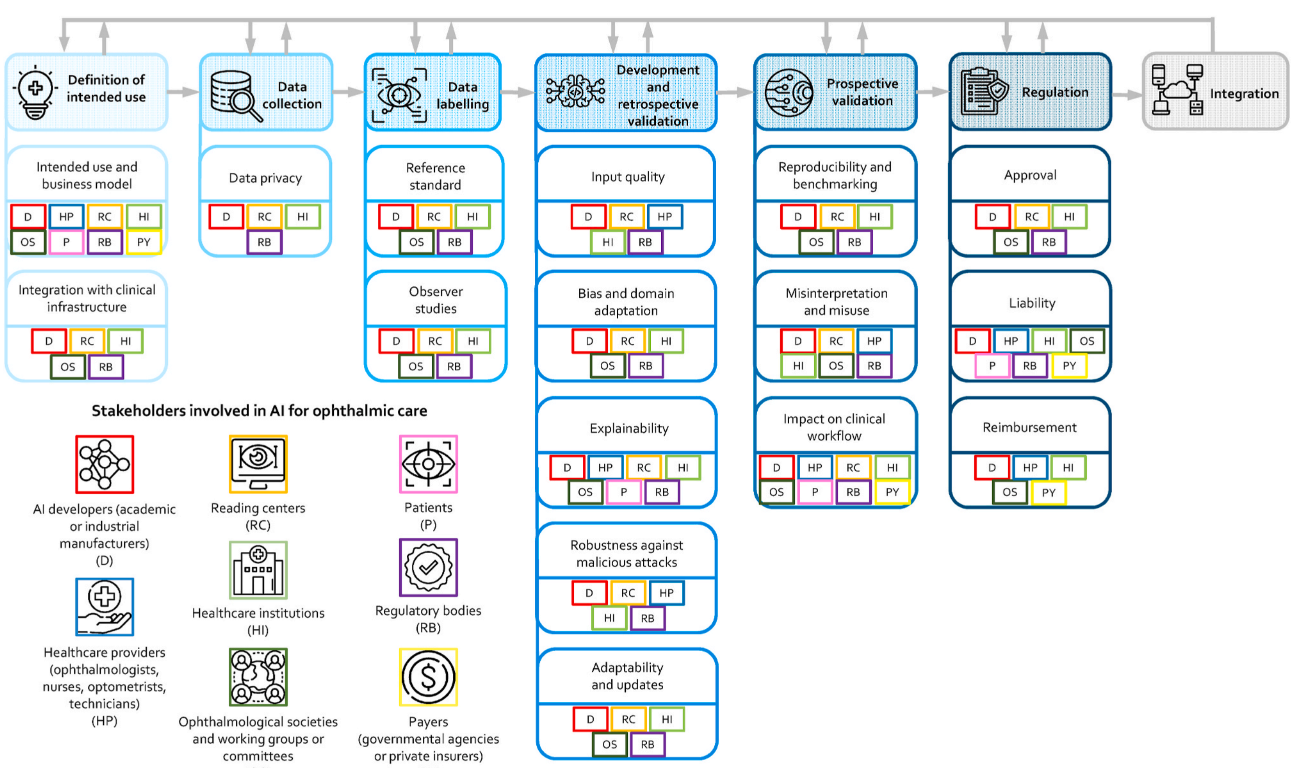


Fig. 1. The diagram illustrates the stages in the AI design pipeline and the stakeholders involved in AI for ophthalmic care. For each stage, it shows the identified aspects or challenges to address in order to generate trustworthy AI systems in ophthalmology. For each aspect or challenge, it indicates which stakeholders have a main role in applying the necessary mechanisms to address them. The AI design pipeline is not meant to be fixed, but cyclic, making it possible to address challenges at previous stages when necessary.

claims, are urged to ensure beneficial societal outcomes from AI (Jacovi et al., 2021). Unfortunately, unverifiable claims provided by the community have been common in the last years. For instance, Liu et al. (2019b) observed that only a small number of studies make direct comparisons between AI systems and healthcare professionals, and an even smaller number validate these findings in an out-of-sample external validation. Unverifiable claims have the potential to encourage undesired reactions to AI (i.e., blind trust or blind rejection), resulting in both overuse and underuse of AI. This hampers AI adoption in clinical practice (Brundage et al., 2020).

In this work, we discuss the importance of trustworthy AI for the development and integration of responsible systems in ophthalmology and propose a set of mechanisms that could support verifiable claims about system's properties, such as accuracy, reliability or resilience, in the various stages of the design pipeline of an ophthalmic AI system. The pipeline established in this manuscript was defined based on the authors' diverse expertise and previous works (Hopkins et al., 2020; Char et al., 2020). As summarized in Fig. 1, for each stage we identify potential challenges to different claims. We provide a better understanding of which properties of AI systems can be verified and through what means. When necessary, we indicate how to adapt them for a specific point of patient care in ophthalmic practice: screening, diagnosis, monitoring, treatment, prognosis. We identify barriers within the design pipeline and discuss the necessary mechanisms to address them, indicating the corresponding role of the different stakeholders involved in AI for ophthalmic care. This allows to anticipate risks and avoid negative consequences during integration or deployment, as well as approach verifiability, safeguards, and best practices in a collaborative fashion.

We believe our work helps to (a) get a better understanding of the properties underlying trustworthy AI, (b) identify the roles and interactions between the different stakeholders to develop responsible AI systems, and (c) contribute mechanisms to develop and promote the uptake of AI-based systems in ophthalmic practice.

2. Definition of intended use

The first stage in the AI design pipeline is the definition of the intended use. Extensive research and business analysis are key to identify an unmet clinical need; for instance, the automation or optimization of a certain clinical task, or the support to clinical decisions to improve personalized ophthalmic care. This is in high contrast with numerous AI systems proposed lately, which are a result of the trend around AI and focus solely on applications where public data are available, disregarding how to bring major positive changes and sustainable solutions into healthcare. As with other types of clinical tools or procedures, early thorough planning of the intended use and the design of a medical AI system is critical. It allows to maximize its alignment with the target clinical application and point in patient care, reducing potential difficulties that might arise in later stages of the pipeline. For instance, regulatory bodies often require a precise and accurate definition of the intended use, since it determines the risk profile and, consequently, the type of approval pathway and post-approval control. The identification and involvement of the different stakeholders is therefore critical, in order to combine different types of expertise (technical, clinical, regulatory…) from the beginning.

In this section, we study the importance and discuss potential solutions for two main aspects to consider at this stage of the pipeline:

- Defining the intended use together with a realistic and sustainable business model for the AI system.
- Analyzing how to integrate the system as seamlessly as possible in the current clinical infrastructure.

2.1. *Intended use and business model*

2.1.1. *Importance and consequences*

It is estimated that AI has the potential to address 20% of unmet clinical needs in the coming five years (Accenture, 2020). AI has the potential to automate certain processes while letting healthcare providers focus on more complex clinical tasks, to improve personalized care and predictive analyses, and to alleviate the increasing shortage of ophthalmologists, accentuated in rural areas (Association of American Medical Colleges, 2020; Royal College of Ophthalmologists, 2020). However, many AI developers remain focused on clinical questions for which public datasets and annotations are already available, without questioning the clinical relevance of the problems and the actual applicability of their solutions (Wiens et al., 2019). Moreover, a realistic business model is required to facilitate integration. A feasible business model considers not only an unmet clinical need, but also the impact of using an AI system on a given clinical workflow and the reimbursement for its adoption and use over time (Hopkins et al., 2020). The lack of engagement of different stakeholders before development contributes to the lack of alignment of AI systems with actual clinical needs and financial viability.

2.1.2. *Proposed solutions and considerations*

The first factor to consider is the creation of an interdisciplinary team from the beginning of the AI design pipeline. Successful AI integration in clinical settings requires the engagement of all relevant stakeholders from different areas, including knowledge experts (AI, clinical, and implementation experts), decision-makers (healthcare institutions, reading centers, ophthalmologic societies, regulatory bodies, governmental agencies, private insurers), and users (ophthalmologists, nurses, optometrists, technicians, patients, graders) (Wiens et al., 2019).

A collaboration between AI developers and clinical stakeholders from the start would allow for a robust interrogation and identification of unmet clinical needs that would benefit from AI, improving the alignment of AI solutions with the clinical problems to solve. It would also allow to maximize the utility of AI in the clinic. For example, in the context of automated screening of eye diseases, most proposed AI solutions focus solely on the detection of diabetic retinopathy (DR) in color fundus photographs (CFP), while other diseases, such as age-related macular degeneration (AMD) or glaucoma, co-exist in the screened subjects. Such patients might not be referred if there is no DR present (Abràmoff et al., 2018). The development of AI systems that perform joint detection of co-existing eye diseases could increase clinical utility and facilitate software centralization (González-Gonzalo et al., 2020b). Nevertheless, this might depend on the intended use and clinical setting, since national diabetic screening programs are set up to detect DR while other diseases may be in part filtered out by visual acuity measure, and the health economics of screening of other diseases are still unclear.

A key factor to promote the involvement of stakeholders is transparency. AI developers should provide transparency when declaring their intentions and the goals of their application, so that other stakeholders have the autonomy to evaluate these intentions and decide to support them (Char et al., 2020). It is also important to identify potential conflicts of interest, which can be individual financial interests, such as payment for services, as well as operational interests that might differ, for instance, between AI manufacturers and healthcare providers regarding data or system ownership. We believe *win-win collaborations* are crucial for this matter. When it comes to collaborations between AI manufacturers and healthcare institutions or reading centers, there is usually hesitation from both sides regarding data and algorithm sharing. It is thus important to establish the basis of balanced collaborations from the beginning.

Transparency also requires that developers ensure auditability along the AI design pipeline, facilitating inspection of the followed processes by the involved stakeholders, from the development stage to the initial clinical deployment (Char et al., 2020). Raji et al. (2020) have recently proposed a framework for end-to-end internal algorithmic auditing. Its goal is to ensure the compliance of internal and external policies and ethical values and promote accountability during the design of AI systems. Internal audits would involve multidisciplinary teams, including internal stakeholders (a dedicated audit team, and the development, product, and management teams) and other stakeholders, which could include external healthcare representatives in the case of AI medical devices. Such internal auditing during AI design would be complementary to post-surveillance auditing, discussed in Section 7.1.

To achieve transparency, general aspects of the AI system to be developed need to be clarified. For example, whether the system will be assistive or autonomous. This aspect is important considering its impact on different elements of patient care (Abràmoff et al., 2021), including liability (Section 7.2). There is also a need to clarify whether the system will be locked or continuously learning, which has a direct impact on the system's maintenance and the required financial and human resources (Section 5.5). Another relevant aspect is whether the system is meant to be used in isolation, in combination with other diagnostic elements, or used as an add-on or as a replacement of a current process or tool (Faes et al., 2020). It is also crucial to clarify the intended use of the system within the clinical pathway (screening, diagnosis, monitoring, treatment, prognosis), since this conditions the data to be used for development and the required validation. Both data and validation need to be representative of the point of care where the system is aimed to be deployed. Relevant pre-specifications regarding data include technical aspects (e.g., device used for the extraction of development data) and contextual and cohort information (e.g., demographics, time period, clinical setting, disease prevalence, inclusion/exclusion criteria). For example, an AI system intended for automated screening of DR should ideally be validated using data acquired at screening settings, i.e., with a very low prevalence of referable and advanced DR cases, instead of using data from hospital-based clinics. It is important to consider that highly curated datasets may be useful to develop AI systems in research, but do not suffice to validate systems for clinical use in real-world settings. Developers must also ensure transparency on data privacy (Section 3.1) and the reliability of the reference standard (Section 4.1). Pre-specifications regarding system validation, internal and external, include indication of primary and secondary outcomes to validate and the statistical analysis planned (Faes et al., 2020). It is important to ensure that the output of the system will be aligned with internationally accepted disease classifications or quantification standards used in current guidelines and therapeutic management of ophthalmic diseases. Similarly, it is important to align performance goals with those in clinical practice, that is, to define when and how the AI system will be considered good enough, by means of pre-specified performance metrics when possible (Abràmoff et al., 2018), and/or by setting up a robust observer study to compare the performance of the system with that of human experts (Section 4.2).

Interdisciplinary collaborations are also necessary to establish a realistic business model around the AI system. Addressing an unmet clinical need and achieving good performance are not the only factors that matter for successful clinical integration. It is crucial to establish a sustainable business model that considers the specific interests of the different stakeholders, long-term implications in the clinical workflow, such as improvements in efficiency and cost-effectiveness (Section 6.3), and reimbursement (Section 7.3) (Hopkins et al., 2020).

2.2. *Integration with clinical infrastructure*

2.2.1. *Importance and consequences*

Current clinical infrastructures present multiple difficulties for the integration of AI systems, mainly due to the limited interoperability between medical devices and healthcare settings. The increasing adoption of electronic health records (EHRs) has allowed great improvements in this regard and can facilitate AI integration. However, there is still large variability in the use of EHRs and the completeness of data entry across clinical settings, as well as in the interoperability between different providers (Panch et al., 2019). Ophthalmology is particularly backwards in this aspect, with a widespread lack of interoperability between EHRs and the various imaging devices. Most ophthalmic imaging devices make use of vendor-specific file formats and data storage software, and often vary within and between primary, secondary, and tertiary care settings, creating an environment of devices and settings that cannot communicate with each other (Li et al., 2020). This can be easily ascertained when compared to other medical specialties, including radiology, where universally-accepted protocols and technologies such as DICOM and PACS facilitate data accessibility and communication. This has allowed a wider and faster proliferation of AI development and integration in these respective fields (Litjens et al., 2017; Muehlematter et al., 2021).

The current lack of uniformly accepted standards for data formatting, storage, and transfer in ophthalmology challenges the integration of AI systems, which would require a seamlessly flow of patient data to use as input, and the transfer and storage of the generated output to make it accessible with other available data in the patient record. It also hinders the adaptability of AI systems to different clinical settings. If AI systems cannot be embedded properly in existing clinical infrastructures and workflows, ophthalmologists, nurses, optometrists, and technicians will be reluctant to include AI in their daily practice. Interruptions in the clinical workflow take time, decrease efficiency, and cause frustration. Additionally, the lack of interoperability across devices and settings challenges the collection of data from multiple sources and populations for AI development, which is an important factor to prevent bias and lack of generalization in AI systems (Section 5.2). It can be observed that without efforts to optimize standardization and interoperability, the practical applicability of AI and its benefits will remain severely limited.

2.2.2. *Proposed solutions and considerations*

In order to circumvent the current lack of standardization and interoperability in ophthalmology, AI manufacturers can make significant efforts to maximize AI integration, starting with a thorough analysis of the target setting/s prior to development. The objective should be to generate an AI system that aligns as much as possible with the target setting's conditions, which may greatly vary with those of the settings used for system's development and validation. This is especially the case of target settings in rural areas and low-resource countries, where there is often great variation and/or lack of protocols for image acquisition, used equipment, and personnel's experience (Beede et al., 2020).

Therefore, it is important to define an appropriate model of care for the AI system, considering the available infrastructure and clinical workflow of the target setting/s (Ting et al., 2019a). In cloud-based solutions, the acquired data are transferred from a client application to the AI manufacturer's cloud via Internet connection. Data are processed by the AI system in the cloud and the output is transferred back to the client application. Cloud-based solutions are common in the context of automated screening of eye diseases due to their ease of integration and use, such as the one proposed by González-Gonzalo et al. (2020b). This model of care becomes especially useful in teleophthalmology settings (Li et al., 2020). Cloud-based solutions are also key for home-based monitoring, which will allow for improved personalized eye care, for instance, in the context of retinal fluid changes in AMD patients (Notal Vision, 2018). Although cloud-based solutions provide high scalability, a stable, secured Internet connection cannot be assured in certain settings, especially in low-resource countries. In a recent prospective validation of a cloud-based AI system for automated screening of DR in rural areas in Thailand, Beede et al. (2020) showed that Internet speed and connectivity were a limiting factor, causing delay or appointment rescheduling. A more suitable alternative for settings in rural and low-resource areas can be the integration of office-based solutions, where the AI system is deployed as part of an application that can be installed in a desktop, laptop, tablet or smartphone, processing the acquired data offline (Natarajan et al., 2019). An important consideration is that office-based solutions might be harder to synchronize with the latest available version and to perform updates to the AI system, which may limit the ability of future continuously learning algorithms once allowed by regulators (Section 5.5). They might also require the implementation of compressed or lightweight models that enable the deployment of deep learning in mobile devices and whose performance might not be at the same level of that of state-of-the-art models (Owen et al., 2021).

Independently of the model of care, vendor-neutral solutions will be key to increase interoperability in ophthalmic infrastructures. AI systems that are not integrated with a given camera model, or that do not require data to be acquired by a specific device, are easier to adapt and to use within the current clinical infrastructure. Their integration is also more economically viable, since acquiring a new device to use an AI system is not as feasible as just acquiring access to a client application or installing a software package. For these reasons, the use of vendor-neutral AI solutions will also be more generally accepted by ophthalmic societies and working groups or committees (Lee et al., 2021a; Royal College of Ophthalmologists, 2021). Current efforts on the development and deployment of vendor-neutral archives of ophthalmic data will also lead to increase interoperability and, consequently, facilitate AI integration at large scale (Swiss Personalized Health Network, 2018). These archives would provide centralized storage and access to raw ophthalmic data from different modalities and vendors. With updated data transfer agreements and ethical approval by the corresponding institutional review board (IRB), vendor-neutral archives would also enhance the communication and data sharing between devices and healthcare institutions. Such archived data could be more easily used as input of AI systems for different clinical applications, as well as collected for AI development and validation.

The development and implementation of standards in ophthalmology becomes necessary to yield vendor neutrality, interoperability, and a successful integration of AI. Imaging standards will facilitate the harmonization of the mentioned vendor-neutral archives, and standards for interfaces and AI outputs will enable widespread adoption and integration of AI in ophthalmic infrastructure. The Fast Healthcare Interoperability Resources (FHIR) framework is a promising advance in this regard, and it is anticipated to become critical for the integration of AI systems. It consists of a set of standards for exchange of healthcare data that will facilitate interoperability within EHRs and mobile-based apps, as well as cloud-based communications (He et al., 2019). There are ongoing efforts at the American Academy of Ophthalmology (AAO) to support the implementation of existing DICOM standards in ophthalmology, while engaging with the broader healthcare standards community (Lee et al., 2021a). The AAO's call for standardization has been endorsed by several ophthalmological societies, including the Royal College of Ophthalmologists (UK) (Royal College of Ophthalmologists, 2021). It is crucial that healthcare institutions, vendors, and AI developers support these efforts in order to incentivize the adoption of standards (Baxter and Lee, 2021). Different stakeholders can have nonetheless different interests, and these interests might not always aim to maximize interoperability (Lehne et al., 2019). To overcome these situations in single-payer healthcare systems, where new bodies are being designated responsible for the implementation of AI and digital medicine (e.g., NHSX in the UK), these could set the standards for interoperability as an essential requirement prior to the commission of a given ophthalmic imaging device or AI system, the same way an MRI or CT scanner must be DICOM-compatible to be adopted. For both single-payer and multi-payer systems, this aspect could also be supervised within a given healthcare institution. In addition, we believe that actions taken by the corresponding regulatory bodies might be necessary to efficiently incentivize and enforce interoperability.

Substantial updates in the clinical infrastructure could also enhance healthcare and facilitate AI integration, as well as updates in protocols concerning data privacy (Section 3.1) and cybersecurity (Section 5.4). However, it is important to acknowledge that certain settings will be limited by the prevalent socio-economic context of their healthcare system rather than by technology, which means that special efforts might be necessary to ensure the benefits of AI reach low-resource settings as well (Burton et al., 2021; Panch et al., 2019).

3. Data collection

Data collection is a critical stage in the design of AI systems. It requires thoughtful preparation and alignment with the defined intended use. In ophthalmology, imaging datasets have been used to develop AI systems for the automated diagnosis, prediction, and prognosis of common eye diseases, such as DR, AMD, and glaucoma. While most healthcare institutions hold imaging data at a sufficient scale for AI development, these data are often inaccessible to AI developers due to barriers of governance, cost, time, format, and privacy. Consequently, AI developers often make use of publicly-available imaging datasets, which have been powerful enablers of AI development in ophthalmology (Khan et al., 2020). Nevertheless, public datasets are not available for all relevant clinical applications, and those that are available may not always be aligned with the identified intended use or be representative enough of the target settings, for instance, in terms of population or imaging device. Importantly, health data poverty, i.e., the inability for individuals, groups or populations to benefit from innovation due to the scarcity of representative data, has been recognized as an important barrier to equitable healthcare (Ibrahim et al., 2021; Burton et al., 2021). As such, joint efforts from AI developers, healthcare institutions, as well as regulatory bodies, are essential in building well-curated inclusive datasets, so as to ensure an appropriate alignment with the intended use and patients' safety.

In this section, we focus on a key aspect at this stage of the pipeline, data privacy, analyzing its influence when it comes to successful AI integration and discussing different procedures and directions to ensure patients' data are safely collected, stored, and used.

3.1. Data privacy

3.1.1. Importance and consequences

Data privacy in healthcare is becoming more critical with clinical and imaging data being digitized, stored, sent, and downloaded for several analytical purposes. Diagnosis of ophthalmic diseases requires large and heterogeneous datasets of images that contain personally-identifiable information. Current laws and regulations for privacy concerns related to medical data use are complex and differ per jurisdiction, such the Health Insurance Portability and Accountability Act (HIPAA) in the US (Cohen and Mello, 2018) and the General Data Protection Regulation (GDPR) in the EU (European Commission, 2016). Moreover, they are not synchronous with the latest technological developments, which hampers implementation of innovative AI-based healthcare solutions globally. With the current regulations, data breaches, and violations of privacy with distribution of medical data still remain an important concern among healthcare providers and patients. Furthermore, ongoing technological developments have led to new privacy concerns, such as patient re-identification from anonymized data that might still contain visually unique identifiers. This is the case of the unique patterns of a patient's retinal blood vessels, which can be used for biometric identification (Moore and Frye, 2019) and can be observed even if imaging data have been previously anonymized. The ongoing privacy concerns call for updated, detailed regulations, and data protection methods that can evolve along with big data and AI.

3.1.2. Proposed solutions and considerations

Several data protection methods that address privacy concerns with AI development are underway and have taken advantage of current technology. Methods currently being investigated for ophthalmic analyses include that of federated learning and distributed models (Tom et al., 2020). Both allow model training across separate servers and datasets, while minimizing the risk of possible data breaches. In line with these methods, Mehta et al. (2020) used a model-to-data approach to develop a DL model that segments intraretinal fluid in OCT B-scans. The model was initially trained on a central system at one institution, then transferred to a distinct computer at another location/institution for re-training and testing with data from that institution. Apart from the DL model, no clinical data were shared, and the study was successfully completed without the developers being exposed to the other institutions' clinical data (Fig. 2). It is important to note that for a successful use of federated learning in practice, personnel with certain technical and implementation expertise at the healthcare institutions would be required.

Other data protection methods currently being investigated are generative methods, specifically generative adversarial networks (GANs) (Bellemo et al., 2018). GANs can be used to generate synthetic images that are different from the original images, but that still capture disease-relevant features. Use of such synthetic data may not only address the current lack of large annotated datasets in ophthalmology, but may additionally address privacy concerns such as membership, attribute inference, and also re-identification of anonymized data (Paul et al., 2021). Nevertheless, an important challenge with GANs and synthetic data is its trade-off in model performance. If the generated images are too different from the original images, the diagnostic performance of the model may suffer in real-world settings (Goodfellow et al., 2014a).

An important consideration is that while the various data protection methods have potential to address privacy concerns with regard to data sharing and re-identification of anonymized data, they are still relatively new and need to be investigated further. Two recent reports showed that federated learning methods could be overcome through adversarial attacks (Bagdasaryan et al., 2020) (discussed in Section 5.4). Therefore, similar methods should not be used as end-solutions, but as a part of a larger scale of solutions to address privacy concerns.

Apart from technical solutions for privacy concerns, ethical and legal frameworks, as well as contractual safeguards, need to be considered (Tom et al., 2020). Adopters of AI systems in healthcare have reported skepticism about a system's output when they are not given adequate information about its origins (Klarenbeek et al., 2021; Morgenstern et al., 2021). Transparency in the way that data are collected and used is key. To increase transparency, it was proposed that AI developers should have traceable written agreements on authorization relating to specific use of data. They should also implement auditable processes and security controls to ensure data privacy is maintained and data are being used according to made agreements (Abràmoff et al., 2020). Hutchinson et al. (2021) proposed a rigorous framework based on software engineering best practices, where in each stage of the data lifecycle (from the analysis of requirements to maintenance) documentation practices, oversight processes such as audits and reviews, and maintenance mechanisms are provided. The goal is to increase transparency and accountability in the decision-making concerning the data used for AI development and validation. These practices can also help prevent future bias and domain adaptation issues, discussed in Section 5.2.

Another point of attention is informed consent. For the use of data containing personal health information (PHI), including retinal imaging data, and data on age and sex, patients need to sign an informed consent. Importantly, patients may need to sign multiple consent forms for different projects, and depending on the project, data may only be used within a certain region or time period. Patients may be given access to dynamic consent management, in line with personalized clinical care; the processes of adaptability and updates (discussed in Section 5.5) could be used to ensure the compliance of an AI system with patient consent over time. Derivation of patient consent can be therefore challenging, particularly when using large prospective datasets for AI development and/or validation. To further ensure that the rights and welfare of patients are protected, AI-related projects need to be approved by a local IRB.

In summary, data protection methods for privacy concerns with regard to ophthalmic analyses are currently underway, but some unaddressed challenges still require innovative and dynamic solutions. Solutions will require transparent conversations between AI developers, healthcare institutions, regulatory bodies, and patients.

4. Data labelling

There are two main approaches within AI: *supervised learning* and *unsupervised learning*. The main difference is that supervised systems are trained using labelled data to generate outcomes based on provided human annotations, while unsupervised systems process data without labels and are trained to find patterns and generate unbiased associations without requiring human annotations (Litjens et al., 2017). The

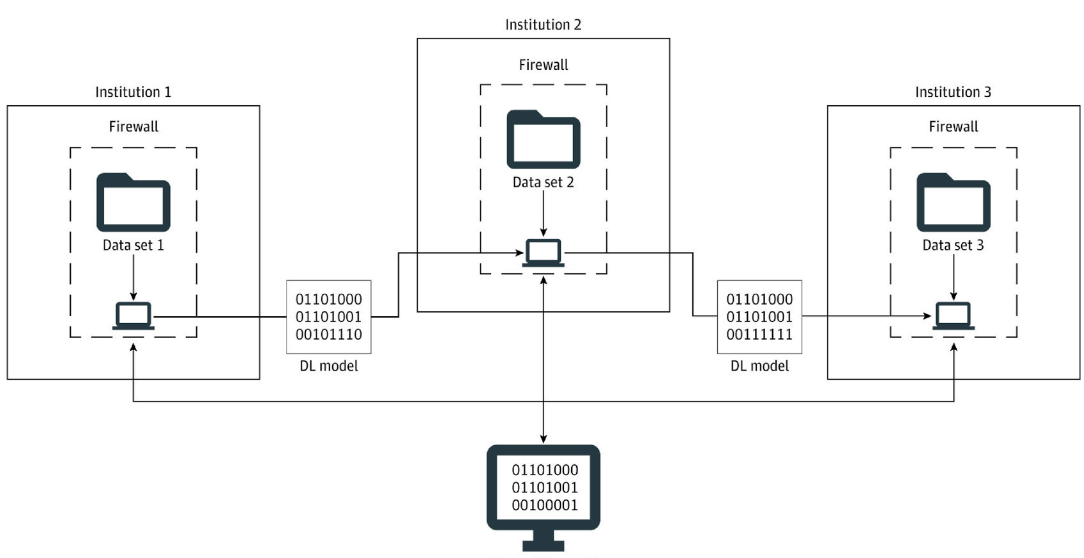


Fig. 2. Schematic description of a model-to-data approach to address concerns about clinical data privacy in AI development and validation (Mehta et al., 2020). In this approach, a trained deep learning model is transferred to a new institution housing its own unique data, which remain within its firewall for model re-training and/or testing. Once the re-training and/or testing using the data available at one institution is complete, the updated model (without clinical data) can be transferred to another institution, allowing for rapid iterative model sharing without transfer of any clinical data.

application of AI in ophthalmology is dominated by supervised systems, since they are less computationally complex and more effective and reliable when there is a well-defined task to solve.

Once the necessary data for development and/or validation have been collected, data labelling can be performed. In ophthalmology, data labels are often obtained from human graders who diagnose disease at image- or lesion-level on different modalities of ophthalmic imaging, most commonly on CFP and OCT scans. Human graders can be ophthalmologists, retinal specialists, optometrists, ophthalmic technicians, or graders at a reading center. So far, AI systems have mainly focused on diagnosing eye diseases on CFP, where data labels are generally based on international classification systems, such as the International Clinical Diabetic Retinopathy (ICDR) Disease Severity Scale for DR, or the Age-Related Eye Disease Study (AREDS) classification for AMD (Thee et al., 2020). There is also increasing attention for the quantification of individual lesions to monitor disease progression or treatment response (Liefers et al., 2021). For disease quantification, data labels are generally based on manual delineations by human graders. Several factors, including the protocols followed for labelling or the graders' expertise, can condition the reliability of the labels used for development and validation of AI systems and, consequentially, have an important impact on the gap to clinical integration.

In this section, we analyze two main aspects to be considered during data labelling:

- The importance and directions for establishing solid reference standards to generate effective and reliable AI systems. We define *reference standard* as the set of labels that are used for training and optimizing an AI system to solve a given clinical task and/or validating its performance. While the term *ground truth* has been used in other studies, we consider reference standard to be a more appropriate term for clinical tasks: ground truth refers to information that is known to be real or truth and that is often not the case for human annotations, which may contain inherent subjectivity due to the lack of a (well-defined) gold standard (Tufail et al., 2016).
- The importance and potential solutions for setting up robust observer studies, whose goal is to enable a fair and meaningful comparison of AI systems with healthcare providers and current clinical practice.

4.1. Reference standard

4.1.1. Importance and consequences

Graders that perform annotations for a reference standard need to be as accurate and objective as possible. Yet, a known issue of human interpretation of medical images is *grader variability*. Grader variability refers to the variability in data labelling between graders (*inter-grader variability*) and the variability in labelling by the same grader (*intra-grader variability*). Several studies have demonstrated substantial inter-grader variability for image-level classification tasks, such as grading DR (Schaekermann et al., 2019) and retinopathy of prematurity (ROP) (Tsai et al., 2021) in CFP. Such levels of disagreement or "lack of consensus" in reference standards can make it challenging to reliably develop AI systems.

The choice of labelling protocol or classification system also matters. Protocols may differ per healthcare institution or reading center, hampering the adequate validation of a system's performance, as well as the comparison of performances of various systems (Thee et al., 2020). In addition, internationally established protocols may not be available for certain tasks, such as the segmentation of individual lesions. This can lead to higher grader variability and makes it more difficult to determine the reliability of the reference standard for these tasks.

Currently, there are no clear guidelines to adequately evaluate the quality of reference standards; there is no established minimum quality for the labels, instructions for choosing the most adequate labelling protocol, or reporting guidelines. In the meantime, efforts are required to reduce grader variability, facilitate the labelling process and, consequently, increase the effectiveness and reliability of AI systems.

4.1.2. Proposed solutions and considerations

Regulatory bodies have highlighted the importance of aligning labelling protocols or classification systems for the training and validation of AI systems (Abràmoff et al., 2016). This will allow for an adequate interpretation of the system's performance and ability to generalize to external data. Additionally, developers should consider the performance and applicability of labelling protocols. Thee et al. (2020) demonstrated that frequently-used AMD classification systems differ markedly in their prognosis of progression towards end stages. Subsequently, they pointed out the advantages and disadvantages of classification systems for different clinical applications. Some classification systems will be more suited for screening and diagnosis, as they only require grading of limited features that are quick and easy to interpret. Others will be more suited for intervention studies, as they show higher rates of progression towards the end stages of disease. The choice of classification system should also consider the outcome goal per use case, for instance, high sensitivity in the case of integrating AI within a DR screening program, and high specificity in the case of applying AI to find eligible subjects for a clinical trial.

As classification systems are subject to change, reference standards for AI systems may need to be updated over time. Future studies may preferably focus on AI for the automated identification of individual lesions. Individual lesion criteria are less subject to change than disease severity classes, and would more likely serve a wider variety of applications. They also allow for a better quantification for treatment applications and prediction of progression (Liefers et al., 2020, 2021). In addition, increasing the granularity of systems' predictions improves their adaptability across the protocols followed in different clinical units. For example, for anti-VEGF treatment, a system that detects all the features related to advanced AMD (Liefers et al., 2021) would be more adaptable to different protocols than an end-to-end system that outputs directly the need for retreatment. Nevertheless, the development of reference standards for the identification and quantification of disease-specific lesions requires considerable time and effort. As there are no internationally established protocols for these tasks, it is important that reading centers, aided by ophthalmological societies and working groups or committees, focus on developing standardized protocols in order to reap the potential benefits of lesion-specific AI systems.

As described, high levels of grader variability can make it difficult to reliably train and validate AI systems. Some studies have focused on resolving disagreements between graders and deriving consensus through supervision (Schlegl et al., 2018), majority decisions (Gulshan et al., 2016), or adjudication by experts (Krause et al., 2018). A study by Krause et al. (2018) compared three different protocols for the labelling of DR. They showed that, compared to individual ophthalmologists and to a majority decision by ophthalmologists, in-person adjudication by retinal specialists allowed to generate a more solid reference standard and, consequently, improve AI performance. Adjudication started with independent grading of a subset of images by retinal specialists and ended with a review of disputed matters in a combination of asynchronous and live adjudication sessions. It particular, it helped reduce errors due to imaging artifacts and omission of small lesions, such as microaneurysms. In-person adjudication to reach consensus is considered highly effective, as it allows the resolution of gray areas in labelling protocols. A consideration for in-person adjudication is that it can be challenging to coordinate. As multiple experts need to be present, the process can take several months due to clinical scheduling conflicts. Some studies have investigated remote adjudication and demonstrated that the quality of the reference standard can be maintained (Schaekermann et al., 2019). Adjudication can also provide benefits beyond developing trusted consensus grades. Through surveys and qualitative feedback from graders, Schaekermann et al. (2020) showed that adjudication provided an effective training intervention, as it significantly improved the graders' accuracy and understanding of the rationale behind the correct diagnosis.

On the other hand, other studies involve a large number of graders and a large amount of annotated data, with the aim of averaging out grader variability in the reference standard, thereby mitigating part of the subjectivity (Son et al., 2020). There are two main aspects to consider when following this approach. First, it becomes relevant to integrate the differences in expertise between human graders in the labelling process. This can be done by establishing tiers of graders with increasing experience, as done by Liu et al. (2019) to create a reference standard for automated detection of glaucomatous optic neuropathy in CFP. Although applied later in the AI design pipeline, technical methods to integrate variability in grader expertise during training also show promise. For instance, the method proposed by Guan et al. (2018) aims to predict the labels of each grader in the training data so that more weight is given to the labels from more reliable experts, while exploiting the unique strengths of individual experts. Second, it is important to provide a training protocol prior to labelling so as to ensure a base quality in the annotations and contribute to reduce subjectivity. In a recent study by Liefers et al. (2021), clear instructions and examples of all abnormalities of interest were provided and discussed with the graders prior to the labelling process. Lesion-level annotations from different graders were then collected to generate the reference standard to train an AI system for automated segmentation of AMD-related features. Graders could additionally be required to pass a certification test before starting the grading process (Phene et al., 2019).

Dependable grading software is necessary so that graders can accurately perform data labelling. Several studies have developed software applications or workstations that allow graders to analyze simultaneously multimodal images, label data at image- and lesion-level, and review the outputs of AI systems, such as the one proposed by van Zeeland et al. (2019) (Fig. 3). Since it is web-based, such workstation enables asynchronous grading and remote consensus grading, while providing anonymization for the graders, which makes it possible to perform an unbiased review of labels.

Regulatory bodies have advocated for transparency in the reporting of labelling protocols used to develop reference standards (He et al., 2019). Such transparency is essential for healthcare institutions and reading centers to determine whether AI systems will have a good performance for their intended use. Improving clarity of the clinical character of the reference standard that underlies the advice of an AI system will also increase the trust and adoption of such advice. Clear reporting of the labelling protocols used in published studies can encourage and allow for re-usability in future studies, accelerating the creation of international protocols for those tasks where they are not available yet, such as the segmentation of retinal lesions. Reading centers should play an important role establishing protocols for the creation of reference standards and reporting, aided by ophthalmological societies and working groups or committees. To further improve the quality of reference standards, it may be beneficial to confirm the presence of disease features on other imaging modalities, or to use available clinical patient data (De Fauw et al., 2018). In addition, reference standards may benefit from measures for grading uncertainty. The monitoring of such measures will additionally allow graders to perform consensus gradings and revisions where necessary. As AI systems for the automated identification and quantification of individual lesions are gaining more

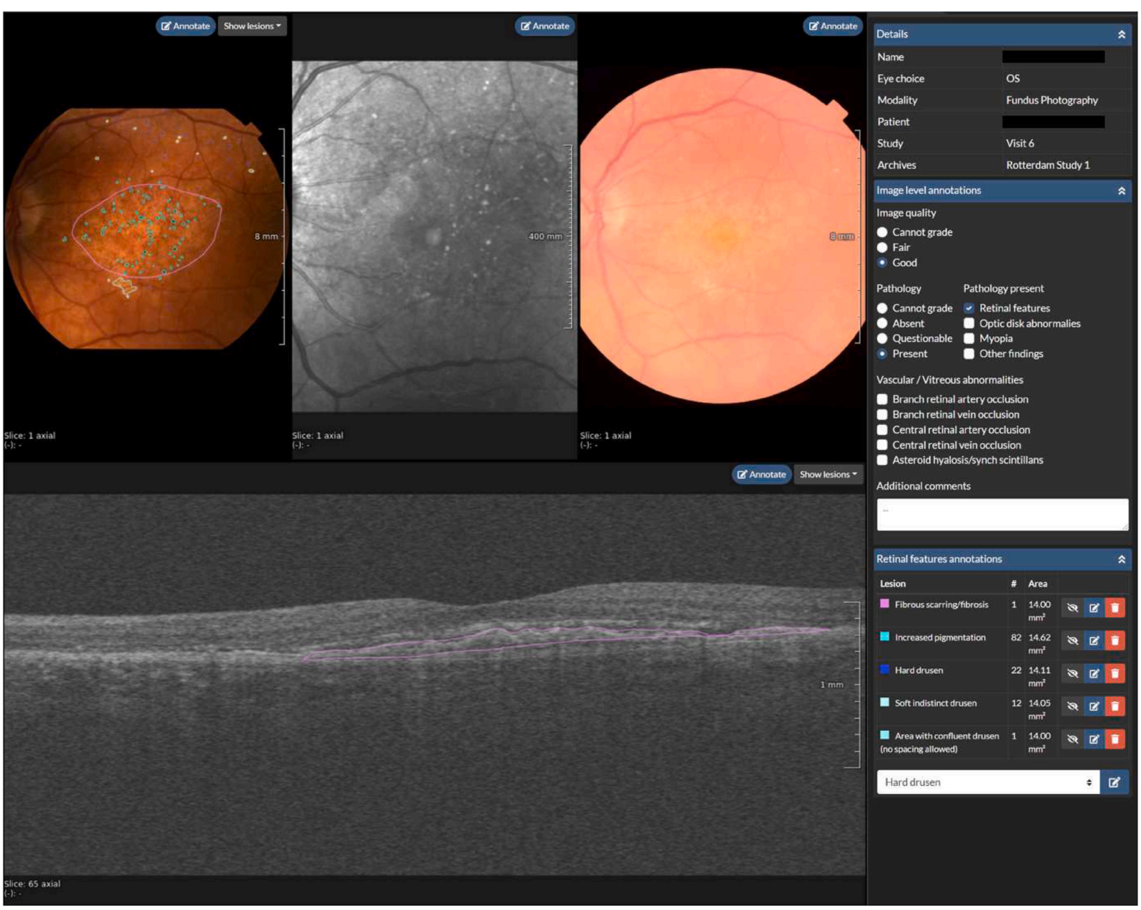


Fig. 3. Screenshot of a multimodal, vendor-independent, web-based software application, initially proposed in van Zeeland et al. (2019). The workstation allows to visualize multimodal retinal images, perform spatial alignment across modalities, perform classification and annotation tasks, and review the outputs of AI systems. It facilitates the grading of images in large studies by multiple users, the posterior comparison of annotations and creation of a consensus grading, as well as the collection of annotations for AI development and validation.

attention, future efforts should also be focused towards improving reference standards for these tasks.

4.2. Observer studies

4.2.1. Importance and consequences

Data labelling can also be performed to carry out an observer study, where the performance of an AI system is compared to that of independent graders or observers, compensating for the absence or lack of a well-determined gold standard. While only few observer studies have been performed in ophthalmology, they have shown that the performance AI systems can exceed that of observers in different tasks (Liefers et al., 2020, 2021; Chen et al., 2021; González-Gonzalo et al., 2020b; Venhuizen et al., 2017; van Grinsven et al., 2013, 2015). Several factors must be considered when performing an observer study, including the selection of observers and labelling protocol, the labels used as reference standard, the setting (i.e., daily clinics or reading center), and the design (i.e., retrospective or prospective). Careful analysis of these factors will allow for a fair and meaningful comparison of AI systems with the current standard of care, particularly if the observer study is performed prospectively within a real clinical or screening workflow.

4.2.2. Proposed solutions and considerations

Studies have highlighted the importance of selecting observers who reflect the standard of care in the application of AI systems (Verbraak et al., 2019; Tufail et al., 2016). For instance, AI systems developed for DR screening should be compared with observers who routinely grade DR in a screening population, including regional differences in the standard of care. A large observer study by Ruamviboonsuk et al. (2019) applied a protocol where observers were only assigned images from their own region, which allowed to evaluate the robustness of an AI system in different regions in Thailand.

Including observers at different levels of expertise is also important in observer studies, since it provides additional insight into system-to-observer comparisons. An observer study by Phene et al. (2019) included 3 glaucoma specialists, 4 ophthalmologists, and 3 optometrists to evaluate an AI system that detects glaucoma on CFP. The AI system reached a higher sensitivity for glaucoma detection than 7 out of 10 observers, including 2 out of 3 glaucoma specialists, but only a higher specificity than 3 observers (including 1 glaucoma specialist), while remaining comparable to others. Chen et al. (2021) included 13 observers at 3 different levels of seniority to evaluate a system that detects reticular pseudodrusen, a key feature in AMD strongly associated to disease progression, on CFP and FAF. They were able to show that the AI system was superior to the highest level of seniority.

As emphasized previously, the use of similar labelling protocols for the training and validation of AI systems should be ensured. This also applies to the labelling protocols used in observer studies (Thee et al., 2020; Verbraak et al., 2019). It should be kept in mind that healthcare providers or graders who act as independent observers or as reference standard for observer studies can deviate from the protocols that were used to develop AI systems. Therefore, it is important to align the methods for observation with those of the original standards to avoid conflicts that are based on differences in definitions.

An important challenge in observer studies is the establishment of a solid reference standard for the AI system and the observers. As discussed in Section 4.1, high-quality reference standards often require expert consensus gradings, which are time-consuming and challenging to organize (Wilson et al., 2021). To bypass this issue, developers have used different methods. In Liefers et al. (2021), an observer study included 4 graders to analyze the performance of an AI system to quantify 13 key retinal features in early and late AMD, including different types of fluid and constituent features of geographic atrophy. 3 out of 4 graders were used to create a reference standard, based on the overlap of their annotations; the 4th grader acted as an independent observer to obtain an estimate of human performance. The set of 3 reference graders was then rotated, allowing the developers to use the gradings of all 4 human observers for system-to-observer comparison. Hence, the reliability of the reference standard was enhanced by using the combined output of observers, and variability in observer grading was obtained by rotating the reference standard (Fig. 4). Using a similar methodology, Lee et al. (2017b) involved 4 observers for the manual annotation of intraretinal fluid in OCT scans, where each observer served as the reference standard for the other 3 observers, again allowing developers to rotate the reference standard and compare 4 observers against the AI system.

The annotations performed in observer studies also allow to analyze the reliability of human gradings for the included features and identify those that have an inter-grader variability low enough to be used as biomarkers for disease progression and therapeutic effects in future interventional clinical trials. For instance, Müller et al. (2021) studied the reliability of human gradings for different features associated with AMD progression, based on the observer study carried out by Liefers et al. (2021). Features with low inter-grader agreement might be inherently problematic for humans to detect and quantify, and their utility as surrogate biomarkers in clinical studies is therefore limited. In this regard, AI systems may allow for a higher consistency in performance than human graders (when developed with a solid reference standard), making it possible to include a wider variety of reliable surrogate biomarkers (Müller et al., 2021).

As more observer studies for AI systems in ophthalmology will become available in the future, other challenges with regard to system-to-observer comparisons will come to light. For instance, it is likely that observers adjust their performance differently in experimental settings, where their assessment will not directly affect patients' outcomes, than in real-world settings. This might be a challenge for extrapolating results from experimental to real-world settings. However, it can be overcome by performing prospective observer studies within existing clinical or screening workflows. In any case, it is imperative that AI developers, reading centers, and healthcare institutions are transparent in their selection of observers, labelling protocols, and reference standards to evaluate system robustness against the standard of care. This will help reveal for which ophthalmic applications AI systems have already reached clinically-acceptable performance and for which they are not ready yet.

5. Training and retrospective validation

Once the necessary data have been collected and labelled, they can be used for the training and/or validation of the AI system. AI systems are based on mathematical models whose aim is to mimic human capabilities by following certain rules. Within AI, *machine learning* (ML) models automatically learn these rules by analyzing the examples in known, labelled data. In this process, known as *training*, models are taught to extract important *features* in the training data to perform a given task. In traditional ML approaches, feature extraction was usually done manually. On the other hand, in *deep learning* (DL), which is a special type of ML, models are able to automatically discover important features using a *deep neural network* (DNN), governed by thousands or millions of parameters. During training, a DNN sequentially optimizes its parameters to reduce the error in its predictions using an objective or loss function, until the performance converges. *Convolutional neural networks* (CNN) are a type of DNN most commonly applied to imaging data.

Once the model (a DNN in the case of using DL) is optimized to solve a certain task, it can be applied for inference, i.e., to predict outcomes on similar, new data. If labels are available for these data, they can be used to validate or test the performance of the model as well as compare it to that of human experts as part of a *retrospective validation*. It is key to use validation data that are collected from different sources, external to the source of the training data, in order to analyze the capability of the model to *generalize*, i.e., to achieve the same performance or as close as

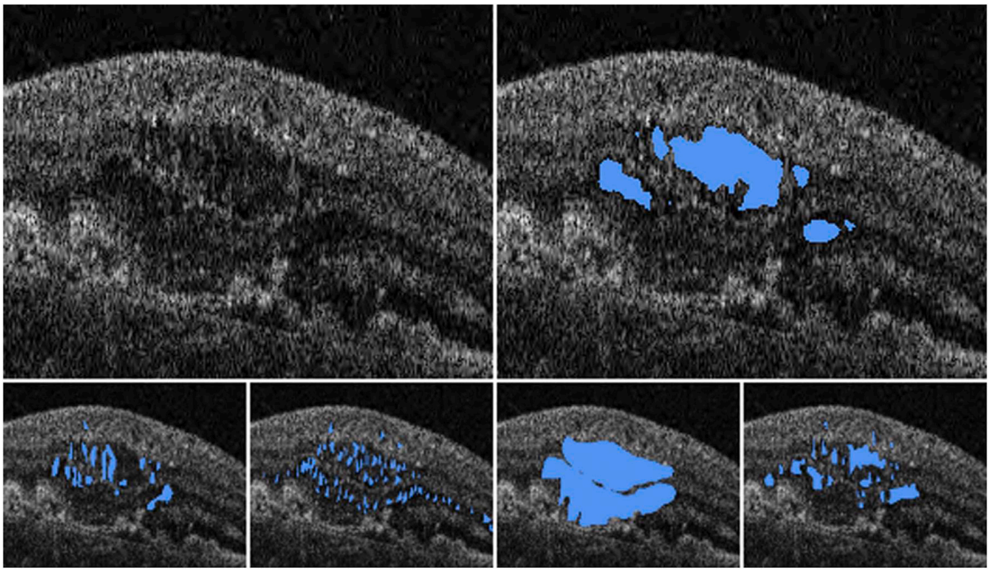


Fig. 4. Observer study to compare the performance of an AI system with that of 4 independent graders for the quantification of key retinal features in early and late age-related macular degeneration (AMD) in optical coherence tomography (OCT) scans; in the figure: segmentation of intraretinal fluid (Liefers et al., 2021). The figures shows the original B-scan (top left), the output of the proposed AI system (top right), and the annotations of the 4 graders (bottom row). 3 out of 4 graders were used to create a reference standard for the observer study, based on the overlap of their annotations; the 4th grader acted as an independent observer to obtain an estimate of human performance. The set of 3 reference graders was then rotated, allowing the developers to use the gradings of all 4 human observers for system-to-observer comparison. By comparing the system's output against the combined output of multiple observers, the reference standard was more reliable than that obtained from a single observer. By rotating the reference standard, the performance of the system was not just compared against a single observer who might, for example, grade more conservatively than others.

possible to that achieved on the training data.

In this section, we focus on five aspects that need to be considered during the training and validation of trustworthy AI systems in ophthalmology:

- The quality of the input data across sources and the completeness of such data, for instance, when making use of multimodal information.
- The impact of algorithmic bias and domain adaptation on a system's capability to generalize across populations, ophthalmic settings, and data acquisition protocols.
- The importance of explainability to open AI's "black box" and its impact on trust and clinical usability.
- The robustness of systems against malicious attacks and the importance of defining their actual threat.
- The capability of systems to adapt to users' feedback, new clinical practices, or data shift, and be updated to maintain their performance over time.

For each aspect, we identify mechanisms and considerations to lessen their potential negative effects in later stages and facilitate AI integration in ophthalmic practice.

5.1. Input quality

5.1.1. Importance and consequences

The quality of the imaging data used as input to an AI system can affect its ability to provide a valid and correct decision, the same way an image of suboptimal quality can affect an expert's ability to interpret it and guarantee a valid diagnosis. Input quality can have an important impact on AI systems' performance (Lee et al., 2021), but also on healthcare providers' and patients' experience, as recently shown in a prospective setting (Beede et al., 2020).

It is therefore crucial to consider input quality during AI development. Two viewpoints can be currently distinguished. The first one considers that only high-quality data should be used as input to the AI system, thereby excluding low-quality or ungradable images for the training and/or validation of the system (Abràmoff et al., 2016; Gulshan et al., 2016). This way, a minimum quality standard is required by the system, in the same way that human experts require a certain level of image quality for extracting valid conclusions. The second viewpoint advocates for including real-world data from the very early stages of development. This viewpoint considers that across settings and patients there is an inevitable variability in image quality due to light conditions, technicians' experience, mydriasis, presence of cataracts, etc. (Lee et al., 2021). Following this notion, systems are developed to expect different standards of input quality. What is clear is that the way input quality is handled during AI development can condition performance and usability during deployment. For this reason, recent guidelines require to specify the procedure to assess and handle poor-quality input data, contributing to transparency and trustworthiness in AI (Rivera et al., 2020).

The quality of the input can also refer to its "completeness", i.e., how well it reflects the clinical reality and exploits available patient information. Despite the rich variety of data sources used in ophthalmic practice, the integration of multimodal data for AI development is highly limited, especially when compared to other medical specialties (Saha et al., 2021; Chen et al., 2020). For example, a system trained to detect AMD in standard CFP will be able to provide correct predictions only on CFP; it will not work if provided with images from other fundus technologies, such as FAF, ultra-widefield images or multicolor scanning laser images. However, clinical tasks at different points of patient care often require multiple contexts, such as imaging data from different modalities, the social background of the patient, and his/her medical record (van Dijk and Boon, 2021). The use of multimodal data in ophthalmology presents different potential benefits. On one hand, training AI models based on data from different sources would allow the automatic extraction of complementary information, as usually done in the clinic. For example, intraocular pressure measurements, visual fields information, and structural parameters derived from CFP and OCT scans are generally combined by ophthalmologists to diagnose and monitor progression of glaucoma. Information from multimodal images is also usually combined to better assess the severity of DR and AMD, including the confirmation on the presence of small or subtle disease features. Automated analyses of multimodal data creates the possibility of generating more powerful models, making them also more reliable as they approach the way human experts work in practice. On the other hand, combining ophthalmic data could help reveal more than just "signals" appearing in some patients and could potentially increase the chances of identifying real biomarkers for Alzheimer's disease, for instance. As long as the use of multimodal ophthalmic data is not assimilated in AI development, certain disease factors and associations will remain disregarded and their potential value for AI decisions unknown.

5.1.2. Proposed solutions and considerations

Automated assessment of the input quality is often integrated in proposed AI systems. Some systems assess the quality of the input image prior to processing it, rejecting those images for which the system cannot guarantee a strong prediction. For instance, in Abràmoff et al. (2018), a module with independent detectors is integrated to evaluate different quality characteristics in CFP (focus, color balance, exposure). This allows to determine whether the input quality is sufficient to be used for automated DR screening, and in case it is not, whether this is due to the field of view or image quality, based on the criteria previously proposed by Niemeijer et al. (2006). Prior quality assessment can be performed similarly for OCT scans, such as in the workflow proposed by Wang et al. (2019), where it is used to discard cases that are off-centered, have total or partial signal loss, or have other types of artifacts before performing automated detection of retinopathy (Fig. 5). Prior quality assessment becomes especially useful in "online" settings where immediate feedback about image quality is beneficial, such as in screening settings. It may motivate camera operators to re-image a patient for higher-quality images (Wang et al., 2019), and cases with insufficient quality can be referred to a specialist, ensuring patient safety (Beede et al., 2020). It is important to consider an adequate threshold to avoid unnecessary referrals and an increase in experts' workload (Beede et al., 2020). Immediate quality feedback also contributes to establish and safeguard image quality standards at clinical settings, not only for AI systems, but also for human assessment. This can be especially helpful for new imaging modalities, such as OCT angiography (OCTA) (Hormel et al., 2021).

Other proposed systems perform AI-based image quality assessment in parallel to their main application. In González-Gonzalo et al. (2020b), an additional image quality score is provided for a system that performs joint automated detection of DR and AMD in CFP. In De Fauw et al. (2018), they include three additional classes related to OCT input quality in a segmentation model, in order to detect mirror, clipping, and blink artifacts, together with different types of healthy and pathological tissue. They also make use of an ensemble of five segmentation networks, which allows to identify ambiguities in the system's decision and could potentially be used for automated quality control. This aspect is related to approaches based on uncertainty quantification of systems' decisions (discussed in Section 5.3), which could be used to flag low-quality images during inference (de Vente et al., 2020). Parallel quality assessment is especially helpful in "offline" settings where there is no chance to retake images and data are analyzed in a retrospective manner, such as in teleophthalmology settings or in clinical studies. The additional feedback on input quality improves the context of AI-based outputs, and can be used to better judge the reliability of such outputs when making posterior clinical decisions.

Certain aspects need to be considered when it comes to automated quality assessment. Firstly, AI-based approaches require human annotations on image quality or gradability that can be subjective even when a labelling protocol is provided, resulting in high rates of disagreement across graders (Ruamviboonsuk et al., 2019). As seen in Section 4.1, the reliability of the reference standard can have an impact on the system's performance and usability. Secondly, generalizability across settings becomes particularly difficult when it comes to quality assessment. Different types of image artifacts from the ones considered during development could be found in new data, due to different acquisition conditions (such as the available camera, the level of expertise of the camera operator, light conditions, mydriasis…), and different quality standards across settings (Lee et al., 2021). These factors need to be analyzed in order to integrate automated quality assessment in AI systems, preferably in a prospective way (Beede et al., 2020). It is possible to analyze the impact of different image quality factors on the system's performance (Yip et al., 2020). Factors can include the resolution of the input, the rate of image compression (especially relevant in teleophthalmology settings), the presence of cataracts, mydriasis, or the type of camera used, while comparing the performance of the system on high-quality data with that achieved on mixed data (high and low-quality images) (Wang et al., 2019).

To generate systems that are more agnostic to changes in input quality, commonly used techniques can also be helpful. Pre-processing pipelines, such as contrast enhancement in CFP (González-Gonzalo et al., 2020b), help with the standardization of the input images. Data augmentation during training helps increase robustness against noise and image artifacts, such as adding artificial speckle noise in OCT scans (Venhuizen et al., 2018). Although less mature at the moment, denoising techniques based on GANs have also potential to increase usability of AI systems with lower-quality inputs (Yoo et al., 2020). Future techniques could also explore the way human experts deal with suboptimal input quality. For example, in settings where multiple images from one patient visit are available, a system could consider the whole batch of images, with possible varying quality, as input and automatically extract valuable information from each image while discarding noisy information,

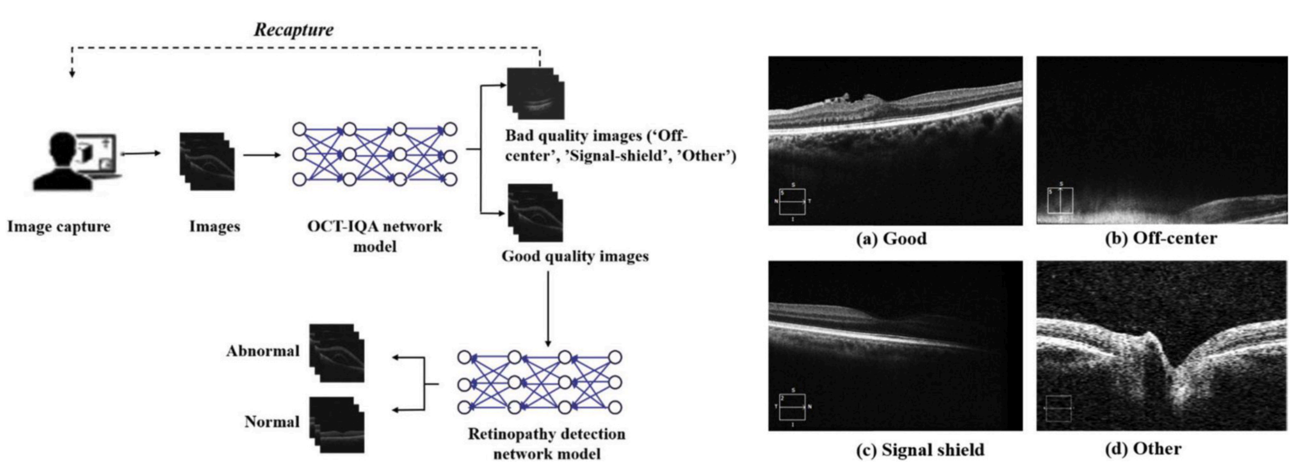


Fig. 5. Prior quality assessment integrated in an AI system for automated detection of retinopathy in optical coherence tomography (OCT) scans (Wang et al., 2019). On the left, the proposed pipeline is shown: AI-based quality assessment is performed to discard input images with bad quality that can affect the reliability of the system's output (images that are off-centered, have total or partial signal loss, or have other types of artifacts), prior to apply automated detection of retinopathy. Immediate quality feedback may motivate camera operators to re-image a patient for higher-quality images. On the right, examples of OCT scans with different qualities.

instead of expecting one high-quality image.

Regarding the completeness in the data used by AI systems, recent solutions have shown the potential of extracting and exploiting information from different modalities of ophthalmic imaging. In van Grinsven et al. (2015), a ML model was proposed to detect the presence of reticular pseudodrusen using multimodal information from CFP, FAF, and NIF, and showed that it achieved higher performance than that obtained when using images from each modality separately. Additionally, the authors showed that the multimodal ML model performed within the same range as human graders, who also achieved higher performance and better agreement in a multimodal grading approach. In Yoo et al. (2019), two separate CNNs, one for CFP and one for OCT scans, were used to extract features from paired images; the extracted features were then concatenated and used as input to a traditional classifier to detect the presence of AMD. A recent study by Chen et al. (2021) proposed a system to detect reticular pseudodrusen consisting of three models: one for CFP, one for FAF, and one for CFP-FAF image pairs based on the features extracted by the other two models (Fig. 6). The three models were trained simultaneously while using attention mechanisms per model and cross-model to combine features from the different modalities, instead of concatenating them as done in previous approaches. This resulted to be advantageous for the system's performance and generalization. Another advantage of this approach is that each model in the system was then fine-tuned separately using only images of the corresponding modality, which allowed to optimize the system and make it useable when only one of the modalities is available. Other applications have explored the integration of patient metadata with different imaging modalities. For instance, Mehta et al. (2021) proposed a system for glaucoma detection based on an ensemble of one CNN for CFP, another CNN for OCT scans, and one ML model for demographic, systemic, and ocular metadata. With the help of interpretability techniques, discussed in Section 5.3, they observed that their system suggested distinct sources of information from each imaging modality and the clinical variables that were relevant for the automated detection of glaucoma.

There are different elements of consideration when it comes to multimodal solutions in ophthalmology. A multimodal dataset needs to be accessible for development, but may be harder to curate; multimodal data may not be available for all the subjects of interest and most publicly-available datasets are based on a single imaging modality. As for the labelling process, the annotations required to establish the reference standard in classification tasks can be performed in one imaging modality and transferred to the other modalities (Chen et al., 2021; Mehta et al., 2021). However, in order to obtain aligned and consistent annotations for segmentation tasks, it might be necessary to perform image registration, which is a field that still requires further research. Most proposed multimodal solutions require all modalities to be available, also during deployment. This limits their accessibility especially when advanced imaging modalities, such as FAF or OCTA, are required. It is thus key to generate systems that benefit from combining information from multimodal data during development, while being optimized to perform well in settings where perhaps only a single source of data might be available (Chen et al., 2021), such as screening settings in rural areas, where only CFP is available. Additionally, multimodal solutions might be less practical due to longer acquisition times, increased mydriasis, and due to the requirement of having experienced technicians and patients who are fully compliant (Yip et al., 2020). Healthcare providers and institutions, as well as ophthalmological societies and working groups or committees, can help define the clinical applications, settings, and conditions under which the use of multimodal data will be most beneficial in order to have the best alignment of AI systems with clinical practice.

5.2. Bias and domain adaptation

5.2.1. Importance and consequences

While AI systems can easily achieve state-of-the-art performance on many computer vision tasks, they often lack the ability to generalize well to images that are outside the distribution of the training domain (Bengio et al., 2017). Even when the training set may be balanced for the

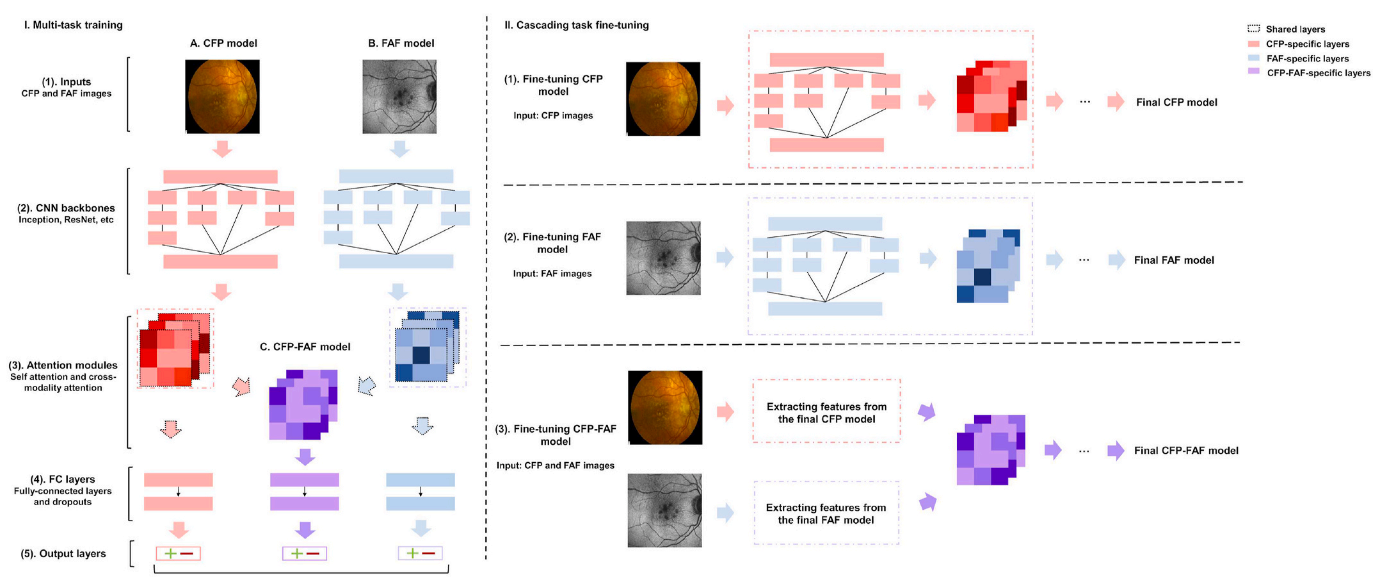


Fig. 6. Multimodal solution for AI-based detection of reticular pseudodrusen, a key feature in age-related macular degeneration (AMD) strongly associated to disease progression, from color fundus photographs (CFP) alone, their corresponding fundus autofluorescence (FAF) images alone, or CFP-FAF image pairs (Chen et al., 2021). The proposed system is composed of three models: one for CFP, one for FAF, and one for CFP-FAF image pairs. In the first stage of development, the three models (CFP-alone, FAF-alone, and CFP-FAF models) are trained simultaneously. The CFP model takes a CFP as input, and the FAF model takes the corresponding FAF image as input. Each image is processed by a convolutional neural network backbone, followed by an attention module to capture important image features. The important features captured from the CFP and FAF images form the basis of the CFP-FAF model, using cross-modality attention. In the second stage, the three models are fine-tuned separately, using the corresponding modality as input. The CFP and FAF models are trained first; the attention modules of the finalized CFP and FAF models are used to extract features for fine-tuning the CFP-FAF model. This makes the system useable when only one of the modalities is available.

intended target labels, the distribution of the images may be imbalanced with respect to other conditions. For example, a dataset for training a binary classifier to distinguish cats vs dogs may have equal number of both animals, but since pictures of cats are found mainly indoors, a trained classifier may generalize to perform poorly when given a picture of a cat outdoors at test time. While this may seem a trivial example, this effect translates to algorithmic bias in the medical domain that is important to quantify and protect against. Algorithmic bias has important medical ethics considerations that are directly related to inequities in healthcare (Obermeyer et al., 2019; Abràmoff et al., 2021).

For instance, a dataset that is used to train an AI system on CFP for the classification of AMD will most likely be unbalanced on at least two different axes: age and race. This stems directly from the epidemiological nature of AMD in that the more severe forms of AMD happen later in life and that Caucasians are more likely to have AMD compared to other races (Klein et al., 2004). AI systems have two possible fallacies when trained on such a dataset. First, they may generalize poorly to patients who are of a different race since the pigmentation of the retina is of a different color distribution than in Caucasians. This is referred to as *source bias*. The second is that they may be dependent on a correlated spurious feature and thus have a *context bias* and lead to shortcut learning. For example, prevalence of AMD is higher in older patients, but other changes due to aging can be observed in these subjects as well, such as cataracts or certain vascular abnormalities. If an end-to-end model is trained on a dataset where a subset of the images contain both AMD as well as other features unrelated to the disease, the resulting classifier may not be extracting features relevant to the disease in question and instead learning spurious correlated features. These two types of biases can work towards directly harming patients in clinical care where AI systems are deployed. The decreased performance may lead to incorrect automated diagnoses or segmentations and, eventually, inappropriate clinical management of the patient.

Another aspect that conditions the generalization of AI systems in ophthalmology is the lack of domain adaptation across devices for a given imaging modality. Given the fragmentation in the market and the proliferation of different vendors and devices (seen in Section 2.2), this weakness in AI systems manifests in a critical manner. In the context of OCT imaging, most clinical centers have adopted one particular OCT scanner and thus the large training sets derived from these centers are heavily represented by a single imaging device. The resulting trained models generalize poorly to other devices (De Fauw et al., 2018; de Vente et al., 2021). Even though visually OCT B-scans from different vendors may appear similar to clinicians' eyes, each device model performs different post-processing of the OCT signal leading to differences in texture, speckle noise, brightness, and contrast. These differences can lead to poor performance of deep learning models even between different versions of the same OCT scanner (Owen et al., 2021).

In line with the generation of trustworthy AI systems, fairness and generalizability are increasingly included in official guidelines, which urge to assess the communication of perceived biases and the measures taken for identification and mitigation (European Commission, 2019). Related policies are also under consideration, such as a data quality control for AI systems to be introduced in the NHS (Harwich and Laycock, 2018). Addressing bias and domain adaptation is therefore a priority for AI integration.

5.2.2. Proposed solutions and considerations

Solutions to mitigate bias and lack of domain adaptation of an AI system can be generally categorized into *pre-processing* solutions, aimed to modify the development data before training, *in-processing* solutions, aimed to modify the system during training, and *post-processing* solutions, applied after training the system (Mehrabi et al., 2021).

Pre-processing solutions try to modify the development data so that underlying bias and domain adaptation issues are removed before training the system, while keeping data usability. The main goal of these solutions is to align the development data with the data found at the point of care where a system is aimed to be applied, so that it is able to generalize across domains and adapt to real-world disease prevalence and population distributions (Faes et al., 2020). Some approaches can be applied during data collection, for instance, by collecting data uniformly from different population groups or domains (different camera devices or scanners, etc.), with the possibility of oversampling unrepresented groups (Parikh et al., 2019), and including such data in the training stage (and not only when performing an external validation) (Phene et al., 2019). If bias is still observed, it might be possible to inject supplements of missing data or to perform distributional shifts (Esteva et al., 2021). These solutions are nevertheless hindered due to the hurdles of (multi-centered) data collection (as seen in Section 3), and become unfeasible when using retrospective data for development. Consequently, solutions based on GANs have gained an increasing interest in different applications, with the aim of generating synthetic ophthalmic images that are anatomically consistent and indiscernible from real data (Bellemo et al., 2018). GANs can be used to synthesize data of group minorities or less frequent anomalies or disease stages, as done in Burlina et al. (2021) and Joshi and Burlina (2021), where racial bias is addressed in the context of automated screening of DR and AMD in CFP, respectively (Fig. 7). However, it is currently hard for generative models to capture all biological variability of a protected factor (such as the presentation of a low-prevalence disease across ethnic groups, age groups, image acquisition protocols, etc.), or to control a specific feature while keeping other attributes unchanged (Joshi and Burlina, 2021). More work is necessary until solutions based on GANs are mature enough.

In the case of lack of domain adaptation across OCT scanners, pre-processing solutions are commonly applied, such as image size and intensity standardization across scans from different vendors or domain-specific data augmentation protocols (Bogunovic et al., 2019; Venhuizen et al., 2018). Solutions based on GANs have also been explored to ensure generalization across OCT devices, specifically, those based on CycleGANs, which allow to perform unsupervised image translation from one domain (i.e., vendor) to another. For instance, Romo-Bucheli et al. (2020b) applied this technique to improve the generalization of AI-based segmentation of fluid and the photoreceptor layer in OCT scans.

In-processing solutions tackle bias and domain adaptation during the training process of the system by modifying the learning algorithm. Some approaches aim to mitigate algorithmic bias in the objective/loss function through regularization or imposing constraints that force the algorithm to account for protected factors (ethnicity, age, sex…) (Zhang et al., 2018). Other in-processing approaches are based on alternative learning techniques. Low-shot learning, a type of machine learning aimed to maintain algorithmic performance when there are limited data for development, has shown potential to address AI bias due to development data that may have few examples from certain population groups or low-prevalence ophthalmic diseases (Burlina et al., 2020). Multi-task learning has also been shown beneficial to increase generalizability and usefulness in this scenario, by training models that are able to perform a variety of tasks, from which more examples are available, rather than one narrowly defined task (Chen et al., 2021; Robinson et al., 2021; Asgari et al., 2019). Curriculum learning is also beneficial for the detection of low-prevalence ophthalmic diseases or features; it allows models to exploit hierarchical information available in the training data, enabling knowledge transfer from general, higher-prevalence features to specific, low-prevalence features (González-Gonzalo et al., 2021). An important consideration is that a trade-off between algorithmic performance and fairness/generalizability arises when applying these approaches. Some of them make use only of "traditional" performance metrics, such as accuracy or F1 score, which do not capture completely such trade-off nor different fairness requirements (Mehrabi et al., 2021), which are then disregarded but remain relevant regarding applicability of AI systems in real-world clinical settings. The inclusion of fairness metrics (Gajane and Pechenizkiy, 2017) will be therefore key to

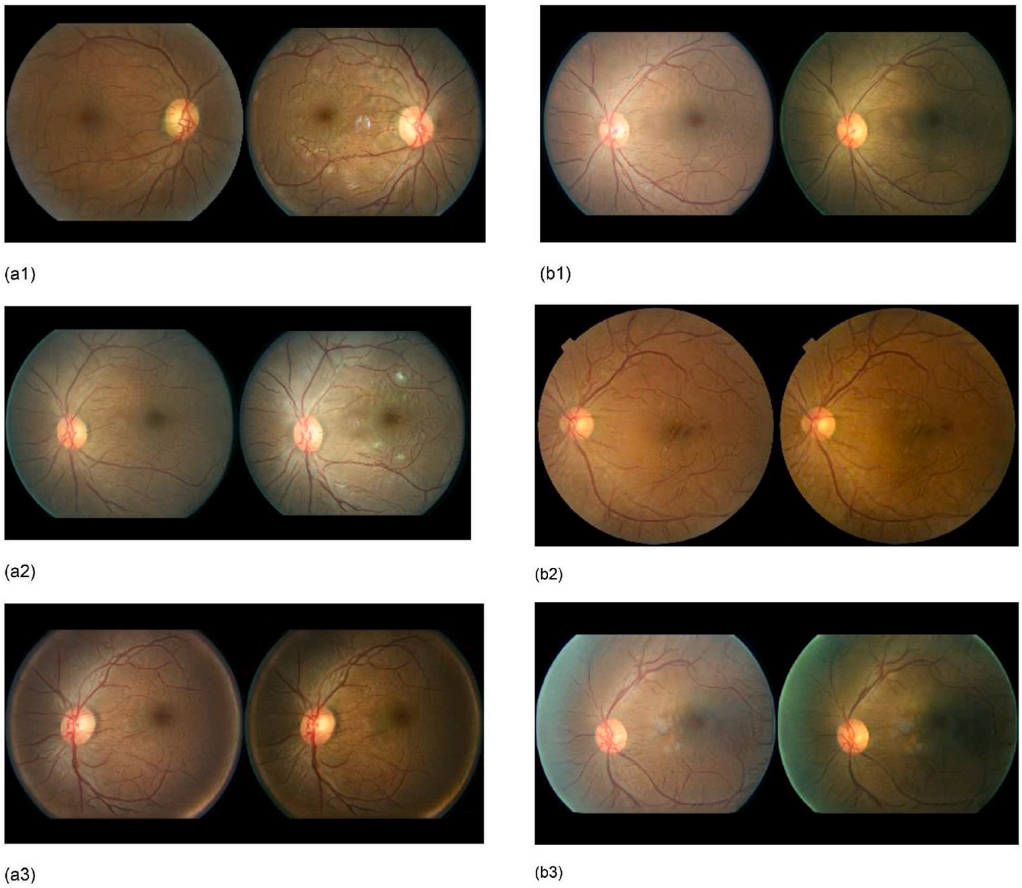


Fig. 7. Generation of synthetic ophthalmic images via generative adversarial networks (GANs) as preprocessing solution to mitigate bias in the context of AI-based diabetic retinopathy (DR) screening in color fundus photography (CFP) (Burlina et al., 2021). In this study, GANs are used to generate synthetic images and perform a subsequent fine manipulation on them to accentuate specific factors that are originally missing in the development data (i.e., images of referable DR from darker-skin individuals). From a1 to a3: a synthetic image corresponding to a darker-skin individual (left) is used as input to generate a new image that accentuates the attribute "referable DR" (right), leaving the amount of coloration due to the melanin concentration and all other markers unchanged. From b1 to b3: a synthetic image corresponding to an individual with referable DR (left) is used as input to generate a new image that accentuates the attribute "darker-skin individuals" (right), while preserving the DR lesions as well as the vasculature.

determine whether the outputs of an AI system are fair or not for protected groups.

When it comes to domain adaptation across OCT scanners, one common in-processing solution is to train different models using development data from each different vendor (Schlegl et al., 2018). When fewer scans from a given vendor are available for development, it is possible to use the parameters of a model trained with data from another vendor to initialize and fine-tune a new model for that given vendor. For instance, in De Fauw et al. (2018), a segmentation model able to segment different types of healthy and pathological tissue in scans from vendor A is re-trained using a mixed dataset with scans from vendors A and B, in order to generate a new segmentation model able to reach high performance in scans from vendor B (Fig. 8). In the case of automated classification in OCT, labels are commonly available at volume level, which supposes an additional hurdle to generalize across vendors, due to not only differences in appearance but also in B-scan spacing. In order to overcome this, de Vente et al. (2021) proposed a multiple instance learning approach where intermediate information is extracted from each B-scan separately, being invariant to B-scan spacing, and then combined to provide a grade of AMD severity for the full OCT volume (Fig. 9).

The integration during development of interpretability and uncertainty techniques, further explored in Section 5.3, does not only provide a better understanding of a system's decisions, but can also be used as a tool to identify bias and lack of domain adaptation. These techniques allow to visually check if a system is focusing on clinically meaningful areas of the input data to reach its decisions (González-Gonzalo et al., 2020a), and to detect samples that fall out of the distribution of the data used for development (de Vente et al., 2020).

Post-processing solutions are aimed to be applied once the system has been trained, without modifying the development data or the learning algorithm. Since we consider that preprocessing and/or in-processing solutions should be considered when addressing bias and domain generalization during AI development, we recommend to use post-processing solutions only in addition to other solutions, or when dealing with a system already in deployment. Some post-processing approaches aim to alter a system's predictions to increase fairness, for instance, by keeping the proportion of decisions of protected and unprotected groups or focusing on the predictions that fall close to the system's decision boundary (Mehrabi et al., 2021). Tools such as the one proposed in Wexler et al. (2019) are specifically designed for probing implemented models under different hypotheses and fairness definitions, as explored in Singh et al. (2021) in the context of automated prediction of visual acuity in diabetic macular edema patients.

A correct preparation and setup of a system's validation are also essential for the detection and mitigation of bias and lack of domain generalization. A multi-institutional and multi-vendor setup is important to determine the generalizability of a system (Lee et al., 2021). Healthcare institutions and reading centers, potentially aided by ophthalmological societies and working groups or committees, can help identify and determine which bias and domain adaptation factors need to be acknowledged during validation. For this purpose, it becomes imperative to count with a diverse team and provide the necessary education for bias awareness to all stakeholders involved.

5.3. *Explainability*

5.3.1. *Importance and consequences*

Although AI systems have achieved expert-level performance, they are often referred to as "black boxes" due to the lack of interpretability or explainability of their predictions and decision-making processes. This results in an important challenge for their integration in clinical

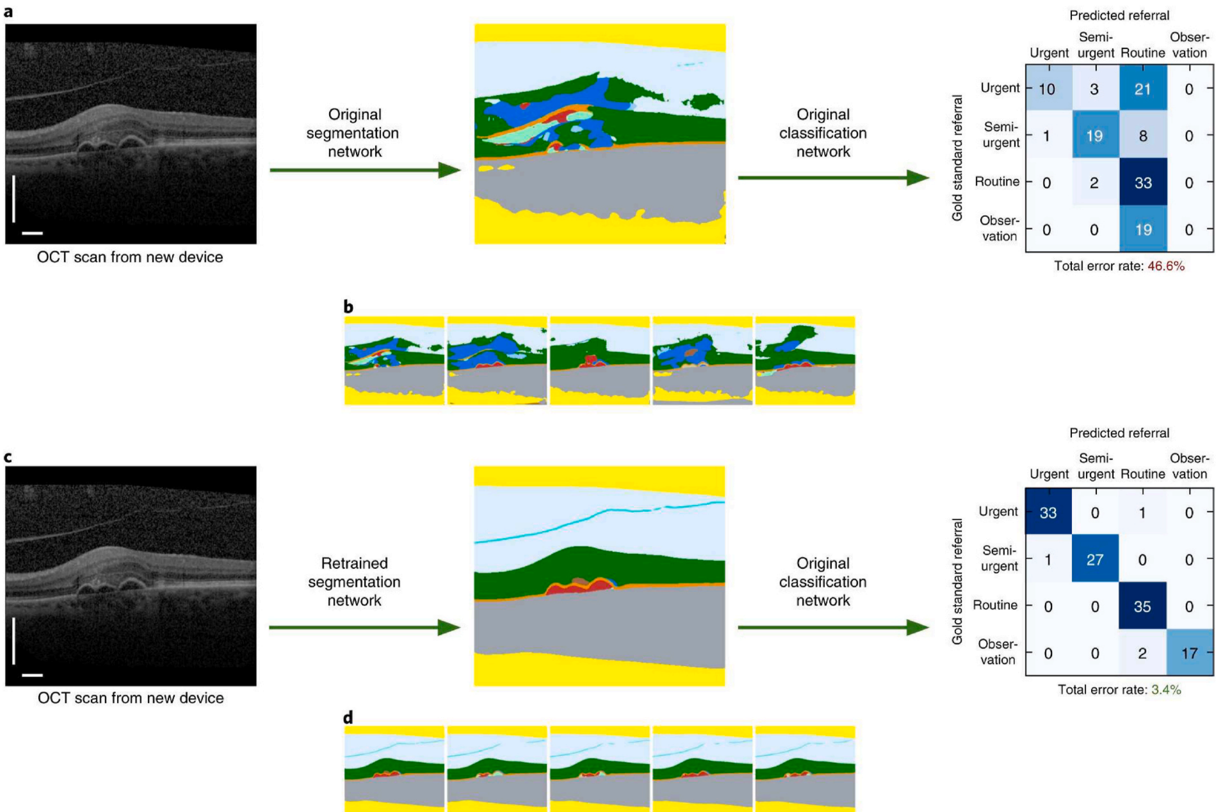


Fig. 8. Re-training of a segmentation model and vendor-agnostic classification as in-processing solutions to mitigate lack of domain adaptation across optical coherence tomography (OCT) scanners from different vendors (De Fauw et al., 2018). In this study, a two-stage approach is proposed: first, they use a segmentation model to generate intermediate representations of different healthy and pathological tissues for each B-scan in an OCT volume; second, they use these representations as input to a vendor-agnostic classification model that outputs diagnosis probabilities for different lesions and a referral suggestion at volume level. The original segmentation model, based on the combination of the output of five separate segmentation networks, is able to segment different types of healthy and pathological tissue in scans from vendor A. The figure shows how the segmentation model fails when scans from a new device of vendor B (a) are used as input. The poor quality segmentation maps generated by the five networks in the original segmentation model (b) lead to failure of the original classification model as well. However, after re-training the segmentation model with OCT scans from both vendor A and vendor B, the new segmentation model is able to reach high performance on scans from vendor B (c and d). The classification network is vendor-agnostic and, therefore, unchanged.

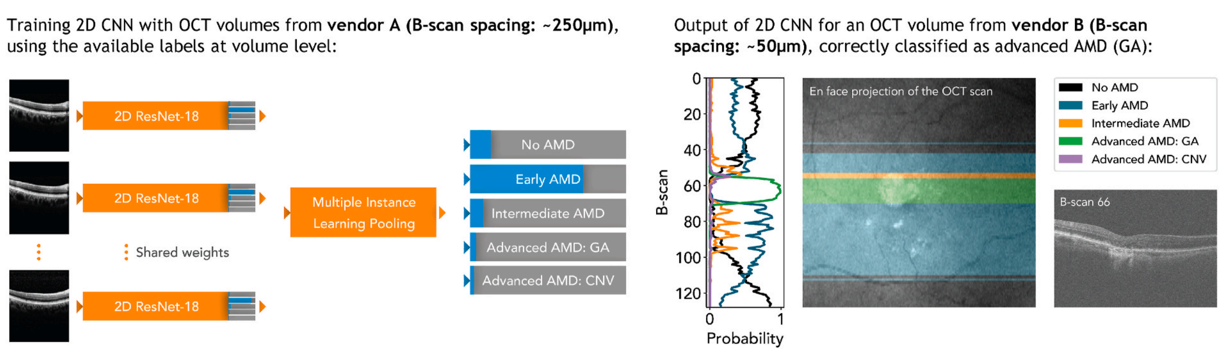


Fig. 9. Multiple instance learning approach to mitigate lack of domain adaptation across optical coherence tomography (OCT) scanners from different vendors in AI-based age-related macular degeneration (AMD) grading (de Vente et al., 2021). In this task, as well as other classification tasks using OCT, labels are commonly available at volume level, which hinders a system's generalization across vendors, due to not only differences in appearance but also in B-scan spacing. Large differences in B-scan spacing make 3D features learned on data from one vendor inappropriate for data from another vendor. As observed in the pipeline (left), the proposed approach makes use of a 2D convolutional neural network (ResNet-18) to process each B-scan separately. This aspect of the approach makes it invariant to B-scan spacing and, consequently, robust to the variability of scanning protocols across vendors. Even though volume-level annotations are used for training, grades at B-scan-level can be obtained. Then, a multiple instance learning pooling layer combines the intermediate output scores related to each B-scan and outputs an AMD grade estimation for the full OCT volume, equivalent to the grade predicted with a highest score. An AI system was trained using scans from vendor A (B-scan spacing: ~250 µm), following the proposed approach. When the system was applied to OCT scans from vendor B (B-scan spacing: ~50 µm), its performance was well maintained, as observed in the example (right).

settings (Litjens et al., 2017). On one hand, it hinders trust from healthcare providers, who might be reluctant to trust an output if provided without further reasoning, such as visual evidence of which features were discriminant for the decision. It can lead to misinterpretation of a system's output, if there is no additional information on how certain the system was about its decision. Counseling and education of patients about their diagnoses become hampered as well (Amann et al., 2020), affecting their trust and consent. Lack of interpretability also hinders sanity checks along development and deployment: the identification of errors, biases, and confounders, as well as validating whether a system is applying clinical recommendations correctly (Kelly et al., 2019).

Besides the relevance of the technical and clinical dimensions, explainability also has strong ethical and legal components. Some studies highlight that omitting explainability in clinical decision support poses a threat to core ethical values (Amann et al., 2020). Consequently, explainability is becoming increasingly required by regulatory bodies: EU's GDPR currently governs a "right to explanation" (European Commission, 2016), and the FDA requires an appropriate level of transparency aimed at users (Food and Drug Administration, 2021). Providing explanations that align with human reasoning contributes to generating trustworthy AI while increasing the intrinsic trust of users on AI systems (Jacovi et al., 2021).

5.3.2. Proposed solutions and considerations

Most focus has been put on answering "why this decision?". Among the proposed solutions, those based on visual attribution have become very popular. Attribution techniques allow to highlight features in the input image that contribute to the output prediction of an image-based classification system (Ancona et al., 2017). When applied to medical imaging, this generates a "heatmap" that highlights the areas the system considered relevant for diagnosis. These techniques have been widely applied in AI-based screening or grading of different eye diseases such as DR, AMD, and glaucoma in CFP and OCT (González-Gonzalo et al., 2020a; Mehta et al., 2021). Visual attribution has also been applied as a tool to unveil new visual features or biomarkers in diagnosis (e.g., for the estimation of retinal sensitivity from OCT scans (Kihara et al., 2019), or the estimation of refractive error from CFP (Varadarajan et al., 2018)), and prognosis (e.g., for the prediction of future DR development using baseline CFP (Bora et al., 2021), or the prediction of treatment requirement in neovascular AMD using OCT (Romo-Bucheli et al., 2020a)). Visual attribution has also been used when establishing links between ophthalmic imaging and systemic medical conditions or specific target variables, such as the prediction of cardiovascular risk factors (Poplin et al., 2018), anemia (Mitani et al., 2020), or chronic kidney disease (Sabanayagam et al., 2020) from CFP.

There are several considerations when generating explainability using visual attribution. Firstly, attribution techniques were developed and optimized with natural images, and it has been shown that they localize only the most discriminative regions (Singh and Lee, 2017). As a consequence, in ophthalmic images, abnormal areas that have less influence on the output prediction are ignored, although they could be still important for diagnosis (González-Gonzalo et al., 2020a). Additionally, interpretability of abnormal predictions requires the localization of different types of lesions of varying appearance and histologic

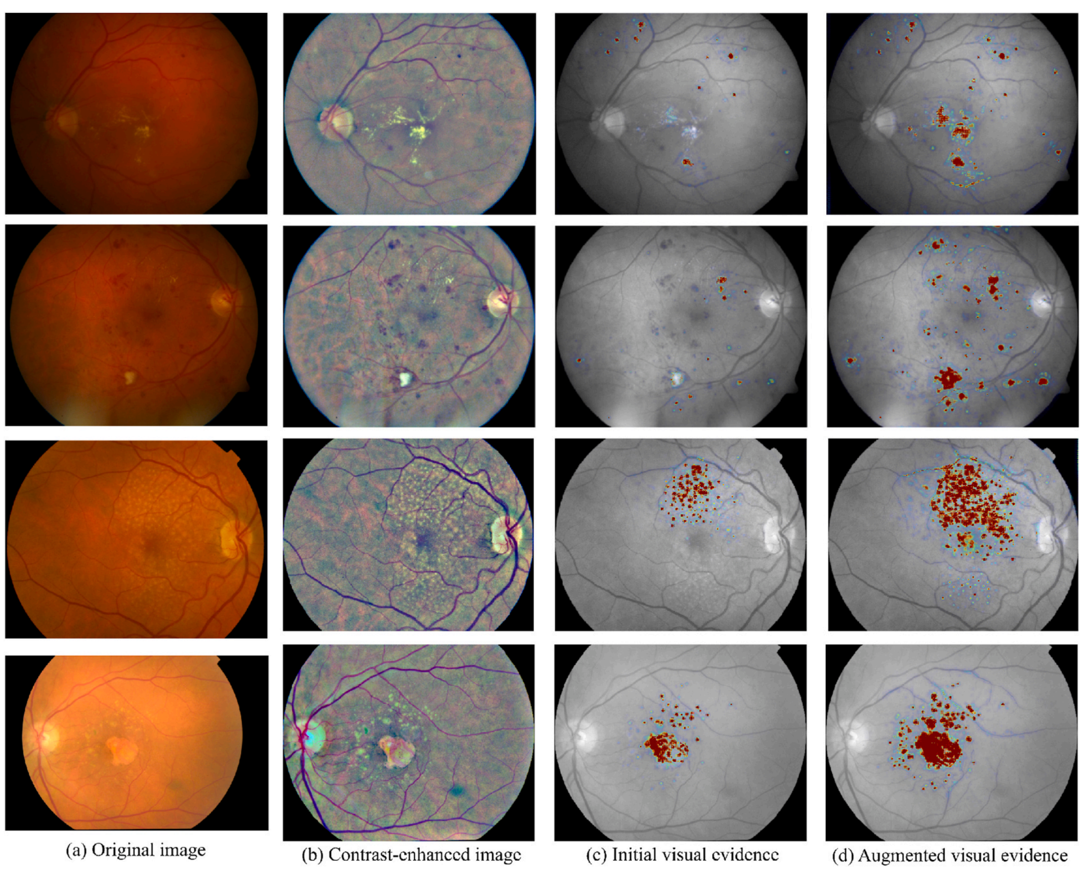


Fig. 10. Visual evidence for AI-based grading of diabetic retinopathy (DR) and age-related macular degeneration (AMD) in color fundus photography (CFP) (González-Gonzalo et al., 2020a). The figure shows the initial visual evidence, generated with visual attribution, and the augmented visual evidence, generated after an iterative process that combines visual attribution and selective inpainting. The augmented visual evidence maps highlight less discriminative areas that might also be relevant for the final diagnosis, including abnormalities of different types, shapes and sizes, and improving the system's performance for weakly-supervised lesion localization. This is shown for images correctly classified as moderate non-proliferative DR (first row), severe non-proliferative DR (second row), intermediate AMD (third row), and advanced AMD (fourth row).

composition that can be simultaneously present and be responsible for the predicted diagnosis. In González-Gonzalo et al. (2020a), it was shown that visual evidence can be iteratively augmented by combining visual attribution and selective inpainting, in a process where the abnormal regions highlighted by visual attribution are modified using healthy, surrounding local information, guiding the attention of the system to new relevant areas, including less discriminative lesions and different types of lesions relevant for diagnosis. Improved interpretability was achieved for automated grading of DR and AMD in CFP when using different attribution techniques (Fig. 10).

Secondly, only a few approaches that use visual attribution for interpretability perform further analysis to better understand if the generated visual evidence includes the biomarkers considered by experts for diagnosis. Qualitative validation is possible by having a subset of heatmaps rated by experts. In Rim et al. (2020), a survey with ten retina specialists was included to rate the correlation between actual biomarkers and the areas highlighted by a system that identifies neovascular AMD in OCT scans, using a Likert scale for different related lesions. The specialists showed strong agreement that the biomarkers the system was focusing on were correct for pigment epithelial detachment and subretinal fluid; however, for intraretinal fluid and mixed pathology, the heatmaps were not deemed satisfactory by the specialists. Quantitative validation can be performed by using a small set of images with expert lesion-level annotations to compare with the highlighted areas in the heatmaps, and to evaluate the *weakly-supervised localization* performance of a classification system. This has been done for different lesion types relevant to DR and AMD grading (González-Gonzalo et al., 2020a; Quellec et al., 2017). Although the heatmaps allowed to locate pathological areas in the classified images, the number of missed lesions and false positives might be still large so as to be deemed clinically dependable in applications where precise quantitative measures are required for clinical management. However, they can be a useful tool in diagnostic applications, where qualitative support in lesion recognition is often enough. Nevertheless, these types of validations are not possible when unveiling new visual biomarkers or when trying to establish a link between the fundus and a given systemic condition or variable. It becomes unclear how to interpret "non-traditional features" identified by visual attribution: as novel biomarkers or erroneous correlations learned by the system (Waldstein, 2020).

There is an increasing interest in approaching explainability by answering the question "how certain is the system about this decision?". The same way the confidence of a diagnosis made by a medical expert is not the same in common cases as in ambiguous or complicated cases, AI systems also make decisions with different levels of uncertainty (Kendall and Gal, 2017). Integrating uncertainty quantification has important benefits during development and deployment. During the system's development stage, it allows to check for well-calibrated uncertainties, i.e., the system is neither overconfident nor insufficiently confident in ambiguous cases (Kendall and Gal, 2017). During deployment, confidence scores provide an additional context to the system's decisions. It particularly provides context on how to interpret and how to make a correct use of the output, thereby increasing experts' trust. In de Vente et al. (2020), an ensemble of models was used to provide uncertainty estimation for AMD grading in OCT. They showed that in a well-calibrated system, high confidences were associated with correct predictions, whereas lower confidences corresponded to ambiguous cases, such as scans with questionable grading or with high levels of noise (Fig. 11). As such, uncertainty quantification can be very useful to reduce the workload in screening settings, where only those cases with high uncertainty estimates would need to be referred for subsequent expert analysis. Segmentation tasks can also benefit from integrating uncertainty techniques, so as to identify potential system's errors and ambiguous areas in the quantification of lesions.

In line with recent analyses, it is important to disprove the myth that there is necessarily a trade-off between system's performance and explainability (Rudin, 2019). This belief can lead stakeholders to forgo the attempt to generate and require explainable systems. Considering available techniques and proposed solutions, as well as the increasing efforts in this aspect, different forms of explainability can be considered and integrated in the system without damaging performance. Simultaneously, as mentioned by Schmidt-Erfurth et al. (2018), the black-box phenomenon is often also intrinsic to daily routine with ophthalmic imaging, where some decisions are currently based on experience or detection of features that go beyond clinically visible correlates.

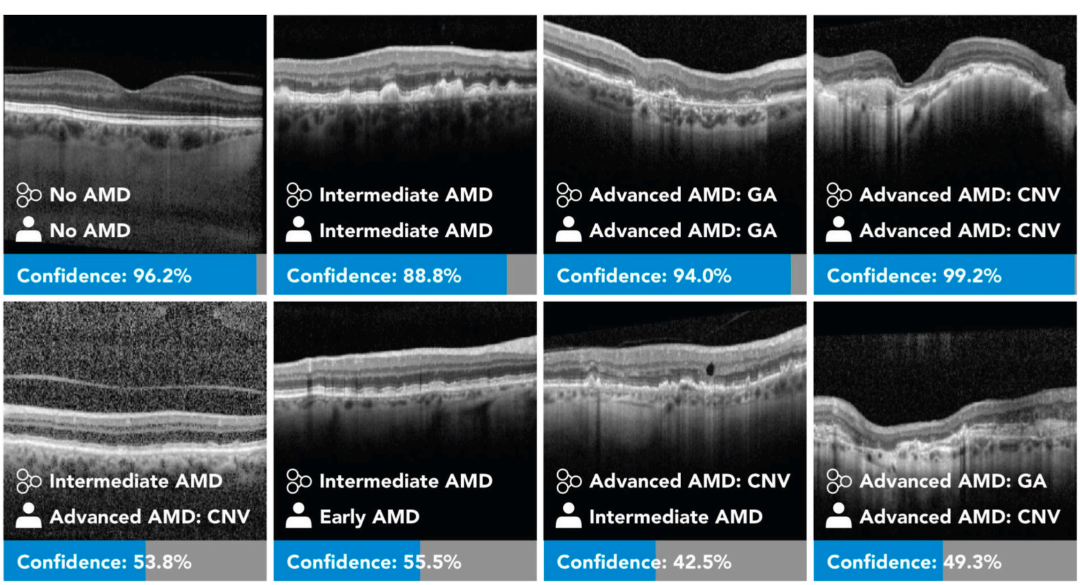


Fig. 11. Uncertainty estimation for AI-based age-related macular degeneration (AMD) grading in optical coherence tomography (OCT) (de Vente et al., 2020). The system makes a prediction based on the full OCT volume. Each image represents a B-scan extracted from an OCT volume. The prediction of the model, the reference standard, and the confidence of the system are shown below each image. The first row includes scans with correct predictions, associated to high confidence scores. The second row includes scans with incorrect predictions and lower confidence scores. The lower confidence of the first image in the second row could be due to the high amount of noise in the scan. In the second image, it is ambiguous whether the patient has intermediate or early AMD. The third image does not have choroidal neovascularization (CNV), but the system could have been confused by the small intraretinal cyst. In the last image, there is both geographic atrophy and CNV; however, the correct prediction according to the reference standard should have been advanced AMD with CNV.

Generalizability and repeatability of a system's performance should be considered as relevant as explainability in those cases, if not a substitute. We consider the intrinsic trust provided by explainable decisions to be complementary to the extrinsic trust provided by verifiable behavior, such as demonstrating repeatable and generalizable performance during a system's validation (Jacovi et al., 2021).

Although there is an increasing demand on explainability by regulatory bodies, as mentioned previously, there is not a clear definition of the right level of explainability required in a given context (Beaudouin et al., 2020). The creation of a framework that considers all the dimensions around explainability (clinical, technical, ethical, legal), as well as associated social and economic costs and benefits, would help to define the right level of explainability in each setting and clinical application. A baseline for such frameworks in clinical settings could be established by regulatory bodies. However, in order to define explainability requirements for specific points of care, applications, and settings, which vary greatly across countries, the intervention, when possible, of ophthalmological societies and working groups or committees will be key. Similarly, different levels of explainability might be required by different stakeholders, i.e., yielding interpretable AI systems might mean opening the black box in different ways for developers, healthcare providers, patients, or regulatory bodies (Bhatt et al., 2020).

5.4. Robustness to malicious attacks and manipulation

5.4.1. Importance and consequences

Recent reports predict an increasing trend of cyberattacks targeted at systems powered by AI, and the healthcare industry is expected to suffer two to three times more attacks than the average amount for other industries (Cisco and Cybersecurity Ventures, 2019). Limited resources and fragmented governance on cybersecurity, combined with larger consequences at both financial and human levels make healthcare particularly vulnerable to cyberattacks (Martin et al., 2017). There are two main factors that amplify the threat of cyberattacks in the medical domain: financial interests and technical sources of vulnerability (Bortsova et al., 2021; Finlayson et al., 2018).

The first factor is strongly associated with healthcare fraud, which has been shown to be committed by large companies as well as individuals, and translates into a significant economic loss from the global health expenditure (Gee and Button, 2015). Some parties involved in healthcare might have a financial interest in manipulating patient data when it comes to insurance, clinical, or drug/device approval decisions. Cyberattacks can boost current fraudulent behavior in these decisions, and they would be facilitated because the attacker would be already inside the healthcare infrastructure. The second factor of vulnerability is mainly associated to the security of healthcare technological infrastructure. In this case, attacks would be most commonly performed from outside the healthcare infrastructure by means of a breach, implying multiple security risks, such as blackmail, ransomware, and malicious data manipulation. In a recent investigation, more than 45 million medical images and their patient metadata were found to be exposed and freely accessible, without hacking tools required, on over 2000 unprotected medical servers across 67 countries (CybelAngel, 2020). Imaging systems (including systems for image acquisition, viewers, workstations, and servers) have been found to have the most security issues, mainly derived from user practice and outdated infrastructure (Healthcare Innovation, 2018).

Malicious attacks can result in deteriorated quality of healthcare, financial loss, decreased trust in AI systems and hence impediments to their integration into clinical practice. In this manuscript, we set the focus on adversarial attacks (Szegedy et al., 2013), which have become very popular in computer vision in the last years. Such attacks apply a carefully engineered, subtle perturbation to the input of a target model to cause wrong predictions. They have been shown effective against AI-based classification systems (Goodfellow et al., 2014b), and multiple works have exposed their threat in different medical imaging modalities and applications, including automated screening of DR in CFP (Finlayson et al., 2018). However, in these works the attacks are performed in a *white-box setting* (Goodfellow et al., 2014b), where the attacker has full access and/or knowledge about the target system. However, in real-world deployment settings, there is restricted knowledge and access to the target system and design factors, such as the data that was used for development. In realistic scenarios, adversarial attacks would be performed in a *black-box setting* (Papernot et al., 2017), in which the attacker does not have full access to the target model and usually uses another model, commonly referred to as *surrogate model*, to craft adversarial inputs that are then transferred to the target model. This aspect conditions the actual vulnerability of AI systems to adversarial attacks in clinical settings (Bortsova et al., 2021). By defining the actual threat that adversarial attacks pose in ophthalmic settings, it will be possible to adapt the necessary solutions to ensure systems' robustness without damaging trust unnecessarily.

5.4.2. Proposed solutions and considerations

Following the recommendations provided by Bortsova et al. (2021), AI developers, aided by other stakeholders such as healthcare institutions and users, are expected to assess the ophthalmic setting where an AI system is meant to be integrated in order to identify the motivations and the capacity to perform undetected attacks by any user. If there is a significant risk of successful adversarial attacks, proactive measures should be taken.

Numerous defense methods have been proposed to protect AI systems from adversarial attacks either by training networks to increase their adversarial robustness (Goodfellow et al., 2014b) or by detecting or neutralizing adversarial inputs during inference (Lu et al., 2017). Although defense methods are only partially effective (Yuan et al., 2019), applying the most successful methods is likely to increase the difficulty of manipulating a system. In addition, using techniques for interpretability and quantifying uncertainty of the system's predictions, such as the ones included in Section 5.3, may aid in detecting adversarial attacks (Li and Gal, 2017; Tao et al., 2018). However, they also provide only partial protection (Smith and Gal, 2018). The transferability of adversarial attacks in black-box settings have been observed to increase between similar models, i.e., when the attacker makes use of certain design components also used in the target system, such as the same network architecture, the same development data, or the same network initialization with ImageNet pre-training. This has been shown in the setting of automated screening of DR in CFP (Bortsova et al., 2021) (Fig. 12). To further increase the difficulty of emulating the target system for an attacker, the use of standard design components and public datasets should not be the sole basis of deployed systems, considering the threat to patient safety and privacy it represents.

We observe there are different levels in the requirements of information to disclose about a system meant to be deployed in clinical practice, in order to ensure robustness and, consequently, trustworthiness. When it comes to the general public, there exists a trade-off in the amount of disclosed information. Limiting the amount of publicly-available information related to the design of the system is recommendable (Bortsova et al., 2021). However, design details that do not represent a threat to patient privacy and safety and information related to the validation of the system (methods, procedures, results) should be made publicly-available. The use of public datasets to facilitate benchmarking should be encouraged, as indicated in Section 6.1. When it comes to the stakeholders involved in the design of the system, as well as external agents such as regulatory or governmental bodies, transparency is key. A trustworthy system should be reproducible and exhaustive information about the system should therefore be made available by AI developers, for instance, in audit trails (Section 2.1). When it comes to front-line users, certain information about the system's design could be provided to allow for a better context on how to incorporate the system's outputs into clinical decisions. A close collaboration between AI developers, healthcare institutions, and ophthalmological societies and

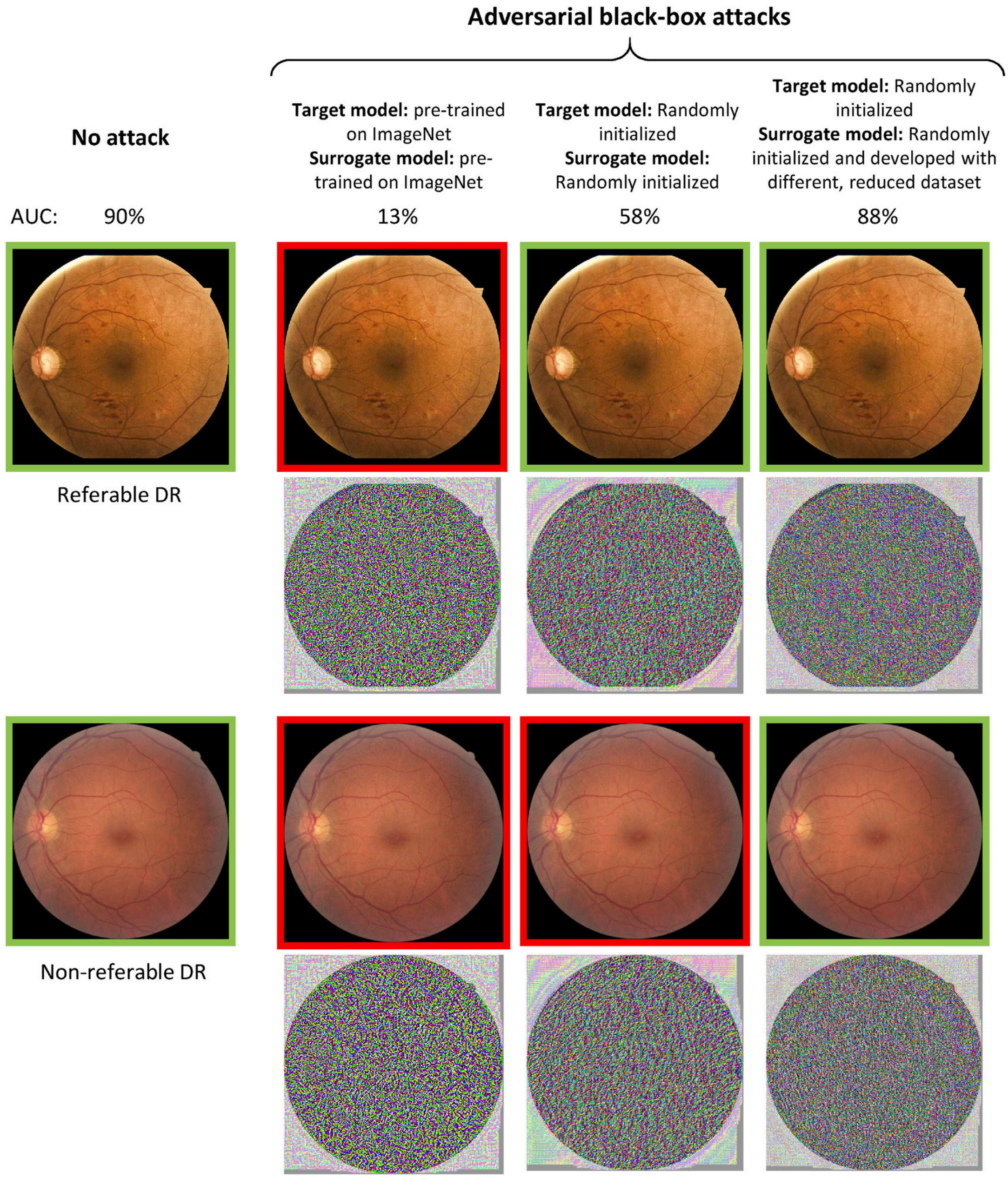


Fig. 12. Adversarial black-box attacks in AI-based screening of diabetic retinopathy (DR) in color fundus photography (CFP) (Bortsova et al., 2021). Original images, adversarial images, and corresponding adversarial noise (difference between original and adversarial image) in different black-box settings, where the attacker does not have full access to the target model and usually uses another model, commonly referred to as surrogate model, to craft adversarial inputs that are then transferred to the target model. Settings from left to right: target and surrogate models both pre-trained on ImageNet; target and surrogate models both randomly initialized; target and surrogate models both randomly initialized plus surrogate developed using a different and reduced dataset. The average area under the receiver operating characteristic curve (AUC) is indicated above of each configuration for the no-attack setting and the black-box settings. Green frame indicates correct classification of referable or non-referable DR; red frame, incorrect classification. It can be observed that attack transferability decreases when target and surrogate models are less similar. Reduced access and knowledge about the target model by the attacker leads to less impact in performance.

working groups or committees will help define the adequate level of required information to prevent misinterpretation and misuse (Section 6.2).

A combination of the mentioned proactive measures would provide the most comprehensive security (Bortsova et al., 2021). It is acknowledged that protective measures could hamper performance (Zhang et al., 2019), however, for AI systems to be deployed in clinical practice, robustness needs to be considered as well, as part of building responsible and trustworthy AI (Leslie, 2019). The decision on how much performance to sacrifice for robustness will differ per case depending on the likelihood of adversarial attacks and the potential consequences. The setup of an evaluation study of adversarial robustness can help on this regard, aimed at obtaining realistic robustness estimates by testing the system under the most likely attack scenarios, considering that the attacker will not have complete access or knowledge about the system. Evaluation practices such as the ones provided by Bortsova et al. (2021) help for the standardization of robustness studies and, consequently, for their inclusion by regulatory bodies when establishing guidelines for trustworthy AI.

5.5. Adaptability and updates

5.5.1. Importance and consequences

Traditionally, AI systems undergo a training phase and are then deployed for inference. The systems learn from a snapshot of the data initially available and the learned concepts and parameters remain unchanged during the deployment phase. Deploying these systems into non-stationary environments, such as clinical settings, is a potential to failure because 1) the behavior of AI systems is unpredictable when the input data are generated within a dynamic setting with shifting characteristics; and 2) the systems cannot seamlessly adapt to evolving clinical and operational practices or feedback from experts. If we look at how healthcare providers learn and improve their skills, the process is very different, for instance, ophthalmologists are continuously trained to become reliable experts, incrementally improving their diagnostic skills and autonomously adjusting their knowledge in consultation with peers.

The adoption of new clinical guidelines or the introduction of new operational procedures would require an update of the AI system's inner working and could be easily identified if properly reported by healthcare institutions. The magnitude of the update and the need of stakeholders' involvement for re-training the system or re-annotating data will depend on the difference between the new and current clinical practice. For example, the adoption of an updated diabetic retinal disease (DRD) staging system to incorporate relevant advances in the field would require automatic methods for DR screening to jointly analyze systemic health measures (e.g., measures of glycemic control or blood pressure) and additional aspects of functional vision, such as visual fields, or low-luminance acuity, together with the currently used CFP (Sun et al., 2021). Training flexible models that, for example, can process as input different type of information (as seen in Section 5.1) would facilitate a quick update of the AI systems and, consequently, the rapid adoption of these new guidelines.

Subtle, progressive shifts in the input data characteristics are, however, more difficult to identify and adapt to. For example, the implementation of new treatment regimens can slowly shift the definition and prevalence of disease stages, invalidating adopted operating points of AI systems and increasing the risk for errors and misuse. The introduction of a new predictive algorithm may cause also changes in practice, resulting in a new distribution compared to that used to train the system (Kelly et al., 2019). Therefore, the development of robust methods to identify data shift and monitor performance over time to proactively identify problems, alongside easily implemented strategies for re-training are crucial to prevent unacceptable harm during the deployment phase.

One important aspect to consider during the design and development phase of AI algorithms is how the communication with the clinical team will take place. The report of errors and incidents as well as suggestions for improvements should be facilitated and systems should be adapted accordingly to address them. AI developers often fail to anticipate all the potential risks associated with the systems they develop, including both inadvertent failures and deliberate misuse. Although ophthalmology has a poor representation of qualitative implementation research relative to other specialties, a recent study by Beede et al. (2020) reflected the importance of this aspect in the context of DR screening. The possibility to create a communication channel between users and AI developers to notify and address these incidents would provide a collaborative way to mitigate these risks. Working towards AI incident sharing platforms for healthcare, similar to the ones developed for other fields (e.g., Partnership on AI's AI Incident Registry), could provide an objective way for reporting. Systems would then need to be designed to handle reported errors and feedback and deploying fixes.

It should be noted that AI systems without appropriate safeguards to monitor their activity and act consequently would not only become obsolete and unreliable but could represent a potential risk for patients' lives, promoting their withdrawal from clinical workflows and the indirect reduction of trust on this technology. Adaptability should then be considered during AI development as part of the strategy for risk mitigation (Rivera et al., 2020) and as a way to maintain claims about AI accuracy and reliability over time.

5.5.2. Proposed solutions and considerations

The adaptability of an AI system to users' feedback, new clinical practices, or data shift, is closely related to a machine learning technique called *continual learning* (also known as lifelong, incremental, or online learning), in which a model continuously learns from new patient data and the generated outcomes, fine-tuning its current task, or even incrementally learning new tasks, while retaining previously learned knowledge (Parisi et al., 2019). Continual learning models have therefore similar ways of learning to that of healthcare providers, since they are able to incrementally learn from their mistakes and fine-tune their performance with progressively more data. In diagnostic tasks, a continual learning system would first perform inference when new patient data become available, the new data would also be manually annotated, and then the annotations and the system's predictions would be used to update the system (Fig. 13). In the context of error analysis and integrating users' feedback, Liu et al. (2019) proposed an online learning system to improve the automated detection of glaucomatous optic neuropathy and glaucoma diagnosis in CFP with a human-computer interaction loop. The loop consists of three iterative steps: first, the system diagnoses glaucoma with a high sensitivity rate on new patient data; second, ophthalmologists manually confirm the positive samples predicted by the system; third, the confirmed samples are used to fine-tune the system for the next iteration. The online learning system was implemented in a teleophthalmology setting, sequentially collecting new samples on a weekly basis, and its performance was shown to increase with each iteration. This shows the potential of online learning systems to improve with the help of human

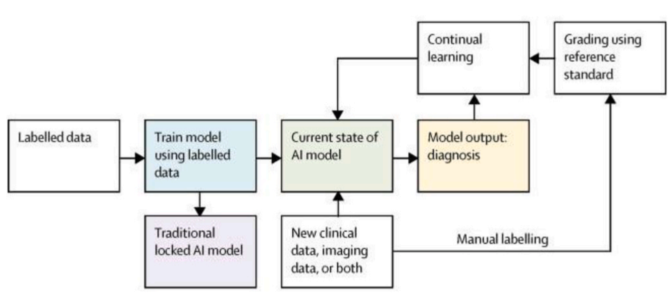


Fig. 13. Continual learning pipeline for diagnostic tasks (Lee and Lee, 2020). In this setting, an AI system is trained to provide a diagnosis and is continually updated with new patient data. When the new data are fed into the current system, the model provides its diagnosis output. Meanwhile, the same new data are manually annotated. Both sources of information are fed into the continual learning system, updating its current state to integrate users' feedback after error analysis, or to adapt to new clinical practices or data shift.

experts, while the systems help human experts to be more efficient in diagnosing negative samples. Related to the adaptability of systems to new clinical practices, He et al. (2021) explored incremental learning to integrate new required tasks in the context of automated segmentation of DR-related lesions. Their system is able to distill the knowledge of a previous model to learn a new task and improve the performance of the current model, which also helps to reduce the required amount of new data to annotate. When it comes to adaptability due to data shift, it becomes critical to count with methods that identify and quantify the shift, monitoring performance over time to then define the necessary periodic updates to the system (Kelly et al., 2019). Data-driven testing approaches have been proposed to recommend the most appropriate update, from simple re-calibration to full model retraining (Davis et al., 2019).

An important consideration regarding the use of continual learning in diagnostic tasks is that manual labelling of new data is time-consuming and limits the overall usability of the system. However, using the system's inference predictions directly as reference standard could negatively affect patients' outcomes and cause an unwanted data shift (Lee and Lee, 2020). Reliable, manual labels are therefore indispensable. A subsample of the new data could be selected in order to reduce the workload of manual labelling. This could be done, for instance, by using the data that fall close to the system's decision boundary for (re-)annotation and fine-tuning, or by integrating specific approaches such as active learning, where the system would proactively select the most beneficial set of images to (re-)annotate (Sánchez et al., 2010). When performing error analysis, it is important that the human graders are not given only system's outputs to correct/re-annotate, since that could bias graders towards minimal corrections on an output that would be already "good enough". It is therefore also advisable to involve only the most experienced graders, who would be more likely to keep their own grading criteria when correcting the output of the system.

Continual learning could also be applied for prognostic tasks, where a system would predict a clinical outcome from new data and then be updated after comparing its prediction with the actual outcome. In this case, there would be a waiting period between using the new clinical data as input and the extraction of the actual clinical outcome. However, this setting would not require manual annotations and the standard of care would not change, making it a safer scenario to test continual learning systems. It is important to note that before the system's predictions are used to change clinical decisions, a prospective randomized clinical trial should be done to compare against the standard of care (Lee and Lee, 2020).

Continual learning systems could therefore be advantageous in real-world ophthalmic settings, in comparison to traditional, locked AI systems. Nevertheless, the technique is not completely mature yet and several considerations arise in its deployment (Lee and Lee, 2020). One important consideration is related to the way new patient data are incorporated to update the system. On one hand, if only the new data are used for fine-tuning the system, *catastrophic forgetting* can happen, leading to an overwrite of the model's previous knowledge. On the other hand, completely re-training the system every time new data are available can be computationally expensive and restrict real-time inference. Additionally, this would require access to retrospective training data, which might have use constraints and might not be accessible after the first stage of development, and merging clinical data from a larger number of patients, which might generate privacy and liability concerns. A question also arises regarding whether a system should be continuously updated exclusively on a local level (i.e., per healthcare institution) or integrate updates on a more general level as well (i.e., per region, per country…). On one hand, local updates are desired to optimize the AI system specifically for a given clinical setting; on the other hand, regional or national updates could be beneficial to improve the system's knowledge to generalize. What is clear is that centralized updates, equivalent to version updates in a software or in an operating system, are necessary for integrating global adjustments such as security fixes or new features. It will be thus necessary to define the most adequate level of updates and how to synchronized them with centralized changes. Another consideration is the absence of established methods for assessing the quality of continuous learning systems. Traditional metrics used to measure the performance of the original system would not suffice, since other factors need to be accounted for, such as the automated collection of data for (re-)annotation, the knowledge transfer between original and new patient data, or the required update to avoid overfitting the system, i.e., ensure it is able to generalize to the new data while keeping the original performance.

Regulatory challenges are also significant. Traditional AI systems are locked for safety to prevent post-approval changes. In contrast, continual learning systems would be incrementally learning and updating their state once deployed. Consequently, the original validation results would not be valid anymore. Regulation policies, further discussed in Section 7, are currently outdated in this aspect and are not prepared for continuous adaptability and updates of AI systems. Nevertheless, progress towards lifecycle regulation is being made. The FDA recently proposed an action plan covering modifications in medical AI systems, in order to facilitate updates without going through the entire pre-market review process again (Food and Drug Administration, 2021). Recently implemented, the new EU Medical Device Regulation increases post-market surveillance requirements, including the preparation of a plan that ensures continuous assessment of risks and indicates how users' feedback will be collected and integrated (European Commission, 2017). Potential data shifts could also be planned ahead and included as regulatory standard. Besides the external periodic audits performed by regulatory bodies, internal audits could also be performed to check if the system is performing as expected. The same way healthcare institutions are expected to study the developments of newly applied clinical guidelines, they should also be involved when auditing an AI system during deployment in their clinical setting. Internal audits would help define the most adequate level of updates and how to synchronize them with centralized updates, as previously mentioned. The obtained conclusions would be then used by developers to perform the required modifications to the system. Regulatory bodies and AI manufacturers should work together to generate a list of allowable modifications that can be applied to an AI system that would be subject to a "safe harbor" and thus not necessarily require a new pre-market review for approval (Hwang et al., 2019).

6. Prospective validation

The vast majority of existing validation studies are carried out in a retrospective manner, i.e., the performance of a given AI system is validated using historically labelled data (Kelly et al., 2019). Although this type of validation can help to demonstrate claims regarding the trustworthy behavior of AI systems, the performance is likely to be worse when encountering real-world data in uncontrolled clinical settings, as shown by Abràmoff et al. (2018). A prospective validation or pre-market validation consists in establishing documented evidence, prior to deployment, that a system does what is proposed to do based on predefined protocols. A prospective validation is therefore key to understand the true utility of AI systems when introduced in clinical practice. However, prospective studies for AI systems are generally scarce; a recent study by Wu et al. (2021) showed that only 3% of 130 FDA-approved AI devices had undergone a prospective validation at their submission. Prospective studies are also scarcely reported for AI systems targeted to ophthalmic practice (Heydon et al., 2020; Abràmoff et al., 2018; Beede et al., 2020; Gulshan et al., 2019). Due to its importance, prospective studies, as well as randomized controlled/clinical trials involving AI, will be increasingly considered prior to approval and integration in clinical practice. Publicly available evidence of claims, following established guidelines such as CONSORT-AI (Liu et al., 2020) and SPIRIT-AI (Rivera et al., 2020), will be also needed to transparently report the benefits of AI systems for patient care.

In this section, we focus on three different aspects to consider while planning and performing prospective studies that can facilitate the integration of trustworthy AI systems in ophthalmic settings:

- Reproducibility of systems and benchmarking, so as to obtain additional evidence on performance and its variability in diverse settings, allowing comparison with other systems.
- Study interaction between users and AI to detect and prevent misuse and misinterpretation.
- Monitor the impact of systems on different dimensions of the clinical workflow, such as healthcare economics, healthcare providers, and patients.

6.1. Reproducibility and benchmarking

6.1.1. Importance and consequences

Reproducibility of results in AI is a key way of enabling verification of claims about a system's performance and properties, as well as determining its trustworthiness in the improvement of patients' outcomes. Reproducibility describes whether an AI system exhibits the same behavior when repeated under the same conditions, enabling AI developers and regulatory bodies to accurately describe what AI systems do and to process technical auditing and other forms of accountability.

The non-deterministic nature of AI systems, restricted access to the underlying data and code, and the use of expensive computational resources pose a challenge to reproduce AI results (Beam et al., 2020). The randomness inherent to the analysis, especially true for DL models, hampers a rigorous comparison between AI experiments. Even if deterministic behavior is enforced, the reproduced result might not be representative of the initial experiments. Besides the well-known limited access to clinical data, for which solutions are being actively sought and implemented (Section 3), requiring large computing resources for the AI system training and deployment could prevent the independent reproducibility of results. However, most AI models currently available or under research for ophthalmology are smaller and can be easily reproduced on fairly standard computer hardware. For those cases where more resources are required, the use of standardized, publicly-available hardware (sometimes called "commodity hardware") could overcome this issue.

A main problem for reproducibility is that AI systems lack traceable logs of steps taken in problem-definition, design, development, and operation, decreasing the ability of outside parties to verify claims made about AI systems and leading to a lack of accountability for subsequent claims about those systems' properties and impacts (Brundage et al., 2020). The creation of reporting standards and audit trails can improve the verifiability of claims as well as facilitating reproducible systems (Collins and Moons, 2019; European Commission, 2019). However, they are not yet a mature mechanism in the context of healthcare AI.

Benchmarking is a way to foster reproducibility and, if properly arranged, to provide a centralized manner for healthcare institutions and users to compare and select the best AI system for their clinical workflows and patient population. However, rigorous benchmarking requires taking into account realistic scenarios with real-world data. For example, benchmarking AI systems for the quantification of intraretinal fluid for the assessment of treatment response using curated OCT datasets might not reproduce properly the accuracy and reliability of these systems based on OCT scans acquired during the short time frame of a busy clinic.

Note that reporting validation performance scores alone is insufficient for reproducibility; trustworthiness requires additional meaningful reporting about the model and environmental details, such as the data, code, computational resources, and other decisions made along the AI design pipeline, to ensure reliable conclusions. Failing to reproduce the reported claims may limit the validity, comparability, and usefulness of the performed research.

6.1.2. Proposed solutions and considerations

A first step to generate reproducible AI systems and facilitate benchmarking is to adhere to the protocol and reporting guidelines for prospective clinical trials (Liu et al., 2020; Rivera et al., 2020), and make clear whether and how the AI system and/or its code can be accessed or re-used, including details about license and possible access restrictions. Adhesion to official guidelines during AI development and validation (Collins and Moons, 2019; Sounderajah et al., 2020) (currently under development) will help increase transparency about how the AI system is trained and validated; this will also result in greater reproducibility and replicability.

Greater openness and participation of AI manufacturers in independent validation and benchmarking or head-to-head studies, such as the ones carried out by Lee et al. (2021) and Tufail et al. (2016) in the context of automated DR screening in CFP, would facilitate comparison between available systems. In Lee et al. (2021), of 23 companies approached, only 5 agreed to participate, providing in total seven systems for evaluation. One main obstacle for the participation of companies might be their lack of control over the marketing of the results. However, head-to-head studies are necessary to advance the field and benefit the global population.

Validation of AI systems on public datasets also allows for direct benchmarking given a clinical task. Several public datasets containing ophthalmic imaging can be accessed and have been used in previous studies (Khan et al., 2020). Examples are Messidor, for the detection of diabetic eye disease in CFP (Abràmoff et al., 2016; González-Gonzalo et al., 2020b; Gulshan et al., 2016); Diaret-DB1, for the localization of lesions related to DR (González-Gonzalo et al., 2020a; Quellec et al., 2017); or the SD-OCT dataset made available by Farsiu et al. (2014), and used for automated layer and drusen segmentation (Liefers et al., 2019; Asgari et al., 2019), and localization of the fovea in OCT scans (Liefers et al., 2017). The popularity of public challenges, which allow to compare solutions with data and evaluation procedures common to all participants, has also increased for ophthalmic tasks. Challenges have been proposed for the automated detection of DR (Kaggle DR detection, APTOS, IDRiD), glaucoma (REFUGE), or pathological myopia (PALM) in CFP, detection and segmentation of various types of fluids in OCT scans from different vendors (RETOUCH, Bogunovic et al. (2019)), and, more recently, for the detection of multiple ocular diseases in CFP (RIADD), and the use of multimodal data (CFP and OCT) for the grading of glaucoma (GAMMA). Nevertheless, benchmarking based on public datasets and challenges is not always possible, since they do not cover all relevant clinical tasks, disease populations are unevenly represented (DR, AMD, and glaucoma are overrepresented in comparison to other eye diseases), and the vast majority focus currently on CFP. Additionally, certain population groups, mainly regarding ethnicity, are still underrepresented (Khan et al., 2020; Ibrahim et al., 2021; Burton et al., 2021).

Additional tests for reproducibility by regulatory bodies or healthcare institutions might require access to the original data used for system development and/or validation. However, this might not be possible due to privacy concerns and access restrictions to the clinical data. In these cases, a "walled garden" approach might be a solution, where the external agents are given access to a private network subject to a data use agreement for the duration of the reproducibility analysis (Beam et al., 2020).

Another alternative for healthcare institutions in order to check for reproducibility and compare available systems is the curation of an independent local test set constructed using a representative sample of the target population. A supplementary local training set could also be provided to allow fine-tuning or re-calibration of the systems prior to the formal testing (Kelly et al., 2019; Bora et al., 2021), or certain adjustments to the systems could be allowed regarding, for instance, image acquisition protocol (Lee et al., 2021). Centralized platforms with a

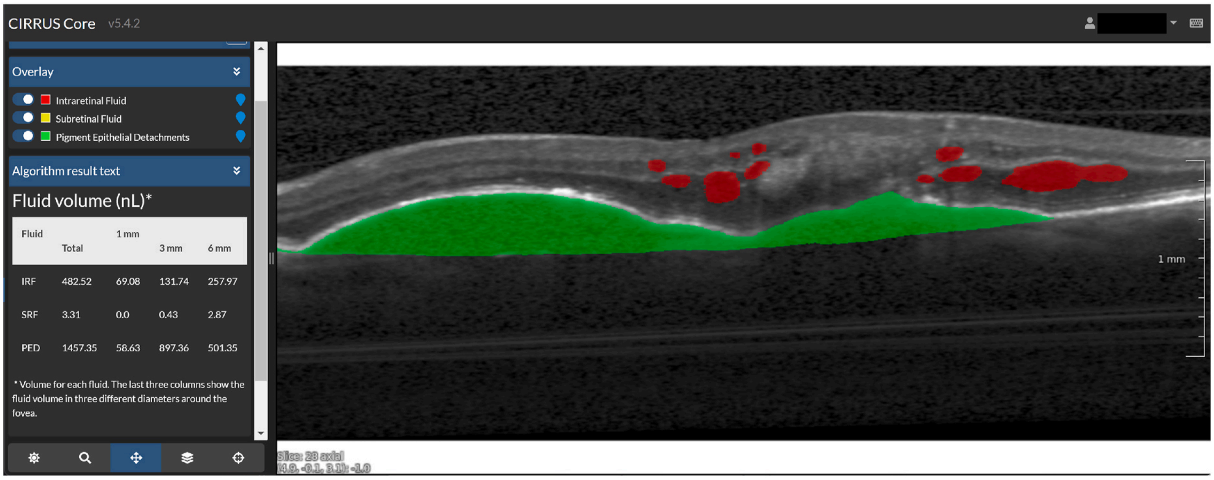


Fig. 14. Example of an algorithm publicly available at Grand Challenge. The AI-based algorithm performs automated segmentation of intraretinal fluid, subretinal fluid, and pigment epithelial detachments in optical coherence tomography (OCT) scans. The algorithm was developed with training data from the RETOUCH challenge (Bogunovic et al., 2019), which contains scans from three different vendors. Users in the platform can upload their OCT volumes and inspect the algorithm's results using the web-based image viewer shown in the figure, or directly download the segmentation masks. As observed, each type of fluid is assigned a different color in the output masks and a quantification of its volume is provided. Although this algorithm is not meant for clinical use, it allows to explore the possibilities of analyzing reproducibility of AI systems in ophthalmology.

user-friendly interface, such as Grand Challenge, can facilitate this process for healthcare institutions. This platform allows to create public or private archives of medical data which can be used to set up a reader study for labelling the data, to set up a public or private challenge to benchmark systems, or to run a specific algorithm (Fig. 14).

6.2. Misinterpretation and misuse

6.2.1. Importance and consequences

The conduct of prospective studies and deployment of AI systems in ophthalmic practice requires clear and open communication between AI developers and end users. Healthcare providers need thus to be well informed about the specific type of setting where a system is meant to be used, as well as the type and quality of data it expects as input so as to provide reliable predictions (Abràmoff et al., 2020). Additionally, they need to be made aware of the potential risks and consequences if AI systems are not deployed properly or are misused in their practice. For instance, users may experience "alert fatigue" due to excessive warnings or over-referrals generated by a system (Singh et al., 2020). There is also the risk of confirmatory or automation bias, where users over-rely on results from AI systems instead of applying their own critical judgement; which could potentially cause harm to patients (Lehman et al., 2015). Eventually, as discussed later in Section 7.2, users can be held liable for possible harm or injury to patients when incorporating AI output into clinical decisions (Abràmoff et al., 2020). While transparency in the reporting of limitations of AI systems has been stressed, guardrails to protect against possible risks and consequences of inappropriate use of AI systems still need to be developed (Sendak et al., 2020).

With regard to the interpretation of results from AI systems, relatively little attention has been given to practical pitfalls of performance metrics. Maier-Hein et al. (2018), however, identified inappropriate selection of metrics as one of the core issues related to performance and quality assessment in challenges for biomedical imaging (mentioned in Section 6.1). For correct interpretation of AI results, it is important that performance metrics are aligned with the clinical task, and that results are reported with a terminology that is understandable to users (Yanagihara et al., 2020; Faes et al., 2020). Otherwise, systems will not provide any actionable guidance.

Overall, the misinterpretation of AI results and the misuse of AI systems will be detrimental to clinical decision making and the quality of care given to patients. Misinterpretation and misuse can also diminish users' trust in AI systems and, in the worst case scenario, cause "algorithm aversion" (Dietvorst et al., 2015). Systematic efforts to ensure that ophthalmologists and other users know how and when to integrate the output of AI systems and solutions to improve the interpretation of such output are needed to allow a safe and more rapid integration of AI systems into ophthalmic practice.

6.2.2. Proposed solutions and considerations

A recent study by Sendak et al. (2020) proposed that AI developers include a "model facts label", i.e., a 1-page of relevant and actionable information to ensure that front-line users know how, when, how not, and when not to incorporate AI output into clinical decisions. The label includes a short summary about the AI system, the working mechanism (including the source and baseline characteristics of data used for AI development), results of validation studies, guidelines for use (including benefits and appropriate decision support), warnings (including potential risks and consequences), and other relevant information related to the AI system. A question arises on how should developers make information accessible, intelligible, and assessable to users. In this regard, a close collaboration with healthcare institutions and ophthalmological societies and working groups or committees would help define the adequate information to be provided and ensure it reaches the final users. Nevertheless, it is not completely clear yet how to deal with potential conflicts of interest. For example, developers or other stakeholders may not want to share certain facts for proprietary or financial reasons. While some questions still need answering, model facts labels might provide a first step towards systematic prevention of misinterpretation of AI output and inappropriate use of AI systems.

Others have highlighted and addressed the limitations of AI performance metrics and their impact on the misinterpretation of AI systems. Recently, a large-scale initiative was started by Reinke et al. (2021) to raise awareness about the limitations of commonly used metrics in the field of medical image analysis, and to provide developers with guidelines and tools to choose the performance metrics in a problem-aware manner. The authors have constructed a document that specifically focuses on segmentation tasks, including fundamental mathematical properties, suitability, aggregation, and combination of segmentation metrics, but a final document will cover classification and detection metrics as well. Yu et al. (2019) highlighted limitations of frequently used metrics for classification tasks. They indicated that metrics such as area under the receiver operating characteristic curve (AUC), sensitivity, and specificity may have limitations for imbalanced datasets. Subsequently, they suggested developers should report additional metrics to improve interpretation, i.e., the positive predictive value and area under the precision–recall curve (AUPRC) for datasets that contain more healthy cases than disease cases, and the negative predictive value for datasets that are dominated by disease cases.

The trade-off between test characteristics is another point of attention. For instance, when computing the AUC, a threshold for sensitivity and specificity needs to be pre-specified. AUC may be calibrated to a higher level of sensitivity, with a corresponding lower level of specificity. Whatever trade-off is selected, the AI system will still generate false positives, false negatives, and indeterminate results, as must be the case with any method of classification. The key consideration would be whether inaccurate results are outweighed by the benefits and whether they are distributed among patients in an equitable manner (Char et al., 2020). To further improve interpretability and applicability of AI output, studies have investigated the use of AI systems across varying *operating points*, i.e., sensitivity/specificity pairs. Yim et al. (2020) used two operating points to demonstrate whether an AI system could predict a fellow-eye conversion to neovascular AMD in OCT scans. Authors included a "conservative" operating point, where the system predicted conversion at 34% sensitivity and 90% specificity, and a "liberal" operating point, where it predicted conversion at 80% sensitivity and 55% specificity. Similarly, Phene et al. (2019) used three different operating points to demonstrate whether an AI system was able to detect referable glaucomatous optic neuropathy. Operating points were chosen based on the system's performance on a tuning (validation) set and included points with a high sensitivity, high specificity, and a balanced sensitivity/specificity. In short, the use of varying operating points can increase the flexibility of AI systems and allows to configure AI systems as required by individual clinical settings, healthcare systems, or therapeutic drug indications. Trade-offs and thresholds should be therefore re-evaluated when AI systems move from internal validations to prospective validations and to deployment stage (Beede et al., 2020). At the same time, pre-specified operating points for the primary endpoints of AI systems can be established by regulatory bodies prior to a prospective validation. For instance, the FDA set mandatory levels of sensitivity (more than 85%) and specificity (more than 82.5%) in the prospective study carried out for the approval of the first autonomous AI system for detection of DR and DME (Abràmoff et al., 2018). In reporting primary endpoints, the computation of confidence intervals (CI) is key. It will show whether the variation of a system's performance, especially at the lower bounds of the CI, is still clinically acceptable. To determine whether the lower bounds of a system's performance will reach a certain level, it is possible to perform a power calculation and compute the minimum number of samples to generate significant results. Tufail et al. (2016) calculated that 24,000 screening episodes in the NHS Diabetic Eye Screening Programme were required in their study to ensure that the lower bound of the 95% CI for sensitivity of severe DR grading would not fall below 97%.

When validating AI systems, it should also be considered that quantitative metrics might not always relate to qualitative assessment by experts. In Wilson et al. (2021), an AI system for macular fluid segmentation performed inferiorly to expert gradings based on quantitative metrics, while qualitatively the AI-based segmentations were assessed as clinically acceptable. Both quantitative and qualitative validations are therefore desirable and should be reported for a complete interpretation of a system's performance.

The terminology used in the reporting of AI output is another point of attention. Metrics used to report on the performance of a system should reflect both its technical and clinical behavior. Stakeholders should therefore be familiarized with the terminology employed for both technical and clinical aspects so as to be able to interpret the system's potential and limitations in all relevant dimensions. There is also a need for consistency in the used terminology for AI output (Kim et al., 2020). Guidelines such as CONSORT-AI (Rivera et al., 2020) and STARD-AI (Sounderajah et al., 2020) (under development) may help to standardize the terminology when reporting about performance of AI systems.

Several steps have been made towards improving interpretation of AI results and appropriate use of AI systems. Nevertheless, it is clear that several hurdles remain to be overcome before their widespread implementation into ophthalmic practice can occur. AI developers, healthcare institutions, and other stakeholders should continue to develop systematic efforts to address the potential risks and consequences of misinterpretation and misuse of AI systems. Ensuring the appropriate use of AI systems will require ongoing monitoring, as discussed later in Section 6.3. An appropriate setup for prospective validations is also key, since they provide a suitable scenario to observe the interaction between system and users and detect and register those situations or cases where users tend to misuse the system and/or misinterpret its output. This will allow to identify and apply the necessary measures towards a more seamless integration. The training of ophthalmologists and other users in early stages could also help prevent misinterpretation and misuse of AI systems. Healthcare institutions and reading centers could provide healthcare providers and graders, respectively, with specific training and/or practice periods to become AI competent in their daily practice, aided by AI experts. Medical schools could prepare students to become AI competent, by learning how to interpret a model's predictions, explainability, and uncertainty measures.

6.3. Impact on clinical workflow

6.3.1. Importance and consequences

Besides numerous potential benefits, the integration of AI systems also involves significant changes in financial and human aspects within the clinical workflow. While performance tends to be the main focus of study in prospective validations, the impact of AI on different dimensions of the clinical workflow has not been properly analyzed yet. We put the focus on the impact of AI on health economics, healthcare providers, and patients.

The integration of AI will suppose a shift in financial resources and clinical workload. For example, in the case of DR screening, AI will allow to increase the number of detections in early stages of the disease, especially in those areas without current coverage when combined with telemedicine. However, this may lead to an increase in the workload of ophthalmologists and a shift in the time available for patient treatment and follow-up. Substantial manpower and funding are required for AI deployment in practice and apply the necessary updates in the current clinical protocols and infrastructure to ensure interoperability (Section 2.2). Additionally, as seen in Section 5.5, optimal performance of AI systems will require ongoing monitoring and maintenance to ensure adaptability to users' feedback after error analyses, new clinical practices, or data shift, and to perform the necessary system updates. All of this monitoring and maintenance activity will require significant effort in human capital as well (He et al., 2019). With increased cost pressure on healthcare systems, AI systems need to be subject to cost-benefit analyses tailored to the healthcare setting and intended use, and compared with the standards followed in current practice. These analyses will be key to be granted reimbursement within a healthcare system (Section 7.3) and for widespread deployment.

The automation of certain tasks by AI systems will also bring a shift in healthcare providers' roles and responsibilities. If the roles of AI and healthcare providers are not fully understood and communicated, this may lead to lack of trust from the users and/or a fear of replacement, as well as to misuse and misinterpretation of the systems once deployed (Section 6.2). This will cause the potential benefits to be lost. Furthermore, even though an AI-based support tool is coherent with clinical grounds, it is often abandoned by adopters if direct or indirect consequences run counter to their clinical routines and values (Beede et al., 2020; Yang et al., 2019). The uncertainty in human–AI interactions may result in significant variation in users' performance and systems' performance (He et al., 2019). It is therefore key to overcome the paucity of implementation studies, particularly scarce in ophthalmology, and invest in a suitable infrastructure for prospective validations that allows to analyze how to best integrate AI within the healthcare providers' workflow.

The integration of AI will also have an impact on patients, who might be concerned about the use of AI in their ophthalmic care and the privacy of their data. Similarly to obtaining consent for undergoing a MRI scan, the patient might not necessarily need to know every detail but certainly has to be informed about the core principles of the procedure, and especially the risks. However, it is difficult to establish to what extent the patient has to be made aware or agree with clinical decisions that were assisted by AI (Amann et al., 2020). Public attitude towards AI is expected to increasingly influence AI policies and, consequently, widespread deployment of AI systems (Zhang and Dafoe, 2020). It becomes thus crucial to incorporate patients' views and predisposition to using AI in prospective validations, while ensuring an adequate management of patient consent.

6.3.2. Proposed solutions and considerations

A traditional way of conceptualizing benefit of a novel intervention in terms of health economics is using a cost-effectiveness model. To populate a cost-effectiveness model, detailed current costings of all aspects of current clinical pathways and modified pathways need to be acquired. Such information gathering should be built into AI system validation studies, preferably in a prospective way. The decision to implement a given AI system in clinical practice will depend on whether the increased cost of per-patient encounter is worth it for the additional benefit, and can be described by means of different outcome variables, mainly as either the incremental cost-benefit ratio or in terms of the net benefit compared to existing practice (Simoens, 2009). For instance, in the context of DR screening, several studies have focused in comparing the cost-effectiveness of current manual grading with that of using AI as a filter prior to level-one human grading and that of replacing level-one human graders. Tufail et al. (2016, 2017) validated several AI systems in the NHS Diabetic Eye Screening Programme in the UK, and showed that AI was cost-saving compared to manual grading, either as a replacement for human grading or as a filter prior to human grading, although the latter approach was less cost-effective (Fig. 15). In Lee et al. (2021), as part of a retrospective validation on US′ veterans' CFP, a cost-effectiveness analysis reported similar per-patient encounter costs among seven AI systems, with significant annual labor savings when used as a filter in the DR screening pathway.

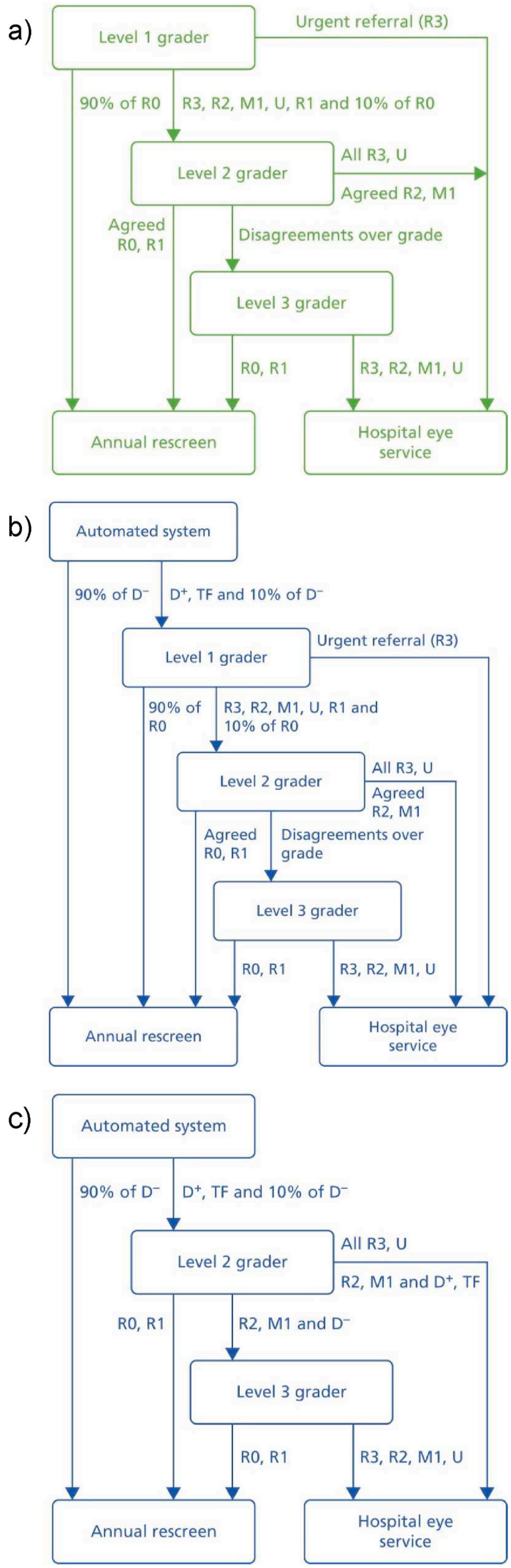
Although the results from retrospective reports are promising, cost-benefit analyses using data from prospective use of AI systems are necessary, as recently done by Heydon et al. (2020). Future prospective studies should also focus on assessing cost-effectiveness of AI systems for other clinical applications and points in patient care, where changes in economic and societal costs when applying AI might differ greatly. Besides system-related cost-saving outcomes, other beneficial outcomes related to cost-effectiveness should be considered, such as reduced hospital visits, reduced costs for patients, early treatment of patients, and reduced burden of over-referral for clinicians (Hopkins et al., 2020). Prospective cost-effectiveness analyses should also be performed in resource-limited settings, where introducing AI systems might require building basic infrastructure and a system to identify all eligible patients. Similarly to what has been discussed concerning bias in Section 5.2, different ethnic groups and low-cost image capture systems would need to be assessed, since they may alter cost-effectiveness in different population subgroups. Validations in resource-limited settings are therefore crucial to ensure access to AI for disadvantaged patients, contributing to the medical ethics principle of equity (Abràmoff et al., 2021).

After extracting the results from performed analyses, depending on the intervention and healthcare system, Health Technology Assessment (HTA) groups in public health bodies or healthcare institutions would need to consider whether the incremental cost-effectiveness is reasonable to support an AI system's introduction over current practice. Payers (government agencies or private health insurers) can also consider these analyses to generate reimbursement agreements over the use of AI (Section 7.3).

To determine the impact of AI systems on healthcare providers and other users in ophthalmic practice, one should consider human-AI interaction analyses. Interaction studies from other medical domains, including pathology, dermatology, cardiology, and radiology, have shown significant synergistic effects of human-AI interaction (Bulten et al., 2021; Tschandl et al., 2020; Yang et al., 2019; Lakhani and Sundaram, 2017). Existing literature on this topic within the ophthalmology domain is limited. Nevertheless, synergy between retinal specialists and an AI-based DR screening system was shown in a study by Sayres et al. (2019). It was observed that AI-assisted retinal specialists graded DR more accurately than unassisted specialists or the standalone AI system. The study also showed that retinal specialists had greater confidence in ratings and that although the time of AI-assisted grading increased for cases of DR, it decreased with more use experience. Human-AI interaction analyses will help clarify and establish the changes in roles and responsibilities that ophthalmologists and other users might experience as AI systems become more prevalent in ophthalmic practice. For instance, while AI systems might perform faster, more accurately, and more sensitive than humans at processing patient data, especially in triaging tasks, clinicians will still have important roles, although more focused on management, knowledge-handling, and communication. AI-supported decisions will require the participation of interdisciplinary personnel with knowledge from both technological and medical disciplines, which is rare at present (Sun and Medaglia, 2019). A role expansion for non-medical staff is to be expected as well (Beede et al., 2020; Gillan et al., 2019). The possibility of workforce deskilling has also been pointed out as a consequence of healthcare providers surrendering to the autonomy of AI (Liberati et al., 2017). On the other hand, a virtuous cycle effect has been suggested, where AI complements providers' competencies and skills instead of threatening their professional autonomy, and incentivizes them to pursue markers of quality on a routine basis. How AI systems reduce or intensify labor needs to be further analyzed (Pope and Turnbull, 2017). Future user-centered studies will help understand better these factors and their impact on healthcare providers' roles and responsibilities.

Opinion, evaluation surveys, and interviews among potential users of AI systems are also important to better understand their expectations, concerns, and needs regarding the implementation of AI systems. Consequently, they can also help identify and prevent certain aspects that could affect negatively AI integration and its impact on the clinical workflow. This will be possible if they are adequately designed in collaboration between AI developers, healthcare institutions, reading centers, and ophthalmological societies and working groups or committees. A study by Al-Khaled et al. (2020) performed a web-based survey among ophthalmologists to identify their perception of AI systems. Among 170 ophthalmologists from a range of subspecialties, 89% reported that they agreed with understanding the concept of AI. Approximately 75% believed that AI will improve the practice of ophthalmology, reported interest in integrating AI into their clinical practice, and indicated that there should be formal instruction in AI during medical school and residency training. However, almost half (45%) reported concerns over the diagnostic accuracy of AI. Furthermore, 22% were concerned that the patient-physician relationship would be impacted by AI and 36% had concerns about AI replacing ophthalmologists.

User-centered analysis should therefore be integrated in prospective validations. Through interviews and observation of human-AI interaction it is possible to identify those environmental factors that will determine the successful implementation of an AI system. For instance, in the context of automated DR screening across local clinics in Thailand, Beede et al. (2020) observed that poor lighting conditions led



(caption on next column)

Fig. 15. Cost-effectiveness analysis of using AI within the NHS Diabetic Eye Screening Programme in the UK (adapted from Tufail et al. (2016)). The figure shows the proposed decision-tree model to calculate the cost-effectiveness of manual grading (a) versus using an AI system as a filter prior to the initial grading performed by level-one human graders (b) and versus replacing level-one human graders with an AI system (c). The authors validated several AI systems and showed that AI was cost-saving compared to manual grading, either as a replacement for human grading or as a filter prior to human grading, although the latter approach was less cost-effective. D-, disease absent classification by AI system; D+, disease present classification by AI system; TF, technical failure; R0, no retinopathy; M0, no maculopathy; R1, background retinopathy; M1, maculopathy; R2, pre-proliferative retinopathy; R3, proliferative retinopathy; U, ungradable images.

to ungradable CFP (Section 5.1) and user frustration, and network connectivity issues caused unnecessary delays for patients and rescheduled appointments (Section 2.2). Furthermore, concerns for potential patient hardship (time, cost, and travel) due to on-the-spot referral recommendations from the system, caused some nurses to discourage patient participation in the study. It is thus important to consider that negative consequences cannot always be controlled despite a careful planning of the study, and that although certain environmental factors can be potentially reduced, the necessary adjustments may be costly and infeasible in low-resource settings. Engagement with real-life challenges of human-AI interaction as early as possible in the AI design pipeline is therefore essential.

Regarding the impact of AI on patients, surveys and interviews are also a useful mechanism to better understand patient acceptability of using AI systems in their care, as done in Keel et al. (2018). In this study, 96 adults with diabetes were prospectively recruited from two outpatient clinics in Australia. Each participant underwent screening by an AI system, as well as manual screening where images were transferred to a reading center and outcomes were shared with patients within 2 weeks. Of the participants, 96% reported that they were satisfied with automated screening, and 78% even reported that they preferred automated over manual screening. Drawbacks of manual screening were highlighted, e.g., delayed communication of the results was prevented with automated screening, as it allowed real-time reporting of results. While more studies are needed to assess the impact of AI systems on patients, this study provides some initial evidence that AI systems in ophthalmology may be feasible and well accepted by patients in a screening setting. Nevertheless, a survey by Ongena et al. (2020) performed within clinical radiology workflows pointed out that the implementation of AI systems will create a general need for patients to be well-informed on various aspects of AI deployment, including image acquisition, processing, and interpretation. As such, stakeholders should consider involving patients in the design and development of AI systems. Most IRB panels already include patient representatives to review AI research studies from an ethical and privacy perspective. Additionally, patient representatives could participate along the AI design pipeline to help define what outcomes are meaningful to patients, and to help determine how patients want outcomes communicated during doctor-patient counseling.

Future studies should also focus on acceptability of AI in non-diseased populations, for example, at general practitioners' or opticians' offices. This will help increase early-stage detection of eye diseases that are currently only detected in the clinic by coincidence. To allow such screenings, solutions for non-patient consent should be developed and implemented with AI systems. People at high risk of developing eye diseases may be a first focus, as they are likely to be more open to consent. In any case, the expectation is that healthcare providers will play a key role in providing patients (and non-patients) with important and transparent information about AI systems prior to consent, as their judgements are generally trusted. Widespread education and training of healthcare professionals to become AI competent, as mentioned in Section 6.2, will be fundamental also in this regard. It will allow healthcare providers to act as educators, counselors, and advisors of patients regarding the use of AI output in their care (Liu et al., 2018). This will also contribute positively to the medical ethics principle of patients' autonomy to decide about their own participation in the use of AI (Abràmoff et al., 2021).

In conclusion, retrospective cost-effectiveness and human-AI interaction analyses have shown good results and have helped to build the understanding of the impact of AI on the clinical workflow, allowing to design potential solutions prior to system development and validation. However, research using prospective data in a contextual environment can provide better opportunities to identify vital factors ahead of widespread deployment. Monitoring ongoing clinical use of AI systems within healthcare institutions during prospective validations, as well as a part of post-market surveillance (covered in Section 7.1), will allow to better understand their impact on healthcare economics, healthcare providers, and patients. It will also help reduce the risk of AI systems failing at deployment, and will increase the likelihood for meaningful improvements in ophthalmic care. Only a few studies have described systematic approaches to monitoring the impact of AI systems in clinical practice. Xie et al. (2020) outlined several methods for safety analysis during monitoring of AI systems to ensure that a system could "fail safely". This could be done via failure mode effects analyses, system-theoretic process analyses, and bowtie analyses. These methods consider clinical, technical, social, and organizational sources of AI systems to identify safety issues and their potential consequences. Although they are time-consuming and resource-intensive, they provide a systematic approach to understand and develop solutions for technical and organizational AI safety risks.

7. Regulation

Regulation plays an essential role for the widespread integration of AI systems in healthcare, as well as for AI acceptance by healthcare providers, patients, and society in general. Currently, regulatory bodies and policy making institutions have not established updated approval pathways covering trustworthy AI aspects, such as those regarding data privacy (Section 3.1), adaptability of systems to ensure reliable behavior over time (Section 5.5), or liability implications regarding the use of medical AI devices (Section 7.2). Although some aspects are being currently considered, more effort on dedicated regulations for trustworthy AI systems is required to accelerate AI integration in healthcare. It is worth noting that a comprehensive set of regulations is required, focused on the highest standards of patient safety while enabling progress in innovation.

In this section, we discuss three aspects relevant for the regulation of trustworthy AI systems:

- The approval process to reach regulatory standards for an adequate assessment of AI claims.
- Liability and identification of responsible actors in case of medical malpractice incurred by the use of AI systems.
- Reimbursement and identification of fair financial incentives to integrate and use AI medical devices in practice.

7.1. Approval

7.1.1. Importance and consequences

One key factor for the widespread integration of AI systems in healthcare is the development of regulatory standards for an adequate assessment of their effectiveness and safety. AI systems are defined as a medical device, under the term Software as a Medical Device (SaMD), by the International Medical Device Regulators Forum (IMDRF), whose members include Australia, Brazil, Canada, China, Europe, Japan, Russia, Singapore, South Korea, and the US. However, no specific regulatory pathway for AI systems has been defined yet. Conventional medical device approval pathways are currently applicable to AI systems, despite their lack of suitability to specific aspects of AI systems, such as the way patient data is handled (Section 3.1), the necessity for periodic updates to ensure optimal performance and adaptability over time (Section 5.5), or new implications regarding liability (Section 7.2).

The regulation and approval of medical devices are handled differently around the world (Muehlematter et al., 2021; He et al., 2019). In the US, the centralized FDA classifies medical devices as class I, II, or III, indicating increasing risk of illness or injury and thus requiring higher regulatory control (He et al., 2019). There are three possible approval pathways in the US: pre-market approval (strictest regulation for class III devices), the 510(k) pathway (for class I, II, and III devices for which pre-market approval is not indicated; devices must be compared to one or more similar devices already marketed), and the de-novo pre-market review (an alternative pathway for novel class I and II devices). In Europe, medical devices are not approved by a centralized regulatory agency. For low-risk devices (class I), the manufacturer has the responsibility to comply with the corresponding regulation without undergoing an approval process. For higher-risk devices (class IIa, IIb, and III, and in-vitro devices), private organizations called Notified Bodies perform a conformity assessment and provide the *Conformité Européenne* (CE) mark; the manufacturer can choose any recognized Notified Body in Europe to undergo the certification process.

A recent study by Muehlematter et al. (2021) identified 222 AI-based medical devices approved by the FDA and 240 AI-based CE-certified devices in Europe, reflecting a steep increase in the number of approved devices between 2015 and 2020 (24.7 times more in the US and 17.5 times more in Europe). However, the number of approved devices varies greatly across medical specialties. While there are 129 (58.5%) FDA-approved and 126 (53%) CE-certified AI systems to be used in radiology, there are only 2 (0.01%) FDA-approved and 12 CE-certified (5%) AI systems for ophthalmology. Two devices for ophthalmology have been approved in both markets: EyeArt, for automated detection of DR, and IDx-DR, for automated DR and DME assessment. IDx-DR was also the first FDA-approved autonomous AI system in any medical specialty, undergoing the de-novo pre-market pathway. For the approval process, the FDA based its clearance on the performance of the system in a pre-registered clinical trial with pre-defined endpoints which included 900 subjects from 10 primary care sites across the US (Abràmoff et al., 2018). Some of the contributing factors to the small number of approved devices in ophthalmology are the reduced amount of curated datasets for AI development compared to other specialties, and the harder integration of AI systems with the current clinical infrastructure (Section 2.2). By any means, based only on the number of approved AI systems, one cannot extrapolate that they are actually being used in the clinic. Similarly, the actual clinical benefits cannot be often defined based on the validation studies that were considered for approval. A recent study by Wu et al. (2021) highlights this aspect. They showed that only 3% of 130 FDA-approved AI devices had undergone a prospective validation at their submission and therefore demonstrated their true utility when used in clinical practice, as seen in Section 6. Additionally, 71.5% of FDA-approved AI devices had not reported a multi-site evaluation, which is key to prevent AI bias and problems of domain adaptation, as seen in Section 5.2.

There has been favorable progress in the development of specific approval pathways for AI-based medical devices across regulatory bodies. In the US, the FDA has expressed the need for an updated regulatory framework. It has recently published an action plan whose goal is to facilitate the FDA and AI manufacturers to evaluate and monitor a product from pre-market assessment to post-market performance, enabling iterative software modifications while ensuring patient safety (Food and Drug Administration, 2021). The European Commission has recently proposed additional regulations for high-risk AI devices, which include medical devices (European Commission, 2021b). They would complement the proposed coordinated plan on AI (European Commission, 2021a) and the guidelines for trustworthy AI (European

Commission, 2019), and stand next to the new Medical Device Regulation (European Commission, 2017) and other key legislation including the GDPR (European Commission, 2016). In other countries with a rapidly growing market for AI medical devices, such as China, there has also been a significant regulatory development in order to accelerate AI deployment (Li, 2020).

7.1.2. Proposed solutions and considerations

We identify two important objectives to facilitate updates in regulations and approval pathways and, consequently, facilitate the deployment of trustworthy AI systems in ophthalmology: collaboration and standardization.

A collaborative approach among stakeholders is key to guarantee new approval pathways that fit both clinical settings and the technological, clinical, ethical, and legal aspects of AI systems. The recently proposed guidelines for clinical trials involving AI systems (Liu et al., 2020; Rivera et al., 2020), as well as the guidelines for AI development and validation, currently under development (Collins and Moons, 2019; Sounderajah et al., 2020), are fruit of multi-stakeholder collaborations. The standardization of reporting guidelines will help streamline and guide the approval pathways set by regulatory bodies (Campbell et al., 2020). The current efforts of professional organizations, such as the Institute of Electrical and Electronics Engineers or the International Organization for Standardization, in the development of international standards for designing and monitoring AI systems will also be crucial for regulatory success (Cihon, 2019). Similarly, the update of standards of care by ophthalmological societies and working groups or committees will also accelerate the update of approval pathways, by acknowledging AI systems and their application from a more precise perspective (Abràmoff et al., 2021). In the context of DR screening, the standards of care in diabetes by the American Diabetes Association recognize from 2020 the use of FDA-approved AI systems as an alternative to traditional screening approaches, indicating that optimal utilization has yet to be fully determined (American Diabetes Association, 2020). Currently at an earlier stage, other ophthalmological societies and working groups or committees have also carried out evidence and cost-effectiveness analyses on the use of AI systems in screening settings, as done in The Netherlands (Nederlands Oogheelkundig Gezelschap, 2021) or the UK (Tufail et al., 2016; UK National Screening Committee, 2021). The acceptance of AI systems by ophthalmological societies and working groups or committees is essential for the improvement of regulatory processes, and, consequently, the successful integration of AI in ophthalmic practice.

The standardization of pre-market assessment or review by regulatory bodies could also facilitate broader and faster access to AI systems. Pre-market assessment should establish normative standards for trustworthy AI, by means of well-defined data quality control, reproducibility and benchmarking tests, and examinations of other aspects such as fairness, explainability, and liability, which should be facilitated by AI manufacturers. Ideally, AI systems would be evaluated in standardized prospective clinical trials using meaningful endpoints for patients, satisfying minimal acceptable criteria conforming to existing procedures in the corresponding clinical application (Rivera et al., 2020; Abràmoff et al., 2021). Clinical evidence extracted from the trials and used to support the initial approval should be summarized, peer-reviewed, and made publicly-available, as in Abràmoff et al. (2018). The definition of minimal acceptable criteria for AI systems is however not straightforward, as there is currently a lack of scientific evidence on this regard (Abràmoff et al., 2021). To improve and accelerate pre-market examinations, regulatory bodies must enable oversight and collaboration from other entities, including healthcare institutions, insurance companies, ophthalmological societies, and patient associations. In case of privacy concerns and access restrictions to the clinical data and the software during the approval process, regulatory bodies could acquire a centralized information-sharing role (Price, 2017).

The standardization of post-market surveillance will also be crucial. As indicated in Section 5.5, the new European Medical Device Regulation (European Commission, 2017) requires a plan for continuous risk assessment and mitigation, where it should be possible to indicate how to plan and integrate system's updates; and the action plan for specific regulation of medical AI devices recently published by the FDA included lifecycle modifications (Food and Drug Administration, 2021). Approval pathways that acknowledge the continual learning nature of AI will allow to exploit the benefits of AI systems' adaptability without going through the whole process of pre-market review again. This will be facilitated by periodic audits, both internal and external. Internal audits and monitoring would ensure a given system is performing as expected and check for necessary updates. Long-term monitoring after deployment will allow to better understand the impact of AI on the clinical workflow, analyzing actual changes in healthcare providers' performance and effectiveness. It will also allow to extract metrics that reflect actual improvements in patients' outcomes (i.e., global improvement of visual acuity in a diabetic population where AI is used to perform DR screening), and allow to check whether a system ensures patient's safety in the long term, committing to the medical ethics principle of non-maleficence (Abràmoff et al., 2021). Protocols for internal audits could be discussed and standardized by the corresponding ophthalmological societies and working groups or committees. External audits would be performed by the corresponding regulatory agency to review accumulated modifications and certify that the risk-benefit profile of the system remains acceptable (Hwang et al., 2019). Sentinel, FDA's monitoring system for medical products, could also be used to monitor AI systems. In Europe, additional efforts to implement such a system would be required, considering the heterogeneity of the judicial and political landscapes and the current decentralization of the approval pathway (Cohen et al., 2020).

More transparency on regulatory and approval processes will improve public trust and increase safety and quality of AI systems to be deployed in clinical practice. Publicly available databases or registries of approved systems are necessary to provide a better overview of what products are currently available (Muehlematter et al., 2021). These databases should contain a summary about each approved system using precise definitions of terms associated to AI from manufacturers and regulatory bodies, the statement of approval and the associated clinical evidence, and, preferably, up-to-date information of the settings where it has been deployed. In the USA, the FDA offers an open database of approved medical devices, although the information provided is currently limited in scope and can be vague in content. In Europe, the European's Commission's database on medical devices (Eudamed2) is not publicly available; however, a more comprehensive database (EUDAMED) will be soon available, with the aim to enhance access to information for the public and healthcare professionals and coordination between the different countries in the EU. Parallel efforts could be done to create an open platform focused on AI in ophthalmology, in order to provide a complete overview of available AI-based software for ophthalmic practice, and aid the comparison, selection, and implementation of such software. This has been done recently for CE-certified AI products in radiology (AI for Radiology), accompanied by a study on the scientific evidence of 100 commercially-available products (van Leeuwen et al., 2021).

7.2. Liability

7.2.1. Importance and consequences

In order to avoid liability for malpractice, healthcare providers must provide care at a competent level, considering available resources. However, the situation becomes more complicated with the integration of AI systems in clinical settings (Price et al., 2019). If an AI system fails at an assigned task, one could propose multiple error sources: the data used for development, the programming code, the input data, the improper operation, or other factors. And who should be thus held responsible: the developers who built the algorithm, the healthcare institution that allowed its deployment, the healthcare provider as final user? (Wang and Siau, 2018; He et al., 2019). In ethics, this is defined as a "problem of many hands", where due to the complexity of the situation and the number of actors involved, it is impossible or very difficult to hold someone solely responsible (Van de Poel et al., 2012).

Current regulations on liability are no longer fit for purpose given the rapid technological change. Current regulations protect clinicians from liability as long as they adhere to the standard of care. To simplify, it can be assumed that: 1) an AI system makes a recommendation either within or without the current standard of care; 2) the AI recommendation could be either correct or incorrect; 3) a clinician could either follow or reject the AI recommendation. As depicted in Price et al. (2019), eight possible scenarios result, treated differently by current regulation. It can be observed that regulation privileges following the standard of care: when clinicians adhere to the standard of care, they will not generally be held liable, regardless of a bad patient outcome. Therefore, the "safest" way to use medical AI is as a confirmatory tool to support existing decision-making processes, rather than as a tool to improve care. As long as the use of AI is not part of the standard of care, current regulations minimize the potential value of AI, since the threat of liability incentivizes clinicians to reject AI recommendations that fall out of the current standard of care, in some cases to patients' detriment (Price et al., 2019). Although AI recently became part of the standards of care for DR screening by the American Diabetes Association (American Diabetes Association, 2020), the process might be slower in different countries, as well as for AI applied to higher-risk clinical tasks, such as critical treatment decisions for blindness prevention.

In order to accelerate AI integration in clinical settings and motivate its use by the final users, it is therefore necessary to update the necessary regulations to provide clear guidance on what entity or entities hold liability. Similarly, the insurance for medical malpractice needs to be clear about liability coverage when decisions are made in part by an AI system (Yu et al., 2018). Only this way, it will be possible to exploit the full potential of medical AI and its benefits for patients and healthcare systems.

7.2.2. Proposed solutions and considerations

Several approaches to liability have been identified in case serious errors occur in relation to AI systems and individuals are harmed. They include the healthcare provider and healthcare institution being held solely responsible, the AI manufacturer being held solely responsible, or a division of responsibility (Smith and Heath Jeffery, 2020). The first approach focuses on whether the healthcare provider acted in accordance with common practice, the second focuses on whether the AI system functioned within acceptable limits. The third approach avoids assigning fault to a specific entity, where each would bear some responsibility in the outcome. This approach would address difficulties in ascertaining how a particular decision was reached by AI, but may mean that AI manufacturers would require medical indemnity insurance to pay claims (Vladeck, 2014).

In dealing with the division of responsibility, one should consider the purpose of the AI system. For an autonomous AI system, Abràmoff et al. (2020) suggested that AI manufacturers would have to assume liability for potential harm, but only if it was used properly. The responsibility for proper use and maintenance of the device lies with the healthcare providers and institution. This view has recently been endorsed by the American Medical Association (AMA) in its 2019 AI Policy (American Medical Association, 2019). Nevertheless, a shift in liability from healthcare providers to AI manufacturers for autonomous AI systems has not yet manifested itself in concrete legal reforms. Courts have been reluctant to assign product liability to AI manufacturers in healthcare, partly due to the fact that most systems have been characterized primarily as assistive AI systems (Price, 2017). Such systems solely provide information or analysis for healthcare providers to help make decisions. As clinicians are able to make an independent evaluation of output by assistive AI, they remain fully liable for such instances (Abràmoff et al., 2020). The authors have also highlighted differential settings in which AI systems can be developed. For an AI system that is privately designed and sold as a finished product, the AI manufacturer would have to bear the responsibility for errors in the output. For a similar AI system that is built in partnership with a healthcare system and will be used in a large-scale setting, the legal responsibility is more diffuse, and likely lies with the healthcare system and other parties for which liability can be divided through a comparative analysis of responsibility.

To facilitate the division of responsibility for potential errors, a study by Smith and Fotheringham (2020) proposed to use *risk pooling* through utilizing insurance. With risk pooling, parties collectively accept and prepare for the potential for harm to arise when AI systems are used, and will pay the insurance company according to their own risk. The concept of risk pooling is based on the following: "If X performs an action which imposes an unreasonable risk of harm on Y, then X is liable to Y, and therefore obliged to make an ex ante compensation into a social pool that is roughly equivalent to the cost of expected harm (i.e., the probability of actual harm multiplied by the amount of the cost incurred by the harm)" (Song, 2019). As mentioned, certain exceptions need to be made. If the healthcare provider has not used the system in the way intended, it may not be reasonable for the AI manufacturer to subsidize the provider's wrongs. Also, if the healthcare provider has used the AI system appropriately and was unable to detect a system error, the AI manufacturer should subsidize the provider (Smith and Fotheringham, 2020). Nevertheless, risk pooling will allow a better construct for division of responsibility, and will also allow a rapid mechanism of compensation to injured patients via insurance, without necessitating long expensive court battles.

In shaping the liability issue, and to improve AI-related ophthalmology practices in the future, it is also imperative that ophthalmologists and other users learn how to better use and interpret AI systems. This includes how and in what situations available AI systems should be applied, and how much confidence should be placed in recommendations from systems (Price et al., 2019). Ophthalmologists are also encouraged to support their institutions to take steps toward AI evaluation. Healthcare institutions need guidelines in place for validation of systems prior to integration, as seen in Section 6, and for the correct deployment of systems, including clear liability practices. The participation of ophthalmological societies and working groups or committees will help ensure these guidelines truly reflect what is needed in clinical care. Considering all scenarios and the positions of the different stakeholders involved, regulatory bodies may proceed to update the corresponding legislation concerning liability. Transparency and auditability along the design pipeline, ensured by AI developers, will also be key for identifying issues concerning liability on time and prevent errors that could potentially harm patients (Char et al., 2020).

7.3. Reimbursement

7.3.1. Importance and consequences

Healthcare reimbursement is the process by which a payer, i.e., a government agency or a private health insurer, pays for a service provided by a healthcare institution based on a prior billing agreement. With the arrival of new technology, such as AI, there exists the necessity of defining new agreements to ensure that its use by healthcare providers is financially covered and, consequently, incentivized. Payers decide the updates in medical management that will improve healthcare and, simultaneously, are cost-effective. While most AI manufacturers have focused on medical outcomes, such as diagnostic performance, they are not the only relevant variables for having AI systems reimbursed and used in real-world ophthalmic settings. Cost-effectiveness and outcomes in terms of efficiency and healthcare quality have an impact as big or bigger when it comes to reimbursement decisions. Without enough evidence of the return on investment, i.e., without knowing if the clinical workflow can be updated without a significant financial burden, payers will not define reimbursement agreements and healthcare institutions will not buy and integrate AI-based products. Approval of AI systems by regulatory bodies it is thus not enough without a financial incentive to use them.

An important step towards reimbursement of medical AI systems was made recently, when the USA's Centers for Medicare and Medicaid Services (CMS) accepted a new Current Procedural Terminology (CPT) code to allow the use of autonomous AI in a reimbursable primary care setting. The new code was submitted by the American Academy of Ophthalmology with the support of Digital Diagnostics to facilitate the correct billing of IDx-DR, the first FDA-authorized system for automated diagnostic assessment for DR and DME (Digital Diagnostics, 2020). Nevertheless, there is still significant uncertainty about financial reimbursement in other countries and reimbursement models. Such uncertainty affects whether healthcare institutions choose to be early adopters of AI systems (Singh et al., 2020).

7.3.2. Proposed solutions and considerations

First, it is key to study each healthcare system where an AI system is meant to be deployed, since healthcare systems in different countries have different payers, structures, and buying incentives. It is important to align the benefits of the AI system with the local incentives, such as having a preference for efficiency and/or for quality gains, and the different roles of the involved stakeholders (healthcare providers, healthcare institutions, public health bodies, government agencies, private insurers…) (van Duffelen, 2021).

When applying for being granted reimbursement, the medical outcomes of an AI system should not be the only evidence to provide, but an extensive study on cost-effectiveness and return of investment, preferably performed in the target healthcare system. These studies can be performed in a retrospective setting, as done in the context of DR screening (Tufail et al., 2016, 2017; Lee et al., 2021). However, as indicated in Section 6.3, it is preferable to study cost-effectiveness as part of a prospective validation, as done in Heydon et al. (2020) when evaluating the performance of an AI system to triage retinal images within the NHS Diabetic Eye Screening Programme. Currently, better guidance on what makes a positive business case within a given healthcare system is necessary. The corresponding institutions and payers need to create or update their guides for buyers of medical AI systems (such as the NHSX report "A Buyer's Guide to AI in Health and Care" (NHSX, 2020)) and include cost-effectiveness requirements, considering their influence on reimbursement decisions.

A re-definition of reimbursement decisions is expected to happen with the integration of AI in ophthalmic practice. It could happen that certain payers start to consider the recommendations provided by AI systems as a precondition for reimbursement and refuse to cover procedures or treatments when the AI recommended against them (Vayena et al., 2018). As a consequence, healthcare institutions might end up prioritizing profitable low-cost and low-risk patients in order to ensure reimbursement while reducing liability complexities (Section 7.2). It is thus crucial to avoid AI reimbursement decisions to take over the benefits of the patient benefits and the dialogue between healthcare providers and patients. One solution is the shifting of reimbursement models towards value-based rather than volume-based reimbursement, that is, where payers would pay healthcare providers based on the quality rather than the quantity of care given to patients. This becomes especially relevant in fee-for-service models, such as the one in the US, where many payers are already shifting from rewarding providers by treatment volume to rewarding by treatment outcome (Jiang et al., 2017). The generalization of value-based healthcare might be accelerated as a necessary side effect of AI integration (Coiera, 2019).

A communicative and collaborative approach will be key for the re-definition of reimbursement policies in ophthalmology. Payers should assess the value created by AI systems and revise their reimbursement policy to reduce the cost of healthcare while focusing on patient benefits (Yu et al., 2018). In the same way, healthcare institutions and users, such as ophthalmologists and technicians, supported by ophthalmological societies and working groups or committees, should be able to demand changes in reimbursement policies to better accommodate the needs of AI-based healthcare (Price et al., 2019). Changes in reimbursement policies are likely to develop more rapidly in the US because they are a central point within their healthcare system, as observed with the recent acceptance of a CPT code for the use of autonomous AI for DR and DME assessment in primary care settings. Subsequent updates in reimbursement policies will allow not only to standardize costs, but to increase transparency in applied procedures and clinical decisions, and to better track over time patients for whom AI was used in their care.

8. Conclusions and future directions

AI systems that are able to achieve or even exceed expert-level performance are the here and now, and an increasing number of studies are demonstrating their potential to improve patients' outcomes, healthcare providers' workflow, and access to ophthalmic care. Trustworthy AI is the necessary next step to contribute to close the current gap between the development and integration of those systems in ophthalmology. Considering the aspects and challenges studied in this manuscript and integrating the corresponding mechanisms during development, validation, and deployment will allow to generate trustworthy AI systems and, consequently, allow the benefits of AI to reach real-world ophthalmic settings. We acknowledge that important aspects and challenges can also arise after initial deployment of AI, for instance, concerning adoption, scale-up, spread, and sustainability. This manuscript limits its scope to those stages prior to integration and does not elaborate on posterior stages that can also affect trustworthiness. However, we believe that additional, prospective studies applying AI in ophthalmology are necessary for a better understanding and extraction of more solid conclusions about the challenges that can take place after initial integration. This would result in an interesting analysis for future work.

We observe there is a key factor to make trustworthy AI possible, present along the various stages of the AI design pipeline: the necessity for multistakeholder collaborations where the different parties involved in AI for ophthalmic care are represented, i.e., AI developers, reading centers, healthcare providers, healthcare institutions, ophthalmological societies and working groups or committees, patients, regulatory bodies, and payers. We acknowledge nonetheless that multi-stakeholder collaborations are complex and can be hindered by diverse factors, including conflicts of interest and lack of financial incentives. A collaborative approach is time- and resource-intensive, however, is one of the biggest determinants whether an AI system will be successfully integrated in clinical practice (Watson et al., 2020). Forming a successful collaboration will likely be a process, starting with smaller multi-stakeholder groups. An order along the AI design pipeline could be beneficial, for instance, setting a collaboration first between AI developers and healthcare institutions and providers, then involving representatives from regulatory and legal backgrounds, then representatives from patient groups and ophthalmological societies and working groups or committees, then payers and governmental agencies. Organizations where representatives from different stakeholders come together will have an important role regarding the creation of networks and providing an environment that helps find common ground on AI development and integration. The organization of workshops would be an interesting starting point. This seems possible in single-payer healthcare systems, where there is an increasing number of bodies clearly designated as a potential single point for implementation of AI in clinical care and communication between stakeholders (e.g., NHSX in the UK). This might be a greater challenge outside of single-payer systems. However, there are a few bodies like the Collaborative Community on Ophthalmic Imaging (CCOI) that would be prime candidates for this. Bringing the key stakeholders together might represent a similar challenge to that of implementing a new drug into clinical care, where CE-certification and FDA-approval do not necessarily result in

Table 1
List of main action points at the various stages of the AI design pipeline to generate trustworthy AI systems and facilitate their integration in ophthalmology. For each stakeholder involved in AI for ophthalmic care, a symbol was added in those points where his action/s and collaboration are required; the type of symbol indicates the main action required. Types of main actions: ×, definition/identification; ○, development and report; □, verification/supervision; +, integration/adoption; Δ, validation/test. Stakeholders: D, AI developers (academic and industrial manufacturers); RC, reading centers; HP, healthcare providers (ophthalmologists, nurses, optometrists, technicians); HI, healthcare institutions; OS, ophthalmological societies and working groups or committees; P, patients; RB, regulatory bodies; PY, payers (government agencies or private insurers).

Stage in AI design pipeline	Multi-stakeholder action points	D	HP	RC	HI	OS	P	RB	PY
Definition of intended use	**Unmet clinical need** to be covered by the AI system	×	×	×	×	×	×		□
	Sustainable **business model** that considers all potential **conflicts of interests**	×	□	×	×	×	□	×	×
	Model of care that ensures best alignment of AI system with target setting/s	×	×		×	□	Δ		
	Vendor-neutral solutions, **standards**, and necessary **updates** in clinical infrastructure and protocols to enhance **interoperability** in ophthalmology	+		+	+	×		□	
Data collection	**Transparency** in data collection processes and data agreements	○		○	○	□	□	□	
	Technical methods to ensure patients' data privacy	○		+	+	□	□	□	
Data labelling	**Alignment of labelling protocol and observers** with use case, as well as between the data used for training and validation/s	○		○	+	□		□	
	Adequate measures in the labelling process to **reduce grader variability** and subjectivity in the creation of reference standard and observer studies	○		○	+	□			
	Protocols to ensure minimum quality for the creation and reporting of reference standard and observer studies in different clinical applications	+		○	+	○			
Training and retrospective validation	Acceptability and variability in **input quality**, potential factors of **bias and lack of domain adaptation**, necessary level/s of **explainability**, **attack** risks given use case and target setting/s		×	×	×	×	×		
	Internal monitoring of AI system to identify the necessary **updates** related to error analyses and integration of users' feedback, adaptation to new clinical practices, or data shift, to ensure adaptability of AI system over time	□	Δ	○	○	□	□	□	□
	Technical methods to ensure adequate handling of **input quality**, **bias and domain adaptation**, meaningful **explainability**, robustness to malicious **attacks**, and the **application** of the necessary **updates** for the **adaptability** of AI system over time	○	Δ	+	+	□	Δ	□	
Prospective validation	**Setup of head-to-head studies** to **benchmark** available AI systems for a given use case and validate their **reproducibility**			○	○	×			
	Openness and **participation** in **head-to-head studies** and public **challenges** to validate the **reproducibility** of the AI system and facilitate its **benchmarking** with other systems	+		□	□	□		□	
	Necessary **information** about the AI system and **training** of users to prevent **misuse and misinterpretation**	○	Δ	+	+	×	Δ		
	Setup of prospective validation/s that allows to study AI system's performance regarding clinically-meaningful endpoints and the **impact of the AI system** on different dimensions of the **clinical workflow**, such as healthcare economics, healthcare providers, and patients, by means of **internal monitoring**	○	Δ	○	○	×	Δ	□	□
Regulation	**Pre-market approval standardization**, including requirements for **trustworthy AI** (data privacy control, proof of reproducibility, fairness examination…) and clinical evidence extracted from **prospective validation/s**	×		□	□	×		○	
	Lifecycle regulations for AI medical devices, and **standardization of post-market surveillance** based on external and internal monitoring of the AI system	×		+	+	×		○	×
	Liability for malpractice considering the type of AI system, liability coverage division, **auditability** along AI design pipeline	×	□		×	×	□	□	×
	Cost-effectiveness requirements for the use of medical AI devices per use case and shift to value-based **reimbursement** practices	□	+		+	○			○

implementation into clinical care without the approval of ophthalmological societies, HTA groups, patient groups, etc. The process can be facilitated if there is a clear body in a particular country or region and a transparent process that ensures that AI cannot be clinically deployed until input from all key stakeholders is achieved. We believe that action taken by the corresponding regulatory bodies is necessary to incentivize and, eventually, ensure multi-stakeholder collaborations along the design of AI in healthcare.

Table 1 summarizes the main action points discussed in this manuscript to generate a trustworthy AI system in ophthalmology, indicating when the participation of each stakeholder is required and the type of action/s to be taken. As observed, there are action points that require a general representation of all stakeholders, such as the establishment of a sustainable business model at the beginning of the design process, where potential conflicts of interests coming from all involved stakeholders need to be communicated and acknowledged. This is also the case for the update and development of specific regulations for medical AI devices, where the input and diverse actions taken by multiple stakeholders, including AI developers and ophthalmological societies and working groups or committees, will be essential for a realistic and clinically meaningful assessment of AI systems' quality and maintenance over time. On the other hand, other action points are more specific and require the attention of a smaller number of stakeholders. For instance, the action by reading centers, aided by ophthalmological societies and working groups or committees, will be indispensable to create reliable protocols to define the minimum quality of reference standards and observer studies. Similarly, AI developers are the most responsible stakeholder for algorithm training and the implementation of technical methods that address the different aspects concerning trustworthy AI. Table 1 is not meant to be fixed, but to be updated whenever new action points need to be added. Collaborative efforts among stakeholders will allow to identify new action points and assign new responsibilities. At its current state, the table already hints that the integration of AI in ophthalmic practice will bring changes in the tasks, responsibilities, and roles currently assimilated by the different stakeholders.

Nevertheless, the fragmented governance in the healthcare sector might hinder the division of tasks and responsibilities, as well as the coordination of efforts (Panch et al., 2019). For instance, in several healthcare systems, there is a generalized lack of clarity over the responsibility to promote and carry out the necessary updates in the clinical infrastructure for AI integration, as well as in the associated protocols relating, for instance, data privacy and cybersecurity. Collaborative efforts and the support from all involved stakeholders at high- and low-level are required. At high-level, it becomes crucial to present a compelling use case around the integration of AI in ophthalmic settings and its benefits, defining and justifying the required updates in the current infrastructure. Subsequent updates in regulation are also necessary to define the responsibilities and align the financial incentives from the different parties. Although not recognized by regulatory bodies yet, promising efforts have been made by multi-stakeholder groups formed by AI and clinical experts in the creation of guidelines that facilitate the standardization of practices and transparent reporting along the AI design pipeline, from the development and retrospective validation of AI systems (Sounderajah et al., 2020; Collins and Moons, 2019), to prospective clinical trials involving AI (Rivera et al., 2020; Liu et al., 2020). Efforts at a more specific level by ophthalmological societies and national working groups or committees are also necessary to make such guidelines more applicable to the actual clinical applications, generate a clear definition of tasks and responsibilities, and accelerate the update of regulations (Abràmoff et al., 2021).

In terms of low-level efforts to overcome the obstacles due to fragmented governance and yield a successful integration of AI in ophthalmic settings, we observe that the creation of interdisciplinary teams at healthcare institutions and reading centers will be essential. Building upon the concept of "algorithmic stewardship" introduced by Eaneff et al. (2020), such team or department would be responsible for AI-related activities within the institution. We would like to emphasize the team should be composed of experts with a diverse skills set (technical, clinical, ethical, regulatory), e.g., not exclusively clinicians, or methodologists or AI experts, so as to cover the domain expertise necessary to ensure an optimal implementation and use of medical AI over time. The team of "AI stewards" could therefore represent the institution within the multi-stakeholder collaborative efforts along the AI design pipeline. Among the main responsibilities, observed in Table 1, the team would be responsible to verify that the different aspects concerning trustworthy AI are being considered by developers, and participate in the necessary mechanisms to address them. For instance, they should ensure a correct handling of patients' data privacy and participate in the application of data protection methods, such as federated learning. The team would have an important role in the setup of validation studies and would be responsible for the necessary internal monitoring during prospective validations and post-market surveillance. Internal monitoring of the use of AI at the institution will be essential to analyze the impact of AI on the clinical workflow and to identify the necessary updates to ensure systems' adaptability and patients' safety over time. They would also be responsible for the selection, acquisition, and benchmarking of AI systems to select the most adequate system given a use case and the clinical infrastructure at the institution, prioritizing interoperability and potentially organizing head-to-head studies to generate publicly available clinical evidence. The team would thus work in constant collaboration with AI developers, patient societies, and payers, and would be advised and updated with the latest developments by the corresponding regulatory bodies and ophthalmological societies and working groups or committees.

Table 1 clearly shows that generating trustworthy AI is not a responsibility of a sole stakeholder. However, algorithm aversion, opposition to AI displacement effect, or, simply, the competition to lead the emerging technological race have led to attempts to centralized AI development under one stakeholder. A monopolistic control of AI development should be discouraged as it can restrain the creation of responsible AI solutions and slow down AI integration in healthcare. A unilateral AI design by one stakeholder might disregard the multifaceted nature of trustworthiness, promoting self-interest over ethics, for example. As the European Commission stated, "*the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of society as a whole*" (European Commission, 2020). Likewise, a unilateral development would naturally reduce variety in domain expertise and specialization, impeding as well the proper identification of responsibilities and leading to suboptimal regulations and liability approaches. Dysfunctional meta-control can also inhibit progress in innovation and hamper efforts towards a socially beneficial AI (Parasuraman and Riley, 1997). Therefore, the AI for healthcare community could benefit from safeguards against the potential for this type of monopolistic behavior.

It should be noted that trust (an attitude of the trustor, e.g., ophthalmologist) and trustworthiness (a property of the trustee, i.e., AI system) are entirely distinct. A trustworthy system does not necessarily gain trust and trust can exist in a system that is not trustworthy (Jacovi et al., 2021). For example, a doctor might have more confidence on an AI system for DR screening because it is embedded in a high-quality visual interface, independently of the model's performance ability, as shown in Ghassemi et al. (2018). A misalignment between trust and trustworthiness can cause issues of abuse, disuse or misuse of AI solutions, and harm AI adoption in practice (Parasuraman and Riley, 1997). Discovering the causes of that misalignment is a necessary step towards an optimal human-AI interaction and a well-founded acceptability of AI solutions in healthcare.

CRediT author statement

Cristina González-Gonzalo: conceptualization, methodology, data curation, formal analysis, investigation, writing – original draft, writing

– review & editing, visualization, project administration. **Eric F. Thee:** investigation, writing – original draft, writing – review & editing. **Caroline C. W. Klaver:** investigation, writing – original draft, writing – review & editing. **Aaron Y. Lee:** investigation, writing – original draft, writing – review & editing. **Reinier O. Schlingemann:** investigation, writing – original draft, writing – review & editing. **Adnan Tufail:** investigation, writing – original draft, writing – review & editing. **Frank Verbraak:** investigation, writing – original draft, writing – review & editing. **Clara I. Sánchez:** conceptualization, methodology, investigation, writing – original draft, writing – review & editing, visualization, project administration, funding acquisition. All listed authors meet the criteria for authorship agreed upon by the International Committee of Medical Journal Editors and are in agreement with the content of the manuscript.

Declaration of interest

Prof. dr. Klaver is a consultant for Bayer, Laboratoires Théa, Novartis, and CooperVision. Dr. Lee reports grants from Santen, personal fees from Genentech, personal fees from US FDA, personal fees from Johnson and Johnson, grants from Carl Zeiss Meditec, personal fees from Topcon, personal fees from Gyroscope, non-financial support from Microsoft, grants from Regeneron, outside this work. Prof. dr. Schlingemann is a consultant for Bayer, Novartis, IDx/Digital Diagnostics, Oxurion, Apellis, and Ciana Therapeutics. Prof. Tufail is an Advisory Board member at Appellis, Allergan, Bayer, Genetech/Roche, IVERIC bio, Heidelberg Engineering, Kanghong, Novartis, and cofounder of the following companies that have no products or products in development that relate to the contents of this manuscript: Oculogics, Vision AI.

Funding

This work was supported by the Deep Learning for Medical Image Analysis (DLMedIA) research program by The Dutch Research Council (project number P15-26). Prof. Sánchez received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 116076. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA and Carl Zeiss Meditec AG. Dr. Lee's funding sources: Unrestricted and career development award from RPB, NEI/NIH K23EY029246 and NIA/NIH U19AG066567.

Acknowledgement

We would like to acknowledge Bart Liefers (Erasmus University Medical Center, The Netherlands; Moorfields Eye Hospital NHS Foundation Trust, UK), Coen de Vente (University of Amsterdam, The Netherlands), Corina Brussee, Amal Hamimida, Jeanette Noordzij, Irene van Zeijl, Daniël Luttikhuizen (EyeNED Reading Center, The Netherlands), and Bram Harder (MaculaVereniging, The Netherlands) for their contribution to this work.

References

Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M., 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investig. Ophthal. Vis. Sci. 57, 5200–5206.

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C., 2018. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. 1, 1–8.

Abràmoff, M.D., Tobey, D., Char, D.S., 2020. Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. Am. J. Ophthalmol. 214, 134–142.

Abràmoff, M.D., Cunningham, B., Patel, B., Eydelman, M.B., Leng, T., Sakamoto, T., Blodi, B., Grenon, S.M., Wolf, R.M., Manrai, A.K., et al., 2021. Foundational considerations for artificial intelligence utilizing ophthalmic images. Ophthalmology.

Accenture, 2020. Artificial intelligence: Healthcare's new nervous system. Accessed: 2021, December 13. https://www.accenture.com/au-en/insights/health/artificial-intelligence-healthcare.

Al-Khaled, T., Valikodath, N., Cole, E., Hallak, J., Campbell, J.P., Chiang, M.F., Chan, R. V.P., 2020. Evaluation of physician perspectives of artificial intelligence in ophthalmology: a pilot study. Investig. Ophthal. Vis. Sci. 61, 2023–2023.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inf. Decis. Making 20, 1–9.

American Diabetes Association, 2020. Microvascular complications and foot care: standards of medical care in diabetes- 2020. Diabetes Care 43, S135–S151.

American Medical Association, 2019. 2019.Augmented Intelligence in Health Care. Accessed: 2021, August 12. https://www.ama-assn.org/system/files/2019-08/ai-2018-board-policy-summary.pdf.

Ancona, M., Ceolini, E., Oztireli, C., Gross, M., 2017. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks arXiv preprint arXiv:1711.06104.

Asgari, R., Orlando, J.I., Waldstein, S., Schlanitz, F., Baratsits, M., Schmidt-Erfurth, U., Bogunovic, H., 2019. Multiclass segmentation as multitask learning for drusen segmentation in retinal optical coherence tomography. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer, pp. 192–200.

Association of American Medical Colleges, 2020. New AAMC report confirms growing physician shortage. Accessed: 2021, December 13. https://www.aamc.org/news-insights/press-releases/new-aamc-report-confirms-growing-physician-shortage.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V., 2020. How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2938–2948.

Baxter, S.L., Lee, A.Y., 2021. Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. Curr. Opin. Ophthalmol.

Beam, A.L., Manrai, A.K., Ghassemi, M., 2020. Challenges to the reproducibility of machine learning models in health care. Jama 323, 305–306.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., Parekh, J., 2020. Identifying the "Right" Level of Explanation in a Given Situation.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., Vardoulakis, L.M., 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12.

Bellemo, V., Burlina, P., Yong, L., Wong, T.Y., Ting, D.S.W., 2018. Generative adversarial networks (GANs) for retinal fundus image synthesis. In: Asian Conference on Computer Vision. Springer, pp. 289–302.

Bengio, Y., Goodfellow, I., Courville, A., 2017. Deep Learning, vol 1. MIT Press Massachusetts, USA.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., Eckersley, P., 2020. Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648–657.

Bogunovic, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al., 2019. RETOUCH: the retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imag. 38, 1858–1874.

Bora, A., Balasubramanian, S., Babenko, B., Virmani, S., Venugopalan, S., Mitani, A., de Oliveira Marinho, G., Cuadros, J., Ruamviboonsuk, P., Corrado, G.S., et al., 2021. Predicting the risk of developing diabetic retinopathy using deep learning. Lancet Digit. Health 3, e10–e19.

Bortsova, G., González-Gonzalo, C., Wetstein, S.C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J.P., Veta, M., Sánchez, C.I., de Bruijne, M., 2021. Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. Medical Image Analysis, p. 102141.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al., 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims arXiv preprint arXiv:2004.07213.

Bulten, W., Balkenhol, M., Belinga, J.J.A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farre, X., Geronatsiou, K., Molinié, V., et al., 2021. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod. Pathol. 34, 660–671.

Burlina, P., Paul, W., Mathew, P., Joshi, N., Pacheco, K.D., Bressler, N.M., 2020. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. JAMA Ophthalmol. 138, 1070–1077.

Burlina, P., Joshi, N., Paul, W., Pacheco, K.D., Bressler, N.M., 2021. Addressing artificial intelligence bias in retinal diagnostics. Transl. Vis. Sci. Technol. 10, 13–13.

Burton, M.J., Ramke, J., Marques, A.P., Bourne, R.R., Congdon, N., Jones, I., Tong, B.A. A., Arunga, S., Bachani, D., Bascaran, C., et al., 2021. The Lancet Global Health Commission on global eye health: vision beyond 2020. Lancet Glob. Health 9, e489–e551.

Campbell, J.P., Lee, A.Y., Abràmoff, M., Keane, P.A., Ting, D.S., Lum, F., Chiang, M.F., 2020. Reporting guidelines for artificial intelligence in medical research. Ophthalmology 127, 1596–1599.

Char, D.S., Abràmoff, M.D., Feudtner, C., 2020. Identifying ethical considerations for machine learning healthcare applications. Am. J. Bioeth. 20, 7–17.

Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F., 2020. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans. Med. Imag.

Chen, Q., Keenan, T.D.L., Allot, A., Peng, Y., Agron, E., Domalpally, A., Klaver, C.C.W., Luttikhuizen, D.T., Colyer, M.H., Cukras, C.A., Wiley, H.E., Teresa Magone, M., Cousineau-Krieger, C., Wong, W.T., Zhu, Y., Chew, E.Y., Lu, Z., For the AREDS2 Deep Learning Research Group, 2021. Multimodal, multitask, multiattention (M3) deep learning detection of reticular pseudodrusen: toward automated and accessible classification of age-related macular degeneration. J. Am. Med. Inform.
Cihon, P., 2019. Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development. Future of Humanity Institute. University of Oxford.
Cisco and Cybersecurity Ventures, 2019. 2019. Press release: 2019/2020 cybersecurity almanac: 100 facts, figures, predictions and statistics. https://cybersecurityventures.com/cybersecurity-almanac-2019/. Accessed: 2021, April 26.
Cohen, I.G., Mello, M.M., 2018. HIPAA and protecting health information in the 21st century. Jama 320, 231–232.
Cohen, I.G., Evgeniou, T., Gerke, S., Minssen, T., 2020. The European artificial intelligence strategy: implications and challenges for digital health. Lancet Digit. Health 2, e376–e379.
Coiera, E., 2019. The price of artificial intelligence. In: Yearbook of medical informatics, 28, pp. 14–15.
Collins, G.S., Moons, K.G., 2019. Reporting of artificial intelligence prediction models. Lancet 393, 1577–1579.
CybelAngel, 2020. 2020. 45M medical images accessible online. Accessed: 2021, April 26. https://cybelangel.com/blog/medical-data-leaks/.
Davis, S.E., Greevy Jr., R.A., Fonnesbeck, C., Lasko, T.A., Walsh, C.G., Matheny, M.E., 2019. A nonparametric updating method to correct clinical prediction model drift. J. Am. Med. Inf. Assoc. 26, 1448–1457.
Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. 144, 114.
Digital Diagnostics, 2020. 2020. Press release: historic proposed CMS rule will allow first-ever reimbursement of autonomous AI in a healthcare setting. Accessed: 2021, December 13. https://www.digitaldiagnostics.com/newsroom/historic-proposed-cms-rule-will-allow-first-ever-reimbursement-of-autonomous-ai-in-a-healthcare-setting/.
van Dijk, E.H., Boon, C.J., 2021. Serous business: delineating the broad spectrum of diseases with subretinal fluid in the macula. Prog. Retin. Eye Res. 84.
van Duffelen, J., 2021. Making a case for buying medical imaging AI: how to define the return on investment. Accessed: 2021, June 30. https://www.aidence.com/articles/medical-imaging-ai-roi/.
Eaneff, S., Obermeyer, Z., Butte, A.J., 2020. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. Jama 324, 1397–1398.
Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R., 2021. Deep learning-enabled medical computer vision. NPJ Digit. Med. 4, 1–9.
European Commission, 2016, 2016. 2016. Regulation (EU) 2016/679 (general data protection regulation). Accessed: 2021, November 30. https://eur-lex.europa.eu/eli/reg/2016/679/oj.
European Commission, 2017. 2017. Regulation (EU) 2017/745 on medical devices (MDR). Accessed: 2021, November 26. https://eur-lex.europa.eu/eli/reg/2017/745/2017-05-05.
European Commission, 2019. 2019. Ethics guidelines for trustworthy AI. Accessed: 2021, August 12. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
European Commission, 2020. 2019. White paper on artificial intelligence: a European approach to excellence and trust. Accessed: 2021, December 13. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.
European Commission, 2021, 2021a. Coordinated plan on artificial intelligence 2021 review. Accessed: 2021, December 13. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.
European Commission, 2021, 2021b. Proposal for a regulation laying down harmonised rules on artificial intelligence. Accessed: 2021, December 13. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
Faes, L., Liu, X., Wagner, S.K., Fu, D.J., Balaskas, K., Sim, D.A., Bachmann, L.M., Keane, P.A., Denniston, A.K., 2020. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl. Vis. Sci. Technol. 9, 7–7.
Farsiu, S., Chiu, S.J., O'Connell, R.V., Folgar, F.A., Yuan, E., Izatt, J.A., Toth, C.A., Group, A.R., et al., 2014. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology 121, 162–172.
De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24, 1342–1350.
Finlayson, S.G., Chung, H.W., Kohane, I.S., Beam, A.L., 2018. Adversarial Attacks against Medical Deep Learning Systems arXiv preprint arXiv:1804.05296.
Food and Drug Administration, 2021. Artificial Intelligence/machine Learning (AI/ML)-based Software as a Medical Device (SaMD) Action Plan. Accessed: 2021, December 13. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
Gajane, P., Pechenizkiy, M., 2017. On Formalizing Fairness in Prediction with Machine Learning arXiv preprint arXiv:1710.03184.
Gee, J., Button, M., 2015. The Financial Cost of Healthcare Fraud 2015: what Data from Around the World Shows.
Ghassemi, M., Pushkarna, M., Wexler, J., Johnson, J., Varghese, P., 2018. Clinicalvis: Supporting Clinical Task-Focused Design Evaluation arXiv preprint arXiv:1810.05798.
Gillan, C., Milne, E., Harnett, N., Purdie, T.G., Jaffray, D.A., Hodges, B., 2019. Professional implications of introducing artificial intelligence in healthcare: an evaluation using radiation medicine as a testing ground. J. Radiother. Pract. 18, 5–9.
González-Gonzalo, C., Liefers, B., van Ginneken, B., Sánchez, C.I., 2020a. Iterative augmentation of visual evidence for weakly-supervised lesion localization in deep interpretability frameworks: application to color fundus images. IEEE Trans. Med. Imag. 39, 3499–3511.
González-Gonzalo, C., Sánchez-Gutiérrez, V., Hernández-Martínez, P., Contreras, I., Lechanteur, Y.T., Domanian, A., van Ginneken, B., Sánchez, C.I., 2020b. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. Acta Ophthalmol. 98, 368–377.
González-Gonzalo, C., Thee, E.F., Liefers, B., de Vente, C., Klaver, C.C., Sánchez, C.I., 2021. Hierarchical curriculum learning for robust automated detection of low-prevalence retinal disease features: application to reticular pseudodrusen. Investig. Ophthal. Vis. Sci. 62, 86–86.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014a. Generative Adversarial Networks arXiv preprint arXiv:1406.2661.
Goodfellow, I.J., Shlens, J., Szegedy, C., 2014b. Explaining and Harnessing Adversarial Examples arXiv preprint arXiv:1412.6572.
van Grinsven, M.J., Lechanteur, Y.T., van de Ven, J.P., van Ginneken, B., Hoyng, C.B., Theelen, T., Sánchez, C.I., 2013. Automatic drusen quantification and risk assessment of age-related macular degeneration on color fundus images. Investig. Ophthal. Vis. Sci. 54, 3019–3027.
van Grinsven, M.J., Buitendijk, G.H., Brussee, C., van Ginneken, B., Hoyng, C.B., Theelen, T., Klaver, C.C., Sánchez, C.I., 2015. Automatic identification of reticular pseudodrusen using multimodal retinal image analysis. Investig. Ophthal. Vis. Sci. 56, 633–639.
Guan, M., Gulshan, V., Dai, A., Hinton, G., 2018. Who said what: modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316, 2402–2410.
Gulshan, V., Rajan, R.P., Widner, K., Wu, D., Wubbels, P., Rhodes, T., Whitehouse, K., Coram, M., Corrado, G., Ramasamy, K., et al., 2019. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. 137, 987–993.
Harwich, E., Laycock, K., 2018. Thinking on its Own: AI in the NHS. Reform Research Trust.
He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., Zhang, K., 2019. The practical implementation of artificial intelligence technologies in medicine. Nat. Med. 25, 30–36.
He, W., Wang, X., Wang, L., Huang, Y., Yang, Z., Yao, X., Zhao, X., Ju, L., Wu, L., Wu, L., et al., 2021. Incremental learning for exudate and hemorrhage segmentation on fundus images. Inf. Fusion 73, 157–164.
Healthcare Innovation, 2018. 2018. IoT report: imaging systems present biggest security risk in healthcare. https://www.hcinnovationgroup.com/cybersecurity/news/13029895/iot-report-imagingsystems-present-biggest-security-risk-in-healthcare. Accessed: 2021, April 26.
Heydon, P., Egan, C., Bolter, L., Chambers, R., Anderson, J., Aldington, S., Stratton, I.M., Scanlon, P.H., Webster, L., Mann, S., et al., 2020. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br. J. Ophthalmol.
Hopkins, J.J., Keane, P.A., Balaskas, K., 2020. Delivering personalized medicine in retinal care: from artificial intelligence algorithms to clinical application. Curr. Opin. Ophthalmol. 31, 329–336.
Hormel, T.T., Hwang, T.S., Bailey, S.T., Wilson, D.J., Huang, D., Jia, Y., 2021. Artificial intelligence in OCT angiography. Prog. Retin. Eye Res. 100965.
Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., Mitchell, M., 2021. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 560–575.
Hwang, T.J., Kesselheim, A.S., Vokinger, K.N., 2019. Lifecycle regulation of artificial intelligence–and machine learning–based software devices in medicine. Jama 322, 2285–2286.
Ibrahim, H., Liu, X., Zariffa, N., Morris, A.D., Denniston, A.K., 2021. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit. Health.
Jacovi, A., Marasovic, A., Miller, T., Goldberg, Y., 2021. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635.
Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 2.
Joshi, N., Burlina, P., 2021. AI Fairness via Domain Adaptation arXiv preprint arXiv:2104.01109.
Keel, S., Lee, P.Y., Scheetz, J., Li, Z., Kotowicz, M.A., MacIsaac, R.J., He, M., 2018. Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: a pilot study. Sci. Rep. 8, 1–6.
Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17, 1–9.

Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? arXiv preprint arXiv:1703.04977.
Khan, S.M., Liu, X., Nath, S., Korot, E., Faes, L., Wagner, S.K., Keane, P.A., Sebire, N.J., Burton, M.J., Denniston, A.K., 2020. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit. Health.
Kihara, Y., Heeren, T.F., Lee, C.S., Wu, Y., Xiao, S., Tzaridis, S., Holz, F.G., Issa, P.C., Egan, C.A., Lee, A.Y., 2019. Estimating retinal sensitivity using optical coherence tomography with deep-learning algorithms in macular telangiectasia type 2. JAMA Netw. 2, e188029.
Kim, D.W., Jang, H.Y., Ko, Y., Son, J.H., Kim, P.H., Kim, S.O., Lim, J.S., Park, S.H., 2020. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. PLoS One 15, e0238908.
Klarenbeek, S.E., Schuurbiers-Siebers, O.C., van den Heuvel, M.M., Prokop, M., Tummers, M., 2021. Barriers and facilitators for implementation of a computerized clinical decision support system in lung cancer multidisciplinary team meetings—a qualitative assessment. Biology 10, 9.
Klein, R., Peto, T., Bird, A., Vannewkirk, M.R., 2004. The epidemiology of age-related macular degeneration. Am. J. Ophthalmol. 137, 486–495.
Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R., 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125, 1264–1272.
Lakhani, P., Sundaram, B., 2017. Deep learning at chest radiography: auto-mated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284, 574–582.
Lee, C.S., Lee, A.Y., 2020. Clinical applications of continual learning machine learning. Lancet Digit. Health 2, e279–e281.
Lee, A., Taylor, P., Kalpathy-Cramer, J., Tufail, A., 2017a. Machine learning has arrived. Ophthalmology 124, 1726–1728.
Lee, C.S., Tyring, A.J., Deruyter, N.P., Wu, Y., Rokem, A., Lee, A.Y., 2017b. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. Biomed. Opt Express 8, 3440–3448.
Lee, A.Y., Yanagihara, R.T., Lee, C.S., Blazes, M., Jung, H.C., Chee, Y.E., Gencarella, M. D., Gee, H., Maa, A.Y., Cockerham, G.C., et al., 2021. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 44, 1168–1175.
Lee, A.Y., Campbell, J.P., Hwang, T.S., Lum, F., Chew, E.Y., 2021a. Recommendations for standardization of images in ophthalmology. Ophthalmology.
van Leeuwen, K.G., Schalekamp, S., Rutten, M.J., van Ginneken, B., de Rooij, M., 2021. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur. Radiol. 31, 3797–3804.
Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., Miglioretti, D. L., Consortium, B.C.S., et al., 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Med. 175, 1828–1837.
Lehne, M., Sass, J., Essenwanger, A., Schepers, J., Thun, S., 2019. Why digital medicine depends on interoperability. NPJ Digit. Med. 2, 1–5.
Leslie, D., 2019. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. Available at SSRN 3403301.
Li, J., 2020. AI Regulatory Developments Issued by China Regulator. Accessed: 2021, December 13. https://www.cisema.com/en/ai-regulatory-developments-and-planning-issued-by-china-regulator/.
Li, Y., Gal, Y., 2017. Dropout inference in Bayesian neural networks with alpha-divergences. In: International Conference on Machine Learning. PMLR, p. 2052, 2061.
Li, J.P.O., Liu, H., Ting, D.S., Jeon, S., Chan, R.P., Kim, J.E., Sim, D.A., Thomas, P.B., Lin, H., Chen, Y., et al., 2020. Digital technology, telemedicine and artificial intelligence in ophthalmology: a global perspective. Prog. Retin. Eye Res. 100900.
Liberati, E.G., Ruggiero, F., Galuppo, L., Gorli, M., González-Lorenzo, M., Maraldi, M., Ruggieri, P., Friz, H.P., Scaratti, G., Kwag, K.H., et al., 2017. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement. Sci. 12, 1–13.
Liefers, B., Venhuizen, F.G., Schreur, V., van Ginneken, B., Hoyng, C., Fauser, S., Theelen, T., Sánchez, C.I., 2017. Automatic detection of the foveal center in optical coherence tomography. Biomed. Opt Express 8, 5160–5178.
Liefers, B., González-Gonzalo, C., Klaver, C., van Ginneken, B., Sánchez, C.I., 2019. Dense segmentation in selected dimensions: application to retinal optical coherence tomography. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 337–346.
Liefers, B., Colijn, J.M., González-Gonzalo, C., Verzijden, T., Wang, J.J., Joachim, N., Mitchell, P., Hoyng, C.B., van Ginneken, B., Klaver, C.C., et al., 2020. A deep learning model for segmentation of geographic atrophy to study its long-term natural history. Ophthalmology 127, 1086–1096.
Liefers, B., Taylor, P., Alsaedi, A., Bailey, C., Balaskas, K., Dhingra, N., Egan, C.A., Rodrigues, F.G., Gonzalo, C.G., Heeren, T.F., et al., 2021. Quantification of key retinal features in early and late age-related macular degeneration using deep learning. Am. J. Ophthalmol. 226, 1–12.
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.
Liu, X., Keane, P.A., Denniston, A.K., 2018. Time to regenerate: the doctor in the age of artificial intelligence. JRSM (J. R. Soc. Med.) 111, 113–116.
Liu, H., Li, L., Wormstone, I.M., Qiao, C., Zhang, C., Liu, P., Li, S., Wang, H., Mou, D., Pang, R., et al., 2019. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. JAMA Ophthalmol. 137, 1353–1360.
Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., Mahendi-ran, T., Moraes, G., Shamdas, M., Kern, C., et al., 2019b. 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 1, e271–e297.
Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K., 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat. Med. 26, 1364–1374.
Lu, J., Issaranon, T., Forsyth, D., 2017. Safetynet: detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 446–454.
Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9, 1–13.
Martin, G., Martin, P., Hankin, C., Darzi, A., Kinross, J., 2017. Cybersecurity and healthcare: how safe are we? BMJ 358.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54, 1–35.
Mehta, N., Lee, C.S., Mendonça, L.S., Raza, K., Braun, P.X., Duker, J.S., Waheed, N.K., Lee, A.Y., 2020. Model-to-data approach for deep learning in optical coherence tomography intraretinal fluid segmentation. JAMA Ophthalmol. 138, 1017–1024.
Mehta, P., Petersen, C.A., Wen, J.C., Banitt, M.R., Chen, P.P., Bojikian, K.D., Egan, C., Lee, S.I., Balazinska, M., Lee, A.Y., et al., 2021. Automated detection of glaucoma with interpretable machine learning using clinical data and multi-modal retinal images. Am. J. Ophthalmol.
Meskó, B., Görög, M., 2020. A short guide for medical professionals in the era of artificial intelligence. npj Digit. Med. 3, 1–8.
Mitani, A., Huang, A., Venugopalan, S., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Liu, Y., Varadarajan, A.V., 2020. Detection of anaemia from retinal fundus images via deep learning. Nat. Biomed. Eng. 4, 18–27.
Moore, W., Frye, S., 2019. Review of HIPAA, part 1: history, protected health information, and privacy and security rules. J. Nucl. Med. Technol. 47, 269–272.
Morgenstern, J.D., Rosella, L.C., Daley, M.J., Goel, V., Schünemann, H.J., Piggott, T., 2021. "AI's gonna have an impact on everything in society, so it has to have an impact on public health": a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. BMC Publ. Health 21, 1–14.
Muehlematter, U.J., Daniore, P., Vokinger, K.N., 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit. Health.
Müller, P.L., Liefers, B., Treis, T., Rodrigues, F.G., Olvera-Barrios, A., Paul, B., Dhingra, N., Lotery, A., Bailey, C., Taylor, P., et al., 2021. Reliability of retinal pathology quantification in age-related macular degeneration: implications for clinical trials and machine learning applications. Transl. Vis. Sci. Technol. 10, 4–4.
Natarajan, S., Jain, A., Krishnan, R., Rogye, A., Sivaprasad, S., 2019. Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. JAMA Ophthalmol. 137, 1182–1188.
Nederlands Oogheelkundig Gezelschap, 2021. 2021. Onderwerp: geautomatiseerde screening op DR met klasse IIa EC-geregistreerde medical devices. Accessed: 2021, December 13. https://www.oogheelkunde.org/richtlijn/diabetische-retinopathie-screening-geautomatiseerd-met-klasse-iia-ec-geregistreerde.
NHSX, 2020. 2020. A buyer's guide to AI in health and care. Accessed: 2021, November 30. https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/.
Niemeijer, M., Abramoff, M.D., van Ginneken, B., 2006. Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening. Med. Image Anal. 10, 888–898.
Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453.
Ongena, Y.P., Haan, M., Yakar, D., Kwee, T.C., 2020. Patients' views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire. Eur. Radiol. 30, 1033–1040.
Owen, J.P., Blazes, M., Manivannan, N., Lee, G.C., Yu, S., Durbin, M.K., Nair, A., Singh, R.P., Talcott, K.E., Melo, A.G., et al., 2021. Student becomes teacher: training faster deep learning lightweight networks for automated identification of optical coherence tomography b-scans of interest using a student-teacher framework. Biomed. Opt Express 12, 5387–5399.
Panch, T., Mattie, H., Celi, L.A., 2019. The "inconvenient truth" about AI in healthcare. NPJ Digit. Med. 2, 1–3.
Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519.
Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Hum. Factors 39, 230–253.
Parikh, R.B., Teeple, S., Navathe, A.S., 2019. Addressing bias in artificial intelligence in health care. Jama 322, 2377–2378.
Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S., 2019. Continual lifelong learning with neural networks: a review. Neural Network. 113, 54–71.
Paul, W., Wang, I.J., Alajaji, F., Burlina, P., 2021. Unsupervised discovery, control, and disentanglement of semantic attributes with applications to anomaly detection. Neural Comput. 33, 802–826.
Phene, S., Dunn, R.C., Hammel, N., Liu, Y., Krause, J., Kitade, N., Schaekermann, M., Sayres, R., Wu, D.J., Bora, A., et al., 2019. Deep learning and glaucoma specialists:

the relative importance of optic disc features to predict glaucoma referral in fundus photographs. Ophthalmology 126, 1627–1639.
Van de Poel, I., Fahlquist, J.N., Doorn, N., Zwart, S., Royakkers, L., 2012. The problem of many hands: climate change as an example. Sci. Eng. Ethics 18, 49–67.
Pope, C., Turnbull, J., 2017. Using the concept of hubots to understand the work entailed in using digital technologies in healthcare. J. Health Organisat. Manag.
Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2, 158–164.
Price, W.N., 2017. Artificial Intelligence in Health Care: Applications and Legal Implications.
Price, W.N., Gerke, S., Cohen, I.G., 2019. Potential liability for physicians using artificial intelligence. Jama 322, 1765–1766.
Quellec, G., Charriere, K., Boudi, Y., Cochener, B., Lamard, M., 2017. Deep image mining for diabetic retinopathy screening. Med. Image Anal. 39, 178–193.
Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P., 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 33–44.
Reinke, A., Eisenmann, M., Tizabi, M.D., Sudre, C.H., Radsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M.J., Cheplygina, V., et al., 2021. Common Limitations of Image Processing Metrics: A Picture Story arXiv preprint arXiv:2104.05642.
Rim, T.H., Lee, A.Y., Ting, D.S., Teo, K., Betzler, B.K., Teo, Z.L., Yoo, T.K., Lee, G., Kim, Y., Lin, A.C., et al., 2020. Detection of features associated with neovascular age-related macular degeneration in ethnically distinct data sets by an optical coherence tomography: trained deep learning algorithm. Br. J. Ophthalmol.
Rivera, S.C., Liu, X., Chan, A.W., Denniston, A.K., Calvert, M.J., 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat. Med. 26, 1351–1363.
Robinson, C., Trivedi, A., Blazes, M., Ortiz, A., Desbiens, J., Gupta, S., Dodhia, R., Bhatraju, P.K., Liles, W.C., Lee, A., et al., 2021. Deep Learning Models for COVID-19 Chest X-Ray Classification: Preventing Shortcut Learning Using Feature Disentanglement. medRxiv.
Romo-Bucheli, D., Erfurth, U.S., Bogunovic, H., 2020a. End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging. IEEE J. Biomed. Health Inform. 24, 3456–3465.
Romo-Bucheli, D., Seebock, P., Orlando, J.I., Gerendas, B.S., Waldstein, S.M., Schmidt-Erfurth, U., Bogunovic, H., 2020b. Reducing image variability across OCT devices with unsupervised unpaired learning for improved segmentation of retina. Biomed. Opt Express 11, 346–363.
Royal College of Ophthalmologists, 2020. New RCOphth workforce census Illustrates the severe shortage of eye doctors in the UK. Accessed: 2021, December 13. https://www.rcophth.ac.uk/news-views/new-rcophth-workforce-census-illustrates-the-severe-shortage-of-eye-doctors-in-the-uk/.
Royal College of Ophthalmologists, 2021. The Royal College of Ophthalmologists Endorses the AAO's Call for Standardisation of Digital Imaging. Accessed: 2021, November 17. https://www.rcophth.ac.uk/2021/04/the-royal-college-of-ophthalmologists-endorses-the-aaos-call-for-standardisation-of-digital-imaging/.
Ruamviboonsuk, P., Krause, J., Chotcomwongse, P., Sayres, R., Raman, R., Widner, K., Campana, B.J., Phene, S., Hemarat, K., Tadarati, M., et al., 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. NPJ Digit. Med. 2, 1–9.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.
Sabanayagam, C., Xu, D., Ting, D.S., Nusinovici, S., Banu, R., Hamzah, H., Lim, C., Tham, Y.C., Cheung, C.Y., Tai, E.S., et al., 2020. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. Lancet Digit. Health 2, e295–e302.
Saha, A., Hosseinzadeh, M., Huisman, H., 2021. End-to-end prostate cancer detection in bpMRI via 3D CNNs: effects of attention mechanisms, clinical priori and decoupled false positive reduction. Med. Image Anal. 73, 102155.
Sánchez, C.I., Niemeijer, M., Abràmoff, M.D., van Ginneken, B., 2010. Active learning for an efficient training strategy of computer-aided diagnosis systems: application to diabetic retinopathy screening. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 603–610.
Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al., 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 126, 552–564.
Schaekermann, M., Hammel, N., Terry, M., Ali, T.K., Liu, Y., Basham, B., Campana, B., Chen, W., Ji, X., Krause, J., et al., 2019. Remote tool-based adjudication for grading diabetic retinopathy. Transl. Vis. Sci. Technol. 8, 40-40.
Schaekermann, M., Cai, C.J., Huang, A.E., Sayres, R., 2020. Expert discussions improve comprehension of difficult cases in medical image assessment. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13.
Schlegl, T., Waldstein, S.M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.M., Podkowinski, D., Gerendas, B.S., Langs, G., Schmidt-Erfurth, U., 2018. Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 125, 549–558.
Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunovic, H., 2018. Artificial intelligence in retina. Prog. Retin. Eye Res. 67, 1–29.
Schmidt-Erfurth, U., Reiter, G.S., Riedl, S., Seeböck, P., Vogl, W.D., Blodi, B.A., Domalpally, A., Fawzi, A., Jia, Y., Sarraf, D., et al., 2021. AI-based monitoring of retinal fluid in disease activity and under therapy. Prog. Retin. Eye Res. 100972.
Sendak, M.P., Gao, M., Brajer, N., Balu, S., 2020. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit. Med. 3, 1–4.
Simoens, S., 2009. Health economic assessment: a methodological primer. Int. J. Environ. Res. Publ. Health 6, 2950–2966.
Singh, K.K., Lee, Y.J., 2017. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 3544–3553.
Singh, R.P., Hom, G.L., Abramoff, M.D., Campbell, J.P., Chiang, M.F., 2020. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl. Vis. Sci. Technol. 9, 45–45.
Singh, R., Abbas, A., Beqiri, S., Korot, E., Struyven, R., Keane, P., et al., 2021. Exploring the What-If-Tool as a solution for machine learning explainability in clinical practice. Investig. Ophthal. Vis. Sci. 62, 79– 79.
Smith, H., Fotheringham, K., 2020. Artificial intelligence in clinical decisionmaking: rethinking liability. Med. Law Int. 20, 131–154.
Smith, L., Gal, Y., 2018. Understanding Measures of Uncertainty for Adversarial Example Detection arXiv preprint arXiv:1803.08533.
Smith, M., Heath Jeffery, R.C., 2020. Addressing the challenges of artificial intelligence in medicine. Intern. Med. J. 50, 1278–1281.
Son, J., Shin, J.Y., Kim, H.D., Jung, K.H., Park, K.H., Park, S.J., 2020. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology 127, 85–94.
Song, F., 2019. Regarding a risk-pooling system of compensation. Ratio 32, 139–149.
Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Denniston, A.K., Greaves, F., Karthikesalingam, A., King, D., Liu, X., Markar, S.R., et al., 2020. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. Nat. Med. 26, 807–808.
Sun, T.Q., Medaglia, R., 2019. Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. Govern. Inf. Q. 36, 368–383.
Sun, J.K., Aiello, L.P., Abràmoff, M.D., Antonetti, D.A., Dutta, S., Pragnell, M., Levine, S. R., Gardner, T.W., 2021. Updating the staging system for diabetic retinal disease. Ophthalmology 128, 490–493.
Swiss Personalized Health Network, 2018. Swiss Ophthalmic Imaging Network (SOIN). Accessed: 2021, November 29, 2018. https://health2030.ch/project/soin-swiss-ophthalmic-imaging-network/.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing Properties of Neural Networks arXiv preprint arXiv:1312.6199.
Tao, G., Ma, S., Liu, Y., Zhang, X., 2018. Attacks meet interpretability: attribute-steered detection of adversarial samples. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 7728–7739.
Thee, E.F., Meester-Smoor, M.A., Luttikhuizen, D.T., Colijn, J.M., Enthoven, C.A., Haarman, A.E., Rizopoulos, D., Klaver, C.C., 2020. Performance of classification systems for age-related macular degeneration in the Rotterdam study. Transl. Vis. Sci. Technol. 9, 26–26.
Ting, D.S., Peng, L., Varadarajan, A.V., Keane, P.A., Burlina, P.M., Chiang, M.F., Schmetterer, L., Pasquale, L.R., Bressler, N.M., Webster, D.R., et al., 2019a. Deep learning in ophthalmology: the technical and clinical considerations. Prog. Retin. Eye Res. 72, 100759.
Ting, D.S.W., Pasquale, L.R., Peng, L., Campbell, J.P., Lee, A.Y., Raman, R., Tan, G.S.W., Schmetterer, L., Keane, P.A., Wong, T.Y., 2019b. Artificial intelligence and deep learning in ophthalmology. Br. J. Ophthalmol. 103, 167–175.
Tom, E., Keane, P.A., Blazes, M., Pasquale, L.R., Chiang, M.F., Lee, A.Y., Lee, C.S., 2020. Protecting data privacy in the age of AI-enabled ophthalmology. Transl. Vis. Sci. Technol. 9, 36–36.
Tsai, A.S., Chou, H.D., Ling, X.C., Al-Khaled, T., Valikodath, N., Cole, E., Yap, V.L., Chiang, M.F., Chan, R.P., Wu, W.C., 2021. Assessment and management of retinopathy of prematurity in the era of anti-vascular endothelial growth factor (VEGF). Prog. Retin. Eye Res. 101018.
Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al., 2020. Human–computer collaboration for skin cancer recognition. Nat. Med. 26, 1229–1234.
Tufail, A., Kapetanakis, V.V., Salas-Vega, S., Egan, C., Rudisill, C., Owen, C.G., Lee, A., Louw, V., Anderson, J., Liew, G., et al., 2016. An observational study to assess if automated diabetic retinopathy image assessment software can replace one or more steps of manual imaging grading and to determine their cost-effectiveness. Health Technol. Assess. 20, 1–72.
Tufail, A., Rudisill, C., Egan, C., Kapetanakis, V.V., Salas-Vega, S., Owen, C.G., Lee, A., Louw, V., Anderson, J., Liew, G., et al., 2017. Automated diabetic retinopathy image assessment software: diagnostic accuracy and cost-effectiveness compared with human graders. Ophthalmology 124, 343–351.
UK National Screening Committee, 2021. Automated Grading in the NHS Diabetic Eye Screening Programme, p. 2021. Accessed: 2021, December 13. https://www.gov.uk/government/consultations/automated-grading-in-diabetic-eye-screening-rapid-review-and-evidence-map.
Varadarajan, A.V., Poplin, R., Blumer, K., Angermueller, C., Ledsam, J., Chopra, R., Keane, P.A., Corrado, G.S., Peng, L., Webster, D.R., 2018. Deep learning for predicting refractive error from retinal fundus images. Investig. Ophthal. Vis. Sci. 59, 2861–2868.
Vayena, E., Blasimme, A., Cohen, I.G., 2018. Machine learning in medicine: addressing ethical challenges. PLoS Med. 15, e1002689.
Venhuizen, F.G., van Ginneken, B., van Asten, F., van Grinsven, M.J., Fauser, S., Hoyng, C.B., Theelen, T., Sánchez, C.I., 2017. Automated staging of age-related macular degeneration using optical coherence tomography. Investig. Ophthal. Vis. Sci. 58, 2318–2328.
Venhuizen, F.G., van Ginneken, B., Liefers, B., van Asten, F., Schreur, V., Fauser, S., Hoyng, C., Theelen, T., Sánchez, C.I., 2018. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomed. Opt Express 9, 1545–1569.

de Vente, C., van Grinsven, M., De Zanet, S., Mosinska, A., Sznitman, R., Klaver, C., Sanchez, C.I., et al., 2020. Estimating uncertainty of deep neural networks for age-related macular degeneration grading using optical coherence tomography. Investig. Ophthal. Vis. Sci. 61, 1630– 1630.

de Vente, C., González-Gonzalo, C., Thee, E.F., van Grinsven, M., Klaver, C.C., Sánchez, C.I., 2021. Making AI transferable across OCT scanners from different vendors. Investig. Ophthal. Vis. Sci. 62, 2118–2118.

Verbraak, F.D., Abramoff, M.D., Bausch, G.C., Klaver, C., Nijpels, G., Schlingemann, R.O., van der Heijden, A.A., 2019. Diagnostic accuracy of a device for the automated detection of diabetic retinopathy in a primary care setting. Diabetes Care 42, 651–656.

Notal Vision, 2018. 2018. Notal Vision announces FDA grants breakthrough device designation for pioneering patient operated home Optical Coherence Tomography (OCT) system. Accessed: 2021, December 13. https://notalvision.com/assets/press-releases/Notal-Vision-Announces-FDA-Grants-Breakthrough-Device-Designation-for-Pioneering-Patient-Operated-Home-Optical-Coherence-Tomography-OCT-System.pdf.

Vladeck, D.C., 2014. Machines without principals: liability rules and artificial intelligence. Wash. Law Rev. 89, 117.

Waldstein, S.M., 2020. Opportunistic deep learning of retinal photographs: the window to the body revisited. Lancet Digit. Health 2, e269–e270.

Wang, W., Siau, K., 2018–2019. Ethical and moral issues with AI: a case study on healthcare robots. In: Twenty-fourth Americas Conference on Information Systems.

Wang, J., Deng, G., Li, W., Chen, Y., Gao, F., Liu, H., He, Y., Shi, G., 2019. Deep learning for quality assessment of retinal OCT images. Biomed. Opt Express 10, 6057–6072.

Watson, J., Hutyra, C.A., Clancy, S.M., Chandiramani, A., Bedoya, A., Ilango-van, K., Nderitu, N., Poon, E.G., 2020. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA open 3, 167–172.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., Wilson, J., 2019. The what-if tool: interactive probing of machine learning models. IEEE Trans. Visual. Comput. Graph. 26, 56–65.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al., 2019. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340.

Wilson, M., Chopra, R., Wilson, M.Z., Cooper, C., MacWilliams, P., Liu, Y., Wulczyn, E., Florea, D., Hughes, C.O., Karthikesalingam, A., et al., 2021. Validation and clinical applicability of whole-volume automated segmentation of optical coherence tomography in retinal disease using deep learning. JAMA Ophthalmol. 139, 964–973.

World Health Organization, 2021. 2021. Ethics and governance of artificial intelligence for health: WHO guidance. Accessed: 2021, August 12. https://www.who.int/publications/i/item/9789240029200.

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J., 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat. Med. 27, 582–584.

Xie, Y., Gunasekeran, D.V., Balaskas, K., Keane, P.A., Sim, D.A., Bachmann, L.M., Macrae, C., Ting, D.S., 2020. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. Transl. Vis. Sci. Technol. 9, 22–22.

Yanagihara, R.T., Lee, C.S., Ting, D.S.W., Lee, A.Y., 2020. Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review. Transl. Vis. Sci. Technol. 9, 11–11.

Yang, Q., Steinfeld, A., Zimmerman, J., 2019. Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11.

Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., et al., 2020. Predicting conversion to wet age-related macular degeneration using deep learning. Nat. Med. 26, 892–899.

Yip, M.Y., Lim, G., Lim, Z.W., Nguyen, Q.D., Chong, C.C., Yu, M., Bellemo, V., Xie, Y., Lee, X.Q., Hamzah, H., et al., 2020. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. NPJ Digit. Med. 3, 1–12.

Yoo, T.K., Choi, J.Y., Seo, J.G., Ramasubramanian, B., Selvaperumal, S., Kim, D.W., 2019. The possibility of the combination of OCT and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Med. Biol. Eng. Comput. 57, 677–687.

Yoo, T.K., Choi, J.Y., Kim, H.K., 2020. CycleGAN-based deep learning technique for artifact reduction in fundus photography. Graefe's Arch. Clin. Exp. Ophthalmol. 258, 1631–1637.

Yu, K.H., Beam, A.L., Kohane, I.S., 2018. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2, 719–731.

Yu, M., Tham, Y.C., Rim, T.H., Ting, D.S., Wong, T.Y., Cheng, C.Y., 2019. Reporting on deep learning algorithms in health care. Lancet Digit. Health 1, e328–e329.

Yuan, X., He, P., Zhu, Q., Li, X., 2019. Adversarial examples: attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. 30, 2805–2824.

van Zeeland, H., Meakin, J., Liefers, B., González-Gonzalo, C., Vaidyanathan, A., van Ginneken, B., Klaver, C.C., Sánchez, C.I., 2019. EyeNED workstation: development of a multi-modal vendor-independent application for annotation, spatial alignment and analysis of retinal images. Investig. Ophthal. Vis. Sci. 60, 6118–6118.

Zhang, B., Dafoe, A., 2020. US public opinion on the governance of artificial intelligence. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 187–193.

Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning, PMLR, pp. 7472–7482.