European Parliament

Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 729.512 – June 2022

EN

Artificial intelligence in healthcare

Applications, risks, and ethical and societal impacts

In recent years, the use of artificial intelligence (AI) in medicine and healthcare has been praised for the great promise it offers, but has also been at the centre of heated controversy. This study offers an overview of how AI can benefit future healthcare, in particular increasing the efficiency of clinicians, improving medical diagnosis and treatment, and optimising the allocation of human and technical resources.

The report identifies and clarifies the main clinical, social and ethical risks posed by AI in healthcare, more specifically: potential errors and patient harm; risk of bias and increased health inequalities; lack of transparency and trust; and vulnerability to hacking and data privacy breaches.

The study proposes mitigation measures and policy options to minimise these risks and maximise the benefits of medical AI, including multi-stakeholder engagement through the AI production lifetime, increased transparency and traceability, in-depth clinical validation of AI tools, and AI training and education for both clinicians and citizens.

AUTHORS

This study has been written by the following authors at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

Karim Lekadir, University of Barcelona Department of Mathematics and Computer Science, Artificial Intelligence in Medicine Lab, Barcelona, Spain; Gianluca Quaglio, Panel for the Future of Science and Technology (STOA), European Parliament, Brussels, Belgium; Anna Tselioudis Garmendia, School of Public Health, Faculty of Medicine, Imperial College London, UK; Catherine Gallin, University of Barcelona Department of Mathematics and Computer Science, Artificial Intelligence in Medicine Lab, Barcelona, Spain.

ADMINISTRATOR RESPONSIBLE

Gianluca Quaglio, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in May 2022.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

Brussels © European Union, 2022.

PE 729.512
ISBN: 978-92-846-9456-3
doi:10.2861/568473
QA-07-22-328-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

Objectives

In recent years, a burgeoning interest in and concern over the use of artificial intelligence (AI) in medicine and healthcare has stood at the centre of interdisciplinary scientific research, political debate, and social activism. The goal of this report is to explain the areas in which AI can contribute to the medical and healthcare field, pinpoint the most significant risks relating to its application in this high-stakes and quickly-changing field, and present policy options to counteract these risks, in order to optimise the use of biomedical AI. Not only will this ensure the safety and respectful treatment of patients receiving AI-mediated healthcare, it should also aid the clinicians and developers involved in implementing it.

Methodology

This study employs an interdisciplinary methodology based on a comprehensive (but non-systematic) literature review and analysis of existing scientific articles, white papers, recent guidelines and regulations, governance proposals, AI studies, and online publications. The multi-disciplinary resources examined for this report include works from the fields of computer science, biomedical research, the social sciences, biomedical ethics, law, industry, and government reporting. This report explores a wide range of technical obstacles and solutions, clinical studies and results, as well as government proposals and consensus guidelines.

Specific applications of AI in medicine and healthcare

This study first outlines the potential for AI in medicine to address pressing issues, in particular the ageing population and the rise of chronic diseases, a lack of health personnel, inefficiency of health systems, lack of sustainability, and health inequities. The report also details the different fields in which biomedical AI could make the most significant contributions: 1) clinical practice, 2) biomedical research, 3) public health, and 4) health administration.

In the realm of clinical practice, the report goes into further detail concerning specific contributions – both realised and potential – to particular medical areas such as radiology, cardiology, digital pathology, emergency medicine, surgery, medical risk and disease prediction, adaptive interventions home care, and mental health. In biomedical research, the report details the potential contributions of AI to clinical research, drug discovery, clinical trials, and personalised medicine. Lastly, the report presents potential contributions of AI at the public health level as well as to global health.

Risks of AI in healthcare

This study identified and clarifies seven main risks of AI in medicine and healthcare: 1) patient harm due to AI errors, 2) the misuse of medical AI tools, 3) bias in AI and the perpetuation of existing inequities, 4) lack of transparency, 5) privacy and security issues, 6) gaps in accountability, and 7) obstacles in implementation. Each section, as summarised below, not only describes the risk at hand, but also proposes potential mitigation measures.

Patient harm due to AI errors

The study explains the main causes of AI errors: noise and artefacts in AI clinical inputs and measurements, data shift between AI training data and real-world data, and unexpected variations in clinical contexts and environments. The medical consequences of such errors may include missed

diagnosis of life-threatening conditions as well as false diagnosis, leading to inadequate treatment and incorrect scheduling or prioritisation of intervention.

Misuse of biomedical AI tools

AI tools, even when accurate and robust, are dependent on how human beings use them in practice and how the results they produce are used; in the healthcare context, these human actors include clinicians, healthcare professionals and patients. Incorrect usage of AI tools can result in incorrect medical assessment and decision making, and subsequently in potential harm for the patient.

Potential causes of AI misuse include limited involvement of clinicians and citizens in AI development, a lack of AI training in medical AI among healthcare professionals, lack of awareness and literacy among patients and the general public, and the proliferation of easily accessible online and mobile AI solutions without sufficient explanation and information.

Risk of bias in medical AI and perpetuation of inequities

Systemic human biases often make their way into AI models, including widespread and rooted bias based on sex and gender, race and ethnicity, age, socioeconomic status, geographic location, and urban or rural contexts. The most common causes of AI biases in the healthcare sphere are due to biased and imbalanced datasets which may be based on structural bias and discrimination (systemic discrimination that is imbedded in the ways that data is collected or the ways in which doctors treat their patients) and disparities in access to quality equipment and digital technologies, as well as lack of diversity and interdisciplinarity in technological, scientific, clinical, and policymaking teams.

Lack of transparency

A significant risk for AI is a lack of transparency concerning the design, development, evaluation, and deployment of AI tools. AI transparency is closely linked to the concepts of traceability and explainability, which correspond to two distinct levels at which transparency is required: 1) transparency of the AI development and usage processes (traceability), and 2) transparency of the AI decisions themselves (explainability).

Specific risks associated with a lack of transparency in biomedical AI include a lack of understanding and trust in predictions and decisions generated by the AI system, difficulties in independently reproducing and evaluating AI algorithms, difficulties in identifying the sources of AI errors and defining who and/or what is responsible for them, and a limited uptake of AI tools in clinical practice and in real-world settings.

Privacy and security

The increasingly widespread development of AI solutions and technology in healthcare, recently underscored by a reliance on big data during the Covid-19 pandemic, has highlighted the potential risks of a lack of data privacy, confidentiality and protection for patients and citizens. The main risks for data privacy and security in AI for healthcare, including personal data sharing without fully informed consent, data repurposing without the patient's knowledge, data breaches that could expose sensitive or personal information, and the risk of harmful – or even potentially fatal – cyberattacks on AI solutions, at both individual and hospital or health-system level.

Gaps in accountability

'Algorithmic accountability' is a crucial aspect of trustworthy and applicable AI in the field of healthcare. However, legal lacunae continue to exist in current national and international regulations concerning who should be held accountable or liable for errors or failures of AI systems, especially in medical AI. It is difficult to define the roles and responsibilities due to the multiplicity of actors involved in the process of medical AI, from design to deployment (e.g. healthcare professionals or AI developers). This lack of definition can leave clinicians and other healthcare

professionals in a particularly vulnerable position, especially if the AI model they are using is not entirely transparent.

Obstacles to implementation in real-world healthcare

Many medical AI tools have been developed recently; however, obstacles abound in the path towards implementation, integration and use of these tools in real-world clinical settings. Such obstacles include limited data quality, structure, and interoperability across heterogeneous clinical centres and electronic health records; potential alterations in the physician-patient relationship owing to the introduction of AI medical tools; increased and under-regulated access to patient data; and a lack of clinical and technical integration and interoperability of AI tools with existing clinical workflows and electronic health systems.

Risk assessment methodology

There is a need for a structured approach to risk assessment and management that specifically addresses the technical, clinical and ethical challenges of AI in healthcare and medicine.

Regulatory frameworks for AI

AI risks can be characterised and classified according to the severity of the harm they may induce, as well as to the probability and frequency of the harm induced. Currently, the applicable regulations for medical AI tools in the EU are the 2017/745 Medical Devices Regulation (MDR) and the 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR), which were passed in 2017. However, because they were derived at a time when AI was at an early stage in its development, many aspects specific to AI are not considered, such as continuous learning of AI models or the identification of algorithmic biases.

In 2021, the European Commission published a long-awaited proposal for an AI regulation and to harmonise the rules governing AI technologies across Europe. The highest category corresponds to AI tools that contradict EU values and hence should be prohibited. The intermediate category, which corresponds to high-risk AI and comprises medical AI technologies, can be permitted only when the tools comply with specific requirements and obligations for adequate risk management, such as ensuring human oversight and conducting post-market monitoring.

The European Commission proposal for AI regulation is general for all domains of society and does not take into account the specificities and risks of AI in the healthcare domain, contrary to the MDR and IVDR regulations. Furthermore, the European Commission proposal retains of some of the limitations of the MDR and IVDR, such as the lack of mechanisms to address the dynamic nature and continuous learning of medical AI technologies.

Risk minimisation through risk self-assessment

For risk identification in AI, several stakeholders have suggested a self-assessment, structured approach composed of specified checklists and questions. For example, the independent High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, published an assessment checklist for trustworthy AI called ALTAI. The checklist is structured around seven categories: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability.

The ALTAI model is general and does not address AI in healthcare specifically. This has motivated the recent development of consensus guidelines for trustworthy AI in medicine by a network of European Commission funded research projects together with international inter-disciplinary experts. Entitled FUTURE-AI, these guidelines are organised according to six principles (fairness, universality, traceability, usability, robustness, explainability) and comprise concrete

recommendations and a self-assessment checklist to enable AI designers, developers, evaluators and regulators to develop trustworthy and ethical AI solutions in medicine and healthcare.

Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions

While identifying and mitigating risks in medical AI by means of adequate evaluation studies is crucial, existing scientific literature focused mostly on evaluating model accuracy and robustness of the AI tools in laboratory settings. Other aspects of medical AI, such as clinical safety and effectiveness, fairness and non-discrimination, transparency and traceability, as well as privacy and security, are more challenging to evaluate in controlled environments and have thus received far less attention in scientific literature.

There is a need for a more holistic, multi-faceted evaluation approach for future AI solutions in healthcare. Best practices to enhance clinical evaluation and deployment include: (i) employing standard definitions of clinical tasks (e.g. disease definition) to enable objective community-driven evaluations; (ii) defining performance elements beyond accuracy, such as for fairness, usability, explainability and transparency; (iii) subdividing the evaluation process into stages of increasing complexity (i.e. to assess feasibility, then capability, effectiveness and durability); (iv) promoting external evaluations by independent third-party evaluators; and (v) employing standardised guidelines for reporting the AI evaluation results to increase reproducibility, transparency and trust.

Policy options

1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements

In order to tailor existing frameworks and AI practices specifically to the medical field, multi-faceted risk assessment should be an integral part of the medical AI development and certification process. Furthermore, risk assessment must be domain-specific, as the clinical and ethical risks differ in different medical fields (e.g. radiology or paediatrics). In the future regulatory framework, the validation of medical AI technologies should be harmonised and strengthened to assess and identify multi-faceted risks and limitations by evaluating not only model accuracy and robustness but also algorithmic fairness, clinical safety, clinical acceptance, transparency and traceability.

2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms

For the future acceptability and implementation of medical AI tools in the real world, many stakeholders beyond AI developers – such as clinicians, patients, social scientists, healthcare managers and AI regulators – will play an integral role. Hence, new approaches are needed to promote inclusive, multi-stakeholder engagement in medical AI and ensure the AI tools are designed, validated and implemented in full alignment with the diversity of real-world needs and contexts. Future AI algorithms should therefore be developed by AI manufacturers based on co-creation, i.e. through strong and continuous collaboration between AI developers and clinical end-users, as well as with other relevant experts such as biomedical ethicists.

Integrating human- and user-centred approaches throughout the whole AI development process will enable the design of AI algorithms that better reflect the needs and cultures of healthcare workers, while also enabling potential risks to be identified and addressed at an early stage.

3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI

New approaches and mechanisms are needed to enhance the transparency of AI algorithms throughout their lifecycle. From this need can emerge the concept of an 'AI passport' for standardised description and traceability of medical AI tools. Such a passport should describe and

monitor key information about the AI technology, covering at least five categories of information: 1) model-related information; 2) data-related information; 3) evaluation-related information; 4) usage-related information; and 5) maintenance-related information.

The AI passport should be standardised to enable consistent traceability across countries and healthcare organisations. Furthermore, the concept of traceability must go beyond the mere documentation of the development process or the phase of testing the AI model; instead, it should also comprise the process of monitoring and maintaining the AI model or system in the real world by continually tracking how it functions after deployment in clinical practice and identifying potential errors or changes in performance. Hence, it is important that algorithms are developed together with live interfaces that will be intended for continuous surveillance and auditing of the AI tools after their deployment in their respective clinical environments.

4. Develop frameworks to improve the definition of accountability and monitoring of responsibilities in medical AI

Frameworks and mechanisms are needed to assign responsibility adequately to all actors in the AI workflow in medical practice, including the manufacturers, thus providing incentives for applying all measures and best practices to minimise errors and harm to the patient. Such expectations are already an integral part of the development, evaluation and commercialisation of medicines, vaccines and medical equipment, and need to be extended to future medical AI products.

Another way to bolster accountability is through periodic audits and risk assessments, which can be used to evaluate how much regulatory oversight a certain AI tool might need. To this end, the assessments must be conducted through the whole AI pipeline, from data collection, to development, to pre-clinical stages, to deployment, but also when the tools are in use.

5. Introduce education programmes and campaigns to enhance the skills of healthcare professionals and the literacy of the general public in medical AI

To increase adoption and minimise error, future medical professionals should be adequately trained in medical AI, including its advantages in terms of improving care quality and access to healthcare, and its limitations and risks. It is therefore time to update educational programmes in medicine and increase their interdisciplinarity.

Furthermore, there is an urgent need to increase the AI literacy of the public so that citizens and patients can empower themselves and thus better take advantage of the benefits of emerging medical AI tools; increased AI literacy will also help minimise the potential risk of misuse of the AI tools, especially during remote monitoring and care management.

6. Promote further research on clinical, ethical and technical robustness in medical AI

There is a need for further research on the interrelated areas of medical AI to address the current clinical, socio-ethical and technical limitations. Examples of areas for future research include explainability and interpretability, bias estimation and mitigation, and secure and privacy-preserving AI.

More research is also needed to develop adaptation methods that can ensure a high level of generalisability of future AI tools across population groups, clinical centres and geographical locations. Future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions.

7. Implement a strategy for reducing the European divide in medical AI

While the EU has made significant investments in AI in recent years, inequalities persist between different European countries. The AI divide can be explained by structural differences in research programmes and technological capacities, as well as by varying levels of investment from the public and private sectors. The disparities in AI development and implementation between EU countries are particularly marked in medical AI. In this context, the EU can act as an umbrella to coordinate an EU-wide strategy for reducing the gaps in medical AI between European countries. This strategy should include concrete actions to boost the technological, research and industrial capacities of emerging EU countries in the field of AI for healthcare.

The EU Member States, in particular those in eastern Europe, could develop specific programmes to further support future AI in healthcare. The European Commission could implement specific coordination and support programmes of activities implemented in this sector by different Member States, thereby supporting the implementation of common guidelines and approaches. Furthermore, infrastructure projects should be established specifically for those EU countries that have limited research infrastructures and data availability. Existing education-focused programmes such as the Marie-Curie training networks could be strengthened to enhance training capacities and human capital in medical AI.

Table of contents

Executive summary I

1. Introduction 1

1.1. Objectives of this study 1

1.2. Methodology and resources used 1

1.3. Definitions 2

2. Artificial intelligence applications in healthcare 4

2.1. Artificial intelligence and healthcare needs 4

2.1.1. Main challenges for EU's healthcare systems 4

2.1.2. Main application domains for AI in healthcare 5

2.2. AI in clinical practice 5

2.2.1. Radiology 6

2.2.2. Digital pathology 6

2.2.3. Emergency medicine 6

2.2.4. Surgery 7

2.2.5. Risk prediction 7

2.2.6. Adaptive interventions 7

2.2.7. Home care 8

2.2.8. Cardiology 8

2.2.9. Nephrology 9

2.2.10. Hepatology 9

2.2.11. Mental health 10

2.3. AI in biomedical research 10

2.3.1. Clinical research 10

2.3.2. Drug discovery 11

2.3.3. Clinical trials 11

2.3.4. Personalised medicine 12

2.4. AI for public and global health 12

2.4.1. Public health 12

2.4.2. Global health 13

2.5. AI in healthcare administration 13

2.5.1. Coding 13

2.5.2. Scheduling 14

2.5.3. Detection of fraudulent activity 14

2.5.4. Patient flow management 14

2.5.5. Healthcare audits 14

3. Risk of AI in healthcare 15

3.1. Patient harm due to AI errors 15

3.2. Misuse of medical AI tools 17

3.3. Risk of bias in medical AI and perpetuation of inequities 20

3.4. Lack of transparency 22

3.5. Privacy and security issues 23

3.6. Gaps in AI accountability 25

3.7. Obstacles to implementation in real-world healthcare 27

4. Risk assessment methodology 30

4.1. Regulatory frameworks for AI 30

4.2. Risk minimisation through risk self-assessment 33

4.3. Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions 36

4.3.1. Standardised definition of clinical tasks 37

4.3.2. Multi-faceted evaluation of performance beyond accuracy 37

4.3.3. Subdivision of the evaluation process into discrete phases. 40

4.3.1. Promotion of external evaluations by third-party evaluators 42

4.3.2. Standardised and comprehensive reporting of the AI evaluation procedure and results __ 43

5. Policy options **46**

5.1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements 46

5.2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms 47

5.3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI 48

5.4. Develop frameworks to better define accountability and monitor responsibilities in medical AI49

5.5. Introduce education programmes to enhance the skills of healthcare professionals and the literacy of the general public 50

5.6. Promote further research on clinical, ethical and technical robustness in medical AI 51

5.7. Implement a strategy for reducing the European divide in medical AI 52

References **53**

List of figures

Figure 1 – Relationship between artificial intelligence, machine learning and deep learning ___ 3

Figure 2 – Main classes of AI tools reviewed in this report ___ 5

Figure 3 – Summary of causes and consequences of errors and failures of medical AI algorithms, together with some recommendations for potential mitigation ___ 16

Figure 4 – Main factors that can lead to incorrect use of medical AI algorithms by clinicians and citizens and potential mitigation measures to improve usability of future algorithms ___ 18

Figure 5 – Most common biases and their causes in medical AI, and potential mitigation measures to develop AI algorithms with increased fairness and equity ___ 20

Figure 6 – Main risks resulting from the current lack of transparency associated with AI algorithms followed by possible mitigation measures ___ 22

Figure 7 – Main privacy and security risks associated with big data and AI, and some mitigation measures ___ 24

Figure 8 – Current limitations in accountability and recommendations to fill in these gaps ___ 26

Figure 9 – Obstacles for clinical implementation and integration of new AI tools in real-world healthcare practice, together with potential mitigation measures ___ 28

Figure 10 – AI risk classification according to the 2021 EU proposal on AI legislation ___ 32

Figure 11 – Recommendations for improved evaluation of algorithm performance and risks in medical AI ___ 36

Figure 12 – Example of a multi-stage approach for medical AI evaluation ___ 41

Figure 13 – Summary of policy options suggested in this report ___ 46

Figure 14 – Example of a possible AI passport that can be used to improve traceability and transparency in medical AI, by documenting all key details about the AI tools, their intended use, model and data details, evaluation results, and information from continuous monitoring and auditing ___ 49

List of tables

Table 1 – Main definitions and concepts in medical AI 2

Table 2 – Examples of performance elements for imaging AI algorithms 38

Table 3 – Excerpts of subdivided evaluation process for medical AI, based on processes implemented in the drug development sector 41

Table 4 – Reporting elements from the MINMAR reporting guidelines 44

1. Introduction

1.1. Objectives of this study

In recent years, there has been growing interest in the application of artificial intelligence (AI) in healthcare. From drug discovery to healthcare provision, artificial intelligence (AI) has the potential to revolutionise the field of health. Precisely, AI will likely improve access to healthcare and how patients are treated, but it also optimises the way resources are allocated, thus helping health systems function more effectively and efficiently (EIT Health, 2020).

The potential for AI to reshape the field of healthcare – to help improve diagnosis and enable an increasingly personalised precision approach to medicine – may seem boundless. Some of the main applications of AI in medicine include medical image quantification, automated analysis of genetic data, disease prediction, medical robotics, telemedicine and virtual doctors. The coronavirus pandemic has accelerated the development and deployment of AI applications in the medical and clinical areas, as AI-related technologies lay at the main core of the response to this worldwide health crisis.

However, as with other technological advances, AI in the domain of healthcare comes with its specific benefits and risks, and needs its own set of regulatory frameworks that address the socio-ethical implications of its use. While the implementation of AI in healthcare holds great promise, this rapidly developing field also raises concerns for patients, healthcare systems and society; these concerns include issues of clinical safety, equitable access, privacy and security, appropriate use and users, as well as liability and regulation. Hence, researchers, the general public, and policymakers have all pointed to important bioethical issues, including how to evaluate the risks and benefits of AI in healthcare, how to establish accountability in the sphere of biomedical AI and how to regulate its use in this particularly high-stakes context. Another important question at the heart of the field is whether AI might increase inclusion and fairness in the treatment of traditionally underrepresented communities, or whether it runs the risk of perpetuating and augmenting pre-existing health disparities and inequities.

The study will provide an overview of AI health-related applications and an analysis of the potential of AI to transform the provision of healthcare. The study will also define, assess and clarify risks in the current and potential applications of AI in the domain of healthcare. At the same time, it will consider major clinical, socio-ethical and regulatory aspects of AI in its various health applications. Finally, the study will also propose a series of policy options aimed at minimising the risks of medical AI, enhancing governance at the EU level and strengthening its responsible development.

1.2. Methodology and resources used

The methodology implemented in this study is based on a comprehensive interdisciplinary (but non-systematic) literature review and analysis of existing scientific articles, white papers, recent guidelines, governance proposals, AI studies and results, news articles and online publications. These have been generated by AI developers, public agencies, expert leaders, clinical researchers, healthcare professionals and social scientists that have been actively working in the field of AI for medicine and healthcare in recent years, especially in the last two to three years.

A highly interdisciplinary body of literature was examined for this report, including works from the fields of computer science, biomedical research, the social sciences, biomedical ethics, law, industry, and government reporting. Hence, this report examines a wide range of technical obstacles and solutions, clinical studies and results, as well as government proposals and consensus guidelines.

A wide range of key phrase searches were performed in literature databases, in particular in Google Scholar, PubMed and Web of Science. Depending on the different themes investigated in this study, examples of key phrase searches include 'medical AI', 'AI risks', 'ethical challenges of AI', 'clinical safety', 'AI fairness', 'AI bias', 'AI inequities', 'AI accountability', 'data privacy', AI explainability', 'AI transparency', 'risk management', 'AI evaluation'.

In addition to summarising the considerations, findings and recommendations that apply to each of the themes examined in this report, concrete examples from a wide range of medical domains and applications (e.g. in radiology, cardiology, digital pathology, surgery, emergency medicine, etc.) are provided whenever possible to illustrate the challenges and potential future directions in medical AI.

1.3. Definitions

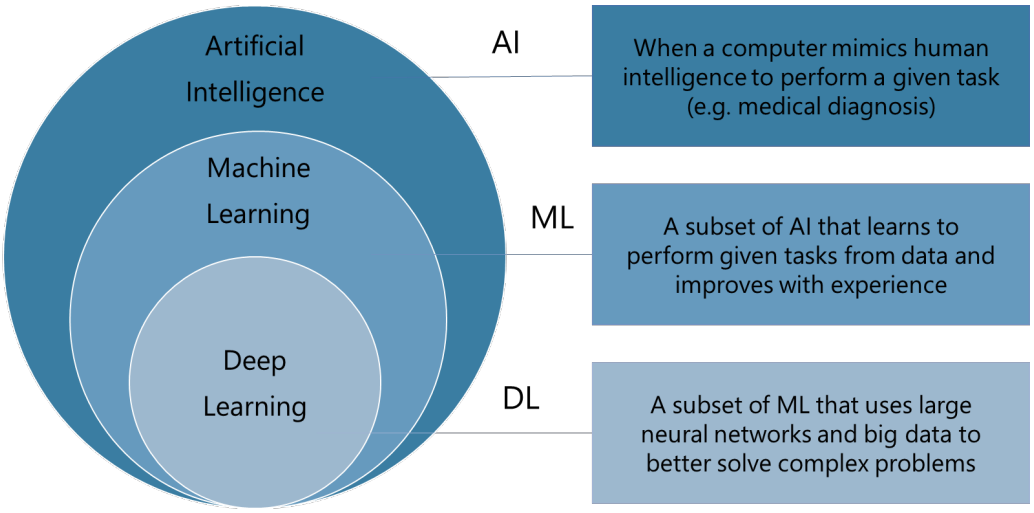
To introduce the readers to the field of AI, the table below provides a list of definitions of the main terms and concepts used throughout this report.

Table 1 – Main definitions and concepts in medical AI

Term	Definition
Artificial intelligence (AI)	Here we will first use the historical definition of AI, i.e. when a machine is able to mimic human intelligence or even surpass it to perform a given task such as prediction or reasoning. However, in this report, we will mostly focus on one subfield of AI that is dominant in the healthcare area, namely machine learning (ML).
Machine learning (ML)	ML is a subfield of AI and concerns the methods that learn to perform given tasks, such as prediction or classification, based on existing data.
Big data	The term big data is used in instances in which the data samples are too large to be adequately analysed with traditional AI methods. In this case, new methods such as deep neural networks (otherwise known as deep learning) can be used (Raghupathi et al., 2014).
Neural networks (NNs)	NNs, technically known as artificial NNs, are circuits composed of a set number of interconnected neurons organised hierarchically in layers and which are capable of learning to perform highly complex tasks from data. Each neuron acts as a type of specialised processing unit which transforms input data into output signals. These transformations are application specific and learned from available application-specific data. Progressively, the neurons combine their outputs, layer by layer, approximating the processing of a large complex function, until the network outputs a final result, such as the prediction of a disease (Esteva et al., 2019).
Deep learning	DL refers to NNs with more than three layers; in this case, the availability of big data is needed to estimate the optimal values of the parameters for this larger, more complex type of deep neural network (Goodfellow et al., 2016). Note that not all AI and ML tools are based on deep learning or NNs. Other techniques such as decision trees or support vector machines are widely used, especially when the data sample is not sufficiently large to build NNs or deep NNs (Figure 1).
AI model, AI algorithm or AI tool	Technically, in the specialised AI literature, an AI algorithm is the procedure used to build an AI model for a specific application, hence the AI model is the output of the machine learning algorithm. In other words, the same AI algorithm can be used to build models (e.g. predictive models) for many different applications, but

	the AI model is specific to a given application (e.g. predicting the patient's response to a given cancer treatment). However, the terms AI algorithms and AI models (or ML algorithms and ML models) are often used interchangeably. AI tools are AI models that are packaged to be used by end-users, so they contain more than just the AI model, such as user interfaces. In non-specialised literature, AI models, algorithms, tools, solutions and software are used interchangeably, especially in medical circles.
Training, validation & testing data	Training data are datasets that are used by AI developers to train their AI models. Validation data are also used by AI developers. However, the latter is used to optimise the parameters of the AI models so that they can be applied to new data other than the training data. In other words, validation data are used to fine-tune the AI models to make them generalisable (to use a terminology from the technical literature). Testing data are new data that are distinct from those used for training and optimising the AI models. They are used to evaluate the AI models, ideally by evaluators that did not take part in the AI development phase (in other words by external independent evaluators, though in practice AI models are still widely evaluated by the same teams that developed them in the first place).
Medical AI or healthcare AI	This is a type of AI which is focused on specific applications in medicine or healthcare.
AI design, development, evaluation & deployment	These are roughly the main steps of the AI lifecycle in healthcare. First the AI tools are designed, generally in a co-creation approach and through collaborations between AI developers and clinical experts in the field (and sometimes by also involving patients and other experts such as healthcare managers). The AI developers write some code to build and optimise the AI models from the training and validation data they have at their disposal. Subsequently, the AI model is evaluated using testing data that is distinct from the training and validation data. The AI tool (AI model with a user interface) is also evaluated with end-users (e.g. doctors and/or patients). If the evaluation is successful and convincing for the relevant stakeholders (e.g. patients, clinicians, healthcare managers, regulatory authorities), the AI tool is validated, approved, and then deployed in practice. The forementioned pipeline is of course an ideal scenario, and in practice there is some degree of variation in the AI development lifecycle.

Figure 1 – Relationship between artificial intelligence, machine learning and deep learning



2. Artificial intelligence applications in healthcare

Information generated by medical science currently spans a very wide scope; it is rapidly growing and will continue to do so both in volume and variety. In parallel, the potential for AI in medicine and health is massive and is constantly expanding as AI technologies are being developed by industry, academia, government, and individuals. It is expected that the integration of AI-based technologies into medical practice will produce substantial changes in many areas of medicine and healthcare (Roski et al., 2019; Fihn et al., 2019).

2.1. Artificial intelligence and healthcare needs

2.1.1. Main challenges for EU's healthcare systems

Before reviewing the most recent developments in medical AI in this chapter, it is important to first detail the main healthcare challenges and unmet needs that could benefit from the deployment of AI in future medical care:

Ageing population and chronic diseases. In 2017, approximately 37% of the ageing population of the EU member states reported having at least two chronic diseases, on average. Among people aged 80 and over, 56% of women and 47% of men reported multiple chronic diseases on average across EU countries (OECD/European Union, 2020).

Lack of health personnel. European countries suffer from gaps in the supply and skill level of health personnel. An estimated overall shortfall of 1.6 million healthcare workers in the EU was reported in 2013; in order to compensate for this shortage, an annual exponential growth greater than 2% would be needed. However, as this rate of increase has not been reached, the expected shortage is anticipated to reach 4.1 million by 2030 (0.6 million physicians, 2.3 million nurses and 1.3 million other healthcare professionals) (WHO, 2016; Michel, 2020).

Inefficiency. There is ample evidence of widespread inefficiency in EU healthcare systems (OECD, 2017). While the relative ability of a particular healthcare system to transform resources into outcomes differs across countries, there is considerable waste of health-related resources, which contributes to excessive expenditure (Medeiros, 2015).

Sustainability. The issue relating to health-systems sustainability is rapidly growing in the EU. According to the OECD 'Health at a glance: Europe 2020' report, the EU spends 8.3% of its GDP on healthcare, with marked differences in spending across regions: in Germany and France, it is 11% and in Luxembourg and Romania, less than 6%. Health expenditure is projected to continue to escalate, mainly due to sociodemographic changes – the ageing population and the subsequent increase in chronic diseases and long-term care needs – as well as the impact of new technologies. In addition to the aforementioned challenges, in recent years EU healthcare systems have also been under significant pressure due to economic difficulties (Quaglio, 2020). The COVID-19 pandemic in particular is expected to increase the health spending share of GDP in multiple countries.

Healthcare inequities. Healthcare inequities and inequalities persist among the EU member states and their populations. The right of every EU citizen to timely access to affordable, preventive, and curative care of high quality is one of the key principles of the newly proclaimed European Pillar of Social Rights (European Commission. The European Pillar, 2021). A recent report identified several challenges and inequalities related to healthcare access, namely: (a) inadequate public resources invested in the healthcare system; (b) fragmented population coverage; (c) gaps in the range of benefits covered; (d) prohibitive user charges, in particular for pharmaceutical products; (e) lack of protection of vulnerable groups from user charges; (f) lack of transparency on how waiting list priorities are set; (g) inadequate availability of services, particularly in rural areas; (h) problems with

attracting and retaining health professionals; (i) difficulties in reaching particularly vulnerable communities who have limited access to qualitative healthcare such as ethnic minorities and socioeconomically disadvantaged people; (j) racial bias and unequal healthcare provision (European Commission. A study of national policies 2018; Hamed, 2020).

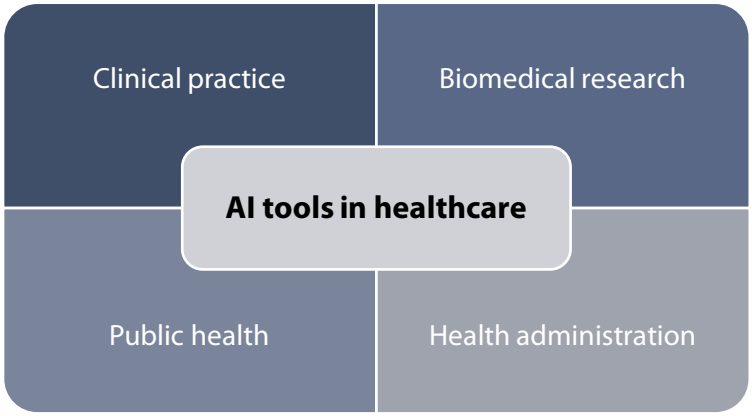
2.1.2. Main application domains for AI in healthcare

To date, AI has progressively been developed and introduced into virtually all areas of medicine, from primary care to rare diseases, emergency medicine, biomedical research and public health. Many management aspects related to health administration (e.g. increased efficiency, quality control, fraud reduction) and policy are also expected to benefit from new AI-mediated tools (Gómez-González, 2020).

Healthcare AI tools have been often classified according to the stakeholder user groups, i.e. 1) patients and citizens; 2) clinicians and caregivers; 3) healthcare administrators; and 4) public health professionals and policy makers. Classification of biomedical AI tools can also be based on the setting in which the tools are used: 1) clinical settings (hospitals, primary care centres, emergency care centres); 2) clinical processing and managing settings (laboratory, pharmacy, radiology, etc); and 3) administrative settings.

For the purpose of this study, we adopt a more comprehensive classification of AI applications, dividing them into four practices: 1) clinical; 2) research; 3) public health; and 4) administrative (Figure 2). The next sections provide a summary of the current developments and applications of AI in these four areas.

Figure 2 – Main classes of AI tools reviewed in this report



2.2. AI in clinical practice

The potential for the application of AI in the clinical setting is enormous and ranges from the automation of diagnostic processes to therapeutic decision making and clinical research. The data necessary for diagnosis and treatment comes from many sources, including clinical notes, laboratory tests, pharmacy data, medical imaging, and genomic information.

AI will play a major role in tasks such as automating image analysis (e.g. radiology, ophthalmology, dermatology, and pathology) and signal processing (e.g. electrocardiogram, audiology, and electroencephalography). In addition to its implementation in test and image interpretation, AI could be used to integrate and array results with other clinical data to facilitate clinical workflows (Topol et al., 2019). Many impressive examples exist in clinical settings where AI tools are applied, a number of which are expounded below. The following sections also touch on the possible application of AI into specific areas of medicine that are more scarcely reported, such as nephrology and personalised medicine.

2.2.1. Radiology

Radiology is among the medical specialities that have seen significant AI developments over the last years. Imaging AI technologies show promise in assisting radiologists in the work of medical image quantification. For example, segmentation with limited human supervision has been achieved by using deep network models, which enable to automatically localise and delineate the boundaries of anatomical structures or lesions (Peng & Wang, 2021). These AI tools can also prioritise and track findings that mandate early attention, and enable radiologists to concentrate on images that are most likely to be abnormal (Lee et al., 2018; Peng & Wang, 2021). A good example of AI tools for medical image segmentation is 'cvi42', a cardiovascular imaging platform commercialised by the Canadian company Circle CVI that has been adopted in over 40 countries (Zange et al., 2019).

Radiomics is another imaging processing technique in which AI has proven useful. Although the term is not strictly defined, radiomics generally aims to extract quantitative information (the so-called radiomic features), from diagnostic and treatment planning images (Gillies, 2016; Mayerhoefer et al., 2020). Radiomic features capture tissue and lesion characteristics, such as heterogeneity and shape, and may be used for clinical problem solving alone or in combination with demographic, histologic, genomic, or proteomic data. The impact of radiomics increases when the wealth of information that it provides is processed using AI techniques (Cook et al., 2019; Mayerhoefer et al., 2020).

A recent meta-analysis compared the performances of deep learning software and radiologists in the field of imaging-based diagnosis (Liu, 2019). According to the study, the diagnostic performance of deep learning models is equivalent to that of healthcare professionals. However, a major finding of the review is that most of the studies analysed have serious limitations: (i) most studies took the approach of assessing deep learning diagnostic accuracy in isolation (many studies were excluded at screening because they did not provide comparisons between the human and the machine); (ii) very few studies reported comparisons with health professionals using the same test dataset; (iii) there were very few prospective studies done in real clinical environments (most studies were retrospective and based on previously assembled datasets); iv) the scrutinised studies showed inconsistencies over key terminology.

2.2.2. Digital pathology

The term digital pathology was initially coined to include the process of digitising whole-slide images using advanced slide-scanning techniques. It now also refers to AI-based approaches for the detection and analysis of digitised images (Bera et al., 2019; Niazi et al., 2019). While the use of standardised guidelines can support the harmonisation of diagnostic processes, histopathological analysis is inherently limited by its subjective nature and by differences in judgement between independent experts (Chi et al., 2016; Evans et al., 2008; Bera et al., 2019).

AI can contribute to the alleviation of some of the challenges faced by oncologists and pathologists, including inter-subject and inter-operator variability. Several studies demonstrate that AI can have a similar level of accuracy to that of pathologists (Ehteshami Bejnordi et al., 2017) and, more significantly, can improve their diagnostic performances when used in tandem (Steiner et al., 2018; Bera et al., 2019). In digital pathology, AI has been applied to a variety of image processing and classification tasks. These include low-level tasks such as detection, focused on object recognition problems (Sornapudi et al., 2018), as well as higher-level tasks such as predicting disease diagnosis and prognosis (Corredor et al., 2018), evaluating disease severity and outcome (Mobadersany et al., 2018) and using assays to predict response to therapy (Bera, 2019).

2.2.3. Emergency medicine

Emergency medicine can benefit from AI in different phases of patient management. For instance, it offers potential value for improved patient prioritisation during triage, and is versatile in analysing

different elements of the patient's clinical history. Currently, patients are assessed with limited information in the emergency department (Berlyand et al., 2019; Kirubarajan et al., 2020). However, there is potential for emergency department flow metrics and resource allocation to be optimised through AI-driven decision making (Berlyand et al., 2018). Nevertheless, concerns remain regarding the use of AI for patient safety considering the limited body of evidence to support its implementation (Challen et al., 2019; Kirubarajan et al., 2020).

A recent scoping review analysed the applications of AI in emergency medicine in a total of 150 studies (Kirubarajan et al., 2020). According to the review, the majority of interventions are centred on: (i) the predictive capabilities of AI; (ii) improving diagnosis within the emergency department; (iii) studies focused on triage of emergent conditions; and iv) studies demonstrating that AI can assist with organisational planning and management within the emergency department.

2.2.4. Surgery

In the area of surgery, decisions sometimes need to be taken under time constraints and conditions of uncertainty regarding an individual patient's diagnoses and predicted response to treatment. Uncertainty may be imposed by unavailability of patient data (e.g. external hospital records or diagnostic tests) or absence of high-level evidence to guide important management decisions. Under such time constraints and uncertainty, clinicians may instead rely on cognitive shortcuts and snap judgments using pattern recognition and intuition (Dijksterhuis et al., 2006; Loftus et al., 2020).

Ultimately, these factors can lead to bias, error and preventable harm. In a number of conditions, traditional decision-support tools appear not to be sufficiently equipped to accommodate time constraints and uncertainty regarding diagnoses and the predicted response to treatment, both of which can impair surgical decision making (Loftus, 2020). These challenges can be overcome by AI models (Loftus et al., 2019). In fact, AI tools provide diverse sources of information (patient risk factors, anatomic information, etc.) that can help in the development of better surgical decisions (Shickel et al., 2019; Hashimoto et al., 2019).

2.2.5. Risk prediction

Risk prediction focuses on assessing the likelihood of individuals experiencing a specific health condition or outcomes. It typically generates probabilities for a wide array of outcomes ranging from death to adverse disease events (e.g. stroke, myocardial infarction, bone fracture). The process involves the identification of individuals with certain diseases or conditions and their classification according to stage, severity, and other characteristics. These individuals may subsequently be targeted to receive specific medical interventions (Miotto et al., 2016; Steele et al., 2018; Fihn et al., 2019).

Risk prediction models have long been available in healthcare. However, these are currently based on regression analysis and subsets of available clinical data, resulting in limited prediction accuracy which renders them less valuable in the clinical setting. Importantly, the advent of large repositories of data and AI techniques has shown promising signs for AI's usefulness in tailoring patient-specific conventional approaches for risk prediction (Islam, 2019). For example, predictive AI-based models in cardiovascular disease risk assessments have shown improved performance when compared to statistically derived predictive risk models (Jamthikar et al., 2019).

2.2.6. Adaptive interventions

Adaptive interventions, also defined as 'just-in-time adaptive interventions', are intervention designs aimed to deliver the right type and level of support by continuously adapting to an individual's changing internal and contextual states (Almirall et al., 2014). In particular, this allows to adjust the frequency, duration and dosage of medicines at different time points throughout the course of care.

AI-driven adaptive interventions can provide support in medical treatment through two different pathways: (i) direct input, via self-assessments by patients; or (ii) via passive data collection, where physiological information is gathered using special sensors. Using mobile technologies to collect self-assessments is referred to as ecological momentary assessment (De Vries et al., 2020). The later helps people to self-monitor behaviours at the time and in the context in which they occur.

For example, ecological momentary assessment has several benefits in substance–use disorders, such as increasing the ability to correlate instances of craving with maladaptive behaviours. Passive data collection often relies on technologies that record patterns of movement within the patient's environment, for example, via global positioning system (GPS) and wireless local area networks (Wi-Fi), which are used to acquire location-based data (Vijayan et al., 2021).

The possibility to gather spatial and temporal information (i.e. where and when the behaviours of the subject occurred) renders these tools highly specific. In addition, physiological information from special sensors (such as those measuring blood pressure, heart rate, temperature or substance concentration levels in blood), can be combined with spatial and temporal data in order to get a more detailed profile of the patient's behaviour, including monitoring physiological responses or precursors to craving (Quaglio et al., 2019).

2.2.7. Home care

In 2019, more than one fifth (20.3%) of the EU-27 population was aged 65 and over. The share of people aged 80 years or above is projected to have a two and a half folds increase between 2019 and 2100, from 5.8% to 14.6% (Eurostat. Statistical expanded, 2020). It is worth noting that the prevalence of dementia increases rapidly with age (Quaglio et al., 2016). In 2018, an estimated 9.1 million people aged over 60 were living with dementia in EU Member States (around 7% of the population aged over 60), compared to 5.9 million in 2000. In fact, the percentage of people living with dementia in EU countries is expected to rise by about 60% over the next two decades and reach 14.3 million by 2040 (OECD/EU, 2018).

Importantly, AI can play a significant role in the self-management of chronic diseases and diseases that affect the elderly. Self-management tasks range from taking medications to adjusting the patient's diet and managing health devices. Home monitoring has the potential to increase independence and improve ageing at home by keeping track of physical space and falls. In particular, tools, software, smartphone and mobile applications can enable patients to manage a large part of their own healthcare and facilitate their interactions with the healthcare system (Sapci et al., 2019).

Nevertheless, smart homes present several inconveniences, namely: 1) changing the lifestyle of users; 2) difficulties in the use of smart home technologies; 3) interoperability between systems; and 4) privacy and security constraints. Despite the current advances, the adoption of these emerging home-based technologies still falls short of end-user needs, prompting the search for new strategies (Azzi et al., 2020).

2.2.8. Cardiology

The most promising application of AI is for the automated processing of cardiac imaging data, which is necessary for the assessment of cardiac structure and function in cardiology (Lopez-Jimenez et al., 2020). Cardiac imaging modalities such as cardiac ultrasound, cardiac computer tomography and cardiovascular magnetic resonance imaging provide complex spatiotemporal data that are tedious and time consuming to process by cardiologists. The availability of new AI-driven cardiac image processing techniques has revolutionised cardiac clinical practice by enabling cardiologists to make more rapid assessment of the patients in their day-to-day practice (Lopez-Jimenez et al., 2020).

Machine learning (ML) models are set to improve the diagnostic capacity of echocardiography which constitutes the predominant cardiac imaging modality but remains heavily reliant on human expertise (Alsharqi et al., 2018). The generation of more accurate and automated echocardiograms with the use of AI is expected to reveal unrecognised imaging features that will facilitate the diagnosis of cardiovascular disease while minimising the limitations associated with human interpretation.

This is already the case in electrocardiography (ECG), for which AI models – such as deep-learning convolutional neural networks – have been generated with the use of large digital ECG datasets derived from clinical records (Siontis et al., 2021). As a result, AI-enabled ECGs are now capable of identifying diseases such as asymptomatic left ventricular dysfunction and silent atrial fibrillation, as well as phenotypic features including sex, age and race (Adedinsewo et al, 2020; Attia et al., 2019a; Attia et al., 2019b; Noseworthy et al., 2020).

Furthermore, AI has been used extensively in nuclear cardiology, which studies non-invasive imaging tools evaluating myocardial blood flow, among other things. ML models have been applied to two techniques in particular; single-photon emission computed tomography (SPECT) and myocardial perfusion imaging (MPI), to ultimately enhance the detection and prognosis of obstructive coronary artery disease (Noseworthy et al., 2020). It is believed that cardiac risk scores (calculating the 10-year risk of presenting with cardiovascular disease) will be assessed more accurately with the use of ML algorithms capable of extrapolating information and delineating unseen patterns in data derived from clinical records (Quer et al., 2021).

Although cardiovascular medicine appears to be at the forefront of AI in health, it will always, to a certain extent, depend on the expertise of cardiovascular specialists. Therefore, it is important for practitioners to be actively involved in this new and emerging field in order for imaging processing techniques to reach their full potential and perhaps revolutionise patient care (Quer et al., 2021).

2.2.9. Nephrology

The application of AI in nephrology is more scarcely reported than in other fields of medicine (Lindenmeyer et al., 2021; Chaudhuri et al., 2021). Nevertheless, its potential is increasingly being recognised by clinicians due to the promising advances made in the last decade. For instance, a novel deep learning model for ultrasound kidney imaging non-invasively classifies chronic kidney disease (CKD) (Kuo et al., 2019). In addition, the digital analysis of histopathological images has been facilitated by the development of a deep neural network capable of annotating and classifying human kidney biopsies (Hermsen, 2019). In an attempt to ameliorate early treatment of acute kidney injury (AKI), scientists took advantage of the widespread increase in data found in electronic healthcare records to develop an AI model enabling up to 48h prediction of inpatient episodes of AKI (Tomašev, 2019). On the other hand, the so-called 'Intraoperative Data Embedded Analytics' (*IDEA*) algorithm has been trained to predict the risk of developing postoperative AKI by integrating physiological data derived before and after an operation (Adhikari et al., 2019).

AI also holds potential in the computer-aided diagnosis of kidney cancer. As algorithms are becoming more robust and generalisable, they are increasingly better at identifying renal masses and distinguishing between benign and cancerous ones (Giulietti et al., 2021). Overall, the implementation of AI models in nephrology will likely facilitate prognosis, reinforce personalised medicine and reduce the global burden of kidney diseases (Park et al., 2021).

2.2.10. Hepatology

AI research is steadily progressing in many areas of medicine, and hepatology is no exception (Ahn et al., 2021). ML models have been used extensively to facilitate the diagnosis of multiple types of liver disease, most of which are life threatening. Interest has been primarily focussed on the automated detection of non-alcoholic fatty liver disease (NAFLD), as most patients remain

asymptomatic until the development of liver cirrhosis. A recently developed AI neural network shows 97.2% accuracy in diagnosing NAFLD (Okanoue et al., 2021).

Importantly, the same model is capable of distinguishing between patients with NAFLD and those with its more advanced form, NASH (non-alcoholic steato-hepatitis). Predictive models have also been developed to estimate the severity and prognosis of chronic viral hepatitis, as well as acute-on-chronic liver failure (Ahn et al., 2021). Despite the considerable progress in AI and hepatology, a number of conditions remain under-researched in this aspect, such as alcohol-associated liver disease and genetic/autoimmune liver disease, which calls for a more widespread adoption of AI in hepatology (Ahn et al., 2021).

2.2.11. Mental health

The EU suffers from a significant mental health burden. Neuropsychiatric disorders constitute 26% of diseases in EU Member States. Up to 40% of years lived with disability in the EU can be attributed to these types of mental health disorders, and especially to depression (WHO, 2021a). The cost of mood disorders and anxiety in the EU is about €170 billion per year (WHO, 2021a). In addition, it has been shown that depression and anxiety contribute greatly to chronic sick leave from the workplace and that these disorders – especially major depression – are often left untreated.

There is potential for AI to lend support to mental health patients and to mitigate the effects of a paucity of health personnel dedicated to mental health conditions. In fact, various tools are currently under development. These include digital tracking of depression and mood via keyboard interaction, speech, voice, facial recognition, sensors, and the use of interactive chatbots (Firth et al, 2017; Fitzpatrick et al., 2017; Mohr et al., 2018).

The computational power harnessed by AI systems could be leveraged to reveal the complex pathophysiology of psychiatric disorders and thus better inform therapeutic applications (Graham 2019; Lee, 2021). Machine learning has been explored to predict the efficacy of antidepressant medication (Chekroud et al., 2016), characterising depression (Wager et al., 2017), predicting suicide (Walsh et al., 2017) and psychosis in schizophrenics (Chung et al., 2018).

AI can help to differentiate between diagnoses with overlapping clinical presentations but with different treatment options (Dwyer et al., 2018). Examples include the identification of bipolar versus unipolar depression (Redlich et al., 2014), or the differentiation between types of dementia (Lee et al., 2021).

Nowadays, social media represent a form of daily communication for an extensive part of the population. Therefore, examining the content and language patterns of social media can provide insights and create new opportunities for predictive psychiatric diagnosis. Mental conditions may become observable in online contexts, while social media information analysed with machine learning has already been leveraged to predict diagnoses and relapses (Reece et al., 2017; Birnbaum et al., 2019; Yazdavar et al., 2020; Lee et al., 2021).

2.3. AI in biomedical research

2.3.1. Clinical research

Biomedical research seems to benefit more from AI-derived solutions compared to clinical applications, with recent advances also showing promising applications of AI in clinical knowledge retrieval. For example, mainstream medical knowledge resources are already using ML algorithms to rank search results, including algorithms that continuously learn from users' search behaviour (Fiorini et al., 2018a).

One example is PubMed, a widely used search engine for biomedical literature (Fiorini et al., 2018b). The AI technologies implemented by PubMed to optimise its search function include machine learning and natural language processing algorithms that are trained on patterns found in users' activities in order to improve a user's search (Fiorini et al., 2018b). For instance, Best Match is a new search algorithm for PubMed that leverages the intelligence of PubMed users and cutting-edge ML technology as an alternative to the traditional date sort order. The Best Match algorithm is trained using past user searches with dozens of relevance-ranking signals (factors), with the most important being the past usage of an article, publication date, relevance score, and type of article. This algorithm has significantly improved the finding of relevant information over the default time order in PubMed and has increased usage of relevance search over time (Fiorini et al., 2018b). Through techniques such as information extraction, automatic summarisation, and deep learning, AI has the potential to transform static narrative articles into patient-specific clinical evidence (Elliott et al., 2014).

2.3.2. Drug discovery

Drug designers frequently apply ML techniques to extract chemical information from large compound databases and to design new drugs. Central to this shift is the development of AI approaches to implement innovative modelling based on the large nature of drug datasets. As a result, recently developed AI approaches provide new solutions to enhance the efficacy and safety evaluation of candidate drugs based on big data modelling and analysis.

AI models such as these can facilitate greater understanding of a wide range of types of drugs and the clinical outcomes that they may offer (Zhu et al., 2020). For example, researchers recently trained a deep learning algorithm to predict molecules' potential antimicrobial activity. The algorithm screened over one billion molecules and virtually tested over 107 million, identifying eight antibacterial compounds that were structurally distant from known antibiotics (Stokes et al., 2020).

Compared to traditional animal models, both in vitro and in silico testing have great potential in lowering the cost of drug discovery. The application of in vitro and in silico approaches in the early stages of drug research and development procedures can reduce the number of drug attritions (Zhang et al., 2017). AI holds great potential as a method to assess compounds according to their biological capacities and toxicities. Existing AI models, such as those based on quantitative structure-activity relationship (QSAR) approaches (Golbraikh et al., 2016), can be used to predict large numbers of new compounds for various biological end points.

However, the resulting QSAR model predictions of new compounds are characterised by a number of limitations (Zhao et al., 2017; Zhu et al., 2020). Over the past decade, new efforts have stimulated the development of high-throughput screening (HTS) techniques (Zhu et al., 2014). HTS is a process that screens thousands to millions of compounds using standardised protocol. Facilitated by the combined efforts of HTS and combinatorial chemical synthesis, modern screening programmes can produce enormous amounts of biological data (Zhu et al., 2020).

2.3.3. Clinical trials

Randomised controlled trials (RCT) are the most robust method of assessing the risks and benefits of any medical intervention. However, undertaking an RCT is not always feasible. Common difficulties of unsuccessful RCTs include poor patient selection, inadequate randomisation, insufficient sample size, and poor selection of end points (Lee et al., 2020). AI models can be trained to better select the study participants with advanced statistical methods, and to assess study end points in a data-driven method. The application of AI will generate more efficient execution and greater statistical power than the one expected from traditional RCTs (Lee et al., 2020).

In addition to the efficient selection process, having a sufficiently large sample size is critical to enable detection of statistically significant differences between groups. Many RCTs require a considerable sample size because the effect of the treatment in question can be small. AI has the potential to select the right patients for RCTs. Furthermore, AI may enable more sensitive quantification of key study end points compared to the way they are usually measured. AI will also improve and complement RCTs significantly in the future. However, enhanced collaboration and synergy among clinicians, researchers, and industries is required for AI algorithms to be used to their full potential in RCTs (Lee et al., 2020).

2.3.4. Personalised medicine

Personalised medicine strongly relies on a scientific understanding of how an individual patient's unique characteristics, such as molecular and genetic profiles, make this patient vulnerable to a disease and sensitive to a therapeutic treatment (Strianese et al., 2020). Hundreds of genes have been identified for their contributions to human illness, and genetic variability in patients has also been used to distinguish individual responses to treatments (Zhu et al., 2020; Strianese et al., 2020).

The original concept of personalised medicine has been expanded to include other properties and individual clinical characteristics to ultimately form a new concept called 'extended personalised medicine'. The latter is developed from additional sources of information such as clinical sources, demographic data, social data, lifestyle parameters (sleep hours, physical activity, nutritional habits, etc), environmental conditions, etc. (Gómez-González, 2020).

AI tools may enhance the progress made in personalised medicine by evaluating the clinical benefit of different research methods and multiple data types (Mamoshina et al., 2018). Drug-target predictions (Sydow et al., 2019), metabolic network modelling, and population genetics pattern identifications (Schrider et al., 2018) constitute some of the recent advancements in this field that rely on computational modelling (Lorkowski et al., 2021). To truly impact routine care, however, the data needs to represent the diversity of patient populations (OECD, 2020). Therefore, the shift toward a data-driven personalised medicine system will have far-reaching implications for patients, clinicians, and the pharmaceutical industry (Boniolo et al., 2021).

2.4. AI for public and global health

2.4.1. Public health

Public health has many definitions, but one that is frequently used is that it is 'the science and art of preventing disease, prolonging life and promoting health through the organised efforts and informed choices of society, organisations, public and private, communities and individuals' (Wanless, 2004). Experiments with relevant AI solutions are currently under way within a number of public health areas. A selected number of these areas are discussed below.

AI can help identify specific demographics or geographical locations where the prevalence of disease or high-risk behaviours exist (Maharana & Nsoesie, 2018; Shin et al., 2018). The range of AI solutions that can improve disease surveillance is also considerable. Digital epidemiological surveillance refers to the integration of case- and event-based surveillance (e.g., news and online media, sensors, digital traces, mobile devices, social media, microbiological labs, and clinical reporting) to analyse approaches for threat verification. This has been implemented to build early warning systems for adverse drug events and air pollution (Mooney & Pejaver, 2018).

AI has already made inroads into environmental and occupational health through data generated by sensors and robots. AI has the potential to intensify contact with patients, as well as to target services to patients. An essential component of these initiatives involves contacting large numbers

of patients via a variety of automated, readily scalable methods, such as text messaging and patient portals (Fihn et al., 2019).

2.4.2. Global health

AI may provide opportunities to address health challenges in low-and middle-income countries (LMICs). These challenges include acute health workforce shortages and weak public health surveillance systems. Although not unique to such countries, these challenges are particularly relevant in low- and middle-income settings, given their contribution to morbidity and mortality (Schwalbe & Wahl, 2020). For example, in some instances, AI-driven interventions have supplemented clinical decision making towards reducing the workload of health workers (Guo & Li, 2018). New developments in AI have also helped identify disease outbreaks earlier than traditional approaches (Lake et al., 2019).

AI studies in LMICs have also addressed public health from a broader perspective: more specifically in health policy and management. These studies include AI research aimed at improving the performance of health facilities, improving resource allocation from a systems perspective, and reducing traffic-related injuries in addition to other health system issues (Schwalbe & Wahl, 2020).

Although AI can help in addressing several existing and emerging health challenges in LMICs, many issues warrant further exploration. These issues relate to the development of specific AI-driven health interventions and their real efficacy and effectiveness. Additionally, ethical regulatory standards should be implemented in order to help protect the interests and needs of the local communities and attempt to increase community-based research and engagement (Collins et al., 2019). Finally, the successful deployment of many AI tools in LMICs will require investment to strengthen the underlying healthcare systems (Schwalbe & Wahl, 2020).

2.5. AI in healthcare administration

Healthcare systems are characterised by a heavy administrative workflow with a wide range of actors and institutions, comprising patients (e.g. management of billing), health professionals, healthcare facilities and organisations (e.g. patient flow), imaging facilities, laboratories (e.g. supply chain of consumables), pharmacies, payers, and regulators. A report carried out in a primary care setting identified several potential areas of concern within this heavy administrative setting. These include time spent on reclaiming financial reimbursement, entering data into various unintegrated practice-based information systems, processing information from hospitals and other external providers and helping patients navigate a fragmented health system. The study concluded that over 50% of practice time was spent on bureaucracy, the majority of which was potentially avoidable (Clay & Stern, 2015).

AI can perform these routine tasks in a more efficient, accurate and unbiased fashion. One argument in favour of using AI in administrative practices is that errors in these activities are less serious than errors in the clinical setting. However, the danger of hacking, lack of privacy and security remains (Roski et al., 2019; OECD, 2020). AI applications can be critical in the organisation of patient flow. For example, lack of bed availability is an important cause of surgical cancellations (Kaddoum et al., 2016); however, it is a preventable administrative error in patient flow. This problem occurs frequently and is also associated with delays in discharge in clinical ward (Stylianou et al., 2017).

2.5.1. Coding

Coding is the process of extracting information from clinical records and codifying it using classifications such as the International Classification of Diseases (ICD) or diagnosis-related groups (DRGs). Coding is a complex, labour-intensive process, and coding accuracy is very important for reimbursement, administration and research. While computer-assisted coding has existed for more

than a decade, AI can enhance the accuracy and transparency of this administrative practice (OECD, 2020).

2.5.2. Scheduling

Scheduling is another example in which AI can add value to the administrative process. Algorithms fed on historical data can predict which patients may not attend their appointments, allowing practitioners to take proactive action to manage the situation. Beyond blanket or even targeted reminders, AI can address a patient's needs and queries (OECD, 2020).

2.5.3. Detection of fraudulent activity

Algorithms can also learn to look for fraudulent activity in healthcare, i.e. using a code for a more expensive medical service than the one performed (OECD, 2020).

2.5.4. Patient flow management

The fluent management and transfer of patients through the different stages of care with minimal delays is what defines patient flow (NHS, 2017). Notably, the quality of the services provided by the healthcare systems as well as patient satisfaction should be maintained throughout. Poor patient flow has been shown to negatively affect patients, staff, and the overall quality of care (Tlapa et al., 2020). Technological solutions such as AI are increasingly applied to purposes associated with patient flow (Dawoodbhoy et al., 2021). For example, the fluctuating volume of patient arrivals is a crucial but uncertain variable in hospital emergency departments.

Knowing the patient arrival volume in advance enables the smooth operational planning of emergency departments and improves related decision making (Menke et al., 2014; Ram et al., 2015). By implementing better resource planning and allocation based on predictive outcomes, the probability of overcrowding can be reduced to ultimately improve healthcare quality (Jiang et al., 2018).

2.5.5. Healthcare audits

Healthcare auditing is the process of reviewing patients' records in order to identify recommendations for improvement (NHS England, 2021). This process provides both quantitative information on the current state of affairs as well as recommendations on how to improve clinical outcomes. Audits can be carried out routinely or in the instance of a significant shortcoming in the delivery of a service, such as an increase in infection rates (Nagar et al., 2015) or patient flow concerns (Kamat & Parker, 2015).

3. Risk of AI in healthcare

In an article published more than 50 years ago, William B. Schwartz stated that 'computing science will probably exert its major effects by augmenting and, in some cases, largely replacing the intellectual functions of the physician' (Schwartz, 1970). Despite promising examples of healthcare AI solutions, Schwartz's prediction has not yet been fully realised. Initial results of AI health applications are not as robust as predicted and it is difficult to assess their real impact (Roski et al., 2019; Fihn et al., 2019).

Some players claim that the potential of AI medicine as a whole has been largely overestimated, with virtually no data demonstrating an actual improvement in patient outcomes (Angus, 2020; Parikh, 2019; Emanuel, 2019). Other experts have raised concerns over the last years regarding potential adverse consequences of medical AI, including clinical, technical and socio-ethical risks (Challen et al., 2019; Gerke & Cohen, 2020; Ellahham et al., 2020; Morley & Floridi, 2020; Manne & Kantheti, 2021).

In this chapter, we will describe the main risks that have been identified in the literature as likely to arise from the introduction of AI in future healthcare. We will focus on seven categories of risks and challenges:

1. Patient harm due to AI errors
2. Misuse of medical AI tools
3. Risk of bias in medical AI and perpetuation of inequities
4. Lack of transparency
5. Privacy and security issues
6. Gaps in AI accountability
7. Obstacles to implementation in real-world healthcare

Not only could these risks result in harms for the patients and citizens, but they could also reduce the level of trust in AI algorithms on the part of clinicians and society at large. Hence, risk assessment, classification and management must be an integral part of the AI development, evaluation and deployment processes.

3.1. Patient harm due to AI errors

Despite continuous advances in data availability and machine learning, AI-guided clinical solutions in healthcare may be associated with failures that could potentially result in safety concerns for the end-users of healthcare services (Challen et al., 2019; Ellahham et al., 2020). These AI algorithm errors can lead, for example, to (1) false negatives in the form of missed diagnoses of life-threatening diseases, (2) unnecessary treatments due to false positives (healthy persons incorrectly classified as diseased by the AI algorithm), (3) unsuitable interventions due to imprecise diagnosis, or incorrect prioritisation of interventions in emergency departments (Figure 3).

Figure 3 – Summary of causes and consequences of errors and failures of medical AI algorithms, together with some recommendations for potential mitigation

Causes of AI errors

1. Noise and artefacts in AI's clinical inputs and measurements
2. Data shift between AI training data and real-world data
3. Unexpected variations in clinical contexts and environments

Medical consequences

1. Missed diagnosis of life-threatening conditions
2. False diagnosis and thus inadequate treatments
3. Incorrect scheduling or prioritisation of interventions

Mitigation measures

1. Comprehensive multi-centre evaluation studies to identify instabilities
2. Assistive AI solutions that maintain the clinician as part of the workflow
3. Traceable and dynamic AI solutions that continue to improve over time

Assuming that AI developers have access to large-scale datasets with sufficient quality for training their AI technologies, there are still at least three major sources of error for AI in clinical practice. Firstly, AI predictions can be significantly impacted by noise in the input data during the usage of the AI tool. For example, ultrasound scanning – the most commonly used imaging modality in clinical practice due to its low-cost and portability – is known to be prone to scanning errors (Farina et al, 2012). This depends particularly on the experience of the operator, the cooperation of the patient, and the clinical context (e.g. emergency ultrasound) (Pinto et al., 2013). Even in high-income countries where there is a high level of medical training, such errors are expected to occur in some scans, thus affecting subsequent AI predictions.

Secondly, AI misclassifications may appear due to dataset shift (Subbaswamy et al., 2020), a common problem in machine learning that occurs when the statistical distribution of the data used in clinical practice is shifted, even slightly, from the original distribution of the dataset used to train the AI algorithm. This shift could be due to differences in the population groups, acquisition protocols between hospitals, or the usage of machines from different manufacturers. A recent study (Campello et al., 2020) has shown that AI models trained on cardiac magnetic resonance image (MRI) scans from two scanners (e.g. Siemens and Philips) lose accuracy when applied to MRI data acquired from different machines (e.g. General Electric and Canon).

Another example of dataset shift can be seen in a multi-centre study in the United States that built a highly accurate pneumonia diagnosis AI system based on data from two hospitals (Zech et al., 2018). When tested with data from a third hospital, a significant decrease in accuracy was noticed, suggesting potential hospital-specific biases. In another example, the company DeepMind developed a deep learning model trained on a large dataset for automated diagnosis of retinal diseases from optical coherence tomography (OCT) (De Fauw, et al., 2018). They found that the AI system was confused when applied to images obtained from a machine that is different from the one used for data acquisition at the AI training stage, with the diagnosis error increasing from 5.5% to a staggering 46%. These examples illustrate the current challenges posed in building AI tools that maintain a high level of accuracy even if the data is heterogeneous across populations, hospitals or machines.

Lastly, the predictions can be erroneous due to the difficulty of AI algorithms to adapt to unexpected changes in the environment and context in which they are applied. To illustrate the problem, researchers at Harvard Medical School described a nice example in the domain of AI for medical imaging (Yu & Kohane, 2019). They imagined an AI system that was trained to detect shadows or dense features on a chest X-ray images that are associated with lesions in major diseases such as lung cancer. Then, they listed a number of simple scenarios in which the AI may lead to incorrect predictions, such as if the X-ray technician leaves the adhesive ECG connectors on their patient's chest or if the patient wears a wedding ring and places their hand on their chest during the scan. In these scenarios, it is possible that the AI model could mistake these circular artefacts as one of the known chest lesions, resulting in a false positive.

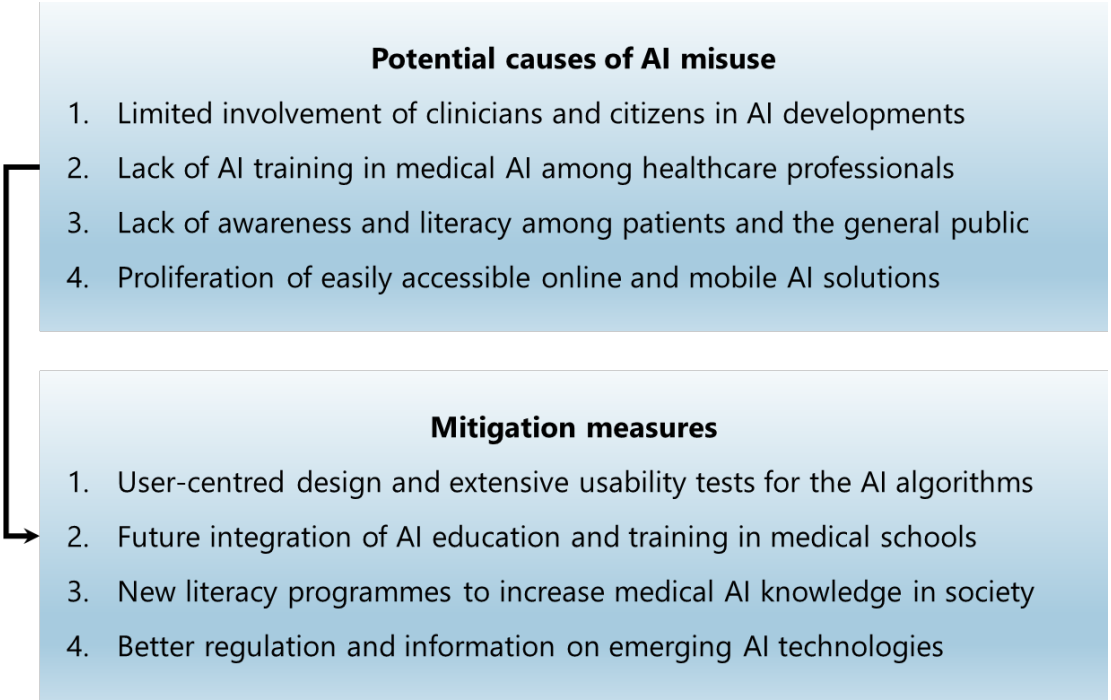
There are at least three avenues to minimise the risk of AI errors and safety issues for patients (Figure 3). First of all, standardised methods and procedures need to be defined for extensive evaluation and regulatory approval of AI solutions, in particular regarding their generalisability to new populations and sensitivity to noise. Second, the AI algorithms should be designed and implemented as assistive tools (as opposed to fully autonomous tools), such that clinicians remain part of the data processing workflow to detect and report potential errors and contextual changes, and hence to minimise harm to patients.

Furthermore, future AI solutions in healthcare must be dynamic, i.e., they should be embedded with mechanisms to continue to learn from new scenarios and mistakes as they are detected in practice. However, this last aspect will still require a certain degree of human control and vigilance to identify problems as they appear; this in turn may increase costs and reduce the initial benefits of AI. Infrastructural and technical developments will also be needed to enable regular AI updates (based on past and new training), and it will be necessary to implement policies that ensure such mechanisms are integrated into healthcare settings.

3.2. Misuse of medical AI tools

As with most health technologies, there is a risk for human error and human misuse with medical AI. Even when the developed AI algorithms are accurate and robust, they are dependent on the way they are used in practice by the end-users, including clinicians, healthcare professionals, and patients. Incorrect usage of AI tools can result in incorrect medical assessment and decision making and subsequently in potential harm for the patient. Hence, it is not enough for clinicians and the general public to have access to medical AI tools, but it is also necessary for them to understand how and when to use these technologies.

Figure 4 – Main factors that can lead to incorrect use of medical AI algorithms by clinicians and citizens and potential mitigation measures to improve usability of future algorithms



There are multiple factors that make existing medical AI technologies prone to human error or incorrect use (Figure 4). First, they have often been designed and developed by computer/data scientists with limited involvement from end-users and clinical experts. As a result, it is the user (i.e., the clinician, the nurse, the data manager or the patient) that is required to learn to use and to adapt to the new AI technology, which can lead to unnatural and complex interactions and experiences. In turn, the clinical user may encounter difficulties in understanding and applying the AI algorithm in day-to-day practice, which will limit the perception of informed decision making, while increasing the chances of human error.

This problem is exacerbated by the fact that existing training programmes in medicine are not yet tailored for medical AI and generally do not equip new clinicians with knowledge and skills in the area of AI. A survey performed in Australia and New Zealand in 2021 with 632 medical trainees (in the areas of ophthalmology, dermatology, and oncology) showed that 71% of the respondents believed AI would improve their field of medicine, especially for improved disease screening and streamlining of monotonous tasks (Scheetz et al., 2021).

However, most respondents indicated that they had never used AI applications in their work as a clinician (>80%) and only 5% viewed themselves as having excellent knowledge of the field. Another study performed in the United Kingdom surveyed 484 students from 19 medical schools and found that none of the students received any AI teaching as part of their compulsory curriculum (Sit et al., 2020). Similar conclusions were reached on knowledge and utilisation of technology-based interventions among health professionals in the European Union in other healthcare domains (Quaglio et al., 2019).

These reflections on AI education and literacy also apply to citizens and patients, who will become active users of future medical AI solutions. A 2021 study performed in five countries (Australia, the United States, Canada, Germany, and the United Kingdom) with over 6,000 citizens showed that the public generally has low awareness and understanding of AI and its use in everyday life (Gillespie et al., 2021). While younger people, men, and the university-educated tend to be more aware and

understand AI better, even these groups report low to moderate AI understanding (Gillespie et al., 2021).

Another cause for potential misuse of medical AI, which could lead to harm for citizens and patients, is the proliferation of easily accessible medical AI applications. For example, commercial mobile apps have been developed by several companies for skin cancer detection with the purpose of enabling individuals to take and upload a picture of their skin through the app, which is then directly analysed and assessed by the app's AI algorithm. Some examples of such apps include Skinvision, MelApp, skinScan and SpotMole.

While these tools are easily accessible to the general public, there is often limited information on how the AI algorithms in question have been developed and validated, while their reliability and clinical efficacy is not always demonstrated. For example, a recent study which evaluated six mobile apps for skin cancer detection demonstrated their lack of efficiency and high risk for bias (Freeman et al., 2020). The authors concluded: '*Current algorithm-based smartphone apps cannot be relied on to detect all cases of melanoma or other skin cancers. The current regulatory process for awarding the CE marking for algorithm-based apps does not provide adequate protection to the public*' (Freeman et al., 2020).

A quick search shows that many AI-powered online/mobile tools have also emerged in a wide range of medical domains and are commercially offered for medical diagnostics and health monitoring, such as Diagnostics.ai, DDXRX Doctor Ai, Symptomate, and Achu Health. While such services can constitute a promising solution for remote diagnosis and disease follow-up, their wide proliferation online can become a public health concern, in the same way that easily accessible online pharmacies have contributed to an abuse of medication by citizens (Bandivadekar , 2020).

Since there is a lot of financial gain to be made from the development and commercialisation of AI-powered web/mobile health applications, this sector will continue to attract a lot of new players and companies with varying standards of ethics, excellence and quality. The companies offering these web or mobile based AI medical tools acknowledge on their websites that their AI products are not certified medical devices and the terms of service often contain disclaimers. One can easily find disclaimers such as '*this site is designed to offer you general health information for educational purposes only*' or '*the health information furnished on this site and the interactive responses are not intended to be professional advice and are not intended to replace personal consultation with a qualified physician, pharmacist or other healthcare professional*'. However, most users may not necessarily come across, read and comprehend these disclaimers, and hence may rely on potentially incorrect information and diagnoses provided by the AI tools, which may negatively impact their decision making regarding their health.

There are several avenues to reduce human error or incorrect use of future medical AI solutions (Figure 4). First of all, end-users such as healthcare professionals, specialists, technicians or patients should be closely involved in the design and development of AI solutions to ensure their points of view, preferences and contexts are well integrated into the final tools that will be deployed and used. Furthermore, education and literacy programmes on AI and medical AI should be developed and generalised across education circles and society to increase the knowledge and skills of future AI end-users and hence reduce human error. Finally, it is important that public agencies help regulate the sector of web/mobile medical AI, such that the citizens are well informed and protected against the misuse and abuse of these emerging, easily accessible AI technologies.

3.3. Risk of bias in medical AI and perpetuation of inequities

Figure 5 – Most common biases and their causes in medical AI, and potential mitigation measures to develop AI algorithms with increased fairness and equity

Most common biases in medical AI

1. According to sex and gender
2. According to age differences
3. According to ethnic groups
4. According to geographic locations
5. According to socioeconomics

Most common causes of AI biases

1. Biased and imbalanced datasets
2. Structural biases and discrimination
3. Disparities in access to quality equipment and digital technologies
4. Lack of diversity and inter-disciplinarity in development teams

Mitigation measures

1. Systematic AI training with balanced, representative datasets
2. Interdisciplinary approaches in medical AI involving social scientists
3. Promotion of more diversity and inclusion in the field of medical AI

Despite continuous advances in medical research and healthcare delivery, there remain important inequalities and inequities in medical care within most countries around the world. The main factors that contribute to these inequalities and inequities include sex/gender, age, ethnicity, income, education and geography. While some of these inequities are systemic, such as due to socioeconomic differences and discrimination, human biases also play an important role. For example, in the United States, existing research has demonstrated that doctors do not take Black patients' complaints of pain as seriously nor do they respond to them as quickly as they do for their White counterparts (Hoffman et al., 2016). Persistent in most countries around the world, to varying degrees, is yet another example of common bias embedded in healthcare systems: gender-based discrimination. Once again, in the domain of pain management, studies have pointed to the increased psychologisation or invisibilisation of female patients when reporting pain (Samulowitz et al., 2018).

Hence, in the recent years, there have been concerns that, if not properly implemented, evaluated and regulated, future AI solutions could embed and even amplify the systemic disparities and human biases that contribute to healthcare inequities. A few examples of algorithmic biases have already made the headlines in recent years, some of which are detailed below.

A study published in *Science* in 2019 showed that an algorithm used in the United States to help in the referral process of patients who need extra or specialist care was shown to discriminate against Black patients (Obermeyer et al., 2019). The authors of the study explained that with the algorithm, *'at a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%'.* A Canadian study in 2020 evaluated the degree of fairness of state-of-the-art deep learning algorithms used to detect abnormalities such as fractures, lung lesions, nodules, pneumonia, etc. in chest X-ray images (Seyyed-Kalantari et al., 2020). The study showed that the highest rate of underdiagnosis was in young females (age: 0-20), in Black patients, and in patients on public health insurance for low-income people and households. Furthermore, patients with intersectional identities (for example, a Hispanic female patient on low-

income health insurance) suffered the highest rates of underdiagnosis. The authors concluded that '*models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modification*' (Seyyed-Kalantari et al., 2020).

It is widely argued that the most common cause for unfairness in medical AI is the bias in the data used to train the machine learning models. As Marzyeh Ghassemi from the University of Toronto stated in a recent presentation on AI in healthcare (Ghassemi, 2021): '*Bias is already part of the clinical landscape. So, it is not as if machine learning is out to get us. It is that when we are training on data that humans make, that humans label, that humans annotate, we might pick up on some of the biases that humans have injected into that data*'.

As an example, in 2002 the National Lung Screening Trial, which compiled datasets from 53,000 smokers to investigate methods for early diagnosis of lung cancer, was found to include only 4% of Black participants in the data (Ferryman & Pitcan, 2018). Machine learning algorithms for skin cancer detection have been all-too-often trained on highly biased datasets – such as the International Skin Imaging Collaboration, one of the most widely used open-access database of skin lesions – which contain images from mostly fair-skinned patients in the United States, Europe, and Australia (Adamson & Smith, 2018). Diagnostic models only trained on fair-skin groups could prove to be detrimental to the diagnostic process of melanoma lesions present on dark-skinned individuals. Similarly, the way COVID-19 appears to affect patients differently according to their sex group means an AI algorithm trained on existing clinical data is likely to suffer from reduced fairness when predicting severity and mortality in men and women (Jin et al., 2020).

Another type of bias that appears in datasets is of a geographic nature. In 2020, researchers from the fields of radiology and biomedical research at Stanford University conducted a review of articles published over a five-year period that had been used in training deep learning algorithms related to patient care (Kaushal et al., 2020). They found that 71% of the United States studies in which geographic location was identified used data only from California, Massachusetts, and New York. In addition, they found the studies did not include any data from 34 of the 50 states in the U.S. Geographic bias can be an important issue in Europe too, as data availability and access to digital equipment are unevenly distributed, particularly in the Eastern European regions (EGA Consortium, 2021).

Another potential source of lack of fairness in medical AI is bias in the data labelling during clinical assessment. For example, existing research has shown that due to gender stereotypes, women are over-diagnosed for some diseases such as depression and under-diagnosed for other diseases such as cancer (Dusenberry, 2018). Furthermore, a large-scale Danish study, which analysed data on hospital admissions for approximately 7 million citizens and 19 disease groups, found that for the vast majority of the diseases, women are diagnosed later than men (Westergaard et al., 2019). Importantly, for many of these medical conditions such as injury, poisoning, congenital malformations and infectious diseases, these discrepancies cannot be explained by anatomical or genetic differences. If the data labels in the health registries are affected by such healthcare disparities, such as in environments where given groups have been systematically misdiagnosed due to stigma or stereotypes, then the AI models will likely learn to perpetuate this disparity (Rajkomar et al., 2018).

In recent years, awareness of algorithmic bias has increased and researchers, particularly in North America, have started to investigate mitigation measures to address the risk of unfairness in medical AI. First, it is evident that AI developers, in collaboration with clinical experts and healthcare professionals, must pay close and continuous attention to the selection and labelling of the data and variables to be used during model training. These should be representative and balanced with respect to key attributes such as sex/gender, age, socioeconomics, ethnicity, as well as geographic location. Furthermore, it is recommended to involve not only data scientists and biomedical researchers in the development teams, but also social scientists, biomedical ethicists, public health

experts, as well as patients and citizens. The latter group must be as diverse as possible to ensure that adequate diversity of backgrounds, experiences and needs are taken into consideration during the AI production lifecycle and that the tools created are truly representative and founded on community-based research.

3.4. Lack of transparency

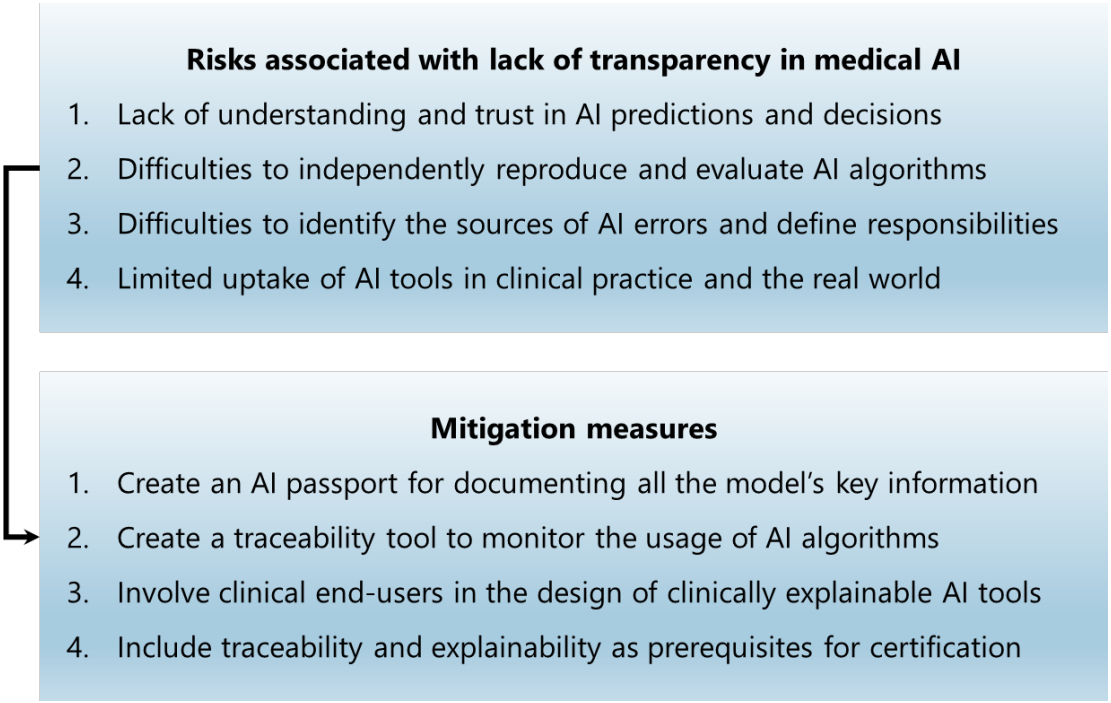
Despite continuous advances in medical AI, existing algorithms continue to be viewed by individuals and experts alike as complex and obscure technologies, which are difficult to fully comprehend, trust and adopt.

A recent AI algorithm developed by Google for breast cancer screening received considerable attention for its promising performance (McKinney , 2020): It was shown to improve the speed and robustness of breast cancer screening, to generalise well to populations in multiple countries beyond those used for training, and it even outperformed radiologists in specific situations. However, this work also received some criticism in the media and in the AI community as it was presented with almost no details on how the algorithm was built and on key technical descriptions. Some critics questioned the usefulness and safety of such an AI tool (Wiggers, 2020; iNews, 2020), while a group of scientists used this algorithm as their central example when they published a call in Nature for more transparency in medical AI (Haibe-Kains et al., 2020).

Lack of transparency is widely regarded as an important issue in the development and use of current AI tools in healthcare (Figure 6). It is expected to result in a great lack of trustworthiness in AI especially in sensitive areas such as medicine and healthcare that are focused on the wellbeing and health of citizens. At the same time, a lack of trustworthiness will evidently impact the level of adoption of emerging AI algorithms by patients, clinicians, and healthcare systems.

AI transparency is closely linked to the concepts of traceability and explainability, which correspond to two distinct levels at which transparency is required, i.e. (1) transparency of the AI development and usage processes (traceability), and (2) transparency of the AI decisions (explainability).

Figure 6 – Main risks resulting from the current lack of transparency associated with AI algorithms followed by possible mitigation measures



Traceability is considered a key requirement for trustworthy AI, and refers to transparently documenting the whole AI development process, including tracking how the AI model functions in real-world practice after deployment (Mora-Cantallops et al., 2021). More specifically, traceability requires maintaining a complete account of (i) model details (intended use, type of algorithm or neural network, hyper-parameters, as well as pre- and post-processing steps), (ii) training and validation data (gathering process, data composition, acquisition protocols and data labelling) and (iii) AI tool monitoring (performance metrics, failures, periodic evaluations) (EU Regulation, 2017; FDA, 2019).

In practice, existing AI tools in healthcare are rarely delivered with full traceability. In fact, companies often prefer not to disclose too much information about their algorithms, which are thus delivered as opaque tools that are difficult to understand and examine by independent parties. This, in turn, reduces the level of trust and adoption into real-world practice.

While traceability addresses the transparency of the AI algorithm's lifecycle, AI explainability is important for providing transparency for each AI prediction and decision. Article 22 of the European Union's General Data Protection Regulation (GDPR) details the 'right to explanation' which requires an explanation to be offered regarding the automated decision-making process (Selbst & Powles, 2017).

However, AI solutions, and specifically deep neural networks lack transparency, and are often described as 'black box AI', referring to the fact that these models learn complex functions that humans struggle to understand (Yang et al., 2021) and whose functions and decision-making processes are not visible or understandable. A lack of transparency makes it difficult for clinicians and other stakeholders to incorporate AI solutions into their real-world practice because in order to work with specific AI solutions, clinicians need to be able to understand the fundamental principles behind each decision and/or prediction, even when the algorithm itself has the potential to enhance the clinician's productivity (Lipton, 2017). Furthermore, the lack of explainability means that it is difficult to identify the source of AI errors and define responsibilities when it goes wrong.

There are numerous avenues available to improve the transparency of AI technologies in healthcare. First of all, there is a need for an 'AI passport' that could be a requirement for each AI algorithm for documenting all the model's key information. There is also a need to develop traceability tools for monitoring the usage of AI algorithms once they are deployed, such as to record potential errors and performance degradation, as well as to perform periodic audits. To improve the explainability of AI algorithms, it is important that AI developers involve clinical end-users from the start of the development process in order to select the best explainability approach for each application and to ensure that the chosen explanations are useful and well accepted in clinical practice. Finally, regulatory entities can play an important role by considering the traceability and explainability of the AI tools as pre-requisites for certification.

3.5. Privacy and security issues

The increasingly widespread development of AI solutions and technology in healthcare, recently highlighted by the COVID-19 pandemic, has shown potential risks for a lack of data privacy, confidentiality and protection for patients and citizens. This could lead to serious consequences (Figure 7), such as the exposure and use of sensitive data which goes against the rights of the citizens or the repurposing of patient data for non-medical gains.

Figure 7 – Main privacy and security risks associated with big data and AI, and some mitigation measures

Risks for privacy and security in big data and AI

1. Risk of personal data being shared and used without informed consent
2. Risk of data re-purposing, without the patient's knowledge
3. Risk of data being exposed, resulting in identify theft or other frauds
4. Risk of harmful and potentially fatal cyberattacks on AI solutions

Mitigation measures

1. Increased awareness of data privacy, consent and cybersecurity
2. Regulations to address accountability and protect citizens
3. Federated approaches for de-centralised privacy-preserving medical AI
4. Continuous research to protect AI algorithms against cyberattacks

These issues are firstly linked to informed consent, i.e., the provision of adequate information for the patients for an informed decision such as for sharing personal health data. Informed consent is a crucial and integral part to the patient's experience in healthcare, which was formalised in the Helsinki Declaration and has since grown as the introduction of digital technology has permeated our daily lives (Pickering, 2021). Informed consent is linked to various ethical issues, including protection from harm, respect for autonomy, privacy protection and property rights concerning data and/or tissue (Ploug & Holm, 2016).

However, the introduction of opaque AI algorithms and complicated informed consent forms limits the level of autonomy and the power of shared patient-physician decision making (Vyas et al., 2020). It has become increasingly difficult for patients to understand the decision-making process and the different ways in which their data can be reused, and to know exactly how they can choose to opt out of sharing their data. Issues of informed consent are also especially prominent in big data research, especially digital platform-based health data research, in which a patient may not be fully aware of or fully understand the extent to which their data is shared and reused (McKeown et al., 2021).

An important example of this occurred in 2016, when records of 1.6 million patients in the United Kingdom were transferred – without patients' informed consent – from the Royal Free NHS Foundation Trust to the Google-owned AI company DeepMind, which at the time was working on developing an app to implement new ways of detecting kidney disease (BBC, 2017). In July 2017, the UK Information Commissioner's Office (ICO) ruled that the Royal Free NHS Trust had breached data protection laws; the Information Commissioner office was famously quoted as saying, 'the price of innovation does not need to be the erosion of fundamental privacy rights' (Gerke et al., 2020).

The use of AI in healthcare also entails a risk of data security breaches, in which personal information may be made widely available, infringing on citizens' rights to privacy and putting them at risk for identity theft and other types of cyberattacks. In July 2020, the New York based AI company Cense AI suffered a data breach that exposed highly sensitive data of upwards of 2.5 million patients who had suffered from car accidents, including such detailed information as names, addresses,

diagnostic notes, dates and types of accident, insurance policy numbers and more (HIPPA Journal, 2020). Although eventually secured, this data was briefly accessible to anyone in the world with an internet connection, underlining the very real danger of personal privacy breaches that patients are exposed to.

Another persistent concern is that of data repurposing, which in certain contexts is also referred to as 'function creep' (Koops, 2021). The World Health Organization has warned against the danger of function creep during the COVID-19 pandemic, highlighting a case in Singapore in which the data from the government's COVID-19 tracing applications was also made available for criminal investigations (WHO, 2021). This is a stark example of health-related data being repurposed for non-healthcare related ends, but repurposing can also occur within the healthcare sphere itself. A 2019 report explored in detail the different ways that patient data is repurposed in the European pharmaceutical industry: Data from electronic health records, registry data and data from health systems are used for pharmaceutical drug development, clinical trial design, marketing and cost-effectiveness analyses, and more (Hocking et al., 2019).

In addition to the issues related to data privacy and security, AI tools are especially vulnerable to cyberattacks, the results of which could be anything from burdensome to fatal, depending on the context. In September 2020, a patient died after having to be redirected to another hospital when the Düsseldorf University Hospital suffered a cyberattack that interfered with the hospital's data and rendered the centre's computer system inoperable (Kiener, 2020). Although it was later argued that it could not be proven that the death was directly caused by the cyberattack, because the patient was already suffering a life-threatening condition, this case brought to the forefront the real physical harms that cyberattacks can cause in the healthcare sphere.

In another example of how technological breaches may affect the physical health of patients, in April 2021 the Swedish oncology software company Elekta suffered a healthcare ransomware attack that affected 170 health systems in the United States, delaying cancer treatment care to patients across the country as well as exposing sensitive patient data (Mulcahy, 2021).

Furthermore, research has shown that personal medical devices controlled by AI are also vulnerable to attacks. For example, researchers discovered that AI-powered insulin pumps for diabetes patients could be hacked and remotely controlled from varying distances, and could even be manipulated to flood the patient's body with excessive insulin (Wired, 2019). While this hack has never been carried out in the real world, researchers' development of the AI attack exposed serious vulnerabilities in the AI system's functionality.

These events garnered enough attention to bring to light the question of how algorithmic security – or lack thereof – can affect human survival in a high-stakes context such as healthcare. Focusing on AI tools as part of the larger technological sphere, it is clear that risks of attacks and hacking must be continually monitored.

To address these important issues, there is a need to increase awareness and literacy on privacy and security risks, as well as on informed consent and cybersecurity. Furthermore, regulations and legal frameworks must be extended to address not only privacy but also accountability, and to protect citizens from data breaches and data repurposing. Decentralised, federated approaches to AI should be promoted to leverage the power of big data from clinical centres without the need for unsafe data transfers. Research must be continued and accelerated to improve security in cloud-based systems and to protect AI algorithms against cyberattacks.

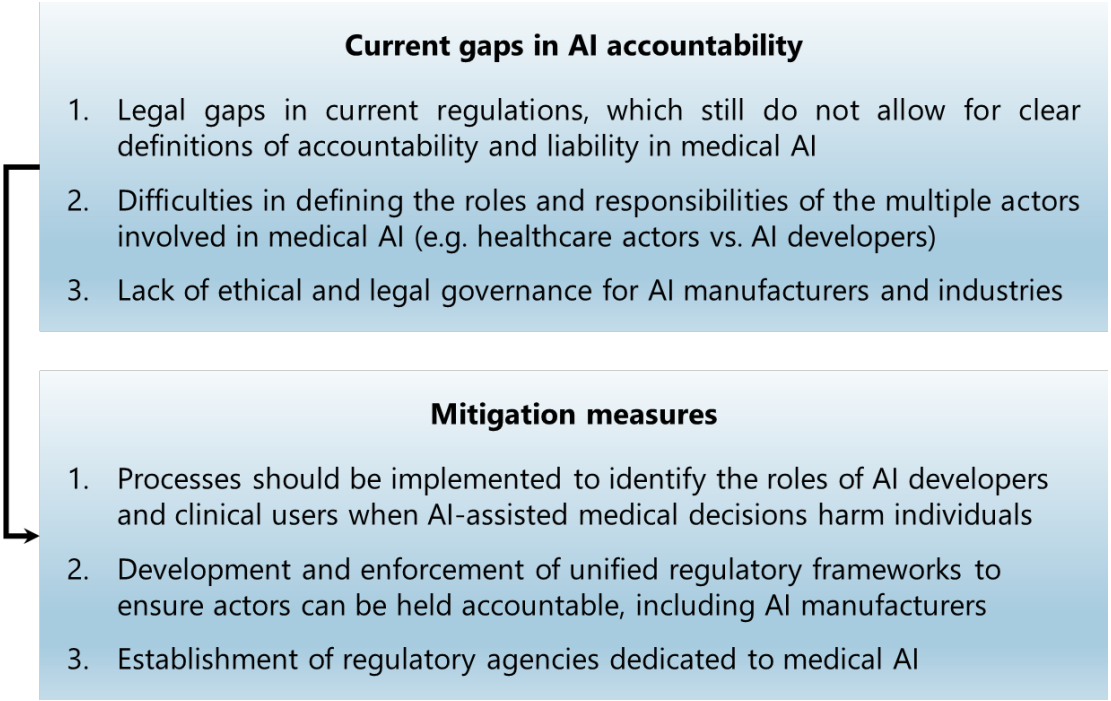
3.6. Gaps in AI accountability

The term 'algorithmic accountability' has garnered increasing importance among researchers and organisations dedicated to addressing the legal impact of the introduction and use of AI algorithms

in different areas of human life. Although the term 'algorithmic accountability' might appear to refer to the task of seeking to hold the algorithm itself accountable, it is actually quite the opposite: It emphasises the fact that algorithms are created through a mixture of machine learning and human design, and that the mistakes or wrongdoings in algorithms come from the humans developing, introducing or using the machines (Kaplan et al., 2018), especially since AI systems themselves cannot be held morally or legally responsible (Raji, 2020).

Accountability is particularly important for medical AI as it will contribute to its acceptability, trustworthiness and future adoption in society and healthcare. For example, clinicians that feel that they are systematically held responsible for all AI-related medical errors – even when the algorithms are designed by other individuals or companies – are unlikely to adopt these emerging AI solutions in their day-to-day practice. Similarly, citizens and patients will lose trust if it appears to them that none of the developers or users of the AI tools can be held accountable for the harm that may be caused. There is a need for new mechanisms and frameworks to ensure adequate accountability in medical AI and to manage reclamations, compensations and sanctions where necessary, as well as to guarantee non-repetition of the acts (WHO, 2021).

Figure 8 – Current limitations in accountability and recommendations to fill in these gaps



Due to the novelty of medical AI and the lack of legal precedence, there is currently a major lack of clarity regarding the definition of responsibilities for AI-related medical errors that could lead to patient harm (Figure 8). The quickly changing and growing field of medical AI poses new challenges for regulators, policymakers and legislators. It pushes current regulations, policies, and laws to adapt their traditional ways of considering responsibility and liability to the new reality of AI-assisted healthcare.

Challenges in applying current law and liability principles to emerging AI applications in medicine include (1) the multi-actor problem in medical AI, which makes it difficult to identify responsibilities among the multiple players involved in the development, implementation and use of medical AI and algorithms (e.g. AI developers, data managers, clinicians, patients, healthcare organisers, etc.); (2) the difficulty in identifying the precise cause of any AI-related medical error, which can be due to the AI algorithm, the data used for training it, or its incorrect use and understanding in clinical

practice; and (3) the multiplicity of governance frameworks and the lack of unified ethical and legal standards in AI industries.

While historically the relationship between the patient and the clinician has stood at the centre of issues concerning medical malpractice and negligence, the introduction of AI tools into healthcare adds a new layer with multiple actors into the patient–physician dynamic (Smith, 2020). These actors may include not only the patient, clinician, healthcare centre, and healthcare system, but also AI developers, researchers, and manufacturers, all of whom are now in some way or another entering into the medical decision-making process. The presence of all these new actors and the lack of clarity – not only on who is responsible for which part of the decision-making process, but also on how the AI tools themselves work – contributes to the complexity of the situation.

While medical professionals are usually under a regulatory responsibility to be able to account for their actions, a requirement that forms an integral part of their professional undertaking, AI developers and technologists generally work under ethical codes (Whitby, 2015). Therefore, for medical professionals the repercussions for not being able to account for their actions and decision-making processes could mean losing their licence to practice medicine; while under the current practice, a lack of accountability for a technologist could mean something much less devastating. Even if an AI manufacturer is found to be responsible for an error, it is often difficult to place blame on one specific person, since so many different developers and researchers work on any given AI system. In addition, the ethical codes and standards of accountability that many private entities use have often been criticised for being vague and difficult to translate into enforceable practice (Raji, 2020).

It is important to note that that the issues of AI accountability and liability in the realm of medicine and healthcare are closely linked to the questions of explainability and transparency. The opaquer an AI algorithm is, the harder it is to find who is accountable for an error involving a patient or a medical decision, and so the burden of responsibility will likely fall more heavily on the clinician who used a non-transparent medical AI tool and is unable to explain their medical decision or the error that occurred (Maliha et al., 2021). This is especially true for assistive AI tools, which are meant to assist the clinician in their decision-making process and may be considered the equivalent of consulting an expert clinical colleague (Harned et al., 2019).

There are avenues to address the current lack of accountability in medical AI. First, processes should be established to identify the roles of AI developers and clinical users when AI-assisted medical decisions harm individuals. There is also a need to establish regulatory agencies dedicated to medical AI. These will develop and enforce regulatory frameworks to ensure specific actors of medical AI can be held accountable, including AI manufacturers.

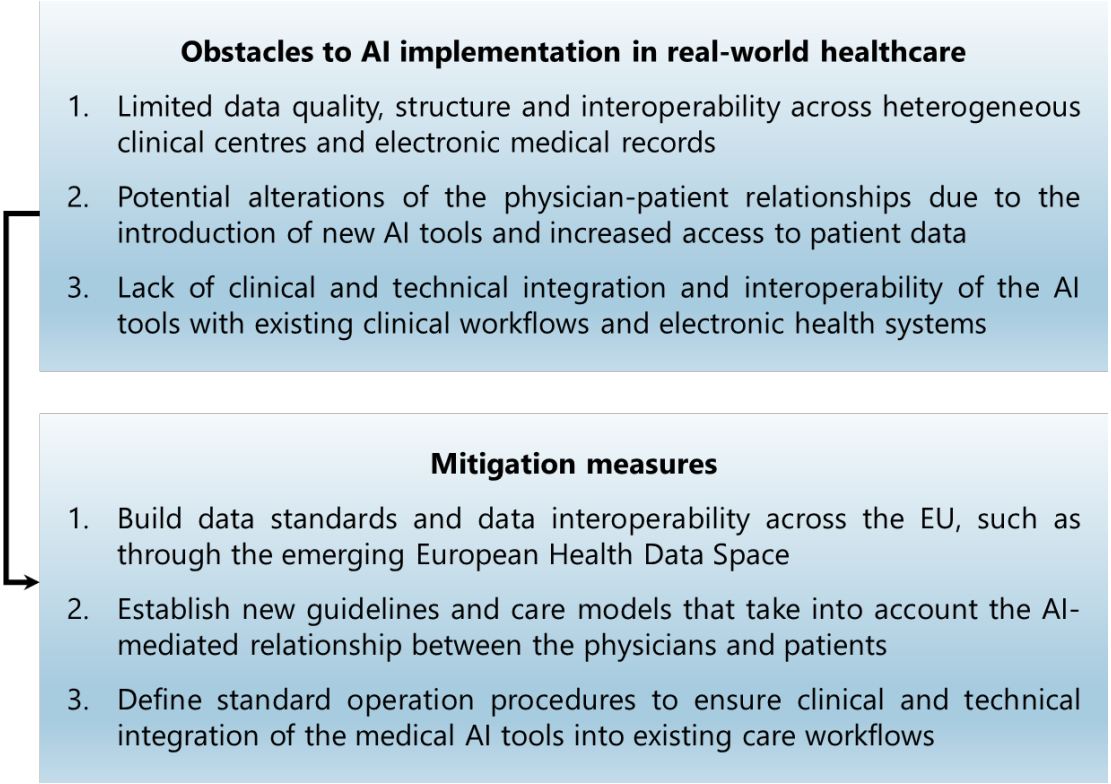
3.7. Obstacles to implementation in real-world healthcare

A large number of medical AI algorithms have been developed and proposed over the last five years, in a wide range of medical applications, as summarised in section 2. However, even when medical AI technologies are well validated and found to be clinically robust and safe, as well as ethically sound and compliant, the road to healthcare implementation, integration and adoption is still laden with specific obstacles in the real world (Shortliffe & Sepúlveda, 2018; Fihn et al., 2019; Nagendran et al., 2020).

Healthcare professionals have traditionally lagged behind other professionals with regards to the adoption of new technologies in their daily activity (Quaglio, 2018). Past experiences in healthcare show that the implementation period is a key stage in the innovation process. In practice, it is not enough to invent and test a new AI technology; other factors which can hinder its implementation in real-world healthcare should also be considered (Arora, 2020), such as (1) the limited data structure and quality in existing electronic health systems, (2) the alteration of the clinician-patient

relationship, as well as (3) the difficulties related to clinical integration and interoperability (Figure 9).

Figure 9 – Obstacles for clinical implementation and integration of new AI tools in real-world healthcare practice, together with potential mitigation measures



First of all, the quality of electronic health data in real-world practice is key to facilitating the implementation of medical AI. However, medical data is notoriously unstructured and noisy, and most existing datasets are not exploitable in AI algorithms. Furthermore, the formats and quality of clinical data vary significantly between clinical centres as well between EU member states (Lehne et al., 2019). Before emerging medical AI tools could be fully implemented and used at large scale, existing data would require significant and costly human revision, quality control, cleaning and re-labelling. To improve data interoperability, the creation of a European Health Data Space was defined as one of the priorities of the European Commission 2019-2025 plan (European Health Data Space). This will promote better re-use of heterogeneous types of health data (electronic health records, genomics data, data from patient registries, etc) across EU countries, including by emerging AI algorithms.

Furthermore, AI technologies are expected to modify the relationship between patients and healthcare professionals in ways that are not yet completely predictable. Certain specialties, particularly those related to image analysis, have already undergone significant transformations due to AI (Gómez-González, 2020 The emergence of patient-centred AI technologies has the potential to transform the historically paternalistic clinician-patient relationship into a join partnership in the decision-making process due to increased transparency and deepened doctor-patient conversations (Aminololama & Lopez, 2019). However, personal and ethical implications of communicating information about AI-derived risks of developing an illness (such as predisposition to cancer or dementia) will need to be elucidated (Fihn et al., 2019; Cohen, 2020). The clinical guidelines and care models will need to be updated to consider the AI-mediated relationships between healthcare workers and patients.

Finally, clinicians and care providers work under established clinical guidelines and technical standards. The introduction of an AI technology into everyday practice will have practical, technical and clinical implications on both clinicians and patients. Secondly, it is not clear that medical AI tools will be systematically interoperable across clinical sites and health systems, and that they will be easily integrated within existing clinical and technical workflows (Meskó & Görög, 2020), without significant modifications to existing clinical practices, care models and even training programmes.

AI manufacturers, in collaboration with healthcare professionals and organisations, will need to establish standard operation procedures for all new AI tools to ensure their clinical interoperability across distinct clinical sites and their integration across heterogeneous electronic healthcare systems. In particular, new AI tools should be developed while ensuring their future integration and communication with already existing technologies, such as genetic sequencing, electronic patient records and e-health consultations (Arora, 2020).

4. Risk assessment methodology

Previous sections of this report have described the main risks that have emerged in recent years concerning the use of AI in healthcare. This calls for a structured approach of risk assessment and management that specifically addresses the technical, clinical and ethical challenges of AI in healthcare and medicine.

4.1. Regulatory frameworks for AI

AI risks can be characterised and classified according to the severity of the harm they may induce, as well as to the probability and frequency of the harm induced. In healthcare, AI risks vary greatly, from infrequent and/or low risks that induce limited and manageable harm to the patients and citizens, to frequent and/or high risks that may cause irreversible damage or harm. For example, an AI algorithm can affect the productivity of the clinicians (e.g. the AI tool fails to accurately delineate the boundaries of the heart in a cardiac image volume, which must be improved manually by the cardiologist), but they can also cause harm to the patient's health and seriously impact the clinical outcomes (e.g. the AI tool fails to diagnose a life-threating condition).

Hence, to minimise the risks of AI and to maximise its benefits in future healthcare, it is important to identify, analyse, understand and monitor the potential risks on a case-by-case basis for each new AI algorithm and application. An important step of the risk assessment procedure should be to devise a methodology for classifying the identified risks into a number of categories representing different levels and types of risk. For each level, a set of tests or regulations must be specified to mitigate and address the AI risks, such that the higher risk classes will require more testing and regulation, while lower risks will result in limited risk mitigation measures. Suitable risk classification of AI according to severity and likelihood will enable manufacturers, care providers and regulators to intervene as much as necessary to ensure the protection of the patients, as well as their rights and values; however, it is also important that these classifications do not –in as much as possible– serve to hamper innovation in healthcare AI.

Currently, the applicable regulations for medical AI tools in the EU are the 2017/745 Medical Devices Regulations (MDR) and the 2017/746 In Vitro Diagnostic Medical Devices Regulation (IVDR), which were established in 2017. The MDR applies to software as medical devices, including AI-based software, while the IVDR applies to in vitro based diagnostics, including AI-based. These regulations included new approaches for stricter pre-market control, increased clinical investigation requirements, reinforced surveillance across the device's lifecycle, and improved transparency by creating a European database of medical devices. However, many aspects specific to AI are not considered, such as continuous learning of the AI models or the identification of algorithmic biases. In particular, the fact that AI is a highly adaptive technology that continues to learn and adjust over time – as more data becomes available – calls for new approaches to monitor the risks of the AI software.

One of the first proposed for risk assessment in the field of AI came in 2018, when the German Data Ethics Commission proposed to classify risks of general decision-making algorithms according to their criticality, i.e., the system's potential to cause harm (German Data Ethics Commission, 2019). A 'criticality pyramid' comprising five levels of risk/criticality was proposed (1: Zero or negligible potential for harm; 2: Some potential for harm; 3: Regular or significant potential for harm; 4: Serious potential for harm; 5: Untenable potential for harm).

Under this proposal, an adapted testing or regulatory system is recommended depending on the risk level, which could include corrective and oversight mechanisms, specifications regarding the transparency of algorithmic systems and the explainability and comprehensibility of the results, or

rules on the assignment of responsibility and liability within the context of the development and use of algorithmic systems.

In 2021, the European Commission (EC) published a long-awaited proposal for AI regulation and for harmonising the rules that govern AI technologies across Europe, in a manner that addresses safety as well as human rights concerns (European Commission, 2021). In a similar fashion to the 2018 proposal of the German Data Ethics Commission, the draft EU framework provided a definition of AI that is risk-based, together with mandatory requirements for high-risk AI systems. Concretely, the document recommended to classify AI tools according to three main levels of risk: (i) unacceptable risk, (ii) high risk, and (iii) low or minimal risk.

The highest category corresponds to AI tools that contradict EU values and hence should be prohibited. The document (Title II, Article 5) provides some examples of such AI tools, e.g. subliminal manipulation resulting in physical/psychological harm; exploitation of vulnerabilities resulting in physical/psychological harm; social scoring; real-time biometric identification in public spaces (with few exceptions).

The intermediate category, and one of particular interest, corresponds to high-risk AI, which can be permitted only when the tools comply with specific requirements. Such high-risk AI tools (Title III, Chapter 1) comprise safety components of regulated products (including medical devices, but also other products such as toys and machinery), and certain stand-alone AI systems in areas such as operation of critical infrastructure, access to private services as well as employment and workers management. It appears that many medical AI tools, especially those that are autonomous, will be categorised as high-risk. The proposal provides concrete requirements and obligations for adequate risk management in high-risk AI, as listed in Box 1:

Box 1 – Requirements and obligations for high-risk AI tools according to the 2021 EC proposal

Requirements for high-risk AI:

- Use high-quality training, validation and testing data (relevant, representative).
- Draw up technical documentation & set up logging capabilities (traceability & auditability).
- Ensure appropriate degree of transparency and provide users with information on capabilities and limitations of the system & how to use it.
- Ensure human oversight (measures built into the system and/or to be implemented by users).
- Ensure robustness, accuracy and cybersecurity.

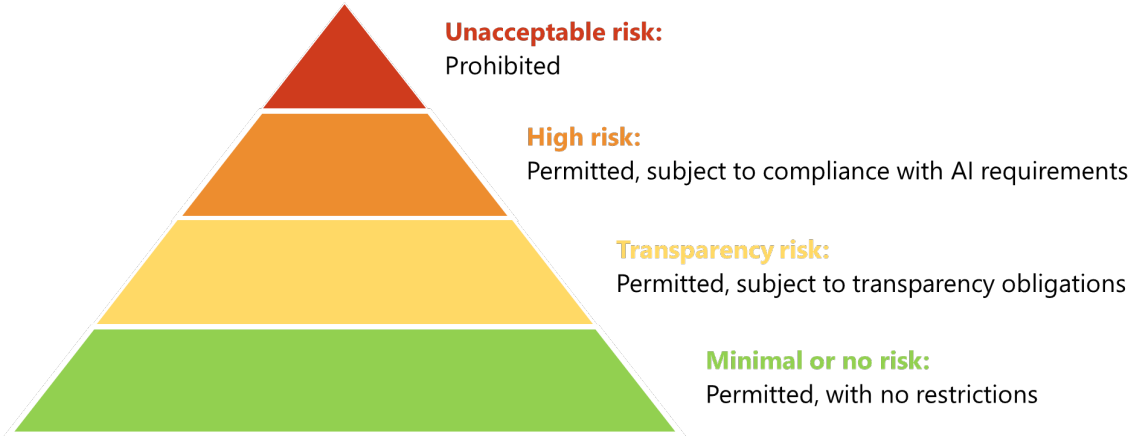
Obligations:

- Establish and implement quality management system in its organisation.
- Draw-up and keep up to date technical documentation.
- Undergo conformity assessment and potentially reassessment of the system (in case of substantial modification).
- Register AI system in EU database.
- Affix CE marking and sign declaration of conformity.
- Conduct post-market monitoring.
- Collaborate with market surveillance authorities.
- Inform the provider or distributor about any serious incident or any malfunctioning.
- Continue to apply existing legal obligations (e.g. under GDPR).

The lowest category refers to AI tools with minimal risk, which have no mandatory obligations but the EC encourages drawing up codes of conduct, as well as voluntary application of requirements for high-risk AI systems or other requirements (Article 69).

In addition to these three categories of risks (unacceptable, high and low), the document (Article 52) discusses an additional category of AI systems, such as those that interact with individuals or expose them to emotional or biometric recognition, for which there is an explicit obligation of transparency. In this case, the individuals must be notified that they are interacting with an AI system (Figure 10).

Figure 10 – AI risk classification according to the 2021 EU proposal on AI legislation



The draft AI regulation does not specifically address AI in healthcare, but it suggests in its current form that AI-driven medical devices will be classified as high-risk, because of the associated safety and privacy concerns. This means future medical AI tools should fulfil all the requirements already established by the Medical Device Regulation, but also those listed in Chapter II of the AI regulation (use of high quality and representative data, technical documentation and traceability, transparency requirement, human oversight, quality management system, conformity assessment, etc).

However, one can argue that not all medical AI tools are systematically high risk. For example, many AI tools have been developed in radiology to accelerate the contouring of organs and lesions on medical images, before quantification and diagnosis (e.g. contouring of the boundaries of the cardiac ventricles or contouring the boundaries of lung tumours). Such AI-powered processing tools are very important and in fact already in use in clinical practice, but they do not necessarily require to be transparent as the clinicians can visually assess the results of the automatic contouring and correct any errors, so the risks are minimal. To continue to promote innovations and investments in medical AI, mechanisms may be needed to discriminate between low- and high-risk AI in healthcare.

With this new regulatory framework, CE marking and regulatory approval in medical AI can take the following form:

- Determine whether the AI tool is classified as high risk under the new AI regulation.
- Ensure AI design, development and quality management systems are in compliance with the AI regulation.
- Undergo conformity assessment procedure to assess and demonstrate compliance.
- Affix the CE marking to the system and sign a declaration of conformity.
- Implement the AI tool in practice or deploy to the market.

It is important to note that the EC proposal for AI regulation is general for all domains of society: it does not take into account the specificities and risks of AI in the healthcare domain. Furthermore, the EC proposal retains of some of the limitations of the MDR and IVDR, such as the lack of

mechanisms to address the dynamic nature of AI technologies. Currently, continuous learning, which is key to medical AI technologies, may be considered as a substantial modification and would require reassessment of the AI technology.

4.2. Risk minimisation through risk self-assessment

For risk identification in AI, several stakeholders have suggested a self-assessment structured approach composed of specified checklists and questions. For example, the independent High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, published an assessment checklist for trustworthy AI called ALTAI. The checklist is structured along seven categories: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability (ALTAI, 2020). In Box 2, some examples of self-assessment questions that were proposed as means to identify potential limitations are provided for reliability, privacy, explainability and fairness:

Box 2 – Examples of self-assessment questions from the ALTAI checklist (ALTAI, 2020)

For reliability:

- Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?
- Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?
- Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?
- Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility?
- Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?
- Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?
- Is your AI system using (online) continual learning?

For data privacy:

- Did you put in place any of the following measures, some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?
  - Data Protection Impact Assessment (DPIA);
  - Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system;
  - Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation);
  - Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system?

For explainability:

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?

For fairness assessment:

- Did you consider diversity and representativeness of end-users and/or subjects in the data?
- Did you test for specific target groups or problematic use cases?
- Did you research and use publicly available technical tools, that are state-of the-art, to improve your understanding of the data, model and performance?
- Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?

The full assessment checklist and questions for all categories can be found online at the Publications Office of the European Union (ALTAI, 2020). It is also available as an online tool for registered users. It is important to note that the list was devised for AI in general and must be tailored to each specific application domain, including healthcare.

To our knowledge, the first self-assessment checklist for AI in healthcare was published by a multi-disciplinary team of researchers from Australia in 2021. Its objective was to help clinicians assess how ready algorithms are for use in routine care and to pinpoint the areas in which further development and finetuning may be necessary before deployment (Scott et al., 2021). This list was put together based on a few narrative reviews on AI in healthcare, which were summarised into a set of assessment questions organised into 10 general questions as listed in Box 3.

Box 3 – Questions from the assessment checklist for medical AI tools, as shown in Scott et al., 2021

- What is the purpose and context of the algorithm?
- How good were the data used to train the algorithm?
- Were there sufficient data to train the algorithm?
- How well does the algorithm perform?
- Is the algorithm transferable to new clinical settings?
- Are the outputs of the algorithm clinically intelligible?
- How will this algorithm fit into and complement current workflows?
- Has use of the algorithm been shown to improve patient care and outcomes?
- Could the algorithm cause patient harm?
- Does use of the algorithm raise ethical, legal or social concerns?

However, this self-assessment list does not contain the same level of detail as the assessment checklist for general AI devised by the AI HLEG. For example, point 10 in Box 3 is rather vague and does enable to pinpoint the exact ethical, legal or social concern (e.g. algorithmic bias). It seems that a combination of both approaches would lead to a detailed and standardised risk assessment checklist for AI in healthcare, generated through consensus and with each category of risk enriched with a detailed set of assessment questions.

This has motivated the recent development of consensus guidelines for trustworthy AI in medicine by a network of EC-funded research projects together with international inter-disciplinary experts. Entitled FUTURE-AI (www.future-ai.eu), these guidelines are organised according to six principles (Fairness, Universality, Traceability, Usability, Robustness, Explainability) and comprise concrete recommendations and a self-assessment checklist to enable AI designers, developers, evaluators and regulators to develop trustworthy and ethical AI solutions in medicine and healthcare (Lekadir et al., 2022). Box 4 lists examples of risk assessment questions included in the FUTURE-AI self-assessment checklist.

Box 4 – Excerpts of risk assessment items from the FUTURE-AI guidelines for trustworthy AI in medicine (version from 27 February 2022)

Fairness:

- Did you design your AI algorithm with a diverse team of stakeholders? Did you collect requirements from a diverse set of end-users?
- Did you define fairness for your specific AI application? Did you ask clinicians about hidden sources of data imbalance?
- Did you thoroughly evaluate the fairness of your AI algorithm? Did you use a suitable dataset and dedicated metrics?

Universality:

- Did you annotate your dataset in an objective, reproducible and standardised way?
- Did you use universal, transparent, comparable, and reproducible criteria and metrics for your model's performance assessment?
- Did you evaluate your model on at least one open-access benchmark dataset that is representative of your model's task and expected real-world data exposure after deployment?

Traceability:

- Did you prepare a complete documentation of the datasets you used? Did you include the relevant metadata?
- Did you keep track, in a structured manner, of the whole pre-processing pipeline of input data? Did you specify input/output, nature, prerequisites and requirements of your pre-processing and data preparation methods?
- Did you record the details of the training process? Did you include a careful description of input predictors?

Usability:

- Did you engage users in the design and development of the AI tool?
- Did you evaluate the usability of your tool after integration in the clinical workflows of the clinical sites?

Robustness:

- Did you train and evaluate your tools with heterogeneous datasets from multiple clinical centres and data protocols?
- Did you evaluate the AI tool under diverse real-world scenarios?
- Did you use any quality control mechanisms to identify potential deviations or artifacts in the input data?

Explainability:

- Did you consult with the clinicians to determine which explainability methods suit them?
- Did you use some quantitative evaluation tests to determine if the explanations are robust and trustworthy? Did you perform some qualitative evaluation tests with clinicians?

The need to further tailor AI risk assessment to specific medical domains have also been stated. For example, in the field radiology, various prominent European and North American radiological associations (American College of Radiology, European Society of Radiology, Radiological Society of North America, Society for Imaging Informatics in Medicine, European Society of Medical Imaging Informatics, Canadian Association of Radiologists, and the American Association of Physicists in Medicine) came together to release a statement on the ethical challenges of using AI in radiology.

They stated that 'the radiology community should start now to develop codes of ethics and practice for AI which promote any use that helps patients and the common good' (Geis et al, 2019).

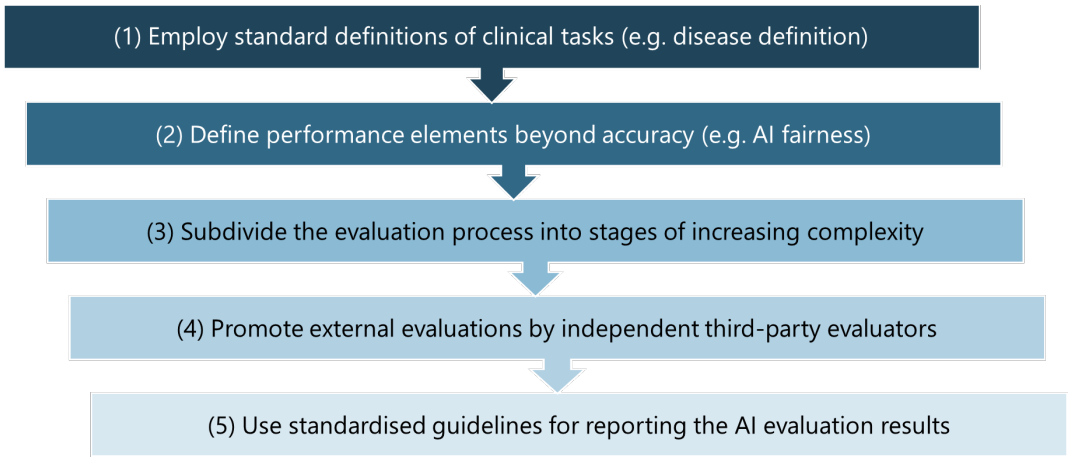
The assessment checklists presented in this section have used different categories of risks, as well as different assessment questions. Standardising, adjusting and validating these approaches through consensus by professional societies and independent groups on a domain-by-domain basis (e.g. radiology vs. surgery) would result in more robust processes for risk identification and management. Furthermore, as more and more healthcare AI algorithms will undergo self-assessment for ethical, legal and technical risks, these checklists should be regularly refined and updated versions will be released for the community taking into account continuous developments in AI methods, processes and regulations.

4.3. Risk identification through comprehensive, multi-faceted clinical evaluation of AI solutions

To identify, anticipate and manage risks in medical AI, adequate procedures for evaluating the AI models are of central importance. Thus far, AI evaluation has been achieved mostly by examining model accuracy and robustness in laboratory settings. Other aspects of AI, such as clinical safety and effectiveness, fairness and non-discrimination, transparency and traceability, as well as privacy and security, are more challenging to evaluate in controlled environments and have received less attention in the scientific literature.

Given the existing gaps, the US Food and Drug Administration (FDA) proposed action plan in 2021 to better regulate and advance the agency's oversight of medical AI software, which promoted '*regulatory science efforts to develop methodology for the evaluation and improvement of machine learning algorithms*' (FDA, 2021). In parallel, several research teams have also investigated and proposed new approaches for improved and comprehensive evaluation of medical AI algorithms, especially in North America (Larson et al., 2021; Park et al., 2020), Europe, and Asia (Park & Han, 2018), as well as by international societies such as the International Association of Medical Informatics (Magrabi et al., 2019). In this section, we will summarise their findings into a set of five main recommendations to enable a multi-faceted and comprehensive evaluation of future AI software in healthcare, as outlined in Figure 11.

Figure 11 – Recommendations for improved evaluation of algorithm performance and risks in medical AI



4.3.1. Standardised definition of clinical tasks

To enable objective and comparative evaluation of medical AI solutions, researchers at Stanford University have recently proposed to standardise the definition of the clinical tasks that the AI algorithms are addressing (Larson et al., 2021). In practice, there are many ways to define a clinical task, such as medical diagnostics. As an illustration, the diagnosis and reporting of COVID-19 severity based on chest imaging scans has been proposed using different schemes (Larson et al., 2021), including:

- Two categories: Radiologist's labelling of presence or absence of the disease.
- Four categories proposed by the Radiological Society of North America (RSNA) (Simpson et al., 2020): (1) typical, (2) indeterminate, (3) atypical appearance, and (4) negative for pneumonia.
- Six categories based on the CO-RADS scale (Prokop et al., 2020): (1) negative, (2) low, (3) indeterminate, (4) high, (5) very high, (6) PCR +.
- Various scoring systems of lesion severity in the lungs, such as (i) a 0 to 4 severity rating for each of six lung zones, for a total score of 0 to 24, (ii) a 0 to 5 severity rating for each of five lung lobes, for a total score of 0 to 25, (iii) a 0 to 7 severity rating for each of five lung lobes, for a total score of 35.

Any of these diagnostic systems could be incorporated into an AI-based algorithm, which makes objective assessment of the algorithm's performance and associated risks more difficult. This also limits the ability to directly compare AI-based algorithms that are originally developed for the same clinical task, given the existence of multiple definitions. To date, clinical task definitions have typically been developed with relatively little oversight and coordination. As these clinical tasks will be increasingly performed based on AI algorithms developed by non-clinical developers, it is important that the definitions, which form part of the AI software specifications, should be developed according to accepted consensus-based standard-setting principles and maintained by nonconflicted entities committed to updating the definitions based on new evidence and input from relevant stakeholders. Medical societies, such as the European Society of Cardiology, the European Society of Radiology, or the European Society for Medical Oncology, could play an important role in standardising the definition of the clinical tasks for medical AI in their respective fields. With this approach, the responsibility of the developers will be limited to optimising the performance of the AI algorithms based on widely accepted and utilised reference diagnostic task definitions, which would help ensure widespread acceptance of AI solutions by relevant stakeholders.

4.3.2. Multi-faceted evaluation of performance beyond accuracy

Given the multiple risks and ethical considerations of medical AI, it is now widely accepted that the evaluation of the algorithms must be extended well beyond existing approaches that have mostly focused on model accuracy. While the empirical evaluation of machine learning algorithms remains a matter of on-going debate among researchers, there is a need for the development of specific performance domains for AI in healthcare. Table 2 shows some examples of performance elements recently proposed for AI-based diagnostic algorithms in radiology (Larson et al., 2021). These include classification accuracy, but also reliability, applicability, transparency, monitorability, usability and more (see Table 2).

Table 2 – Examples of performance elements for imaging AI algorithms (from Larson, et. al., 2021)

Accurate	The algorithm should accurately perform all diagnostic tasks for which it is designed.
Reliable	The algorithm should remain accurate in the setting of reasonably expected variation encountered in the clinical environment, including reasonable variations in image quality.
Applicable	The accuracy of the algorithm should be maintained across all makes and models of image modalities and for all patient populations for which it is designed to function.
Deterministic	The algorithm should give the same answer for the same image when used at different times and in different settings.
Non-distractible	The algorithm should be able to recognise the salient information from the image and not change its assessment based on extraneous, non-contributory image data.
Self-aware of limitations	The algorithm should have the means to detect when it is at or beyond the boundaries of its capabilities, whether due to inherent limitations of the model, limitations of its clinical applicability, or limitations imposed by clinical variation such as unexpected patient anatomy or image quality.
Fail-safe	The algorithm should recognise when it has reached an erroneous conclusion and have the means for ensuring that all errors are caught and stopped before they are propagated into the clinical environment
Transparent logic	The user interface should enable the operator to clearly see the linkage between the input and output, including what data were analysed, what alternatives were considered, and why certain possibilities were excluded, to be able to correctly accept or reject the algorithm's conclusion on any given case.
Transparent degree of confidence	The algorithm should share with the user a level of confidence in its assessment for each case. The accuracy of the model's expression of confidence should be validated as well as the accuracy of the model itself.
Able to be monitored	The algorithm should share performance data with users to enable ongoing monitoring of both individual and aggregated cases, quickly highlighting any significant deviations in performance.
Auditable	An independent means should be provided to monitor the algorithm's ongoing performance in a way that guides appropriate intervention. This may include periodic quality control checks similar to those performed by operators on imaging equipment.
Intuitive user interface	The user interface should enable the operator to intuitively how to use the algorithm with as little training as possible and impose the minimum possible cognitive load on the user.

However, it appears that such a list is incomplete, as some important risks of AI in healthcare, such as algorithmic bias and inequality, have not been considered. Among the few works that have directly investigated AI fairness in medicine, it is worth mentioning a recent study that evaluated the state-of-the-art deep neural networks on large public chest X-ray datasets with respect to patient sex, age, race, and insurance type, the latter as a proxy for socioeconomic status (Seyyed-Kalantari et al, 2020). The study concluded that '*models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modification*'. In this work, the authors used the so-called true positive rates (TPR) as a measure of fairness, but other criteria have also been proposed in the literature, such as statistical parity, group fairness, equalised odds and predictive equality (Barocas et al., 2017).

Given the current lack of literacy and trust in AI, clinical usability is another aspect of medical AI that has been recommended for validation with end-users. To enhance clinical acceptance, perceived utility and future adoption, the AI algorithm and its visual interfaces should enable the operator to intuitively know how to use the tool with as little training as possible, to impose the minimum possible cognitive workload on the user, and to enhance clinical efficiency by decreasing decision-making time. During usability tests, questionnaires can be used to gather quantitative and qualitative information on the user's satisfaction with the AI tool (Lewis, 2018). For example, when assessing the usability of an AI-powered algorithm for depression care, the researchers in (Tanguay-Sela et al., 2020) used specific usability questions, as illustrated in Box 5.

Box 5 – Excerpts of a usability questionnaire for assessing an AI technology for depression care (Tanguay-Sela et al., 2020)

- The probabilities produced by the model, overall, were: too optimistic; reasonable; too pessimistic.
- The application interfered with my patient interview: strongly agree; somewhat agree; unsure; somewhat disagree; strongly disagree.
- Based on your overall experience today, how much do you trust the predictive model to help you choose treatments for depression (1 being 'very little' and 5 being 'very much)?
- The model provided us with more rich information to discuss: strongly agree; somewhat agree; unsure; somewhat disagree; strongly disagree.
- Based on your experience today, do you think using the application would cost you significant time (1 being 'cost you significant time' and 5 being 'save you significant time'):
- You would use the application: For all patients with depression; Only for the most severe patients; only for patients where one treatment has failed; only for patients where more than one treatment has failed; not at all; to review patient info.

Other usability elements that could evaluated in a usability questionnaire include: level of understanding of diagnosis by patients and clinicians; level of understanding of treatment options by patients and clinicians; perceived quality of communication between patient and doctor; degree of interpretability of the AI-driven predictions for the clinicians; level of satisfaction with the technology, user interfaces; understanding of technical terminology by clinicians and patients; usefulness of error messages/alerts; overall ease-of-use; impact on clinician's productivity; level of intention-to-use of the system (e.g. only when needed vs. full use), and so on.

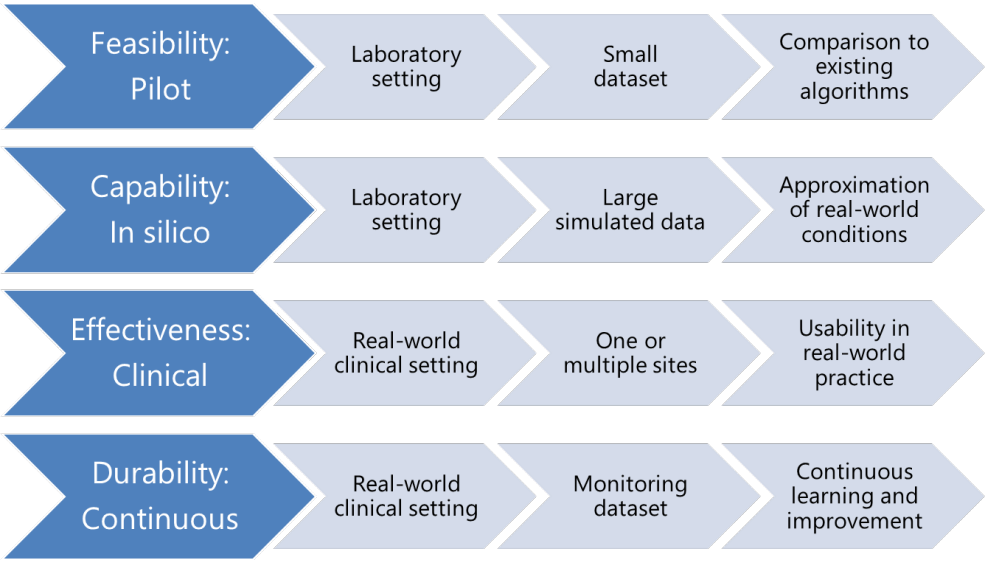
Even if the AI is validated as being accurate, reliable, fair and user-friendly, this may not necessarily lead to patient benefit. Researchers from South Korea suggested assessing impact on patient outcomes to confirm clinical utility and to enable AI technology to be accepted and recommended by clinical experts, academic societies, or independent third-party organisations (Park & Han, 2018). In addition to demonstrating its clinical effectiveness, evaluation of the cost-effectiveness should also be systematically performed, given the huge investments into medical AI with promised efficiencies and cost reductions only being assumed. For example, economic evaluations using decision analytic modelling (Hill et al., 2020) can be used to assess whether additional AI solution costs are justified given the modelled effect, such as on health-related quality of life (e.g. QALY, or quality-adjusted life years). Importantly, the initial investment and operational costs for a given AI infrastructure and service need to be included in the cost-effectiveness analysis (Wolff et al., 2020). Finally, given that AI algorithms continue to learn over time as more data become available, it is important to adapt existing validation frameworks to enable the continuous monitoring of performance throughout the life cycle of the AI tool in the clinical environment.

4.3.3. Subdivision of the evaluation process into discrete phases.

Instead of evaluating medical AI solutions in one single procedure, a few publications have recently recommended implementing a multi-stage approach in which developed algorithms undergo several steps of evaluation of varying goals and increasing complexity. For example, four steps (phase I to phase IV) were proposed for AI validation in the diagnostic imaging field, namely (1) feasibility testing, (2) capability, (3) effectiveness, and (4) durability (Larson et al., 2021) (Figure 12):

- Phase I – Feasibility: The goal is to perform a first/pilot evaluation of the algorithm in the laboratory under ideal conditions, typically on a single small test dataset. This stage will include comparison to existing algorithms that address the same clinical task, or with results obtained directly by expert clinicians. At this stage, the AI algorithms do not need to be fully robust, as the goal is simply to assess feasibility. The resulting findings may be disseminated in a scientific publication, even if the algorithm is not demonstrated for clinical application at this stage.
- Phase II – Capability: In this phase, the goal is to simulate real-world conditions in a laboratory setting and evaluate as well as refine the AI algorithm accordingly to enhance its capabilities. The phase can be also referred to as in-silico validation (Viceconti et al., 2021) (i.e. using computer simulation) or virtual clinical trials (Abadi et al., 2020). In this phase, reliability can be tested by simulating the input data and the clinical conditions under which it may be used. Safety tests will evaluate the algorithm's ability to minimise the risk of harm when deployed and subjected to unanticipated situations, that will be also simulated for testing. Furthermore, this phase should be implemented with end-users, especially clinicians and operators, to evaluate their behaviours and decision making given the simulated conditions and outputs of the AI algorithm.
- Phase III – Effectiveness: At this stage, the validation is moved to the clinical environment to assess real-world performance and to specific clinical sites to perform local validations. The primary objective is to confirm that the real-world performance of the algorithm matches its performance in the test environment. All results and feedback from this stage should be leveraged to update and optimise the AI algorithm, which will be retested in the controlled environment as in previous stages, before another round of local clinical evaluation. This evaluation stage in the clinic may reveal local quality control problems and AI manufacturers should work with local clinical sites to resolve the identified quality issues.
- Phase IV – Durability: At this stage, the manufacturer should put in place a mechanism to enable ongoing performance evaluation and monitoring, with the intent of continuous improvement. They may integrate monitoring or auditing systems within their AI solution to automatically detect, correct, and report errors, and to compile clinical feedback and user feedback. Furthermore, depending on the errors and problems identified over time, the AI algorithms should be updated and improved, such as by using additional training data, and then retested in the controlled environment before they are re-used in the clinic.

Figure 12 – Example of a multi-stage approach for medical AI evaluation



Researchers from IBM Research have proposed an alternative subdivision of the evaluation process by drawing analogies from the drug discovery and testing sectors (Park et al., 2020), as described in Table 3.

Table 3 – Excerpts of subdivided evaluation process for medical AI, based on processes implemented in the drug development sector (Park et al., 2020)

Testing phase of AI algorithm	*Procedures*	*Examples*	*Equivalence in drug discovery*
Phase 1: Technical performance & safety	In silico algorithm performance optimisation Usability tests	Determination of thresholds to balance sensitivity and specificity for a particular clinical use case, scenario-based testing to assess cognitive overload	Determine optimal dose Identify potential toxicities
Phase 2: Efficacy & side effects	Controlled algorithm performance/efficacy evaluation by intended users in medical setting Interface design Quality improvement	Retraining and reassessing model performance with larger real-world data sets, measurement of the efficiency of information delivery and workflow integration with representative users, pilot study of predictive algorithm in a clinical setting	Early efficacy tests Adverse event identification
Phase 3: Therapeutic efficacy	Clinical trial Adverse events identification	Randomised trial to test whether delivery AI-based decision support affects clinical outcomes and/or results in user over-trust	Clinical trial Adverse event identification
Phase 4: Safety & effectiveness	Post-deployment surveillance	Measurement of algorithmic performance drift	Post-marketing surveillance

While there are overlaps between the two subdivisions of the medical evaluation process presented in this section (Figure 12 & Table 2 – Examples of performance elements for imaging AI algorithms (from Larson, et. al., 2021)). The first subdivision (Figure 12) is focused on separating the environments

and populations in which the algorithm is tested (small datasets to demonstrate feasibility, simulated environments to test robustness to contextual changes, clinical setting to demonstrate real-world applicability). The second approach (Table 2) does not necessarily separate the testing environments (e.g. medical settings are used in both phases 2 and 3) but each step is more focused on a particular risk and clinical aspect such as on safety, effectiveness, usability and efficacy.

In both multi-stage evaluation approaches, each of these phases is dependent upon the successful completion of the previous step, which reduces costs. For example, algorithms that do not perform well in a controlled environment are almost certain to not perform well in the real world. While they require to be further developed and adopted by the relevant stakeholders, these multi-stage and multi-faceted evaluation studies are promising as they take into consideration the complexity of AI-guided healthcare delivery, which is compounded by user- and context-dependent applications.

4.3.1. Promotion of external evaluations by third-party evaluators

Evaluating the performance of an AI model with similar datasets than those used to develop and train the model is called internal validation. In the early days of medical AI, this was the most reported approach for algorithm validation as it is easy to implement. However, internal validation – even by developers and manufacturers with a culture of quality and good practices of excellence in medical AI – is likely to be inherently biased and to overestimated performance, while it is limited in its ability to identify all risks associated with changes in the data or clinical environment. A 2019 study reviewed more than 500 research papers in the field of radiology AI and found that only 6% of the AI algorithms reported underwent an external evaluation (Kim et al., 2019). Hence, in recent years many researchers and opinion leaders have recommended promoting the external evaluation of AI algorithms in healthcare (Park & Han, 2018; Larson et al., 2021).

External validation refers to the use of completely separate, external datasets for evaluating AI tools. The external datasets should strongly represent the variability in the population and the usage of the AI solution. Such data will ideally come from different clinical sites and geographical locations to evaluate the generalisability of the given AI algorithm outside of the controlled environment in which it was built. With this approach, it will be possible, for example, to evaluate the AI algorithm when the technical parameters of the data acquisition vary (e.g. differences in imaging scanners and protocols between hospitals). Furthermore, many researchers have recommended the use of common reference datasets, acquired from representative real-world populations, for external evaluation and benchmarking of AI models. These reference datasets can be directly compared to similar algorithms that have been previously evaluated with the same reference dataset. For example, in 2010 the National Cancer Institute in the United States set up the Cancer Imaging Archive (www.cancerimagingarchive.net), which now comprises a wide range of cancer imaging collections from all cancer types, that are extensively and routinely used for external validation and comparison of AI algorithms.

Several research projects have recently been funded by the European Commission to build European repositories of reference cancer imaging datasets, such as the EuCanImage project (https://eucanimage.eu). Furthermore, external validation should ideally be carried out by using third-party evaluators to ensure an objective and exhaustive evaluation of the AI algorithm is performed according to the performance criteria outlined in the previous section, such as accuracy, reliability, fairness and usability. Such third-party evaluators could include clinical research organisations, research laboratories, or independent institutions that develop and maintain reference standard data sets. Such testing organisations would be specialised to enable the highest standards, quality and objectivity in the evaluation and monitoring of AI solutions in healthcare, resulting in reduced undetected risks and increased trust in medical AI for real-world practice. It is worth noting that DIGITAL EUROPE is currently preparing new research initiatives to develop Testing and Experimentation Facilities (TEF) in Europe, which -once established- will greatly facilitate external validation of medical AI tools, especially for companies.

4.3.2. Standardised and comprehensive reporting of the AI evaluation procedure and results

To further enhance trust and usability of the AI tools, transparent documentation and reporting of the validation process is essential. This type of reporting will facilitate the critical appraisal process for developers, researchers, and other stakeholders; in addition, it should help replicate the AI algorithm and results, if necessary. Before the widespread use of AI, researchers had already identified the need for standardised and comprehensive reporting guidelines for predictive models used in healthcare, among which is TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (Collins et al., 2015). The TRIPOD statement was first published in 2015 and shortly afterwards adopted at large in the biomedical community. TRIPOD provides guidance on how to clearly report the development of a predictive model in order to assess its potential bias and usefulness. (Collins et al, 2015) Concretely, and as illustrated in Box 5, the TRIPOD statement includes a checklist of 22 items deemed essential for transparent reporting of a prediction model study.

Box 5 – Essential items to be included when reporting a prediction model, according to TRIPOD

- Title, abstract, background, and objectives.
- Methods: Source of data, participants, predictors, sample size, missing data, type of prediction model and other model-building procedures, etc.
- Results: Participants (number and characteristics), performance measures, confidence intervals, model updating., etc.
- Discussion: Limitations (e.g. non-representative sample, missing data), interpretation (incl. comparison to similar studies), implications (e.g. potential clinical use).
- Other information: Supplementary information, funding.

Although TRIPOD primarily aims to improve reporting, it also facilitates more comprehensive understanding and analysis of prediction models, ensuring that they can be further studied and used to guide the provision of healthcare, thus enhancing reproducible research, trust and clinical translation. While many aspects of the TRIPOD statement are inherently applicable to prediction model studies using machine learning methods, its uptake by AI communities has not been high. Possible reasons for the low level of uptake include subtle differences in terminology or a perceived lack of relevance because TRIPOD – at least in its original definition – focused on regression-based prediction model approaches (and not machine-learning based ones). In response to more AI-specific reporting guidelines, an extension of TRIPOD devoted to health prediction models that use machine learning techniques is currently being developed under the name of TRIPOD-AI (Collins et al, 2021)[1].

Another example of reporting and validation guidelines is the work carried out by the CONSORT consortium (Consolidated Standards of Reporting Trials), which has extended their 2010 reporting guidelines to include AI-specific aspects with their CONSORT-AI statement. While the original guidelines recommended including elements such as title, trial design, participants, interventions, outcomes and sample size, the extended CONSORT-AI statement proposes that researchers 'provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human–AI interaction and provision of an analysis of error cases' (Liu et al, 2020). As shown in Box 6 (Liu et al, 2020), the CONSORT-AI extension enumerates new AI-specific items to be used in

[1] TRIPOD. www.tripod-statement.org

the reporting process, in addition to those included in the original CONSORT guidelines published in 2010.

Box 6 – Reporting elements for medical AI in clinical trials, according to the CONSORT-AI guidelines

- Indication that the intervention involves AI in the title and abstract and specify the type of model.
- Intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, public).
- Description of how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.
- Version of the AI algorithm that was used.
- Description of the input data that were acquired and selected for the AI intervention.
- Description of any human–AI interaction in the handling of the input data, and the level of expertise required from users.
- The output of the AI intervention.
- Explanation on how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.
- Results of any analysis of performance errors and how errors were identified, where applicable.
- Information on how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.

Researchers at Stanford University proposed a new set of standards for reporting AI solutions in healthcare, entitled MINMAR (MINimum Information for Medical AI Reporting) (Hernandez-Boussard et al., 2020). The MINMAR standards describe the minimum information necessary to understand intended predictions, target populations, model architecture, evaluation processes, and hidden biases. The MINMAR guidelines are specifically designed for medical AI and comprise reporting elements in four main categories, as shown in Table 4.

Table 4 – Reporting elements from the MINMAR reporting guidelines

Element	Description
1. Population & setting	
Population	Population from which study sample was drawn
Study setting	The setting in which the study was conducted.
Data source	The source from which data were collected
Cohort selection	Exclusion/inclusion criteria
2. Patient demographic characteristics	
Age	Age of patients included in the study
Sex	Sex breakdown of study cohort
Race/ethnicity	Race/ethnicity breakdown of patients included in the study
Socioeconomic status	A measure or proxy measure of the socioeconomic status of patients included in the study

3. Model properties	
Model task	Classification or prediction
Model architecture	Algorithm type: Machine learning, deep learning, etc.
Data splitting	How data were split for training, testing, and validation
Gold standard	Labelled data used to train and test the model
Features	List of variables used/selected in the AI model
Missingness	How missingness was addressed: reported, imputed, or corrected
Optimisation	Model or parameter tuning applied
Internal model validation	Study internal validation
External model validation	External validation using data from another setting
Transparency	How code and data are shared with the community

Such a reporting model for medical AI evaluation will promote transparency, thoroughness, and trust, by including all the key information from the AI evaluation studies in a single detailed document, as well as by assisting publishing editors, AI developers, clinicians and researchers in understanding, interpreting and critically appraising the quality of the AI study design, validation and results.

5. Policy options

This section describes seven policy options suggested to better develop, evaluate, deploy and exploit technically, clinically and ethically sound AI solutions in future healthcare (Figure 13).

Figure 13 – Summary of policy options suggested in this report

5.1. Extend AI regulatory frameworks and codes of practice to address healthcare-specific risks and requirements

As described in Section 4.1, current medical AI devices are regulated by the MDR and IVDR regulations, which were introduced in 2017. Furthermore, in 2021 the European Commission (EC) proposed a new regulation for AI which provides new requirements and obligations for high-risk applications, including medical AI technologies, such as to establish and implement quality management systems in organisations, undergo conformity assessment and potentially reassessment of AI systems (in the event of substantial modification), as well as conduct post-market monitoring.

While the new proposal has been elaborated for AI technologies in general, the new framework considers medical AI tools as high risk, requiring them to undergo increased scrutinisation. However, the requirements are presented in a generic fashion, while – as seen in this report – AI in healthcare is faced with specific and high-stake technical, clinical and socio-ethical challenges and risks.

It is thus important that regulatory frameworks and codes of practice are extended and put into practice for medical AI (as described in sections 4.2 and 4.3). The need for updating the regulatory approvals of AI-driven medical devices has been voiced worldwide, such as in the United States (Harvey & Gowda, 2020; Allen, 2019), Japan (Chinzei et al., 2018; Ota et al., 2020) and China (Roberts

et al., 2020). Particularly, in 2021 the U.S. Food & Drug Administration (FDA) published the Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan (FDA, 2021), which calls for tailored regulations for medical AI, good machine learning practices, and patient-centred approaches.

For tailoring existing frameworks and AI practices to the medical field, multi-faceted risk assessment (section 4.2) should be an integral part of the medical AI development and certification process. Furthermore, risk assessment must be domain-specific, as the clinical, social and ethical risks and constraints differ between, for example, radiology, surgery, genomics, mental health, child health, and home care.

The validation of medical AI technologies should be harmonised and strengthened to assess and identify multi-faceted risks and limitations by evaluating not only model accuracy and robustness but also algorithmic fairness, clinical safety, clinical acceptance, transparency and traceability.

An important proposal (highlighted in section 4.3) for improved medical AI validation and certification is the introduction and generalisation of third-party external validation by independent entities that will be specialised in this process. This will allow for a more objective and expert validation of medical AI tools in a manner that systematically takes into account variability in real-world clinical practices and socio-ethical contexts.

5.2. Promote multi-stakeholder engagement and co-creation throughout the whole lifecycle of medical AI algorithms

For the future acceptability and implementation of medical AI tools in the real world, many stakeholders beyond AI developers – such as clinicians, patients, social scientists, healthcare managers and AI regulators – will play an important role. Hence, new approaches are needed to promote inclusive, multi-stakeholder engagement in medical AI and ensure the AI tools are designed, validated and implemented in full alignment with the diversity of real-world needs and contexts.

Hence, future AI algorithms should be developed by AI manufacturers based on co-creation (Leone et al., 2021), i.e. through strong and continuous collaborations between the AI developers and the clinical end-users, as well as with other relevance experts such as biomedical ethicists. These collaborations should be present at all stages, from the design and development of the AI solution to its validation and deployment (Filice & Ratwani, 2020).

Integrating human- and user-centred approaches throughout the whole AI development process will enable to design AI algorithms that better reflect the needs and cultures of healthcare workers, but also to identify and address potential risks at an early stage. This will shift the focus towards optimising the clinical performance of the end-users and the health benefits for the citizens, while considering existing social, ethical and legal requirements.

Through strong user engagement, future implementations of medical AI algorithms will take into close consideration the expected interactions between the end-users and the algorithms (otherwise referred to as human-computer interaction) (Xu, 2019). Visual interfaces should be carefully designed based on requirements from the clinical end-users to enable human-centred and clinically meaningful displays of explanations for the machine learning model predictions in healthcare (Barda et al., 2020). This will allow human errors to be reduced and will improve explainability and acceptance of the AI-driven predictions and decisions.

Finally, multi-stakeholder engagement and co-creation will address specific social issues related to equity, equality and fairness, which are application-specific issues that require an understanding of the clinical tasks, possible confounding factors, and relevant group differences; hence continuous

collaboration between the domain experts, healthcare professionals, social scientists, and real-world community members, especially from underrepresented groups, is key.

5.3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI

New approaches and mechanisms are needed to enhance the transparency of AI algorithms throughout their lifecycle. In order to be able to understand the details of what has occurred when something goes wrong in the clinical implementation of medical AI, transparency is essential, including but not limited to documenting the whole AI development process; this type of documentation and transparency helps eliminate potential ambiguities and lack of accountability (Felzmann et al, 2020).

One option is for regulatory bodies for medical AI to introduce an 'AI passport' for standardised description and traceability of medical AI tools (see illustration in Figure 14). Such a passport should describe and monitor key information about the AI technology, covering at least five categories of information:

1. Model related information (e.g. model owners, developers and reviewers, intended clinical uses, applicable licences(s), algorithmic details, hyper-parameters, key assumptions and requirements).
2. Data related information (training vs. testing data, data types e.g. imaging, real vs. simulated datasets, data origins).
3. Evaluation related information (model accuracy, robustness, biases, limitations and extreme cases).
4. Usage related information (e.g. statistical distributions, (dis)agreements with clinicians, identified failures, memory usage, etc.).
5. Maintenance related information (last updates, software versioning, last periodic evaluation, dates, etc.).

The AI passport should be standardised to enable consistent traceability across countries and healthcare organisations.

Figure 14 – Example of a possible AI passport that can be used to improve traceability and transparency in medical AI, by documenting all key details about the AI tools, their intended use, model and data details, evaluation results, and information from continuous monitoring and auditing

- **Main details**
  - Identifier:
  - Owner(s):
  - TRL level:
  - Licence:
  - Data of creation:
- **Intended use**
  - Primary use:
  - Secondary use:
  - Users:
  - Counter-indications:
  - Ethical considerations:
- **Model details**
  - Model design:
  - Model hyperparameters:
  - Objective functions:
  - Fairness constraints:
- **Training data**
  - Data provenance:
  - Population groups:
  - Variables:
  - Pre-processing:
- **Evaluation**
  - Evaluation data:
  - Evaluation metrics:
  - Evaluation results:
  - Identified limitations:

Monitoring
- Last periodic evaluation:
- Identified failures:
- Version number:

Miscellaneous
- Assumptions:

Furthermore, medical AI is a highly dynamic technology with new data, equipment and users regularly introduced into its workflows. It is therefore clear that the concept of traceability must go beyond the mere documentation of the development process or the phase of testing the AI model; instead, it should also comprise the process of monitoring and maintaining the AI model or system in the real world by continually tracking how it functions after deployment in clinical practice and identifying potential errors or changes in performance (Lekadir et al, 2022).

Hence, it is important that the algorithms are developed together with accompanying live interfaces that will be intended for continuous surveillance and auditing of the AI tools after their deployment in their respective clinical environment. Such monitoring tool should include user-friendly capabilities for quality control and detection of errors and extreme cases, a human-in-the-loop mechanism to enable for human oversight and feedback, a system of alerts to inform the clinicians of suspected deviations from previous states or performance degradation (e.g. when new equipment or protocol is introduced), as well as a periodic evaluation system that can be configured to indicate reference test datasets, as well as periodicity of the evaluations (e.g. monthly vs. quarterly).

5.4. Develop frameworks to better define accountability and monitor responsibilities in medical AI

Accountability continues to be a pressing issue in the field of AI, especially in the high-stake areas of medical AI. It is an especially important issue when considering situations in which an AI-based healthcare tool deployed in real clinical settings fails, produces errors, or results in unexpected side effects (Geis et al, 2019). Frameworks and mechanisms are needed to adequately assign responsibility to all actors in the AI workflow in medical practice, including the manufacturers, thus providing incentives for applying all measures and best practices to minimise errors and harm to

the patient. Such expectations are already an integral part in the development, evaluation and commercialisation of medicines, vaccines and medical equipment, and need to be extended to future medical AI products.

Above all, unified legal frameworks are needed to define responsibility and liability and enforce relevant consequences in medical AI across Europe and beyond. Of the existing regulations, the GDPR offers a two-pronged approach to algorithmic accountability – approaching the issue from the perspective of individual rights on the one hand and systemic regulatory frameworks on the other (Kaminski & Malgieri, 2019). In particular, the GDPR establishes transparency as a key principle for data processing and links it with lawfulness (Art. 5 para 1(a) GDPR) which both are important parts of the principle of accountability (Art. 5 para 2 GDPR).

However, while the GDPR is highly variable in terms of outlining the rights to data privacy as well as to explanation, some researchers in the field have stressed that it is not in and of itself sufficient in terms of outlining algorithmic accountability in medical AI (Barocas, 2019). There is a legal gap for medical AI accountability that remains to be addressed; in the face of this challenge, expert leaders in the field have recommended the establishment of a singular new regulatory body for AI (Tuut, 2017; Koene et al., 2019).

It is expected that in 2022 the EC will propose EU-wide measures adapting existing liability frameworks to the challenges of AI in order to ensure that victims who suffer damages to their life, health or property from an AI technology have access to the same compensation as victims of other technologies (Communication to EU Parliament, 2021). This may include a revision of the Product Liability Directive (Council Directive, 1985) and may require sectorial adjustments such as for AI in healthcare.

One important way of increasing accountability of AI tools in healthcare is through periodic audits and risk assessments, which can be used to evaluate how much regulatory oversight a certain AI tool might need (Kaminski & Malgieri, 2019; Reisman et al., 2018). To this end, the assessments must be conducted through the whole AI pipeline, from data collection, to development, to pre-clinical stages, to deployment, but also when the tools are in use. Future AI solutions should maintain an archive of AI-based decisions and have a mechanism for continuous monitorability and traceability over time as described in the previous section. Audits to assess fairness, transparency, accuracy, and safety could be used to hold AI decision-making processes to the same standard as human processes (Caplan et al., 2018). While some companies and agencies lean heavily on internal auditing processes, numerous researchers as well as civil rights organisations call for these audits to be carried out externally by independent auditing organisations.

5.5. Introduce education programmes to enhance the skills of healthcare professionals and the literacy of the general public

To increase adoption and minimise error, future medical professionals need to be adequately trained in this new technology, including its advantages to improve care, quality, and access to healthcare, as well as its limitations and risks (Paranjape et al., 2019). Hence, it is time to update educational programmes in medicine and increase their interdisciplinarity, with dedicated lectures and practical sessions that seamlessly integrate the implications of medical AI in future clinical practice (McCoy et al., 2020; Rampton et al., 2020).

Furthermore, there is an urgent need to increase the AI literacy of the general public to empower citizens and patients, who will better seize the benefits of emerging medical AI tools, while minimising the potential risk of misuse of the AI tools, especially during remote monitoring and care management. Some countries have already invested in providing free AI public literacy courses, such as Finland's 'Elements of AI' course run by the University of Helsinki (www.elementsofai.com).

5.6. Promote further research on clinical, ethical and technical robustness in medical AI

Despite major advances in recent years in AI and machine learning, as well as in their applications to medicine and healthcare, the multitude of risks discussed in this report call for further research and development to realise the full promise of medical AI, while addressing the existing clinical, socio-ethical and technical limitations. Examples of areas for future research include explainability and interpretability, bias estimation and mitigation, as well as secure and privacy-preserving AI.

Explainable AI is a research area that is investigating a new generation of AI algorithms that can be understood by humans, such as by clinicians and patients in medical AI. It has attracted a lot of interest in recent years and various approaches are being developed and tested. However, explainable AI in healthcare remains very challenging due to the complexity and variability of the biomedical and clinical data, and existing methods are yet to find their way to clinical practice. To improve their potential, it is important to assess and ensure that explainability methods produce explanations that are clinically meaningful and accepted by the end-users. There is a need for interdisciplinary approaches during AI developments that start by examining the needs of the clinicians and understanding the types of explanations (visual vs. quantitative methods) that better suit their needs and specific clinical task.

To explicitly mitigate the presence of unwanted bias in the data, methods have already been investigated (Li & Vasconcelos, 2019; Zhang et al., 2018) and some open-source toolkits have already been published, such as those by IBM (AI Fairness 360) and Microsoft (Fairlearn (Bird et al., 2020)). However, the detection of biases, in particular implicit and hidden biases, remains to a great extent an open problem. Qualitative biases such as cognitive biases of clinicians generating, interpreting or annotating the data, require multidisciplinary research and increased diversity in AI development, healthcare, and policy teams to mitigate bias and strengthen the fairness of medical AI algorithms.

There is also need for more research to develop adaptation methods that will ensure a high level of generalisability of future AI tools across population groups, clinical centres and geographical locations. In addition, it is important to develop new validation platforms that can robustly assess AI algorithms for accuracy but also for fairness with respect to sex/gender, age, ethnicity and race, socioeconomic status and other sociodemographic categories.

Furthermore, future AI solutions for healthcare should be implemented by integrating uncertainty estimation, a relatively new field of research that aims to provide clinicians with clinically useful indications on the degree of confidence in AI predictions (Kompa et al., 2021). Ideally, the clinician should receive alerts/warnings when the uncertainty for certain predictions is high. In future settings, the AI system could provide information on the cause of the high uncertainty (e.g. low-quality image scans, insufficient evidence in the data), and even advise the clinicians on the course of action needed to improve the AI predictions (e.g. inclusion of additional lab tests and predictors, re-scanning of the patient).

Finally, current cyberattacks on medical AI technologies remain difficult to detect, as the actual tools themselves may continue to function properly, but the conclusions that the AI system will confidently provide will be erroneous. Further research is needed to develop, validate and deploy medical AI tools that are able to protect themselves against privacy as well as security risks. This will result in a new generation of AI algorithms which can be robustly deployed and used in their real-world environment with maximal resilience and confidence.

5.7. Implement a strategy for reducing the European divide in medical AI

While the EU has made significant investments in AI in recent years, inequalities persist between different European countries when it comes to advancements in the field of AI (Caradaica, 2020). The AI divide – especially between the Western and Eastern regions of the continent – can be explained by structural differences in research programmes and technological capacities, as well as by the varying levels of investments from the public and private sectors (Quaglio, et al., 2020B).

The disparities in AI development and implementation between EU countries are particularly marked in medical AI, since developments and innovations in this field are highly dependent on access to large databases of well-curated biomedical data as well as to technological capacities. At the same time, these AI disparities may exacerbate the existing health inequities and disparities that exist across the EU; for example, studies have shown that there is a gap between Eastern and Western Europe in life expectancy, maternal mortality, and other population health indicators (Forster, 2018; The World Bank, 2019).

In this context, the EU Member States, in particular those of Eastern Europe, could develop specific programmes to support AI in health. These should include concrete actions to boost the technological, research and industrial capacities of emerging EU countries in the field of AI for healthcare. In particular, infrastructure projects should be established by Member States that have limited research infrastructures and data availability. This would build and enhance much-needed capacities in biomedical and health data sharing, storage, curation and security across the entire EU (ECRIN, 2019). Other programmes should be established to increase the technological, clinical and industrial capacities of several European countries for the development, testing and deployment of novel AI tools in medicine and healthcare, including high-performance computing, open cloud services, clinical testing facilities and pre-commercial procurement.

The European Commission could implement specific coordination and support programmes of activities implemented in this sector by different Member States, thereby supporting the implementation of common guidelines and approaches. Such coordination should ensure the development of an inclusive European Health Data Space (EHDS), which takes into close consideration national and regional challenges across Europe (Marschang, 2021). Similarly, existing education-focused programmes such as the Marie-Curie training networks could be strengthened to enhance the training capacities and human capital in medical AI specifically in emerging EU countries.

Lastly, the disparities that exist in medical AI between different European countries – and especially between Eastern and Western Europe – also reflect the broader social, economic, and health inequities across the different regions of Europe. The issue of reducing the European divide in medical AI is one that requires an approach that goes beyond focusing solely on the fields of medicine and/or the fields of AI and instead involves policy actions that will tackle the larger issues of systemic inequality in European society.

References

Abadi, E., Segars, W.P., Tsui, B.M., Kinahan, P.E., Bottenus, N., Frangi, A.F., Maidment, A., Lo, J. and Samei, E., 2020. 'Virtual clinical trials in medical imaging: a review', Journal of Medical Imaging, 7(4), p.042805.

Abdi J, Al-Hindawi A, Ng T, Vizcaychipi MP. 'Scoping review on the use of socially assistive robot technology in elderly care', BMJ Open. 2018;8(2):e018815.

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N. and Folk, J.C.. 'Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices', NPJ digital medicine, 1(1), pp.1-8., 2018

Abràmoff, M.D., Tobey, D. and Char, D.S. 'Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process,' American journal of ophthalmology, 214, pp.134-142, 2020

Adamson, A.S. and Smith, A., 'Machine learning and health care disparities in dermatology' JAMA dermatology, 154(11), pp.1247-1248, 2018.

Adedinsewo D, Carter RE, Attia Z, Johnson P, Kashou AH, Dugan JL, et al. 'Artificial Intelligence-Enabled ECG Algorithm to Identify Patients with Left Ventricular Systolic Dysfunction Presenting to the Emergency Department with Dyspnea'. Circ Arrhythmia Electrophysiol.;13(8), 2020

Adhikari L, Ozrazgat-Baslanti T, Ruppert M, Madushani RWMA, Paliwal S, Hashemighouchani H, et al. 'Improved predictive models for acute kidney injury with IDEA: Intraoperative data embedded analytics' PLoS One. 2019;14(4).

Ahn J, Connell A, Simonetto D, Hughes C and Shah VH. 'Application of Artificial Intelligence for the Diagnosis and Treatment of Liver Diseases,' Hepatology. 2021;73(6):2546-2563.

Alder, S. 'AI Company Exposed 2.5 Million Patient Records Over the Internet', HIPPA Journal. 21 August 2020.

Allen M, Pearn K, Monks T, Bray BD, Everson R, Salmon A, James M, Stein K. 'Can clinical audits be enhanced by pathway simulation and machine learning? An example from the acute stroke pathway', BMJ Open. 2019;9(9):e028296.

Allen, B., 'The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices', Journal of the American College of Radiology, 16(2), pp.208-210, 2019.

Almirall, D., Nahum-Shani, I., Sherwood, N.E. and Murphy, S.A., 'Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research', Translational behavioral medicine, 2014, 4(3), pp.260-274.

Alsharqi M, Woodward WJ, Mumith JA, Markham DC, Upton R, Leeson P. 'Artificial intelligence and echocardiography', Echo Res Pract. 2018;5(4):115–25.

Aminololama-Shakeri S, Lopez E. 'The Doctor-Patient Relationship With Artificial Intelligence', American Journal of Roentgenology. 2019;202(2)

Andrew D Selbst, Julia Powles., 'Meaningful information and the right to explanation', International Data Privacy Law, Volume 7, Issue 4, Pages 233–242, 2017

Angus DC. 'Randomized clinical trials of artificial intelligence', JAMA. 323(11):1043-1045, 2020.

Arora A. 'Conceptualising Artificial Intelligence as a Digital Healthcare Innovation: An Introductory Review', Med Devices (Auckl). 3:223-230. doi: 10.2147/MDER.S262590, 2020.

Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. 'Age and Sex Estimation Using Artificial Intelligence from Standard 12-Lead ECGs', Circ Arrhythmia Electrophysiol. 12(9), 2019

Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. 'An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction', Lancet. 394(10201):861–7, 2019.

Azzi, S.; Gagnon, S.; Ramirez, A.; Richards, G. 'Healthcare Applications of Artificial Intelligence and Analytics: A Review and Proposed Framework', Appl. Sci., 10, 6553. https://doi.org/10.3390/app10186553, 2020.

Baetan, R., Spasova, S., Vanhercke, B., Coster, S., Inequalities in access to healthcare: A study of national policies, European Commission, 2018.

Bandivadekar, S.S., 'Online Pharmacies: Global Threats and Regulations', AAYAM: AKGIM Journal of Management, 10(1), pp.36-42, 2020.

Barda, A.J., Horvat, C.M. and Hochheiser, H., 'A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare.', BMC medical informatics and decision making, 20(1), pp.1-16, 2020.

Barocas, S. Legal and Policy Implications of Model Interpretability. This Week in Machine Learning and AI (TWiMLAI), January 2019. https://twimlai.com/twiml-talk-219-legal-and-policy-implications-of-model-interpretability-with-solon-barocas/.

Barocas, S., Hardt, M. and Narayanan, A., 'Fairness in machine learning', Nips tutorial, 1, p.2, 2017.

BBC, 'Google DeepMind NHS app test broke UK privacy law', BBC News, 3 July 2017.

Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-715.

Berlyand Y, Raja AS, Dorner SC, et al. How artificial intelligence could transform emergency department operations. Am J Emerg Med. 2018;36(8):1515-1517.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. and Walker, K., 'Fairlearn: A toolkit for assessing and improving fairness in AI,' Microsoft, Tech. Rep. MSR-TR-2020-32, 2020.

Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, R Van Meter A, De Choudhury M, Kane JM. 'Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook', NPJ Schizophr.;5(1):17, 2019.

Boniolo F, Dorigatti E, Ohnmacht AJ, Saur D, Schubert B, Menden MP. 'Artificial intelligence in early drug discovery enabling precision medicine', Expert Opin Drug Discov:1-17, 2021.

Campello, V. et al. 'Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge.' Medical Image Computing and Computer Assisted Intervention, 2020.

The Cancer Imaging Archive. www.cancerimagingarchive.net, accessed November 2021.

Caplan, R., Donovan, J., Hanson, L. and Matthews, J., Algorithmic accountability: A primer. Data & Society, 18, 2018.

Caradaica, M., 'Artificial Intelligence and Inequality in European Union. Europolity-Continuity and Change in European Governance', 2020, 14(1), pp.5-31.

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T. and Tsaneva-Atanasova, K., 2019. 'Artificial intelligence, bias and clinical safety', BMJ Quality & Safety, 28(3), pp.231-237.

Chaudhuri S, Long A, Zhang H, Monaghan C, Larkin JW, Kotanko P, et al. 'Artificial intelligence enabled applications in kidney disease', Semin Dial. 34:5–16, 2021.

Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD, Krystal JH, Corlett PR. 'Cross-trial prediction of treatment outcome in depression: a machine learning approach', Lancet Psychiatry.;3(3):243-50, 2016.

Chi AC, Katabi N, Chen HS, Cheng YSL. Interobserver variation among pathologists in evaluating perineural invasion for oral squamous cell carcinoma. Head Neck Pathol. 10, 451–464, 2016.

Chinzei, K., Shimizu, A., Mori, K., Harada, K., Takeda, H., Hashizume, M., Ishizuka, M., Kato, N., Kawamori, R., Kyo, S. and Nagata, K., 'Regulatory science on AI-based medical devices and systems', Advanced Biomedical Engineering, 7, pp.118-123, 2018.

Chung Y, Addington J, Bearden CE, Cadenhead K, Cornblatt B, Mathalon DH, McGlashan T, Perkins D, Seidman LJ, Tsuang M, Walker E, Woods SW, McEwen S, van Erp TGM, Cannon TD; North American Prodrome Longitudinal Study (NAPLS) Consortium and the Pediatric Imaging, Neurocognition, and Genetics (PING) Study Consortium. Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. JAMA Psychiatr,75(9):960-968, 2018.

Clay H, Stern R. 'Making time in general practice', Primary Care Foundation, 1–83, 2015.

Cohen G. 'Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?' Georgetown Law Journal. (108), 2020.

Collins GS, Moons KGM. 'Reporting of artificial intelligence prediction models', Lancet 393: 1577–79, 2019.

Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G. 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement', Circulation, 131(2), pp.211-219, 2015.

Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence. April 2021.

Cook C, Sheets C. 'Clinical equipoise and personal equipoise: two necessary ingredients for reducing bias in manual therapy trials', J Man Manip Ther. 19(1):55-7, 2011.

Cook GJR, Goh V. What can artificial intelligence teach us about the molecular mechanisms underlying disease? Eur J Nucl Med Mol Imaging. 46:2715–2721, 2019.

Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, Rimm DL, Yang M, Romero E, Schalper KA, Velcheti V, Madabhushi A. Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. Clin Cancer Res.;25(5):1526-1534, 2019.

Davenport, T.H., Barth, P. and Bean, R.,. How 'big data' Is different. MIT Sloan Management Review, 2012.

Dawoodbhoy FM, Delaney J, Cecula P, Yu J, Peacock I, Tan J, Cox B. AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. Heliyon. 7(5):e06993, 2021.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D. and van den Driessche, G., 'Clinically applicable deep learning for diagnosis and referral in retinal disease', Nature medicine, 24(9), pp.1342-1350, 2018.

De Vries L, Baselmans B, Bartels M. 'Smartphone-Based Ecological Momentary Assessment of Well-Being: A Systematic Review and Recommendations for Future Studies', Journal of Happiness Studies. 22:2361–2408, 2021.

Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB. On making the right choice: the deliberation-without-attention effect. Science. 2006; 311:1005e1007.

Directive, C., Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. Official Journal L, 210(07/08), pp.0029-0033, 1985.

Du, R., Lee, V. H., Yuan, H., Lam, K. O., Pang, H. H., Chen, Y., Lam, E. Y., Khong, P. L., Lee, A. W., Kwong, D. L., & Vardhanabhuti, V. 'Radiomics Model to Predict Early Progression of Non-metastatic Nasopharyngeal Carcinoma after Intensity Modulation Radiation Therapy: A Multicenter Study. Radiology', Artificial intelligence, 1(4), e180075, 2019.

Dusenbery, M. 'Everybody was telling me there was nothing wrong', The Health Gap, BBC News, 2018. www.bbc.com/future/article/20180523-how-gender-bias-affects-your-healthcare

Dwyer DB, Falkai P, Koutsouleris N. 'Machine learning approaches for clinical psychology and psychiatry', Annu Rev Clin Psychol, 14:91–118, 2018.

ECRIN, 'EFPIA, EATRIS, ELIXIR, BBMRI, ECRIN statement on the role of research infrastructures to boost patient-centred research and innovation in Europe', https://ecrin.org/news/efpia-eatris-elixir-bbmri-ecrin-statement-role-research-infrastructures-boost-patient-centred, 24 July 2019.

EGA Consortium (European Genome-Phenome Archive), https://ega-archive.org/datasets, 2021.

Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, et al. 'Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer' JAMA.;318(22):2199-2210., 2017.

Elements of AI. www.elementsofai.com, accessed November 2021.

Ellahham, S., Ellahham, N. and Simsekler, M.C.E., 'Application of artificial intelligence in the health care safety context: opportunities and challenges', American Journal of Medical Quality, 35(4), pp.341-348, 2020.

Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. 11(2):e1001603, 2014.

Emanuel EJ, Wachter RM. 'Artificial intelligence in health care: will the value match the hype?' JAMA. 321(23):2281-2282, 2019.

EuCanImage, https://eucanimage.eu, accessed November 2021.

European Commission, 'A Proposal for Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', April 2021.

European Commission. Digital Education Action Plan (2021-2027): Resetting Education for the Digital Age, 2020.

European Commission. Employment, Social Affairs & Inclusion Inequalities in access to healthcare. A study of national policies, 2018.

European Commission. The European Pillar of Social Rights in 20 principles. 2021.

European Genome-Phenome Archive, 'Browse datsasets', https://ega-archive.org/datasets, accessed November 2021.

European Health Data Space, https://ec.europa.eu/health/ehealth/dataspace_en, last access November, 2021.

Eurostat. Statistical expanded. Population structure and ageing, 2020.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J. 'A guide to deep learning in healthcare'. Nature medicine, 25(1), 24-29, 2019.

Evans AJ, Henry PC, Van der Kwast TH, Tkachuk DC, Watson K, Lockwood GA, Fleshner NE, Cheung C, Belanger EC, Amin MB, Boccon-Gibod L, Bostwick DG, Egevad L, Epstein JI, Grignon DJ, Jones EC, Montironi R, Moussa M, Sweet JM, Trpkov K, Wheeler TM, Srigley JR. 'Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens', Am J Surg Pathol. 32(10):1503-12, 2008.

Farina, R. and Sparano, A. 'Errors in sonography. In Errors in radiology' (pp. 79-85). Springer, Milano, 2012.

Felzmann, H., et al., 'Towards Transparency by Design for Artificial Intelligence,', Science and Engineering Ethics, 26:3333–3361, 2020.

Fernández García, J., Spatharou, A., Hieronimus, S., Beck, J.P., Jenkins, J. Transforming healthcare with AI: the impact on the workforce and organisations. Executive summary. EIT Health & McKinsey & Company, March 2020.

Ferryman, K. and Pitcan, M., Fairness in precision medicine. Data & Society, 2018.

Fihn SD, Saria S, Mendonça E, Hain S, Matheny M, Shah N, Liu H, Auerbach, A. 'Deploying AI in clinical settings. In artificial intelligence in health care: The hope, the hype, the promise, the peril', Editors: Matheny M, Israni ST, Ahmed M, Whicher D. Washington, DC: National Academy of Medicine, 2019.

Filice, R.W. and Ratwani, R.M. 'The case for user-centered artificial intelligence in radiology', Radiology: Artificial Intelligence, 2020, Vol. 2, No. 3.

Finlayson, S.G., Bowers, J.D. Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 'Adversarial attacks on medical machine learning', Science, 2019.

Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving. PubMed. Nat Biotechnol. 2018a.

Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, Osipov M, Kholodov M, Ismagilov R, Mohan S, Ostell J, Lu Z. 'Best Match: New relevance search for PubMed', PLoS Biol. 2018b;16(8):e2005343.

Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, Sarris J. 'The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials', World Psychiatry. 2017;16(3):287-298.

Fitzpatrick KK, Darcy A, Vierhile M. 'Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial', JMIR Ment Health.;4(2):e19, 2017.

Fleming N. 'How artificial intelligence is changing drug discovery', Nature: 557:S55–57, 2018.

Forster T, Kentikelenis A, Bambra, C, 'Health Inequalities in Europe: Setting the Stage for Progressive Policy Action', Foundation for European Progressive Studies. TASC: Think tank for action on social change. 2018.

Freeman, K, Dinnes, J, Chuchu, N, Takwoingi, Y, Bayliss, SE, Matin, RN, Jain, A, Walter, FM, Williams, HC and Deeks, JJ, 'Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies' BMJ, 2020, 368.

FUTURE-AI: Best practices for trustworthy AI in medical imaging, www.future-ai.eu, accessed November 2021.

Geis JR, Brady A, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R, Kotter E, Gichoya JW, Cook TS, Morgan MB, Tang A, Safdar NM, Kohli M. 'Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement', Insights Imaging, 2019 Oct 1;10(1):101. doi: 10.1186/s13244-019-0785-8. PMID: 31571015; PMCID: PMC6768929.

General Data Protection Regulation (GDPR), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed December 2021.

Gerke, S., Minssen, T. and Cohen, G. 'Ethical and legal challenges of artificial intelligence-driven healthcare'. In Artificial intelligence in healthcare (pp. 295-336). Academic Press, 2020.

German Data Ethics Commission, Opinion of the Data Ethics Commission, July 2019, https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html.

Ghassemi, M. 'Exploring Healthy Models in ML for Health', AI for Healthcare Equity Conference, AI & Health at MIT, 2021. https://www.youtube.com/watch?v=5uZROGFYfcA

Gillespie, N., Lockey, S., & Curtis, C. 'Trust in Artificial Intelligence: A Five Country Study', The University of Queensland and KPMG Australia, 2021.

Gillies RJ, Kinahan PE, Hricak H. 'Radiomics: images are more than pictures, they are data,' Radiology.;278:563–577, 2016.

Giulietti M, Cecati M, Sabanovic B, Scirè A, Cimadamore A, Santoni M, et al. 'The role of artificial intelligence in the diagnosis and prognosis of renal cell tumors', Diagnostics, ;11(2):206, 2021.

Golbraikh A, Wang X, Zhu H, Tropsha A. 'Predictive QSAR modelling: methods and applications in drug discovery and chemical risk assessment', In Handbook of Computational Chemistry, ed. J Leszczynski, A Kaczmarek-Kedziera, T Puzyn, MG Papadopoulos, H Reis, MK, 2012.

Gómez-González E, Gómez E. 'Artificial Intelligence in medicine and healthcare: applications, availability and societal impact', EUR 30197 EN. Publications Office of the European Union, Luxembourg, 2020.

Goodfellow, I., Bengio, Y. and Courville, A., Deep learning. MIT Press, 2016.

Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, Jeste DV. 'Artificial intelligence for mental health and mental illnesses: An overview', Curr Psychiatry Rep;21:116, 2019.

Guo J, Li B. 'The application of medical artificial intelligence technology in rural areas of developing countries', Health Equity, ; 2: 174–81, 2018.

Gupta R, Kleinjans J and Caiment F. 'Identifying novel transcript biomarkers for hepatocellular carcinoma (HCC) using RNA-Seq datasets and machine learning', BMC Cancer.;21(962), 2021.

Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S. and Broderick, T., 'Transparency and reproducibility in artificial intelligence', Nature, 586(7829), pp.E14-E16, 2020.

Hamed S, Thapar-Björkert S, Bradby H, Ahlberg B. 'Racism in European Health Care: Structural Violence and Beyond', Sage Journals.;30(11), 2020.

Harned, Z., Lungren, M.P. and Rajpurkar, P. 'Machine vision, medical AI, and malpractice', Harv. JL & Tech. Dig, 2019.

Harvey, H.B. and Gowda, V., 'How the FDA regulates AI. Academic radiology', 27(1), pp.58-61, 2020.

Hashimoto DA, Rosman G, Witkowski ER, et al. 'Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy', Ann Surg.;270:414e421, 2019.

Hashimoto, D.A., Rosman, G., Rus, D. and Meireles, O.R., 'Artificial intelligence in surgery: promises and perils', Annals of surgery, 268(1), p.70, 2018.

Hermsen M, Bel T, Boer M Den, Steenbergen EJ, Kers J, Florquin S, et al. 'Deep learning-based histopathologic assessment of kidney tissue', J Am Soc Nephrol.;30(10):1968–79, 2019.

Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J.P. and Shah, N.H. 'MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care', Journal of the American Medical Informatics Association, 27(12), pp.2011-2015, 2020.

Hill, N.R., Sandler, B., Mokgokong, R., Lister, S., Ward, T., Boyce, R., Farooqui, U. and Gordon, J., 'Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm', Journal of medical economics, 23(4), pp.386-393, 2020.

Hocking, L., Parks, S., Altenhofer, M. and Gunashekar, S.,. 'Reuse of health data by the European pharmaceutical industry', RAND Corporation, 2019.

Hoffman, K.M., Trawalter, S., Axt, J.R. and Oliver, M.N., 'Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites', Proceedings of the National Academy of Sciences, 113(16), pp.4296-4301, 2016.

Human-Centred Artificial Intelligence Programme, www.dtu.dk/english/Education/msc/Programmes/human-centered-artificial-intelligence, accessed November 2021.

Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. 'Prediction of sepsis patients using machine learning approach: a meta-analysis', Comput Methods Programs Biomed. 170:1-9, 2019.

Jamthikar AD, Gupta D, Saba L, Khanna NN, Viskovic K, Mavrogeni S, Laird JR, Sattar N, Johri AM, Pareek G, Miner M, Sfikakis PP, Protogerou A, Viswanathan V, Sharma A, Kitas GD, Nicolaides A, Kolluri R, Suri JS. 'Artificial intelligence framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound', Comput Biol Med.;126:104043, 2020.

Jiang S, Chin KS, Tsui KL. 'A universal deep learning approach for modeling the flow of patients under different severities', Comput Methods Programs Biomed.;154:191-203, 2018.

Jin, J.M., Bai, P., He, W., Wu, F., Liu, X.F., Han, D.M., Liu, S. and Yang, J.K.,. 'Gender differences in patients with COVID-19: focus on severity and mortality', Frontiers in public health, 8, p.152, 2020.

Kaddoum R, Fadlallah R, Hitti E, El-Jardali F, El Eid G. 'Causes of cancellations on the day of surgery at a Tertiary Teaching Hospital', BMC Health Serv. Res. 16, 2016.

Kaissis, G.A., Makowski, M.R., Rückert, D. and Braren, R.F. 'Secure, privacy-preserving and federated machine learning in medical imaging', Nature Machine Intelligence, 2(6), pp.305-311, 2020.

Kamat AS, Parker A. 'Effect of perioperative inefficiency on neurosurgical theatre efficacy: a 15-year analysis', Br. J. Neurosurg. 29: 565–568, 2015.

Kaminski, M.E. and Malgieri, G. 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations. U of Colorado Law Legal Studies Research Paper', (19-28), 2019.

Kaushal, A., Altman, R. and Langlotz, C. 'Geographic distribution of US cohorts used to train deep learning algorithms', Jama, 324(12), pp.1212-1213, 2020.

Kiener, M. ''You may be hacked' and other things doctors should tell you'. The Conversation. 3 November 2020. https://theconversation.com/you-may-be-hacked-and-other-things-doctors-should-tell-you-148946

Kim, D.W., Jang, H.Y., Kim, K.W., Shin, Y. and Park, S.H. 'Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers', Korean Journal of Radiology, 20(3), p.405, 2019.

Kirubarajan A, Taher A, Khan S, Masood S. 'Artificial intelligence in emergency medicine: A scoping review', J Am Coll Emerg Physicians Open.;1(6):1691-1702, 2019.

Koene, A., Clifton, C., Hatada, Y., Webb, H. and Richardson, R., A governance framework for algorithmic accountability and transparency, EPRS, European Parliament, 2019.

Kompa, B., Snoek, J. and Beam, A.L. 'Second opinion needed: communicating uncertainty in medical machine learning', NPJ Digital Medicine, 4(1), pp.1-6, 2021.

Koops, B.J.. 'The concept of function creep. Law, Innovation and Technology', 13(1), pp.29-56, 2021.

Krittanawong, C. 'The rise of artificial intelligence and the uncertain future for physicians', European Journal of Internal Medicine, 48, pp.e13-e14, 2018.

Kulkarni S, Seneviratne N, Baig MS, Khan AHA. 'Artificial Intelligence in Medicine: Where Are We Now?' Acad Radiol. Jan;27(1):62-70., 2020.

Kuo C-C, Chang C-M, Liu K-T, Lin W-K, Chiang H-Y, Chung C-W, et al. 'Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning', NPJ Digit Med.;2(29), 2019.

Lake IR, Colón-González FJ, Barker GC, Morbey RA, Smith GE, Elliot AJ. 'Machine learning to refine decision making within a syndromic surveillance service', BMC Public Health; 19: 559, 2019.

Larson, D.B., Harvey, H., Rubin, D.L., Irani, N., Justin, R.T. and Langlotz, C.P., 2021. 'Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: Summary and recommendations', Journal of the American College of Radiology, 18(3), pp.413-424, 2021.

Leavy, S. 'Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning', GE '18: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering, May 2018.

Lee CS, Lee AY. 'How Artificial Intelligence Can Transform Randomized Controlled Trials', Transl Vis Sci Technol.;9(2):9, 2020.

Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. 'Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom', Biol Psychiatry Cogn Neurosci Neuroimaging.;6(9):856-864, 2021.

Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. 'Why digital medicine depends on interoperability', NPJ Digit Med.;2:79, 2019.

Lekadir, K. et al. 'FUTURE-AI: Best practices for trustworthy AI in medicine', www.future-ai.org, 2022.

Leone, D., Schiavone, F., Appio, F.P. and Chiao, B. 'How does artificial intelligence enable and enhance value co-creation in industrial markets? An exploratory case study in the healthcare ecosystem', Journal of Business Research, 129, pp.849-859, 2021.

Lewis, J.R. 'The system usability scale: past, present, and future', International Journal of Human–Computer Interaction, 34(7), pp.577-590, 2018.

Li, Y. and Vasconcelos, N. 'Repair: Removing representation bias by dataset resampling', In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9572-9581), 2019.

Lindenmeyer MT, Alakwaa F, Rose M, Kretzler M. 'Perspectives in systems nephrology,' Cell Tissue Res, 2021.

Lipton, Zachary C. 'The doctor just won't accept that!' arXiv preprint arXiv:1711.08037, 2017.

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. '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(6):e271-e297, 2019.

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J. and Denniston, A.K. 'Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension', BMJ, 370, 2020.

Loftus TJ, Filiberto AC, Li Y, Balch J, Cook AC, Tighe PJ, Efron PA, Upchurch GR Jr, Rashidi P, Li X, Bihorac A. 'Decision analysis and reinforcement learning in surgical decision-making', Surgery.168(2):253-266, 2020.

Loftus TJ, Upchurch GR Jr, Bihorac A. 'Use of Artificial Intelligence to Represent Emergent Systems and Augment Surgical Decision-Making', JAMA Surg. 154(9):791-792, 2019.

Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. 'Artificial Intelligence in Cardiology: Present and Future', Mayo Clin Proc; 95(5):1015–39, 2020.

Lorkowski J, Kolaszyńska O, Pokorski M. 'Artificial intelligence and precision medicine: a perspective', Adv Exp Med Biol. Jun 18. Doi. Epub ahead of print. PMID: 34138457, 2021.

Lundberg, S.M., Lee, S.I.. 'A unified approach to interpreting model predictions', in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4768–4777, 2017.

Lv, Z. and Piccialli, F. 'The security of medical data on Internet based on differential privacy technology', ACM Transactions on Internet Technology, 21(3), pp.1-18, 2021.

Maddox TM, Rumsfeld JS, Payne PRO. 'Questions for artificial intelligence in health care', JAMA. 321(1):31-32, 2019.

Madine, M.M., Battah, A.A., Yaqoob, I., Salah, K., Jayaraman, R., Al-Hammadi, Y., Pesic, S. and Ellahham, S. 'Blockchain for giving patients control over their medical records' IEEE Access, 8, pp.193102-193115, 2020.

Magrabi, F., Ammenwerth, E., McNair, J.B., De Keizer, N.F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P.J., Vehko, T., Wong, Z.S.Y. and Georgiou, A. 'Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implication', Yearbook of Medical Informatics, 28(1), p.128, 2019.

Maharana A, Nsoesie EO. 'Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity', JAMA Network Open. 1(4):e181535, 2018.

Maliha, G., Gerke, S., Cohen, I.G. and Parikh, R.B. 'Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation', The Milbank Quarterly, 2021.

Mamoshina P, Ojomoko L, Yanovich Y, Ostrovski A, Botezatu A, Prikhodko P, Izumchenko E, Aliper A, Romantsov K, Zhebrak A, Ogu IO, Zhavoronkov A. 'Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare', Oncotarget.;9:5665-5690, 2017.

Manne, R. and Kantheti, S.C. 'Application of artificial intelligence in healthcare: chances and challenges'. Current Journal of Applied Science and Technology, pp.78-89, 2021.

Marschang S, 'The European Health Data Space: is there room enough for all?' European Public Health Alliance', https://epha.org/the-european-health-data-space-is-there-room-enough-for-all/, 2021.

Mayerhoefer ME, Materka A, Langs G, Häggström I, Szczypiński P, Gibbs P, Cook G. 'Introduction to Radiomics', J Nucl Med. 61:488-495, 2020.

McCoy, L.G., Nagaraj, S., Morgado, F., Harish, V., Das, S. and Celi, L.A. 'What do medical students actually need to know about artificial intelligence?' NPJ Digital Medicine, 3(1), pp.1-3, 2020.

McKeown, A., Mourby, M., Harrison, P., Walker, S., Sheehan, M. and Singh, I. 'Ethical issues in consent for the reuse of data in health data platforms', Science and Engineering Ethics, 27(1), pp.1-21, 2021.

McKinney, S. M. et al. 'International evaluation of an AI system for breast cancer screening', Nature 577, 89–94, 2020.

Medeiros J, Schwierz C. Efficiency estimates of health care systems in the EU. European Commission. Directorate-General for Economic and Financial Affairs. 2015.

Medeiros, J., Schwierz, C., 'Efficiency estimates of health care systems', Economic Papers, European Commission, 2015.

Menke NB, Caputo N, Fraser R, Haber J, Shields C, Menke MN. 'A retrospective analysis of the utility of an artificial neural network to predict ED volume', Am J Emerg Med.32:614-7, 2014.

Meskó B, Görög M. 'A short guide for medical professionals in the era of artificial intelligence', NPJ Digit Med. 3:126, 2020.

Michel JP, Ecarnot F. 'The shortage of skilled workers in Europe: its impact on geriatric medicine', Eur Geriatr Med. 11(3):345-347, 2020.

Miotto, R, Li L, Kidd BA, Dudley JIT. 'Deep patient: An unsupervised representation to predict the future of patients from the electronic health records', Scientific Reports. 6:26094, 2020.

Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, Brat DJ, Cooper LAD. 'Predicting cancer outcomes from histology and genomics using convolutional networks', Proc Natl Acad Sci U S A.; 115(13):E2970-E2979, 2018.

Mohr DC, Riper H, Schueller SM. 'A Solution-Focused Research Approach to Achieve an Implementable Revolution in Digital Mental Health', JAMA Psychiatry; 75(2):113-114, 2018.

Mooney SJ, Pejaver V. 'Big data in public health: terminology, machine learning, and privacy', Annu Rev Public Health;39:95-112, 2018.

Mora-Cantallops, M.; Sánchez-Alonso, S.; García-Barriocanal, E.; Sicilia, M.-A. 'Traceability for Trustworthy AI: A Review of Models and Tool', Big Data Cogn. Comput. 5, 20, 2021.

Morley, J. and Floridi, L. 'An ethically mindful approach to AI for health care', Lancet vol. 395, pp. 254-255, 2020.

Mulcahy, N. 'Recent Cyberattack Disrupted Cancer Care Throughout U.S' WebMD. 20 July 2021. https://www.webmd.com/cancer/news/20210720/recent-cyberattack-disrupted-cancer-care-us

Nagar A, Yew P, Fairley D, Hanrahan M, Cooke S, Thompson I, Elbaz W. 'Report of an outbreak of Clostridium difficile infection caused by ribotype 053 in a neurosurgery unit', J. Infect. Prev. 16: 126–130, 2015.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. 'Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies', BMJ. ;368:m689., 2020.

National Careers Service: The Skills Toolkit, https://nationalcareers.service.gov.uk/find-a-course/the-skills-toolkit, accessed November 2021.

Newman, L.H. 'These Hackers Made an App That Kills to Prove a Point', WIRED. 16 July 2019, https://www.wired.com/story/medtronic-insulin-pump-hack-app/.

NHS England. Clinical audit, https://www. england.nhs. uk/ clinaudit/, 2021.

NHS Improvement, Good Practice Guide: Focus on Improving Patient Flow, 2017. https://improvement.nhs.uk/documents/1426/Patient_Flow_Guidance_2017___13_July_2017.pdf

Niazi MKK, Parwani AV, Gurcan MN. 'Digital pathology and artificial intelligence', Lancet Oncol.; 20(5):e253-e261, 2019.

Noseworthy PA, Attia ZI, Brewer LPC, Hayes SN, Yao X, Kapa S, et al. 'Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis', Circ Arrhythmia Electrophysiol.;13(3), 2020.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S, 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

OECD/European Union, Health at a Glance: Europe 2020: State of health in the EU cycle. OECD Publishing, Paris, 2020.

OECD/European Union. Dementia prevalence. In Health at a Glance: Europe 2018: State of Health in the EU Cycle, OECD Publishing, Paris/European Union, Brussels, 2018.

Okanoue T, Shima T, Mitsumoto Y, Umemura A, Yamaguchi K, Itoh Y, Yoneda M, Nakajima A, Mizukoshi E, Kaneko S, Harada K. 'Artificial intelligence/neural network system for the screening of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis', Hepatol Res 51(5):554–569, 2021.

Ota, N., Tachibana, K., Kusakabe, T., Sanada, S. and Kondoh, M. 'A Concept for a Japanese Regulatory Framework for Emerging Medical Devices with Frequently Modified Behavior', Clinical and translational science, 13(5), pp.877-879, 2020.

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. 'A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images', Chaos, Solitons & Fractals, 140, 110190, 2020.

Paranjape, K., Schinkel, M., Panday, R.N., Car, J. and Nanayakkara, P. 'Introducing artificial intelligence training in medical education' JMIR Medical Education, 5(2), p.e16048, 2019.

Parikh RB, Teeple S, Navathe AS. 'Addressing bias in artificial intelligence in health care', JAMA; 322(24):2377-2378, 2019.

Park S, Park BS, Lee YJ, Kim IH, Park JH, Ko J, et al. 'Artificial intelligence with kidney disease: A scoping review with bibliometric analysis', PRISMA-ScR. Medicine (Baltimore);100(14), 2021.

Park, S.H. and Han, K. 'Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction', Radiology, 286(3), pp.800-809, 2018.

Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J. and Das, A.K. 'Evaluating artificial intelligence in medicine: phases of clinical research', Journal of American Medical Informatics Associations Open, 3(3), pp.326-331, 2020.

Peng J, Wang Y. 'Medical Image Segmentation with Limited Supervision: A Review of Deep Network Models', IEEE Access.; 99:, 2021.

Pérez MJ, Grande RG. 'Application of artificial intelligence in the diagnosis and treatment of hepatocellular carcinoma: A review', World J Gastroenterol. 26(37):5617–5628, 2021.

Pickering B. Trust, but Verify: Informed Consent, AI Technologies, and Public Health Emergencies, Future Internet 13(5):132, 2021.

Pinto, A., Pinto, F., Faggian, A., Rubini, G., Caranci, F., Macarini, L., Genovese, E.A. and Brunese, L. 'Sources of error in emergency ultrasonography', Critical Ultrasound Journal, 5(1), pp.1-, 2013.

Ploug, T, Holm S. 'Meta Consent –A Flexible Solution to the Problem of Secondary Use of Health Data', Bioethics, 30 (9), 2016.

Prokop M, van Everdingen W, van Rees Vellinga T, et al. 'CORADS— a categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation', Radiology, 2020:201473, [E-pub ahead of print, 2020 Apr 27].

Quaglio G, Brand H, Dario C. 'Fighting dementia in Europe: the time to act is now', Lancet Neurol. 15(5):452-4, 2016.

Quaglio GL, Boone R. What if we could fight drug addiction with digital technology?, EPRS, European Parliament, 2019.

Quaglio GL, Pirona A, Esposito G, Karapiperis T, Brand H, Dom G, Bertinato L, Montanari L, Kiefer F, Giuseppe Carrà G. 'Knowledge and utilization of technology-based interventions for substance use disorders: an exploratory study among health professionals in the European Union. Drugs: Education, Prevention and Policy; 26 (5): 437-446, 2018.

Quaglio GL. EU public health policy. 2020. European Parliamentary Research Services (EPRS). European Parliament, Brussels.

Quaglio GL, Millar S, Pazour M, Albrecht V, Vondrak T, Kwiek M, Schuch K. Exploring the performance gap in EU Framework Programmes between EU13 and EU15 Member States. 2020B. European Parliamentary Research Services (EPRS). European Parliament, Brussels.

Quer G, Arnaout R, Henne M, Arnaout R. 'Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review', J Am Coll Cardiol. 77(3):300–13, 2021.

Raghupathi, W. and Raghupathi, V. 'Big data analytics in healthcare: promise and potential. Health information science and systems', 2(1), pp.1-10, 2014.

Raji, I.D. 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', arXiv preprint, arXiv:2001.00973, 2020.

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H., 2018. 'Ensuring fairness in machine learning to advance health equity', Annals of Internal Medicine, 169(12), pp.866-872, 2018.

Ram S, Zhang W, Williams M, Pengetnze Y. 'Predicting asthma-related emergency department visits using big data', IEEE J Biomed Health Inform. 19:1216-23, 2015.

Rampton, V., Mittelman, M. and Goldhahn, J. 'Implications of artificial intelligence for medical education', The Lancet Digital Health, 2(3), pp.e111-e112, 2020.

Reardon, S. 'Rise of robot radiologists', Nature, 576(7787), pp.S54-S54, 2019.

Recht, M.P., Dewey, M., Dreyer, K., Langlotz, C., Niessen, W., Prainsack, B. and Smith, J.J. 'Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations', European Radiology, pp.1-9, 2020.

Redlich R, Almeida JJ, Grotegerd D, Opel N, Kugel H, Heindel W, et al. 'Brain morphometric biomarkers distinguishing unipolar and bipolar depression: A voxel-based morphometry—Pattern classification approach', JAMA Psychiatry; 71:1222–1230, 2014.

Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. 'Forecasting the onset and course of mental illness with Twitter data', Sci Rep. 7(1):13006, 2017.

Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, 2015.

Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. 'Algorithmic impact assessments: A practical framework for public agency accountability', AI Now Institute, pp.1-22, 2018.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L. 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', AI & SOCIETY, pp.1-19, 2020.

Roski J, Chapman W, Heffner J, Trivedi R, Del Fiol G, Kukafka R, Bleicher Estiri OH, Klann J, Pierce J. 'How artificial intelligence is changing health and health care'. In Artificial Intelligence in Health Care: The hope, the hype, the promise, the peril. Editors: Matheny M, Israni ST, Ahmed M, Whicher D. Washington, DC: National Academy of Medicine, 2019.

Samulowitz A, Gremyr I, Eriksson E, Hensing G. ''Brave Men' and 'Emotional Women': A Theory-Guided Literature Review on Gender Bias in Health Care and Gendered Norms towards Patients with Chronic Pain', Pain Res Manag. 2018;2018:6358624, 2018.

Sapci AH, Sapci HA. 'Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review', JMIR Aging ;2(2):e15429, 2019.

Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H.P., Janda, M., Condon, J.J., Oakden-Rayner, L., Palmer, L.J., Keel, S. and van Wijngaarden, P. 'A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology', Scientific Reports, 11(1), pp.1-10, 2021.

Schrider DR, Kern AD. 'Supervised machine learning for population genetics: a new paradigm', Trends Genet. 34:301–12, 2018.

Schwalbe N, Wahl B. 'Artificial intelligence and the future of global health', Lancet; 395(10236):1579-1586, 2020.

Schwartz WB. 'Medicine and the computer: the promise and problems of chang', N Engl J Med. 1970;283(23):1257-1264, 2020.

Scott, I., Carter, S. and Coiera, E. 'Clinician checklist for assessing suitability of machine learning applications in healthcare', BMJ Health & Care Informatics, 28(1), 2021.

Secretary-General of the OECD. Tackling wasteful spending on health, OECD Publishing, Paris, 2017.

Secretary-General of the OECD. Trustworthy AI in health. Background paper for the G20 AI Dialogue, Digital Economy Task Force, 2020.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y. and Ghassemi, M. 'CheXclusion: Fairness gaps in deep chest X-ray classifiers' BIOCOMPUTING 2021: Proceedings of the Pacific Symposium (pp. 232-243), 2021.

Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R. and Bakas, S. 'Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data', Scientific Reports, 10(1), pp.1-12, 2020.

Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. 'DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning' Sci Rep; 9:1879, 2019.

Shin EK, Mahajan R, Akbilgic O, Shaban-Nejad A. 'Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits', NPJ Digit Med; 1:50, 2018.

Shortliffe EH, Sepúlveda MJ. 'Clinical decision support in the era of artificial intelligence', JAMA.;320(21):2199-2200, 2018.

Shukla, 2016; pp. 2303–40. Dordrecht, Neth.: Springer.

Simpson S, Kay FU, Abbara S, et al. Radiological Society of North America Expert consensus statement on reporting chest CT findings related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA [E-pub ahead of print, 2020 Apr 28]. J Thorac Imaging 2020.

Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. 'Artificial intelligence-enhanced electrocardiography in cardiovascular disease management', Nat Rev Cardiol.;18:465–478, 2021.

Sit, C., Srinivasan, R., Amlani, A., Muthuswamy, K., Azam, A., Monzon, L. and Poon, D.S. 'Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey', Insights into Imaging, 11(1), p.14, 2020.

Smith, H. 'Clinical AI: opacity, accountability, responsibility and liability', AI & SOCIETY, pp.1-11, 2020.

Sornapudi S, Stanley RJ, Stoecker WV, Almubarak H, Long R, Antani S, Thoma G, Zuna R, Frazier SR. 'Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels', J Pathol Inform; 9:5, 2018.

Stanford University, Human-Centered Artificial Intelligence, https://hai.stanford.edu/, accessed November 2021.

Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. 'Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease.', PLoS One 13(8):e0202344, 2018.

Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, Thng F, Peng L, Stumpe MC.' Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer', Am J Surg Pathol. ;42(12):1636-1646, 2018.

Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ. 'A Deep Learning Approach to Antibiotic Discovery', Cell. 180(4):688-702.e13, 2020.

Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A, Del Giudice C, Tesorio P, and Rusciano M. 'Precision and Personalized Medicine: How Genomic Approach Improves the Management of Cardiovascular and Neurodegenerative Disease', Genes. 11(7):747, 2020.

Stylianou N, Fackrell R, Vasilakis C. 'Are medical outliers associated with worse patient outcomes? A retrospective study within a regional NHS hospital using routine data', BMJ Open 7. e015676, 2017.

Subbaswamy, A. and Saria, S. 'From development to deployment: dataset shift, causality, and shift-stable models in health AI', Biostatistics, 21(2), pp.345-352, 2020.

Sydow D, Burggraaff L, Szengel A, van Vlijmen HWT, AP IJ, et al.' Advances and challenges in computational target prediction', J. Chem. Inf. Model. 59:1728–42, 2019.

Tanguay-Sela, M., Benrimoh, D., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., Fratila, R., Parikh, S., Karp, J., Heller, K. and Vahia, I. 'Evaluating the Usability and Impact of an Artificial Intelligence-Powered Clinical Decision Support System for Depression Treatment', Biological Psychiatry, 87(9), p.S171, 2020.

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment, ALTAI, European Commission, https://op.europa.eu/es/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1 2020.

The Assessment List for Trustworthy Artificial Intelligence, https://altai.insight-centre.org, accessed November 2021.

The World Bank, 'Maternal mortality ratio (modeled estimate, per 100,000 live births) – European Union', https://data.worldbank.org/indicator/SH.STA.MMRT?locations=EU, last accessed December 2021.

Tjoa, E. and Guan, C. 'A survey on explainable artificial intelligence (xai): Toward medical xai', IEEE Transactions on Neural Networks and Learning Systems, 2020.

Tlapa D, Zepeda-Lugo CA, Tortorella GL, Baez-Lopez YA, Limon-Romero J, Alvarado-Iniesta A, Rodriguez-Borbon MI. 'Effects of Lean Healthcare on Patient Flow: A Systematic Review', Value Health. 23(2):260-273, 2020.

Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. 'A clinically applicable approach to continuous prediction of future acute kidney injury', Nature. 572(7767):116–9, 2019.

Topol, EJ, 'High-performance medicine: the convergence of human and artificial intelligence' Nature Medicine, 25(1), 44–56, 2019.

TRIPOD, www.tripod-statement.org, accessed November 2021.

Tutt, A.. 'An FDA for algorithms', Admin. L. Rev., 69, p.83, 2017.

U.S. Food and Drug Administration (FDA). Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD), 2019.

U.S. Food and Drug Administration (FDA), 'Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback'.

U.S. Food and Drug Administration (FDA). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, 2021

United Nations Educational, Scientific and Cultural Organization (UNESCO). Artificial Intelligence and Gender Equality: Key Findings of UNESCO's Global Dialogue, 2020.

United Nations News. 'More women and girls needed in the sciences to solve world's biggest challenges', February 2019.. https://news.un.org/en/story/2019/02/1032221

Viceconti, M, Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J. Musuamba Tshinanu, F. 'In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products', Methods 185; 120-127, 2021.

Vijayan V, Connolly J, Condell J, McKelvey N and Gardiner P. Review of Wearable Devices and Data Collection Considerations for Connected Health. Sensors. 2021; 21(16): 5589.

Vyas, D.A., et al. 'Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms', The New England Journal of Medicine (383), pp. 874-882., 2020.

Wager TD, Woo CW. 'Imaging biomarkers and biotypes for depression', Nat Med. 23(1):16-17, 2017.

Walsh CG, Ribeiro JD, Franklin JC. 'Predicting risk of suicide attempts over time through machine learning', Clin Psychol Sci; 5, 457–469, 2017.

Wanless D. 'Securing Good Health for the Whole Population', HM Treasury; 2004.

Westergaard, D., Moseley, P., Sørup, F.K.H., Baldi, P. and Brunak, S. 'Population-wide analysis of differences in disease progression patterns in men and women', Nature communications, 10(1), pp.1-14, 2019.

Whitby B. 'Automating medicine the ethical way', In: Pontier M (ed) Rysewyk Machine Medical Ethics (Intelligent Systems, Control and Automation: Science and Engineering). Springer, Switzerland, 2015.

Wiggers, K. 'Google's breast cancer-predicting AI research is useless without transparency, critics say', VentureBeat, 14 October 2020. https://venturebeat.com/2020/10/14/googles-breast-cancer-predicting-ai-research-is-useless-without-transparency-critics-say/.

Williams, R. 'Lack of transparency in AI breast cancer screening study 'could lead to harmful clinical trials', scientists say', iNews UK, 14 October 2020.

Wolff, J., Pauling, J., Keck, A. and Baumbach, J. 'The economic impact of artificial intelligence in health care: systematic review', Journal of Medical Internet Research, 22(2), p.e16866, 2020.

Wood, A., Najarian, K. and Kahrobaei, D. 'Homomorphic encryption for machine learning in medicine and bioinformatics', ACM Computing Surveys (CSUR), 53(4), pp.1-35, 2020.

World Health Organization (WHO). Depression in Europe: facts and figures, 2021a. https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/news/news/2012/10/depression-in-europe/depression-in-europe-facts-and-figures

World Health Organization (WHO). Ethics and governance of artificial intelligence for health: WHO guidance, 2021b.

World Health Organization (WHO). Global strategy on human resources for health: workforce 2030, Geneva, 2016. https://www.who.int/hrh/resources/pub_globstrathrh-2030/en/

Xu, W. 'Toward human-centered AI: a perspective from human-computer Interactions', 26(4), pp.42-46, 2019.

Yang, G., Ye, Q., & Xia, J. 'Unbox the Black box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Mini-Review, Two Showcases and Beyond' ArXiv, abs/2102.01998, 2021.

Yazdavar AH, Mahdavinejad MS, Bajaj G, Romine W, Sheth A, Monadjemi AH, Thirunarayan K, Meddar JM, Myers A, Pathak J, Hitzler P. 'Multimodal mental health analysis in social media', PLoS One; 15(4):e0226248, 2020.

Yu, K.H. and Kohane, I.S. 'Framing the challenges of artificial intelligence in medicine', BMJ Quality & Safety, 28(3), pp.238-241, 2019.

Zange L, Muehlberg F, Blaszczyk E, Schwenke S, Traber J, Funk S and Schulz-Menger J. 'Quantification in cardiovascular magnetic resonance: agreement of software from three different vendors on assessment of left ventricular function, 2D flow and parametric mapping', Journal of Cardiovascular Magnetic Resonance; 21:12, 2019.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and Oermann, E.K. 'Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study', PLoS Medicine, 15(11), p.e1002683, 2018.

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

Zhang L, Tan J, Han D, Zhu H. 'From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today*; 22(11):1680–85, 2017.

Zhao L, Wang W, Sedykh A, Zhu H. 'Experimental errors in QSAR modeling sets: What we can do and what we cannot do', ACS Omega, 2:2805–12, 2017.

Zhu H, Zhang J, Kim MT, Boison A, Sedykh A, Moran K. 'Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants', Chem. Res. Toxicol; 27:1643–51, 2014.

Zhu H. 'Big Data and Artificial Intelligence Modeling for Drug Discover', Annu Rev Pharmacol Toxicol. Jan 6;60:573-589, 2020.

In recent years, the use of artificial intelligence (AI) in medicine and healthcare has been praised for the great promise it offers, but has also been at the centre of heated controversy. This study offers an overview of how AI can benefit future healthcare, in particular increasing the efficiency of clinicians, improving medical diagnosis and treatment, and optimising the allocation of human and technical resources.

The report identifies and clarifies the main clinical, social and ethical risks posed by AI in healthcare, more specifically: potential errors and patient harm; risk of bias and increased health inequalities; lack of transparency and trust; and vulnerability to hacking and data privacy breaches.

The study proposes mitigation measures and policy options to minimise these risks and maximise the benefits of medical AI, including multi-stakeholder engagement through the AI production lifetime, increased transparency and traceability, in-depth clinical validation of AI tools, and AI training and education for both clinicians and citizens.

QA-07-22-328-EN-N

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

ISBN 978-92-846-9456-3 | doi: 10.2861/568473 | QA-07-22-328-EN-N