FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir, Alejandro F Frangi, Antonio R Porras, Ben Glocker, Celia Cintas, Curtis P Langlotz, Eva Weicken, Folkert W Asselbergs, Fred Prior, Gary S Collins, et al.

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir,[1,2] Alejandro F Frangi,[3,4] Antonio R Porras,[5] Ben Glocker,[6] Celia Cintas,[7] Curtis P Langlotz,[8] Eva Weicken,[9] Folkert W Asselbergs,[10,11] Fred Prior,[12] Gary S Collins,[13] Georgios Kaissis,[14] Gianna Tsakou,[15] Irène Buvat,[16] Jayashree Kalpathy-Cramer,[17] John Mongan,[18] Julia A Schnabel,[19] Kaisar Kushibar,[1] Katrine Riklund,[20] Kostas Marias,[21] Lameck M Amugongo,[22] Lauren A Fromont,[23] Lena Maier-Hein,[24] Leonor Cerdá-Alberich,[25] Luis Martí-Bonmatí,[26] M Jorge Cardoso,[27] Maciej Bobowicz,[28] Mahsa Shabani,[29] Manolis Tsiknakis,[21] Maria A Zuluaga,[30] Marie-Christine Fritzsche,[31] Marina Camacho,[1] Marius George Linguraru,[32] Markus Wenzel,[9] Marleen De Bruijne,[33] Martin G Tolsgaard,[34] Melanie Goisauf,[35] Mónica Cano Abadía,[35] Nikolaos Papanikolaou,[36] Noussair Lazrak,[1] Oriol Pujol,[1] Richard Osuala,[1] Sandy Napel,[37] Sara Colantonio,[38] Smriti Joshi,[1] Stefan Klein,[33] Susanna Aussó,[39] Wendy A Rogers,[40] Zohaib Salahuddin,[41] Martijn P A Starmans[33];
on behalf of the FUTURE-AI Consortium

For numbered affiliations see end of the article

Correspondence to: K Lekadir
karim.lekadir@ub.edu
(ORCID 0000-0002-9456-1612)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2025;388:e081554
http://dx.doi.org/10.1136/bmj-2024-081554

Accepted: 10 January 2025

Despite major advances in artificial intelligence (AI) research for healthcare, the deployment and adoption of AI technologies remain limited in clinical practice. This paper describes the FUTURE-AI framework, which provides guidance for the development and deployment of trustworthy AI tools in healthcare. The FUTURE-AI Consortium was founded in 2021 and comprises 117 interdisciplinary experts from 50 countries representing all continents, including AI scientists, clinical researchers, biomedical ethicists, and social scientists. Over a two year period, the FUTURE-AI guideline was established through consensus based on six guiding principles—fairness, universality, traceability, usability, robustness, and explainability. To operationalise trustworthy AI in healthcare, a set of 30 best practices were defined, addressing technical, clinical, socioethical, and legal dimensions. The recommendations cover the entire lifecycle of healthcare AI, from design, development, and validation to regulation, deployment, and monitoring.

SUMMARY POINTS

- Despite major advances in medical artificial intelligence (AI) research, clinical adoption of emerging AI solutions remains challenging owing to limited trust and ethical concerns
- The FUTURE-AI Consortium unites 117 experts from 50 countries to define international guidelines for trustworthy healthcare AI
- The FUTURE-AI framework is structured around six guiding principles: fairness, universality, traceability, usability, robustness, and explainability
- The guideline addresses the entire AI lifecycle, from design and development to validation and deployment, ensuring alignment with real world needs and ethical requirements
- The framework includes 30 detailed recommendations for building trustworthy and deployable AI systems, emphasising multistakeholder collaboration
- Continuous risk assessment and mitigation are fundamental, addressing biases, data variations, and evolving challenges during the AI lifecycle
- FUTURE-AI is designed as a dynamic framework, which will evolve with technological advancements and stakeholder feedback

Introduction

In the field of healthcare, artificial intelligence (AI)—that is, algorithms with the ability to self-learn logic—and data interactions have been increasingly used to develop computer aided models, for example, disease diagnosis, prognosis, prediction of therapy response or survival, and patient stratification.[1] Despite major advances, the deployment and adoption of AI technologies remain limited in real world clinical practice. In recent years, concerns have been raised about the technical, clinical, ethical, and societal risks associated with healthcare AI.[2 3] In particular, existing research has shown that AI tools in healthcare can be prone to errors and patient harm, biases and increased health inequalities, lack of transparency and accountability, as well as data privacy and security breaches.[4-8]

To increase adoption in the real world, it is essential that AI tools are trusted and accepted by patients, clinicians, health organisations, and authorities. However, there is an absence of clear, widely accepted guidelines on how healthcare AI tools should be designed, developed, evaluated, and deployed to be trustworthy—that is, technically robust, clinically safe,

ethically sound, and legally compliant (see glossary in appendix table 1).[9] To have a real impact at scale, such guidelines for responsible and trustworthy AI must be obtained through wide consensus involving international and interdisciplinary experts.

In other domains, international consensus guidelines have made lasting impacts. For example, the FAIR guideline[10] for data management has been widely adopted by researchers, organisations, and authorities, as the principles provide a structured framework for standardising and enhancing the tasks of data collection, curation, organisation, and storage. Although it can be argued that the FAIR principles do not cover every aspect of data management because they focus more on findability, accessibility, interoperability, and reusability of the data, and less on privacy and security, they delivered a code of practice that is now widely accepted and applied.

AI in healthcare has unique properties compared with other domains, such as the special trust relation between doctors and patients, because patients themselves generally do not have the opportunity to objectively assess the diagnosis and treatment decisions of doctors. This dynamic underscores the need for AI systems to be not only technically robust and clinically safe, but also ethically sound and transparent, ensuring that they complement the trust patients place in their healthcare providers. However, compared with non-AI tools, the highly complicated underlying data processing frequently comes with a lack of transparency into the exact working mechanisms. Unlike medical equipment, AI currently lacks universally accepted measures for quality assurance. Compared with chat assistants and synthetic image generators that receive increased public interaction, healthcare is a more sensitive domain where errors can have major consequences. Addressing these specific gaps for the healthcare domain is therefore crucial for trustworthy AI.

Initial efforts have focused on providing recommendations for the reporting of AI studies for different medical domains or clinical tasks (eg, TRIPOD+AI,[11] CLAIM,[12] CONSORT-AI,[13] DECIDE-AI,[14] PROBAST-AI,[15] CLEAR[16]). These guidelines do not provide best practices for the actual development and deployment of the AI tools, but promote standardised and complete reporting of their development and evaluation. Recently, several researchers have published promising ideas on possible best practices for healthcare AI.[17-24] However, these proposals have not been established through wide international consensus and do not cover the whole lifecycle of healthcare AI (ie, from design, development, and validation to deployment, usage, and monitoring).

In other initiatives, the World Health Organization published a report focused on key ethical and legal challenges and considerations. Because it was intended for health ministries and governmental agencies, it did not explore the technical and clinical aspects of trustworthy AI.[25] Likewise, Europe's High-Level Expert Group on Artificial Intelligence established a comprehensive self-assessment checklist for AI developers. However, it covered AI in general and did not address the unique risks and challenges of AI in medicine and healthcare.[26]

This paper addresses an important gap in the field of healthcare AI by delivering the first structured and holistic guideline for trustworthy and ethical AI in healthcare, established through wide international consensus and covering the entire lifecycle of AI. The FUTURE-AI Consortium was started in 2021 and currently comprises 117 international and interdisciplinary experts from 50 countries (fig 1), representing all continents (Europe, North America, South America, Asia, Africa, and Oceania). Additionally, the members represent a variety of disciplines (eg, data science, medical research, clinical medicine, computer engineering, medical ethics, social sciences) and data domains (eg, radiology, genomics,

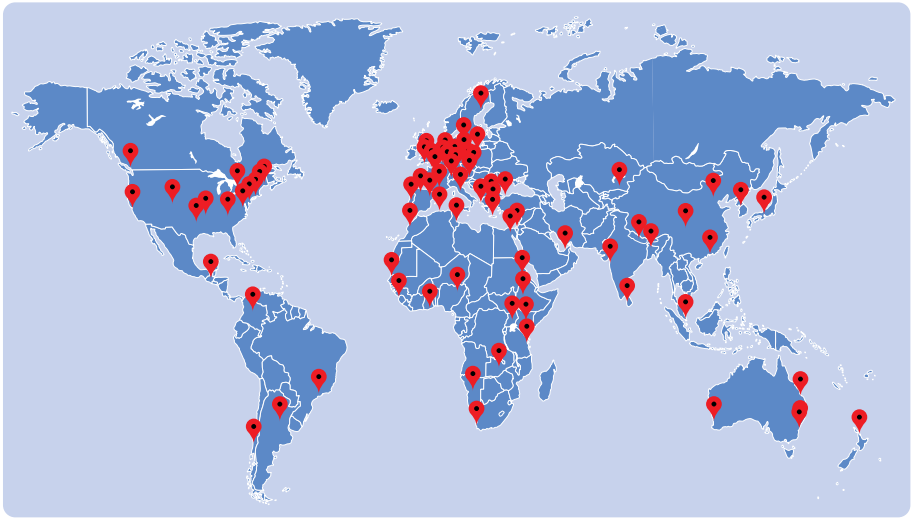

Fig 1 | Geographical distribution of the multidisciplinary experts

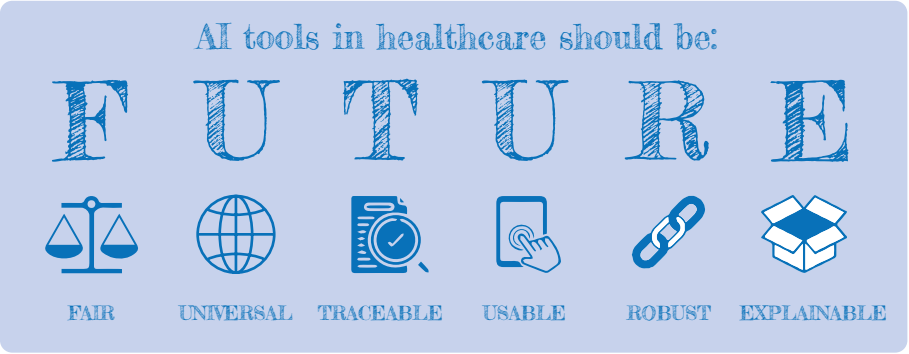

Fig 2 | Organisation of the FUTURE-AI framework for trustworthy artificial intelligence (AI) according to six guiding principles—fairness, universality, traceability, usability, robustness, and explainability

mobile health, electronic health records, surgery, pathology). To develop the FUTURE-AI framework, we drew inspiration from the FAIR principles for data management, and defined concise recommendations organised according to six guiding principles—fairness, universality, traceability, usability, robustness, and explainability (fig 2).

Methods

FUTURE-AI is a structured framework that provides guiding principles and step-by-step recommendations for operationalising trustworthy and ethical AI in healthcare. This guideline was established through international consensus over a 24 month period using a modified Delphi approach.[27 28] The process began with the definition of the six core guiding principles, followed by an initial set of recommendations, which were then subjected to eight rounds of extensive feedback and iterative discussions aimed at reaching consensus. We used two complementary methods to aggregate the results: a quantitative approach, which involved analysing the voting patterns of the experts to identify areas of consensus and disagreement; and a qualitative approach, focusing on the synthesis of feedback and discussions based on recurring themes or new insights raised by several experts.

Definition of FUTURE-AI guiding principles: To develop a user friendly guideline for trustworthy AI in medicine, we used the same approach as in the FAIR guideline, based upon a minimal set of guiding principles. Defining overarching guiding principles facilitates streamlining and structuring of best practices, as well as implementation by future end users of the FUTURE-AI guideline.

To this end, we first reviewed the existing literature in healthcare AI, with a focus on trustworthy AI and related topics in healthcare, such as responsible AI, ethical AI, AI deployment, and terms relating to the six principles identified later. Additional searches were performed for related guidelines, for example, for AI reporting, AI evaluation, and guidelines or position statements from relevant (public) bodies such as the EU, the United States Food and Drug Administration (FDA), and WHO. This review enabled us to identify a wide range of requirements and dimensions often cited as essential for trustworthy AI.[29 30] Throughout the following rounds, the literature review was iteratively expanded based on the advice by experts and widening of the scope, see round 3.

As table 1 shows, these requirements were then thematically grouped, leading to our definition of the six core principles (ie, fairness, universality, traceability, usability, robustness, and explainability), which were arranged to form an easy-to-remember acronym (FUTURE-AI).

Round 1: Definition of an initial set of recommendations

Six working groups composed of three experts each (including clinicians, data scientists, and computer engineers) were created to explore the six guiding principles separately. The experts were recruited from five European projects (EuCanImage, ProCAncer-I, CHAIMELEON, PRIMAGE, INCISIVE), which together formed the AI for Health Imaging (AI4HI) network. By using "AI for medical imaging" as a common use case, each working group conducted a thorough literature review, then proposed a definition of the guiding

Table 1 | Clustering of trustworthy artificial intelligence (AI) requirements and selection of FUTURE-AI guiding principles

Clusters of requirements	Core principles
1. Fairness, diversity, inclusiveness, non-discrimination, unbiased AI, equity	Fairness
2. Generalisability, adaptability, interoperability, applicability, universality	Universality
3. Traceability, monitoring, continuous learning, auditing, accountability	Traceability
4. Human centred AI, user engagement, usability, accessibility, efficiency	Usability
5. Robustness, reliability, resilience, safety, security	Robustness
6. Transparency, explainability, interpretability, understandability	Explainability

principle in question, together with an initial list of best practices (between 6 and 10 for each guiding principle).

Subsequently, the working groups engaged in an iterative process of refining these preliminary recommendations through online meetings and by email exchanges. At this stage, a degree of overlap and redundancy was identified across recommendations. For example, a recommendation to report any identified bias was initially proposed under both the fairness and traceability principles, while a recommendation to train the AI models with representative datasets appeared under fairness and robustness. After removing the redundancies and refining the formulations, a set of 55 preliminary recommendations was derived and then distributed to a broader panel of experts for further assessment, discussion, and refinement in the next round.

Round 2: Online survey

In this round, the FUTURE-AI Consortium was expanded to 72 members by recruiting new experts, including AI scientists, healthcare practitioners, ethicists, social scientists, legal experts, and industry professionals. The same original group took part in rounds 2–5. Experts were identified from the literature, through networks, and an online search, with selection focusing on underrepresented expertise or demographics. Most of the experts were recruited to complement the original consortium based on academic credentials, geographical location, and under-represented expertise, to ensure a representative consortium in terms of geography and (healthcare) disciplines. We then conducted an online survey to enable the experts to assess each recommendation using five voting options (absolutely essential, very important, of average importance, of little importance, not important at all). The participants were also able to rate the formulation of the recommendation ("I would keep it as it is," "I would refine its definition") and propose modifications. Furthermore, they were able to propose merging recommendations or adding new ones. The survey included a section for free text feedback on the core principles and the overall FUTURE-AI guideline.

The survey responses were quantitatively analysed to assess the consensus level. Recommendations that garnered a high level agreement (>90%) were selected for further discussion. Recommendations that attracted considerable negative feedback, which were particularly those that suggested specific methods over general guidelines, were discarded. The written feedback also prompted the merging of some recommendations, aiming to craft a more concise guideline for easier adoption by future users. Consequently, a revised list of 22 recommendations was derived, along with the identification of 16 contentious points for further discussions.

As part of the survey, we also sought feedback from the experts on the adequacy of these guiding principles in capturing the diverse requirements for trustworthy AI in healthcare. While the consensus among experts was largely affirmative, it was suggested a seventh "general" category was introduced to cover broader issues such as data privacy, societal considerations, and regulatory compliance, and to produce a holistic framework. The best practices in this category are overarching, for example, multistakeholder engagement (general 1) is relevant for all six guiding principles, thereby avoiding repetition for each principle.

Round 3: Feedback on the reduced set of recommendations

The updated version of the guideline from round 2 was distributed to all experts for another round of feedback. This involved assessing both the adequacy and the phrasing of the recommendations. Additionally, we presented the points of contention identified in the survey, encouraging experts to offer their insights on these disagreements. Examples of contentious topics included the recommendation to perform multicentre versus local clinical evaluation, and the necessity (or not) to systematically evaluate the AI tools against adversarial attacks.

The feedback received from the experts played a crucial role in resolving several contentious issues, particularly through the refinement of the recommendations' wording. Moreover, the scope was broadened from "AI in medical imaging" more generally to "AI in healthcare" because we realised most of the recommendations hold for healthcare in general, making the guideline more broadly applicable. As a result, this led to the expansion of the FUTURE-AI guideline to a total of 30 best practices, which included six new recommendations within the "general" category. Areas of disagreement that remained unresolved were carefully documented and summarised for future discussions.

Round 4: Further feedback and rating of the recommendations

The updated recommendations were sent out to the experts for additional feedback, this time in written form, to assess each recommendation's clarity, feasibility, and relevance. This phase allowed for more precise phrasing of the recommendations. As an example, the original recommendation to train AI models with "diverse, heterogeneous data" was refined by using the term "representative data" because many experts argued that representative data more effectively capture the essential characteristics of the populations, while the term heterogeneous is more ambiguous.

Furthermore, we implemented a system to rate each best practice depending on the specific needs and goals of each AI project. A key focus was to make a distinction between healthcare AI tools at the research or proof-of-concept stage and those intended for clinical deployment because they require different levels of compliance. Healthcare AI tools in the research or proof-of-concept stage are typically in their experimental phase and require some flexibility as

Table 2 | List of FUTURE-AI recommendations, together with the expected compliance for both research and deployable artificial intelligence (AI) tools (+: recommended, ++: highly recommended)

Recommendations	Research	Deployable
Fairness		
1. Define any potential sources of bias from an early stage	++	++
2. Collect information on individuals' and data attributes	+	+
3. Evaluate potential biases and, when needed, bias correction measures	+	++
Universality		
1. Define intended clinical settings and cross setting variations	++	++
2. Use community defined standards (eg, clinical definitions, technical standards)	+	+
3. Evaluate using external datasets and/or multiple sites	++	++
4. Evaluate and demonstrate local clinical validity	+	++
Traceability		
1. Implement a risk management process throughout the AI lifecycle	+	++
2. Provide documentation (eg, technical, clinical)	++	++
3. Define mechanisms for quality control of the AI inputs and outputs	+	++
4. Implement a system for periodic auditing and updating	+	++
5. Implement a logging system for usage recording	+	++
6. Establish mechanisms for AI governance	+	++
Usability		
1. Define intended use and user requirements from an early stage	++	++
2. Establish mechanisms for human-AI interactions and oversight	+	++
3. Provide training materials and activities (eg, tutorials, hands-on sessions)	+	++
4. Evaluate user experience and acceptance with independent end users	+	++
5. Evaluate clinical utility and safety (eg, effectiveness, harm, cost-benefit)	+	++
Robustness		
1. Define sources of data variation from an early stage	++	++
2. Train with representative real world data	++	++
3. Evaluate and optimise robustness against real world variations	++	++
Explainability		
1. Define the need and requirements for explainability with end users	++	++
2. Evaluate explainability with end users (eg, correctness, impact on users)	+	+
General		
1. Engage interdisciplinary stakeholders throughout the AI lifecycle	++	++
2. Implement measures for data privacy and security	++	++
3. Implement measures to address identified AI risks	++	++
4. Define adequate evaluation plan (eg, datasets, metrics, reference methods)	++	++
5. Identify and comply with applicable AI regulatory requirements	+	++
6. Investigate and address application specific ethical issues	+	++
7. Investigate and address social and societal issues	+	+

their capabilities are being explored and fine-tuned. In contrast, AI tools intended for clinical deployment will interact directly with patient care and therefore should need higher standards of compliance to ensure they are ethical, safe, and effective. At this point of the process, the consortium members were requested to assess all the recommendations separately for both proof-of-concept and deployable AI tools, and categorise them as either "recommended" or "highly recommended."

Round 5: Feedback on the manuscript

At this stage, with a well developed set of 30 recommendations, the first and last authors of the study drafted the first version of the FUTURE-AI manuscript. The draft manuscript was circulated among the experts, starting a series of iterative feedback sessions to ensure that the FUTURE-AI guideline was articulated with precision and clarity. This process enabled incorporation of diverse perspectives, from clinical, technical, and non-technical experts, hence making the manuscript more reader friendly and accessible to a broad audience. Experts were also able to suggest additional resources or references to substantiate the recommendations further. At this stage, examples of methods were integrated to the manuscript where relevant, aiming to demonstrate the practical implementation of the best practices in real world scenarios.

Round 6: New "external" feedback

In round 6 we invited additional experts (n=44) who had not participated in the initial stages of the study to provide independent feedback. This group was carefully selected to ensure a more diverse representation across the experts (eg, patient advocates, social scientists, regulatory experts), as well as wider geographical diversity (especially across Africa, Latin America, and Asia).

These experts were requested to provide written feedback and express their opinion on each recommendation using a voting system (ie, agree, disagree, neutral, did not understand, no opinion). For most of the recommendations on which no clear agreement was reached, again using consensus level, the primary cause was misinterpretation or unclarity. Therefore, this stage was especially helpful in pinpointing any remaining areas of ambiguity or contention that required further discussions, as well as

in identifying the formulations that needed refinement to ensure the entire guideline is clear and accessible to a diverse audience within the medical AI community.

Round 7: Online consensus meetings

Based on the feedback from previous rounds, we identified a few topics that continued to evoke a degree of contention among experts, particularly concerning the exact wording of certain recommendations. Hence, we convened four online meetings in June 2023 specifically aimed at deepening the discussions around the remaining contentious areas and reaching a final consensus on both the recommendations and their formulations.

These discussions resolved outstanding issues, such as the recommendation to systematically validate AI tools against adversarial attacks, which was considered by many experts as a cybersecurity concern and thus grouped with other related concerns; or the recommendation that the clinical evaluations should be conducted by third parties, which was deemed impractical at scale, especially in resource limited settings. As a result of these consensus meetings, the final list of FUTURE-AI recommendations was established, and their formulations were completed as detailed in table 2.

Round 8: Final consensus vote

The very last step of the process involved a final vote on the derived recommendations, which took place through an online survey. At this stage, the final consortium consisted of 117 experts as more replied to the above recruitments: the original 72 experts from round 2, some of the 44 experts who provided feedback in round 6, and several additional experts. By the end of this process, all the recommendations were approved with less than 5% disagreement among all FUTURE-AI members. The little remaining disagreement mostly originated from whether recommendations should be "recommended" or "highly recommended" for research and deployable tools.

FUTURE-AI guideline

In this section, we provide definitions and justifications for each of the six guiding principles and give an overview of the FUTURE-AI recommendations. Table 2 provides a summary of the recommendations, together with the proposed level of compliance (ie, recommended *v* highly recommended). Note that supplementary table 1 in the appendix presents a glossary of the main terms used in this paper, while supplementary table 2 lists the main stakeholders of relevance to the FUTURE-AI framework.

Fairness

The fairness principle states that AI tools in healthcare should maintain the same performance across individuals and groups of individuals (including under-represented and disadvantaged groups). AI driven medical care should be provided equally for all citizens. Biases in healthcare AI can be due to differences in the attributes of the individuals (eg, sex, gender, age, ethnicity, socioeconomic status, medical conditions) or the data (eg, acquisition site, machines, operators, annotators). As, in practice, perfect fairness might be impossible to achieve, fair AI tools should be developed such that potential AI biases are identified, reported, and minimised as much as possible to achieve ideally the same but at least highly similar performance across subgroups to be considered fair.[31] To this end, three recommendations for fairness are defined in the FUTURE-AI framework.

Fairness 1: Define sources of bias

Bias in healthcare AI is application specific.[32] At the design phase, the interdisciplinary AI development team (see glossary) should identify possible types and sources of bias for their AI tool.[33] These might include group attributes (eg, sex, gender, age, ethnicity, socioeconomic, geography), the medical profiles of the individuals (eg, with comorbidities or disability), as well as human and technical biases during data acquisition, labelling, data curation, or the selection of the input features.

Fairness 2: Collect information on individual and data attributes

To identify biases and apply measures for increased fairness, relevant attributes of the individuals, such as sex, gender, age, ethnicity, risk factors, comorbidities, or disabilities, should be collected. This should be subject to informed consent and approval by ethics committees to ensure an appropriate balance between the benefits of non-discrimination and the risks of reidentification. Measuring similarity of medical profiles should also be included to verify equal treatment (eg, risk factors, comorbidities, biomarkers, anatomical properties[34]). Furthermore, relevant information about the datasets, such as the centres where they were acquired, the machine used, the preprocessing and annotation processes, should be systematically collected to address technical and human biases. When complete data collection is logistically challenging, two alternative approaches can be considered: imputing missing attributes or removing samples with incomplete data. The choice between these methods should be evaluated on a case-by-case basis, considering the specific context and requirements of the AI system.

Fairness 3: Evaluate fairness

When possible—that is, the individuals' and data attributes are available—bias detection methods should be applied by using fairness metrics such as true positive rates, statistical parity, group fairness, and equalised odds.[31 35] To correct for any identified biases, mitigation measures should be tested, such as data resampling, bias free representations, and equalised odds postprocessing,[36-40] to verify their impact on both the tool's fairness and the model's accuracy. Importantly, any remaining bias should be

documented and reported to inform the end users and citizens (see traceability 2).

Universality

The universality principle emphasises that a healthcare AI tool should be generalisable outside the controlled environment where it was built. Specifically, the AI tool should be able to generalise to new patients and new users (eg, new clinicians), and when applicable, to new clinical sites. Depending on the intended radius of application, healthcare AI tools should be as interoperable and as transferable as possible so they can benefit citizens and clinicians at scale. To this end, four recommendations for universality are defined in the FUTURE-AI framework.

Universality 1: Define clinical settings

At the design phase, the development team should specify the clinical settings in which the AI tool will be applied (eg, primary healthcare centres, hospitals, home care, low versus high resource settings, one or several countries), and anticipate potential obstacles to universality (eg, differences in end users, clinical definitions, medical equipment or IT infrastructures across settings).

Universality 2: Use existing standards

To ensure the quality and interoperability of the AI tool, it should be developed based on existing community defined standards. These might include clinical definitions of diseases by medical societies, medical ontologies (eg, Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[41]), data models (eg, Observational Medical Outcomes Partnership (OMOP)[42]), interface standards (eg, Digital Imaging and Communications in Medicine (DICOM), Fast Healthcare Interoperability Resources (FHIR) Health Level Seven (HL7)), data annotation protocols, evaluation criteria,[21] and technical standards (eg, Institute of Electrical and Electronics Engineers (IEEE)[43] or International Organisation for Standardization (ISO)[44]).[21 41-44]

Universality 3: Evaluate using external data

To assess generalisability, technical validation of the AI tools should be performed with external datasets that are distinct from those used for model training.[45] These might include reference or benchmarking datasets that are representative for the task in question (ie, approximating the expected real world variations). Except for AI tools intended for single centres, the clinical evaluation studies should be performed at several sites to assess performance and interoperability across clinical workflows.[46] If the tool's generalisability is limited, mitigation measures (eg, transfer learning or domain adaptation) should be applied and tested.

Universality 4: Evaluate local clinical validity

Clinical settings vary in many aspects, such as populations, equipment, clinical workflows, and end users. Therefore, to ensure trust at each site, the AI tools should be evaluated for their local clinical validity.[17] In particular, the AI tool should fit the local clinical workflows and perform well on the local populations. If the performance is decreased when evaluated locally, recalibration of the AI model should be performed and tested (eg, through model fine tuning).

Traceability

The traceability principle states that medical AI tools should be developed together with mechanisms for documenting and monitoring the complete trajectory of the AI tool, from development and validation to deployment and usage. This will increase transparency and accountability by providing detailed and continuous information on the AI tools during their lifetime to clinicians, healthcare organisations, citizens and patients, AI developers, and relevant authorities. AI traceability will also enable continuous auditing of AI models,[47] identify risks and limitations, and update the AI models when needed.

Traceability 1: Implement risk management

Throughout the AI tool's lifecycle, the multidisciplinary development team shall analyse potential risks, assess each risk's likelihood, effects and risk-benefit balance, define risk mitigation measures, monitor the risks and mitigations continuously, and maintain a risk management file. The risks might include those explicitly covered by the FUTURE-AI guiding principles (eg, bias, harm, data breach), but also application specific risks. Other risks to consider include human factors that might lead to misuse of the AI tool (eg, not following the instructions, receiving insufficient training), application of the AI tool to individuals who are not within the target population, use of the tool by others than the target end users (eg, technician instead of physician), hardware failure, incorrect data annotations or input values, and adversarial attacks. Mitigation measures might include warnings to the users, system shutdown, reprocessing of the input data, the acquisition of new input data, or the use of an alternative procedure or human judgment only. Monitoring and reassessment of risk might involve the use of various feedback channels, such as customer feedback and complaints, as well as logged real world performance and issues (see traceability 5).

Traceability 2: Provide documentation

To increase transparency, traceability, and accountability, adequate documentation should be created and maintained for the AI tool,[48] which might include (a) an AI information leaflet to inform citizens and healthcare professionals about the tool's intended use, risks (eg, biases) and instructions for use; (b) a technical document to inform AI developers, health organisations, and regulators about the AI model's properties (eg, hyperparameters), training and testing data, evaluation criteria and results, biases and other limitations, and periodic audits and updates[49-51]; (c) a publication based on existing AI reporting

standards[13 15 52]; and (d) a risk management file (see traceability 1).

Traceability 3: Implement continuous quality control
The AI tool should be developed and deployed with mechanisms for continuous monitoring and quality control of the AI inputs and outputs,[47] such as to identify missing or out-of-range input variables, inconsistent data formats or units, incorrect annotations or data preprocessing, and erroneous or implausible AI outputs. For quality control of the AI decisions, uncertainty estimates should be provided (and calibrated[53]) to inform the end users about the degree of confidence in the results.[54]

Traceability 4: Implement periodic auditing and updating
The AI tool should be developed and deployed with a configurable system for periodic auditing,[47] which should define the datasets and timelines for periodic evaluations (eg, every year). The periodic auditing should enable the identification of data or concept drifts, newly occurring biases, performance degradation or changes in the decision making of the end users.[55] Accordingly, necessary updates to the AI models or AI tools should be applied.[56]

Traceability 5: Implement AI logging
To increase traceability and accountability, an AI logging system should be implemented to trace the user's main actions in a privacy preserving manner, specify the data that are accessed and used, record the AI predictions and clinical decisions, and log any encountered issues. Time series statistics and visualisations should be used to inspect the usage of the AI tool over time.

Traceability 6: Implement AI governance
After deployment, the governance of the AI tool should be specified. In particular, the roles of risk management, periodic auditing, maintenance, and supervision should be assigned, such as to IT teams or healthcare administrators. Furthermore, responsibilities for AI related errors should be clearly specified among clinicians, healthcare centres, AI developers, and manufacturers. Accountability mechanisms should be established, incorporating both individual and collective liability, alongside compensation and support structures for patients affected by AI errors.

Usability

The usability principle states that the end users should be able to use an AI tool to achieve a clinical goal efficiently and safely in their real world environment. On one hand, this means that end users should be able to use the AI tool's functionalities and interfaces easily and with minimal errors. On the other hand, the AI tool should be clinically useful and safe, for example, improve the clinicians' productivity and/or lead to better health outcomes for the patients and avoid harm. To this end, five recommendations for usability are defined in the FUTURE-AI framework.

Usability 1: Define user requirements
The AI developers should engage clinical experts, end users (eg, patients, physicians), and other relevant stakeholders (eg, data managers, administrators) from an early stage to compile information on the AI tool's intended use and end user requirements (eg, human-AI interfaces), as well as on human factors that might affect the usage of the AI tool[57] (eg, digital literacy level, age group, ergonomics, automation bias). Special attention should be paid to the fit with the current clinical workflow, including system level implementation of AI and interactions with other (AI) support tools. Using a majority voting strategy among diverse stakeholders to identify the most relevant clinical issues might help to ensure that solutions are broadly applicable rather than tailored to individual preferences.

Usability 2: Define human-AI interactions and oversight
Based on the user requirements, the AI developers should implement interfaces to enable end users to effectively use the AI model, annotate the input data in a standardised manner, and verify the AI inputs and results. Given the high stakes nature of medical AI, human oversight is essential and increasingly required by policy makers and regulators.[17 26] Human-in-the-loop mechanisms should be designed and implemented to perform specific quality checks (eg, to flag biases, errors, or implausible explanations), and to overrule the AI predictions when necessary. Regulations, the benefits of automation, and patient preferences regarding AI autonomy might vary per use case and over time,[58] therefore requiring use case specific human oversight mechanisms and periodic auditing and updates (see traceability 4).

Usability 3: Provide training
To facilitate best usage of the AI tool, minimise errors and harm, and increase AI literacy, the developers should provide training materials (eg, tutorials, manuals, examples) and/or training activities (eg, hands-on sessions) in an accessible format and language, taking into account the diversity of end users (eg, specialists, nurses, technicians, citizens, or administrators).

Usability 4: Evaluate clinical usability
To facilitate adoption, the usability of the AI tool within the local clinical workflows should be evaluated in real world settings with representative and diverse end users (eg, with respect to sex, gender, age, clinical role, digital proficiency, and disability). The usability tests should gather evidence on the user's satisfaction, performance and productivity, and assess human factors that might affect the usage of the AI tool[57] (eg, confidence, learnability, automation bias).

Usability 5: Evaluate clinical utility
The AI tool should be evaluated for its clinical utility and safety. The clinical evaluations of the AI tool

should show benefits for the patient (eg, earlier diagnosis, better outcomes), for the clinician (eg, increased productivity, improved care), and/or for the healthcare organisation (eg, reduced costs, optimised workflows) compared with the current standard of care. Additionally, it is important to show that the AI tool is safe and does not cause harm to individuals (or specific groups), such as through a randomised clinical trial.[59]

Robustness

The robustness principle refers to the ability of a medical AI tool to maintain its performance and accuracy under expected or unexpected variations in the input data. Existing research has shown that even small, imperceptible variations in the input data might lead AI models into incorrect decisions.[60] Biomedical and health data can be subject to major variations in the real world (both expected and unexpected), which can affect the performance of AI tools. Therefore, it is important that healthcare AI tools are designed and developed to be robust against real world variations, and evaluated and optimised accordingly. To this end, three recommendations for robustness are defined in the FUTURE-AI framework.

Robustness 1: Define sources of data variations

At the design phase, the development team should first define robustness requirements for the AI tool in question by making an inventory of the sources of variation that might affect the AI tool's robustness in the real world. These might include differences in equipment, technical fault of a machine, data heterogeneities during data acquisition or annotation, and/or adversarial attacks.[60]

Robustness 2: Train with representative data

Clinicians, citizens, and other stakeholders are more likely to trust the AI tool if it is trained on data that adequately represent the variations encountered in real world clinical practice.[61] Therefore, the training datasets should be carefully selected, analysed, and enriched according to the sources of variation identified at the design phase (see robustness 1). Training with representative datasets also allows for improvement of other principles, for example, more representative bias estimation and mitigation for fairness.

Robustness 3: Evaluate robustness

Evaluation studies should be implemented to evaluate the AI tool's robustness (eg, stress tests, repeatability tests[62]) under conditions that reflect the variations of real world clinical practice. These might include data, equipment, technician, clinician, patient, and centre related variations. Depending on the results, mitigation measures should be implemented and tested to optimise the robustness of the AI model, such as regularisation,[63] data augmentation,[64] data harmonisation,[65] or domain adaptation.[66]

Explainability

The explainability principle states that medical AI tools should provide clinically meaningful information about the logic behind the AI decisions. Although medicine is a high stake discipline that requires transparency, reliability and accountability, machine learning techniques often produce complex models that are black box in nature. Explainability is considered desirable from a technological, medical, ethical, legal, and patient perspective.[67] It enables end users to interpret the AI model and outputs, understand the capacities and limitations of the AI tool, and intervene when necessary, such as to decide to use it or not. However, explainability is a complex task that has challenges that need to be carefully addressed during AI development and evaluation to ensure that AI explanations are clinically meaningful and beneficial to end users.[68] Two recommendations for explainability are defined in the FUTURE-AI framework.

Explainability 1: Define explainability needs

At the design phase, it should be established with end users and domain experts if explainability is required for the AI tool. If so, the specific requirements for explainability should be defined with representative experts and end users, including (a) the goal of the explanations (eg, global description of the model's behaviour *v* local explanation of each AI decision); (b) the most suitable approach for AI explainability[69]; and (c) the potential limitations to anticipate and monitor (eg, over-reliance of the end users on the AI decision[68]).

Explainability 2: Evaluate explainability

The explainable AI methods should be evaluated, first quantitatively by using computational methods to assess the correctness of the explanations,[70 71] then qualitatively with end users to assess their impact on user satisfaction, confidence, and clinical performance.[72] The evaluations should also identify any limitations of the AI explanations (eg, they are clinically incoherent[73] or sensitive to noise or adversarial attacks,[74] they unreasonably increase the confidence in the AI generated results[75]).

General recommendations

Finally, seven general recommendations are defined in the FUTURE-AI framework, which apply across all principles of trustworthy AI in healthcare.

General 1: Engage stakeholders continuously

Throughout the AI tool's lifecycle, the AI developers should continuously engage with interdisciplinary stakeholders, such as healthcare professionals, citizens, patient representatives, expert ethicists, data managers, and legal experts. This interaction will facilitate the understanding and anticipation of the needs, obstacles, and pathways towards acceptance and adoption. Methods to engage stakeholders might include working groups, advisory boards, one-to-one interviews, cocreation meetings, and surveys.

General 2: Ensure data protection

Adequate measures to ensure data privacy and security should be put in place throughout the AI lifecycle. These might include privacy enhancing techniques (eg, differential privacy, encryption), data protection impact assessment, and appropriate data governance after deployment (eg, logging system for data access, see traceability 5). If deidentification is implemented (eg, pseudonymisation, k-anonymity), the balance between the health benefits for citizens and the risks for reidentification should be carefully assessed and considered. Furthermore, the manufacturers and deployers should implement and regularly evaluate measures for protecting the AI tool against malicious or adversarial attacks, such as by using system level cybersecurity solutions or application specific defence mechanisms (eg, attack detection or mitigation).[76]

General 3: Implement measures to address AI risks

At the development stage, the development team should define an AI modelling plan that is aligned with the application specific requirements. After implementing and testing a baseline AI model, the AI modelling plan should include mitigation measures to address the challenges and risks identified at the design stage (see fairness 1 to explainability 1). These might include measures to enhance robustness to real world variations (eg, regularisation, data augmentation, data harmonisation, domain adaptation), ensure generalisability across settings (eg, transfer learning, knowledge distillation), and correct for biases across subgroups (eg, data resampling, bias free representation, equalised odds post processing).

General 4: Define an adequate AI evaluation plan

To increase trust and adoption, an appropriate evaluation plan should be defined, including test data, metrics, and reference methods. First, adequate test data should be selected to assess each dimension of trustworthy AI. In particular, the test data should be well separated from the training to prevent data leakage.[77] Furthermore, adequate evaluation metrics should be carefully selected, taking into account their benefits and potential flaws.[78] Finally, benchmarking with respect to reference AI tools or standard practice should be performed to enable comparative assessment of model performance.

General 5: Comply with AI regulations

The development team should identify the applicable AI regulations, which vary by jurisdiction and over time. For example, in the EU, the recent AI Act classifies all AI tools in healthcare as high risk, hence they must comply with safety, transparency and quality obligations, and undergo conformity assessments. Identifying the applicable regulations at an early stage enables regulatory obligations to be anticipated based on the AI tool's intended classification and risks.

General 6: Investigate application specific ethical issues

In addition to the well known ethical issues that arise in medical AI (eg, privacy, transparency, equity, autonomy), AI developers, domain specialists, and professional ethicists should identify, discuss, and address all application specific ethical, social, and societal issues as an integral part of the development and deployment of the AI tool.[79]

General 7: Investigate social and environmental issues

In addition to clinical, technical, legal, and ethical implications, a healthcare AI tool might have specific social and environmental issues. These will need to be considered and addressed to ensure a positive impact for the AI tool on citizens and society. Regulatory agencies or independent organisations could provide certifications or marks for AI tools that meet certain sustainability criteria. This approach can encourage transparency, give insight on an AI tool's environmental impact, and highlight those that adopt environmentally friendly practices. Relevant issues might include the impact of the AI tool on the working conditions and power relations, on the new skills (or deskilling) of the healthcare professionals and citizens,[80] and on future interactions between citizens, health professionals, and social careers. Furthermore, for environmental sustainability, AI developers should consider strategies to reduce the carbon footprint of the AI tool.[81] To enable the implementation of the FUTURE-AI framework in practice, we provide a step-by-step guide by embedding the recommended best practices in chronological order across the key phases of an AI tool's lifecycle, as shown in figure 3 and as follows:

- The design phase is initiated with a human centred, risk aware strategy by engaging all relevant stakeholders and conducting a comprehensive analysis of clinical, technical, ethical, and social requirements, leading to a list of specifications and a list of risks to monitor (eg, potential biases, lack of robustness, generalisability, and transparency).
- Accordingly, the development phase prioritises the collection of representative datasets for effective training and testing, ensuring they reflect variations across the intended settings, equipment, protocols, and populations as identified previously. Furthermore, an adequate AI development plan is defined and implemented given the identified requirements and risks, including mitigation strategies and human centred mechanisms to meet the initial design's functional and ethical requirements.
- Subsequently, the validation phase comprehensively examines all dimensions of trustworthy AI, including system performance but also robustness, fairness, generalisability, and explainability, and concludes with the generation of all necessary documentation.

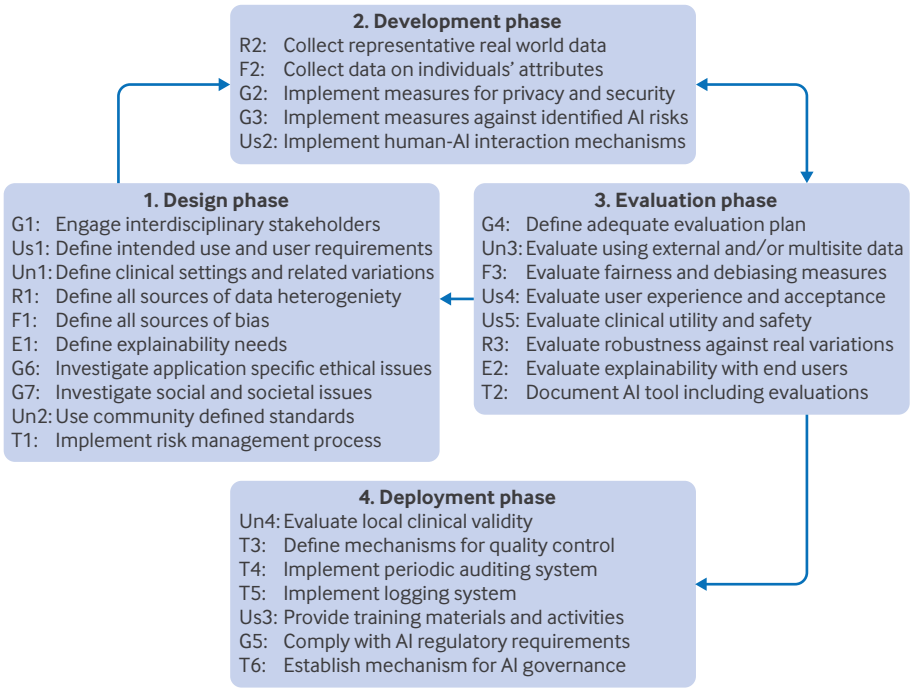

Fig 3 | Embedding the FUTURE-AI best practices into an agile process throughout the artificial intelligence (AI) lifecycle. E=explainability; F=fairness; G=general; R=robustness; T=traceability; Un=universality; Us=usability

- Finally, the deployment phase is dedicated to ensuring local validity, providing training, implementing monitoring mechanisms, and ensuring regulatory compliance for adoption in real world healthcare practice.

Operationalisation of FUTURE-AI

In this section, we provide a detailed list of practical steps for each recommendation, accompanied by specific examples of approaches and methods that can be applied to operationalise each step towards

Table 3 | Practical steps and examples to implement FUTURE-AI recommendations during design phase

Recommendations	Operations	Examples
Engage interdisciplinary stakeholders (general 1)	Identify all relevant stakeholders	Patients, GPs, nurses, ethicists, data managers[82 83]
	Provide information on the AI tool and AI	Educational seminars, training materials, webinars[84]
	Set up communication channels with stakeholders	Regular group meetings, one-to-one interviews, virtual platform[85]
	Organise cocreation consensus meetings	One day cocreation workshop with n=15 multidisciplinary stakeholders[86]
	Use qualitative methods to gather feedback	Online surveys, focus groups, narrative interviews[87]
Define intended use and user requirements (usability 1)	Define the clinical need and AI tool's goal	Risk prediction, disease detection, image quantification
	Define the AI tool's end users	Patients, cardiologists, radiologists, nurses
	Define the AI model's inputs	Symptoms, heart rate, blood pressure, ECG, image scan, genetic test
	Define the AI tool's functionalities and interfaces	Data upload, AI prediction, AI explainability, uncertainty estimation[88]
	Define requirements for human oversight	Visual quality control, manual corrections[89 90]
	Adjust user requirements for all end user subgroups	According to role, age group, digital literacy level[91]
Define intended clinical settings and cross setting variations (universality 1)	Define the AI tool's healthcare setting(s)	Primary care, hospital, remote care facility, home care
	Define the resources needed at each setting	Personnel (experience, digital literacy), medical equipment (eg, >1.5 T MRI scanner), IT infrastructure
	Specify if the AI tool is intended for high end and/or low resource settings	Facilities with MRI scanners >1.5 T *v* low field MRIs (eg, 0.5 T), high end *v* low cost portable ultrasound[92 93]
	Identify all cross settings variations	Data formats, medical equipment, data protocols, IT infrastructure[94]
Define sources of data heterogeneity (robustness 1)	Engage relevant stakeholders to assess data heterogeneity	Clinicians, technicians, data managers, IT managers, radiologists, device vendors
	Identify equipment related data variations	Differences in medical devices, manufacturers, calibrations, machine ranges (from low cost to high end)[95]
	Identify protocol related data variations	Differences in image sequences, data acquisition protocols,[96] data annotation methods, sampling rates, preprocessing standards
	Identify operator related data variations	Different in experience and proficiency, operator fatigue, subjective judgment, technique variability
	Identify sources of artefacts and noises	Image noise, motion artefacts, signal dropout, sensor malfunction
	Identify context specific data variations	Lower data quality acquisition in emergency units, during high patient volume times

(Continued)

Table 3 | Continued

Recommendations	Operations	Examples
Define any potential sources of bias (fairness 1)	Engage relevant stakeholders to define the sources of bias	Patients, clinicians, epidemiologists, ethicists, social carers[97 98]
	Define standard attributes that might affect the AI tool's fairness	Sex, age, socioeconomic status[99]
	Identify application specific sources of bias beyond standard attributes	Skin colour for skin cancer detection,[100 101] breast density for breast cancer detection[34]
	Identify all possible human biases	Data labelling, data curation[99]
Define the need and requirements for explainability with end users (explainability 1)	Engage end users to define explainability requirements	Clinicians, technicians, patients[102]
	Specify if explainability is necessary	Not necessary for AI enabled image segmentation part, critical for AI enabled diagnosis
	Specify the objectives of AI explainability (if it is needed)	Understanding AI model, aiding diagnostic reasoning, justifying treatment recommendations[103]
	Define suitable explainability approaches	Visual explanations, feature importance, counterfactuals[104]
	Adjust the design of the AI explanations for all end user subgroup	Heatmaps for clinicians, feature importance for patients[105 106]
Investigate ethical issues (general 6)	Consult ethicists on ethical considerations	Ethicists specialised in medical AI and/or in the application domain (eg, paediatrics)[107]
	Assess if the AI tool's design is aligned with relevant ethical values	Right to autonomy, information, consent, confidentiality, equity[107]
	Identify application specific ethical issues	Ethical risks for a paediatric AI tool (eg, emotional impact on children)[108 109]
	Comply with local ethical AI frameworks	AI ethical guidelines from Europe,[2] United Kingdom,[110 111] United Sates,[112] Canada,[113] China,[114] India,[115] Japan,[116 117] Australia,[118] etc
Investigate social and environmental issues (general 7)	Investigate AI tool's social and environmental impact	Workforce displacement, worsened working conditions and relations, deskilling,[80] dehumanisation of care, reduced health literacy, increased carbon footprint,[119] negative public perception[107 120]
	Define mitigations to enhance the AI tool's social and environmental impact	Interfaces for physician-patient communication, workforce training, educational programmes, energy efficient computing practices, public engagement initiatives
	Optimise algorithms, energy efficiency	Develop and use energy efficient algorithms that minimise computational demands. Techniques like model pruning, quantisation, and edge computing can reduce the energy required for AI tasks
	Promote responsible data usage	Focus on collecting and processing only the necessary amount of data. Implement federated learning techniques to minimise data transfers. This approach keeps data localised, reducing need for extensive data movement, which consumes energy
	Monitor and report the environmental impact of the AI tool	Regularly monitor and report on the environmental impact of AI systems used in healthcare, including energy usage, carbon emissions, and waste generation
Use community defined standards (universality 2)	Use a standard definition for the clinical task	Definition of heart failure by the American Academy of Cardiology[121]
	Use a standard method for data labelling	BI-RADS for breast imaging[122]
	Use a standard ontology for the AI inputs	DICOM for imaging data,[123] SNOMED for clinical data[41]
	Adopt technical standards	IEEE 2801-2022 for medical AI software[43]
	Use standard evaluation criteria	See Maier-Hein et al[21] for medical imaging applications, Barocas et al[31] and Bellamy et al[35] for fairness evaluation
Implement a risk management process (traceability 1)	Identify all possible clinical, technical, ethical, and societal risks	Bias against under-represented subgroups, limited generalisability to low resource facilities, data drift, lack of acceptance by end users, sensitivity to noisy inputs[124]
	Identify all possible operational risks	Misuse of the AI tool (owing to insufficient training or not following the instructions), application of the AI tool outside of the target population (eg, individuals with implants), use of the tool by others than the target end users (eg, technician instead of physician), hardware failure, incorrect data annotations, adversarial attacks[76 125]
	Assess the likelihood of each risk	Very likely, likely, possible, rare
	Assess the consequences of each risk	Patient harm, discrimination, lack of transparency, loss of autonomy, patient reidentification[126]
	Prioritise all the risks depending on their likelihood and consequences	Risk of bias (if no personal attributes are included in the model) *v* risk of patient reidentification (if personal attributes are collected)
	Define mitigation measures to be applied during AI development	Data enhancement, data augmentation,[127] bias correction techniques, domain adaptation,[66] transfer learning,[128] continuous learning[129]
	Define mitigation measures to be applied after deployment	Warnings to the users, system shutdown, reprocessing of the input data, acquisition of new input data, use of an alternative procedure, or human judgment only
	Set up a mechanism to monitor and manage risks over time	Periodic risk assessment every six months
	Create a comprehensive risk management file	Including all risks, their likelihood and consequences, risk mitigation measures, risk monitoring strategy

AI=artificial intelligence; BI-RADS=breast imaging reporting and data system; DI-COM=Digital Imaging and Communications in Medicine; ECG=electrocardiogram; GP=general practitioner; IEEE=Institute of Electrical and Electronics Engineers; MRI=magnetic resonance imaging; SNOMED=Systematized Nomenclature of Medicine.

trustworthy AI, as shown in table 3, table 4, table 5, and table 6. This approach offers easy-to-use, step-by-step guidance for all end users of the FUTURE-AI framework when designing, developing, validating and deploying new AI tools for healthcare.

Discussion

Despite the tremendous amount of research in medical AI in recent years, currently only a limited number of AI tools have made the transition to clinical practice. Although many studies have shown the huge potential

Table 4 | Practical steps and examples to implement FUTURE-AI recommendations during development phase

Recommendations	Operations	Examples
Collect representative training dataset (robustness 2)	Collect training data that reflect the demographic variations	According to age, sex, ethnicity, socioeconomics
	Collect training data that reflect the clinical variations	Disease subgroups, treatment protocols, clinical outcomes, rare cases
	Collect training data that reflect variations in real world practice	Data acquisition protocols, data annotations, medical equipment, operational variations (eg, patient motion during scanning)[125]
	Artificially enhance the training data to mimic real world conditions	Data augmentation,[127] data synthesis (eg, low quality data, noise addition),[130] data harmonisation,[131 132] data homogenisation[133]
Collect information on individuals' and data attributes (fairness 2)	Request approval for collecting data on personal attributes	Sex, age, ethnicity, socioeconomic status[134]
	Collect information on standard attributes of the individuals (if available and allowed)	Sex, age, nationality, education[135]
	Include application specific information relevant for fairness analysis	Skin colour, breast density,[34] presence of implants, comorbidity[136]
	Estimate data distributions across subgroups	Male *v* female, across ethnic groups
	Collect information on data provenance	Data centres, equipment characteristics, data preprocessing, annotation processes
Implement measures for data privacy and security (general 2)	Implement measures to ensure data privacy and security	Data deidentification, federated learning,[137 138] differential privacy, encryption[139]
	Implement measures against malicious attacks	Firewalls, intrusion detection systems, regular security audits[139]
	Adhere to applicable data protection regulations	General Data Protection Regulation,[140] Health Insurance Portability and Accountability Act[141]
	Define suitable data governance mechanisms	Access control, logging system
Implement measures against identified AI risks (general 3)	Implement a baseline AI model and identify its limitations	Bias, lack of generalisability[142]
	Implement methods to enhance robustness to real world variations (if needed)	Regularisation,[143] data augmentation,[127] data harmonisation,[131] domain adaptation[66]
	Implement methods to enhance generalisability across settings (if needed)	Regularisation, transfer learning,[144] knowledge distillation[145]
	Implement methods to enhance fairness across subgroups (if needed)	Data resampling, bias free representation,[36] equalised odds postprocessing[37 38 146]
Establish mechanisms for human-AI interactions (usability 2)	Implement mechanisms to standardise data preprocessing and labelling	Data preprocessing pipeline, data labelling plugin
	Implement an interface for using the AI model	Application programming interface
	Implement interfaces for explainable AI	Visual explanations, heatmaps, feature importance bars[105 106]
	Implement mechanisms for user centred quality control of the AI results	Visual quality control, uncertainty estimation[147]
	Implement mechanism for user feedback	Feedback interface[148]

AI=artificial intelligence.

of AI to improve healthcare, major clinical, technical, socioethical, and legal challenges persist.

In this paper, we presented the results of an international effort to establish a consensus guideline for developing trustworthy and deployable AI tools in healthcare. To this end, the FUTURE-AI Consortium was established, which provided knowledge and expertise across a wide range of disciplines and stakeholders, resulting in consensus and wide support, both geographically and across domains. Through an iterative process that lasted 24 months, the FUTURE-AI framework was established, comprising a comprehensive and self-contained set of 30 recommendations, which covers the whole lifecycle of medical AI. By dividing the recommendations across six guiding principles, the pathways towards responsible and trustworthy AI are clearly characterised. Because of its broad coverage, the FUTURE-AI guideline can benefit a wide range of stakeholders in healthcare, as detailed in table 2 in the appendix.

FUTURE-AI is a risk informed framework, proposing to assess application specific risks and challenges early in the process (eg, risk of discrimination, lack of generalisability, data drifts over time, lack of acceptance by end users, potential harm for patients, lack of transparency, data security vulnerabilities, ethical risks), followed by implementing tailored measures to reduce these risks (eg, collect data on individuals' attributes to assess and mitigate bias). As the specific measures to be implemented have benefits and potential weaknesses that the developers need to assess and take into consideration, a risk-benefit balancing trade-off has to be made. For example, collecting data on individuals' attributes might increase the risk of reidentification, but can enable the risk of bias and discrimination to be reduced. Therefore, in FUTURE-AI, risk management (as recommended in traceability 1) must be a continuous and transparent process throughout the AI tool's lifecycle.

FUTURE-AI is also an assumption-free, highly collaborative framework, recommending to continuously engage with multidisciplinary stakeholders to understand application specific needs, risks, and solutions (general 1). This is crucial to investigate all possible risks and factors that might reduce trust in a specific AI tool. For example, instead of making any assumptions on possible sources of bias, FUTURE-AI recommends that AI developers engage with healthcare professionals, domain experts, representative citizens, and/or ethicists early in the process to form interdisciplinary AI development teams and investigate in-depth the application specific

Table 5 | Practical steps and examples to implement FUTURE-AI recommendations during evaluation phase

Recommendations	Operations	Examples
Define adequate evaluation plan (general 4)	Identify the dimensions of trustworthy AI to be evaluated	Robustness, clinical safety, fairness, data drifts, usability, explainability
	Select appropriate testing datasets	External dataset from a new hospital, public benchmarking dataset[148]
	Compare the AI tool against standard of care	Conventional risk predictors, visual assessment by radiologist, decision by clinician[149 150]
	Select adequate evaluation metrics	F1 score for classification, concordance index for survival,[21] statistical parity for fairness[151]
Evaluate using external datasets and/or multiple sites (universality 3)	Identify relevant public datasets	Cancer Imaging Archive,[152] UK Biobank,[153] M&Ms,[154] MAMA-MIA,[155] BRATS[156]
	Identify external private datasets	New prospective dataset from same site or from different clinical centre[157 158]
	Select multiple evaluation sites	Three sites in same country, five sites in two different countries
	Verify that evaluation data and sites reflect real world variations	Variations in demographics, clinicians, equipment
	Confirm that no evaluation data were used during training	Yes/no
Evaluate fairness and bias correction measures (fairness 3)	Select attributes and factors for fairness evaluation	Sex, age, skin colour, comorbidity
	Define fairness metrics and criteria	Statistical parity difference defined fairness between −0.1 and 0.1[35]
	Evaluate fairness and identify biases	Fair with respect to age, biased with respect to sex
	Evaluate bias mitigation measures	Training data resampling,[159] equalised odds postprocessing[37 38 146]
	Evaluate impact of mitigation measures on model performance	Data resampling removed sex bias but reduced model performance[160]
	Report identified and uncorrected biases	In AI information leaflet and technical documentation[161] (see traceability 2).
Evaluate user experience (usability 4)	Evaluate usability with diverse end users	According to sex, age, digital proficiency level, role, clinical profile[162 163]
	Evaluate user satisfaction using usability questionnaires	System usability scale[164]
	Evaluate user performance and productivity	Diagnosis time with and without AI tool, image quantification time[165]
	Assess training of new end users	Average time to reach competency, training difficulties[166]
Evaluate clinical utility and safety (usability 5)	Define clinical evaluation plan	Randomised control trial,[59 167] in silico trial[168]
	Evaluate if AI tool improves patient outcomes	Better risk prevention, earlier diagnosis, more personalised treatment[169]
	Evaluate if AI tool enhances productivity or quality of care	Enhanced patient triage, shorter waiting times, faster diagnosis, higher patient intake[169]
	Evaluate if AI tool results in cost savings	Reduction in diagnosis costs,[170 171] reduction in overtreatment[172]
	Evaluate AI tool's safety	Side effects or major adverse events in randomised control trials[173 174]
Evaluate robustness (robustness 3)	Evaluate robustness under real world variations	Using test-retest datasets,[175 176] multivendor datasets[177]
	Evaluate robustness under simulated variations	Using simulated repeatability tests,[148] synthetic noise and artefacts (eg, image blurring)[178]
	Evaluate robustness against variations in end users	Different technicians or annotators
	Evaluate mitigation measures for robustness enhancement	Regularisation,[63] data augmentation,[64 127] noise addition, normalisation,[179] resampling, domain adaptation[66]
Evaluate explainability (explainability 2)	Assess if explanations are clinically meaningful	Reviewing by expert panels, alignment to current clinical guidelines, explanations not pointing to shortcuts[73]
	Assess explainability quantitatively using objective measures	Fidelity, consistency, completeness, sensitivity to noise[180-182]
	Assess explainability qualitatively with end users	Using user tests or questionnaires to measure confidence and affect clinical decision making[183 184]
	Evaluate if explanations cause end user overconfidence or overreliance	Measure changes in clinician confidence,[185 186] performance with and without AI tool[187]
	Evaluate if explanations are sensitive to input data variations	Stress tests under perturbations to evaluate the stability of explanations[74 188]
Provide documentation (traceability 2)	Report evaluation results in publication using AI reporting guidelines	Peer reviewed scientific publication using TRIPOD-AI reporting guideline[15]
	Create technical documentation for AI tool	AI passport,[189] model cards[49] (including model hyperparameters, training and testing data, evaluations, limitations, etc)
	Create clinical documentation for AI tool	Guidelines for clinical use, AI information leaflet (including intended use, conditions and diseases, targeted populations, instructions, potential benefits, contraindications)
	Provide risk management file	Including identified risks, mitigation measures, monitoring measures
	Create user and training documentation	User manuals, training materials, troubleshooting, FAQs (see usability 2)
	Identify and provide all locally required documentation	Compliance documents and certifications (see general 5)

AI=artificial intelligence; FAQ=frequently asked questions.

sources of bias, which might include domain specific attributes (eg, breast density for AI applications in breast cancer).

The FUTURE-AI guideline was defined in a generic manner to ensure it can be applied across a variety of domains (eg, radiology, genomics, mobile health, electronic health records). However, for many recommendations, their applicability varies across medical use cases, even within domains. To this end, the first recommendation in each of the guiding principles is to identify the specificities to be addressed, such as the types of biases (fairness 1), the clinical settings (universality 1), or the need and approaches for explainable AI (explainability 1). This facilitates generalisability across domains, but also ensures sustainability for future use. Furthermore, we recognise that a one-size-fits-all approach is not feasible, as addressal of many of the recommendations is use case specific, and standards do not exist yet or are subject to change. Therefore, we focused on developing best practices for enhancing the trustworthiness of medical AI tools, while consciously avoiding the imposition of specific techniques for the implementation of each recommendation. This flexibility also acknowledges the diversity of methods for tackling challenges and mitigating risks

Table 6 | Practical steps and examples to implement FUTURE-AI recommendations during deployment phase

Recommendations	Operations	Examples
Evaluate and demonstrate local clinical validity (universality 4)	Test AI model using local data	Data from local clinical registry
	Identify factors that could affect AI tool's local validity	Local operators, equipment, clinical workflows, acquisition protocols
	Assess AI tool's integration within local clinical workflows	AI tool's interface aligns with hospital IT system[148] or disrupts routine practice
	Assess AI tool's local practical utility and identify any operational challenges	Time to operate, clinician satisfaction, disruption of existing operations[148 190]
	Implement adjustments for local validity	Model calibration, fine-tuning,[191] transfer learning[192-194]
	Compare performance of AI tool with that of local clinicians	Side-by-side comparison, in silico trial
Define mechanisms for quality control of AI inputs and outputs (traceability 3)	Implement mechanisms to identify erroneous input data	Missing value or out-of-distribution detector,[195] automated image quality assessment[73 196 197]
	Implement mechanisms to detect implausible AI outputs	Postprocessing sanity checks, anomaly detection algorithm[198]
	Provide calibrated uncertainty estimates to inform on AI tool's confidence	Calibrated uncertainty estimates per patient or data point[53 54 199]
	Implement system for continuous quality monitoring	Real time dashboard tracking data quality and performance metrics[200]
	Implement feedback mechanism for users to report issues	Feedback portal enabling clinicians to report discrepancies or anomalies
Implement system for periodic auditing and updating (traceability 4)	Define schedule for periodic audits	Biannual or annual
	Define audit criteria and metrics	Accuracy, consistency, fairness, data security[148]
	Define datasets for periodic audits	Newly acquired prospective dataset from local hospital
	Implement mechanisms to detect data or concept drifts	Detecting shifts in input data distributions[148 190]
	Assign role of auditor(s) for AI tool	Internal auditing team, third party company[190]
	Update AI tool based on audit results	Updating AI model,[56] re-evaluating AI model,[148] adjusting operational protocols, continuous learning[201-204]
	Implement reporting system from audits and subsequent updates	Automatic sharing of detailed reports to healthcare managers and clinicians
	Monitor impact of AI updates	Impact on system performance and user satisfaction[56]
Implement logging system for usage recording (traceability 5)	Implement logging framework capturing all interactions	User actions, AI inputs, AI outputs, clinical decisions
	Define data to be logged	Timestamp, user ID, patient ID (anonymised), action details, results
	Implement mechanisms for data capture	Software to automatically record every data and operation
	Implement mechanisms for data security	Encrypted log files, privacy preserving techniques[205]
	Provide access to logs for auditing and troubleshooting	By defining authorised personnel, eg, healthcare or IT managers
	Implement mechanism for end users to log any issues	A user interface to enter information about operational anomalies
	Implement log analysis	Time series statistics and visualisations to detect unusual activities and alert administrators
Provide training (usability 3)	Create user manuals	User instructions, capabilities, limitations, troubleshooting steps, examples, and case studies
	Develop training materials and activities	Online courses, workshops, hands-on sessions
	Use formats and languages accessible to intended end users	Multiple formats (text, video, audio) and languages (English, Chinese, Swahili)
	Customise training to all end user groups	Role specific modules for specialists, nurses, and patients
	Include training to enhance AI and health literacy	On application specific AI concepts (eg, radiomics, explainability), AI driven clinical decision making
Identify and comply with applicable AI regulatory requirements (general 5)	Engage regulatory experts to investigate regulatory requirements	Regulatory consultants from intended local settings
	Identify specific regulations based on AI tool's intended markets	FDA's SaMD in the United States,[206] MDR and AI Act[207] in the EU
	Identify specific requirements based on AI tool's purpose	De Novo classification (Class III)[208]
	Define list of milestones towards regulatory compliance	MDR certification: technical verification, pivotal clinical trial, risk and quality management, postmarket follow-up
Establish mechanisms for AI governance (traceability 6)	Assign roles for AI tool's governance	For periodic auditing, maintenance, supervision (eg, healthcare manager)
	Define responsibilities for AI related errors	Responsibilities of clinicians, healthcare centres, AI developers, and manufacturers
	Define mechanisms for accountability	Individual *v* collective accountability/liability,[25] compensations, support for patients

AI=artificial intelligence; FDA=United States Food and Drug Administration; MDR=medical device regulation; SaMD= Software as a Medical Device.

in medical AI. For example, the recommendation to protect personal data during AI training can be implemented through data deidentification, federated learning, differential privacy or encryption, among other methods. While such concrete examples are listed in this article, especially in table 3, table 4, table 5, and table 6, the most adequate techniques for implementing each recommendation should be ultimately selected by the AI development team as a function of the application domain, clinical use case, and data characteristics, as well as the advantages and limitations of each method. Similarly, all stakeholders of the AI development team are together responsible for addressing the recommendations, where the role of each party might vary per application, method, domain, project setup, and use case.

While the FUTURE-AI framework offers insights for regulating medical AI, future work is needed to incorporate these recommendations into regulatory procedures. For example, we propose mechanisms to enhance traceability and governance, such as logging. However, the crucial issue of liability is yet

to be addressed, for example, who should perform the audits and who should be accountable for errors. Furthermore, we recommend continuous evaluation and fine tuning of AI models over time. However, current regulations prevent post release modifications because they would formally invalidate the manufacturer's initial validation. Future regulations should address the possibility of local adaptations within predefined acceptance criteria.

On one hand, implementation of the FUTURE-AI guideline might involve substantial costs, which could affect both AI developers and healthcare systems. These financial considerations could potentially exacerbate disparities in AI adoption, particularly affecting smaller developers and resource limited health systems. Collaborative efforts involving stakeholders from various sectors could help to distribute the financial burden and support equitable access to advanced AI tools. On the other hand, early adoption of the FUTURE-AI guideline might save costs. Instead of developing AI tools that do not have clinical added value or having to address various of the outlined principles after developing a tool, early adoption will result in a trustworthy and deployable AI tool by design and can be more cost effective than post development adoption, which, in practice, often requires costly change requests that affect large parts of a tool's solution architecture.

Finally, progressive development and adoption of medical AI tools will lead to new requirements, challenges, and opportunities. For some of the recommendations, no clear standard on how these should be addressed yet exists. Aware of this reality, we propose FUTURE-AI as a dynamic, living framework. To refine the FUTURE-AI guideline and learn from other voices, we set up a dedicated webpage (www.future-ai.eu) through which we invite the community to join the FUTURE-AI network and provide feedback based on their own experience and perspective. On the website we include a FUTURE-AI self-assessment checklist, which comprises a set of questions and examples to facilitate and illustrate the use of the FUTURE-AI recommendations. Additionally, we plan to organise regular outreach events such as webinars and workshops to exchange with medical AI researchers, manufacturers, evaluators, end users, and regulators. Future research includes more in-depth studies of the operationalisation of FUTURE-AI in specific healthcare domains, leading to domain specific methods on the addressal of the recommendations, and of each principle as these have become rapidly evolving fields of their own, for example, Fair ML and Explainable AI (XAI).

AUTHOR AFFILIATIONS

[1]Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[3]Center for Computational Imaging & Simulation Technologies in Biomedicine, Schools of Computing and Medicine, University of Leeds, Leeds, UK

[4]Medical Imaging Research Centre (MIRC), Cardiovascular Science and Electronic Engineering Departments, KU Leuven, Leuven, Belgium

[5]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[6]Department of Computing, Imperial College London, London, UK

[7]IBM Research Africa, Nairobi, Kenya

[8]Departments of Radiology, Medicine, and Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

[9]Fraunhofer Heinrich Hertz Institute, Berlin, Germany

[10]Amsterdam University Medical Centers, Department of Cardiology, University of Amsterdam, Amsterdam, Netherlands

[11]Health Data Research UK and Institute of Health Informatics, University College London, London, UK

[12]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

[13]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

[14]Institute for AI and Informatics in Medicine, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

[15]Gruppo Maggioli, Research and Development Lab, Athens, Greece

[16]Institut Curie, Inserm, Orsay, France

[17]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[18]Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA

[19]Institute of Machine Learning in Biomedical Imaging, Helmholtz Center Munich, Munich, Germany

[20]Department of Radiation Sciences, Diagnostic Radiology, Umeå University, Umeå, Sweden

[21]Foundation for Research and Technology—Hellas (FORTH), Crete, Greece

[22]Department of Software Engineering, Namibia University of Science & Technology, Windhoek, Namibia

[23]Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

[24]Division of Intelligent Medical Systems, German Cancer Research Centre, Heidelberg, Germany

[25]Biomedical Imaging Research Group, La Fe Health Research Institute, Valencia, Spain

[26]Medical Imaging Department, Hospital Universitario y Politécnico La Fe, Valencia, Spain

[27]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK

[28]2nd Division of Radiology, Medical University of Gdansk, Gdansk, Poland

[29]Faculty of Law and Criminology, Ghent University, Ghent, Belgium

[30]Data Science Department, EURECOM, Sophia Antipolis, France

[31]Institute of History and Ethics in Medicine, Technical University of Munich, Munich, Germany

[32]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington DC, USA

[33]Department of Radiology & Nuclear Medicine, Erasmus MC University Medical Centre, Rotterdam, Netherlands

[34]Copenhagen Academy for Medical Education and Simulation Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

[35]BBMRI-ERIC, ELSI Services & Research, Graz, Austria

[36]Computational Clinical Imaging Group, Champalimaud Foundation, Lisbon, Portugal

[37]Integrative Biomedical Imaging Informatics at Stanford (IBIIS), Department of Radiology, Stanford University, Stanford, CA, USA

[38]Institute of Information Science and Technologies of the National Research Council of Italy, Pisa, Italy

[39]Artificial Intelligence in Healthcare Program, TIC Salut Social Foundation, Barcelona, Spain

[40]Department of Philosophy, and School of Medicine, Macquarie University, Sydney, Australia

[41]The D-lab, Department of Precision Medicine, GROW—School for Oncology and Reproduction, Maastricht University, Maastricht, Netherlands

This work has been supported by the European Union's Horizon 2020 under grant agreement No 952159 (ProCAncer-I), No 952172 (CHAIMELEON), No 826494 (PRIMAGE), No 952179 (INCISIVE), No 101034347 (OPTIMA), No 101016775 (INTERVENE), No 101100633 (EUCAIM), No 101136670 (GLIOMATCH), No 101057062 (AIDAVA), No 101095435 (REALM), and No 116074 (BigData@Heart). This work received support from the European Union's Horizon Europe under grant agreement No 101057699 (RadioVal), No 101057849 (DataTools4Heart), and No 101080430 (AI4HF). This work received support from the European Research Council under grant agreement No 757173 (MIRA), No 884622 (Deep4MI), No 101002198 (NEURAL SPICING), No 866504 (CANCER-RADIOMICS), and No 101044779 (AIMIX). This work was partially supported by the Royal Academy of Engineering, Hospital Clinic Barcelona, Malaria No More, Carnegie Cooperation New York, Human frontier science programme, Natural Sciences and Engineering Research Council of Canada (NSERC), the Australian National Health and Medical Research Council Ideas under grant No 1181960, United States Department of Defence W81XWH2010747-P1, 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR-19-P3IA-0002), InTouchAI.eu, IITP grant funded by the Korean government (No 2020-0-00594), A*STAR Career Development Award (project No C210112057) from the Agency for Science, Technology and Research (A*STAR), National Institute for Health and Care Research Barts Biomedical Research Centre, Centre National de la Recherche Scientifique (CNRS), MPaCT-Data. Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología (Exp. IMP/00019) funded by Instituto de Salud Carlos III and the Fondo Europeo de Desarrollo Regional (FEDER, "Una manera de hacer Europa"), Ministry of Science, Technology and Innovation of Colombia project code 110192092354, Gordon and Betty Moore Foundation, Google Award for Inclusion Research, Fraunhofer Heinrich Hertz Institute, US National Institutes of Health, National Council for Scientific and Technological Development (CNPq), European Heart Network, NIBIB/University of Chicago (MIDRC), Hong Kong Research Grants Council Theme-based Research Scheme (TRS) project T45-401/22-N, Young Researcher Project (19PEJC09-03) funded by the Ministry of High Education of Tunisia, Juan de la Cierva with reference number FJC2021-047659-I, Nepal Applied Mathematics and Informatics Institute for Research (NAAMII), Fogarty International Center of the National Institutes of Health under Award No 5U2RTW012131-02, Universidad Galileo, Natural Science Foundation of China under grant 62271465, Israel Science Foundation, National Institutes of Health (NIH), Dutch Cancer Society (KWF Kankerbestrijding) under project No 14449, Netherlands Organisation for Scientific Research (NWO) VICI project VI.C.182.042, National Center for Artificial Intelligence CENIA (ANID-BASAL FB210017), Google Research, Independent Research Fund Denmark (DFF, grant No 9131-00097B), Wellcome Flagship Programme (WT213038/Z/18/Z), Cancer Research UK programme grant (C49297/A27294), the MIDRC (The Medical Imaging and Data Resource Center), made possible by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under contract 75N92020D00021, and the Employee European Heart Network. Also, this work was partially supported by the project FUTURE-ES (PID2021-126724OB-I00) and AIMED (PID2023-146786OB-I00) from the Ministry of Science, Innovation and Universities of the Government of Spain.

FUTURE-AI Consortium authors: Aasa Feragen, Abdul Joseph Fofanah, Alena Buyx, Anais Emelie, Andrea Lara, An-Wen Chan, Arcadi Navarro, Benard O Botwe, Bishesh Khanal, Brigit Beger, Carol C Wu, Daniel Rueckert, Deogratias Mzurikwao, Dimitrios I Fotiadis, Doszhan Zhussupov, Enzo Ferrante, Erik Meijering, Fabio A González, Gabriel P Krestin, Geletaw S Tegenaw, Gianluca Misuraca, Girish Dwivedi, Haridimos Kondylakis, Harsha Jayakody, Henry C Woodruf, Horst Joachim Mayer, Hugo JWL Aerts, Ian Walsh, Ioanna Chouvarda, Isabell Tributsch, Islem Rekik, James Duncan, Jihad Zahir, Jinah Park, Judy W Gichoya, Kensaku Mori, Leticia Rittner, Lighton Phiri, Linda Marrakchi-Kacem, Lluís Donoso-Bach, Maria Bielikova, Marzyeh Ghassemi, Md Ashrafuzzaman, Mohammad Yaqub, Mukhtar ME Mahmoud, Mustafa Elattar, Nicola Rieke, Oliver Díaz, Olivier Salvado, Ousmane Sall, Pamela Guevara, Peter Gordebeke, Philippe Lambin, Pieta Brown, Purang Abolmaesumi, Qi Dou, Qinghua Lu, Rose Nakasi, S Kevin Zhou, Shadi Albarqouni, Stacy Carter, Steffen E Petersen, Suyash Awate, Tammy Riklin Raviv, Tessa Cook, Tinashe EM Mutsvangwa, Wiro J Niessen, Xènia Puig-Bosch, Yi Zeng, Yunusa G Mohammed, Yves Saint James Aquino (web appendix 2 gives full details).

Contributors: KL, RO, NL, KK, GT, SC, SA, LC-A, KM, MT, NP, ZS, HCW, PL, and LM-B conceptualised the FUTURE-AI framework and provided the first set of recommendations. All co-authors participated in the surveys and provided feedback throughout the process. KL organised four online meetings to discuss the final recommendations. AE and XP-B coordinated the last consensus survey. KL and MPAS coordinated the feedback gathering process and wrote the manuscript. All authors and the FUTURE-AI Consortium contributed, reviewed, and approved the manuscript. KL is the guarantor of this work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: Funding for this work was provided by the European Union's Horizon 2020 under grant agreement No 952103 (EuCanImage). The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from European Union's Horizon 2020 for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. GD owns equity interest in Artrya Ltd and provides consultancy services. JK-C receives research funding from GE, Genetech and is a consultant at Siloam Vision. GPK advises some AI startups such as Gleamer.AI, FLUIDDA BV, NanoX Vision and was the founder of Quantib BV. SEP is a consultant for Circle Cardiovascular Imaging, Calgary, Alberta, Canada. BG is employed by Kheiron Medical Technologies and HeartFlow. PL has/had grants/sponsored research agreements from Radiomics SA, Convert Pharmaceuticals SA and LivingMed Biotech srl. He received a presenter fee and/or reimbursement of travel costs/consultancy fee (in cash or in kind) from Astra Zeneca, BHV srl, and Roche. PL has/had minority shares in the companies Radiomics SA, Convert pharmaceuticals SA, Comunicare SA, LivingMed Biotech srl, and Bactam srl. PL is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/059089), licensed to ptTheragnostic/DNAmito; one granted patent on LSRT (PCT/P126537PC00, US patent No 12 102 842), licensed to Varian; one issued patent on Radiomic signature of hypoxia (US patent No 11 972 867), licensed to a commercial entity; one issued patent on Prodrugs (WO2019EP64112) without royalties; one non-issued, non-licensed patents on Deep Learning-Radiomics (N2024889) and three non-patented inventions (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures. ARP serves as advisor for mGeneRX in exchange for equity. JM receives royalties from GE, research grants from Siemens and is unpaid consultant for Nuance. HCW owns minority shares in the company Radiomics SA. JWG serves on several radiology society AI committees. LR advises an AI startup Neurlamind. CPL is a shareholder and advisor to Bunker Hill Health, GalileoCDS, Sirona Medical, Adra, and Kheiron Medical. He serves as a board member of Bunker Hill Health and a shareholder of whiterabbit.ai. He has served as a paid consultant to Sixth Street and Gilmartin Capital. His institution has received grants or gifts from Bunker Hill Health, Carestream, CARPL, Clairity, GE Healthcare, Google Cloud, IBM, Kheiron, Lambda, Lunit, Microsoft, Philips, Siemens Healthineers, Stability.ai, Subtle Medical, VinBrain, Visiana, Whiterabbit.ai, the Lowenstein Foundation, and the Gordon and Betty Moore Foundation. GSC is a statistics editor for the BMJ and a National Institute for Health and Care Research (NIHR) Senior Investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR, or the Department of Health and Social Care. All other authors declare no competing interests.

Patient and public involvement: This study involved extensive input from over 100 authors with diverse expertise, including AI scientists, healthcare practitioners, ethicists, social scientists, legal experts, and industry professionals. To further refine the work, several rounds of feedback were sought from experts in these and related fields. We are confident that this broad, multidisciplinary collaboration encompassed the expertise required for this study. Owing to the absence of dedicated funding for this project, direct involvement of patients and the public was not feasible.

Dissemination to participants and related patient and public communities: The authors plan to disseminate the research widely through presentations at conferences and through social media to interest holders who generate or use evidence, including to consumers in evidence synthesis organisations. They are committed to disseminating the findings in formats accessible to the public and patient communities to promote broader engagement and understanding of the results.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56. doi:10.1038/s41591-018-0300-7
2 Lekadir K, Quaglio G, Tselioudis Garmendia A, Gallin C. *Artificial intelligence in healthcare—Applications, risks, and ethical and societal impacts* . European Parliament, Directorate-General for Parliamentary Research Services, 2022.
3 Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927. doi:10.1136/bmj.l6927
4 Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28:231-7. doi:10.1136/bmjqs-2018-008370
5 Celi LA, Cellini J, Charpignon ML, et al, for MIT Critical Data. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digit Health* 2022;1:e0000022. doi:10.1371/journal.pdig.0000022
6 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30-6. doi:10.1038/s41591-018-0307-0
7 Haibe-Kains B, Adam GA, Hosny A, et al, Massive Analysis Quality Control (MAQC) Society Board of Directors. Transparency and reproducibility in artificial intelligence. *Nature* 2020;586:E14-6. doi:10.1038/s41586-020-2766-y.
8 Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 2021;22:122. doi:10.1186/s12910-021-00687-3
9 High-Level Expert Group on AI, EU Commission. Ethics Guidelines For Trustworthy AI. 2019. https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf
10 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. doi:10.1038/sdata.2016.18.
11 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378
12 Tejani AS, Klontzas ME, Gatti AA, et al, CLAIM 2024 Update Panel. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell* 2024;6:e240300. doi:10.1148/ryai.240300
13 Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020;370:m3164. doi:10.1136/bmj.m3164
14 Vasey B, Nagendran M, Campbell B, et al, DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904. doi:10.1136/bmj-2022-070904
15 Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008. doi:10.1136/bmjopen-2020-048008
16 Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging* 2023;14:75. doi:10.1186/s13244-023-01415-8
17 Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory frameworks for development and evaluation of artificial intelligence-based diagnostic imaging algorithms: summary and recommendations. *J Am Coll Radiol* 2021;18(3 Pt A):413-24. doi:10.1016/j.jacr.2020.09.060
18 Reddy S, Rogers W, Makinen VP, et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform* 2021;28:e100444. doi:10.1136/bmjhci-2021-100444
19 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-9. doi:10.1148/radiol.2017171920
20 Walsh I, Fishman D, Garcia-Gasulla D, et al, ELIXIR Machine Learning Focus Group. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods* 2021;18:1122-7. doi:10.1038/s41592-021-01205-4
21 Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation.*arXiv* 2022. https://arxiv.org/abs/2206.01653v7
22 Bradshaw TJ, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: best practices for algorithm development. *J Nucl Med* 2022;63:500-10. doi:10.2967/jnumed.121.262567
23 Solanki P, Grundy J, Waqar H. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI Ethics* 2023;3:223-40. doi:10.1007/s43681-022-00195-z
24 Amugongo LM, Kriebitz A, Boch A, Lütge C. Operationalising AI ethics through the agile software development lifecycle: a case study of AI-enabled mobile health applications. *AI Ethics* 2023. doi:10.1007/s43681-023-00331-3
25 World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO guidance*. World Health Organization, 2021.
26 Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Shaping Europe's digital future. 2023. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
27 Taylor E. We agree, don't we? The Delphi Method for Health Environments Research. *HERD* 2020;13:11-23. doi:10.1177/1937586719887709
28 Grime MM, Wright G. Delphi Method. Wiley StatsRef: Statistics Reference Online. 2016. https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat07879
29 McGreevey JD3rd, Hanson CW3rd, Koppel R. Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. *JAMA* 2020;324:552-3. doi:10.1001/jama.2020.2724
30 Solomonides AE, Koski E, Atabaki SM, et al. Defining AMIA's artificial intelligence principles. *J Am Med Inform Assoc* 2022;29:585-91. doi:10.1093/jamia/ocac006.
31 Barocas S, Hardt M, Narayanan A. Fairness and machine learning. 2023. https://fairmlbook.org/
32 Ferryman K, Pitcan M. Fairness in Precision Medicine. Data & Society Research Institute. 2018. https://datasociety.net/library/fairness-in-precision-medicine/
33 Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022;28:2232-3. doi:10.1038/s41591-022-01987-w.
34 Garrucho L, Kushibar K, Osuala R, et al. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Front Oncol* 2023;12:1044496. doi:10.3389/fonc.2022.1044496.
35 Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* 2018. https://arxiv.org/abs/1810.01943v1
36 Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med (Lond)* 2021;1:25. doi:10.1038/s43856-021-00028-w.
37 Li X, Cui Z, Wu Y, Gu L, Harada T. Estimating and improving fairness with adversarial learning.*arXiv* 2021. https://arxiv.org/abs/2103.04243v2
38 Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. *arXiv* 2017. https://arxiv.org/abs/1709.02012v2
39 Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. Data handling. *Radiol Artif Intell* 2022;4:e210290. doi:10.1148/ryai.210290
40 Zhang K, Khosravi B, Vahdati S, et al. Mitigating bias in radiology machine learning: 2. Model development. *Radiol Artif Intell* 2022;4:e220010. doi:10.1148/ryai.220010
41 Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;27:129-39. doi:10.1055/s-0038-1667077
42 Lima DM, Rodrigues-Jr JF, Traina AJM, Pires FA, Gutierrez MA. Transforming two decades of ePR data to OMOP CDM for clinical research. *Stud Health Technol Inform* 2019;264:233-7.
43 IEEE Standards Association. IEEE Recommended Practice for the quality management of datasets for Medical Artificial Intelligence. https://standards.ieee.org/ieee/2801/7459/
44 ISO. ISO/IEC JTC 1/SC 42 - Artificial intelligence. https://www.iso.org/committee/6794475.html
45 Cabitza F, Campagner A, Soares F, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed* 2021;208:106288. doi:10.1016/j.cmpb.2021.106288

46 Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res* 2022;6:24. doi:10.1186/s41512-022-00136-8.
47 Oala L, Murchison AG, Balachandran P, et al. Machine learning for health: algorithm auditing & quality control. *J Med Syst* 2021;45:105. doi:10.1007/s10916-021-01783-y
48 Königstorfer F, Thalmann SAI. Documentation: a path to accountability. *J Responsib Technol* 2022;11:100043. doi:10.1016/j.jrt.2022.100043.
49 Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 2019. https://dl.acm.org/doi/10.1145/3287560.3287596
50 Arnold M, Piorkowski D, Reimer D, et al. FactSheets: increasing trust in AI services through supplier's declarations of conformity. *IBM J Res Develop* 2019;63. doi:10.1147/JRD.2019.2942288.
51 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM* 2021;64:86-92. doi:10.1145/3458723.
52 Mongan J, Moy L, Kahn CEJr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020;2:e200029. doi:10.1148/ryai.2020200029
53 Dormann CF. Calibration of probability predictions from machine-learning and statistical models. *Glob Ecol Biogeogr* 2020;29:760-5. doi:10.1111/geb.13070.
54 Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med* 2021;4:4. doi:10.1038/s41746-020-00367-3.
55 Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol* 2023;96:20220878. doi:10.1259/bjr.20220878.
56 Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* 2022;5:66. doi:10.1038/s41746-022-00611-y.
57 Sujan M, Furniss D, Grundy K, et al. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform* 2019;26:e100081. doi:10.1136/bmjhci-2019-100081
58 Kim D, Vegt N, Visch V, De Vos MB. How Much Decision Power Should (A)I Have? Investigating Patients' Preferences Towards AI Autonomy in Healthcare Decision Making. Conference on Human Factors in Computing Systems - Proceedings. 2024. https://dl.acm.org/doi/10.1145/3613904.3642883
59 Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ Digit Med* 2021;4:154. doi:10.1038/s41746-021-00524-2.
60 Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019;363:1287-9. doi:10.1126/science.aaw4399
61 Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019;20:e262-73. doi:10.1016/S1470-2045(19)30149-4
62 Lemay A, Hoebel K, Bridge CP, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. *NPJ Digit Med* 2022;5:174. doi:10.1038/s41746-022-00709-3.
63 Tian Y, Zhang Y. A comprehensive survey on regularization strategies in machine learning. *Inf Fusion* 2022;80:146-66. doi:10.1016/j.inffus.2021.11.005.
64 Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop. https://ieeexplore.ieee.org/document/8388338
65 Gao Y, Wang Y, Yu J. Optimized resolution-oriented many-to-one intensity standardization method for magnetic resonance images. *Applied Sci* 2019;9:5531. doi:10.3390/app9245531
66 Garrucho L, Kushibar K, Jouide S, Diaz O, Igual L, Lekadir K. Domain generalization in deep learning based mass detection in mammography: a large-scale multi-center study. *Artif Intell Med* 2022;132:102386. doi:10.1016/j.artmed.2022.102386
67 Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020;20:310. doi:10.1186/s12911-020-01332-6
68 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021;3:e745-50. doi:10.1016/S2589-7500(21)00208-9
69 Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans Neural Netw Learn Syst* 2021;32:4793-813. doi:10.1109/TNNLS.2020.3027314
70 Arras L, Osman A, Samek W. CLEVR-XAI: a benchmark dataset for the ground truth evaluation of neural network explanations. *Inf Fusion* 2022;81:14-40. doi:10.1016/j.inffus.2021.11.008.
71 Hedström A, Leander W, Bareeva D, et al. Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond.*arXiv* 2022. https://arxiv.org/abs/2202.06861v3
72 Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst* 2021;11. doi:10.1145/3387166.
73 DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell* 2021;3:610-19. doi:10.1038/s42256-021-00338-7
74 Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. *Proc Conf AAAI Artif Intell* 2019;33:3681-8. https://ojs.aaai.org/index.php/AAAI/article/view/4252. doi:10.1609/aaai.v33i01.33013681.
75 Channa R, Wolf R, Abramoff MD. Autonomous artificial intelligence in diabetic retinopathy: from algorithm to clinical application. *J Diabetes Sci Technol* 2021;15:695-8. doi:10.1177/1932296820909900.
76 Kaviani S, Han KJ, Sohn I. Adversarial attacks and defenses on AI in medical imaging informatics: a survey. *Expert Syst Appl* 2022;198:116815. doi:10.1016/j.eswa.2022.116815.
77 Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. *arXiv* 2022;4. https://arxiv.org/abs/2207.07048v1
78 Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med* 2022;5:48. doi:10.1038/s41746-022-00592-y.
79 McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Med Ethics* 2022;23:6. doi:10.1186/s12910-022-00746-3
80 Rafner J, Dellermann D, Hjorth A, et al. Deskilling, upskilling, and reskilling: a case for hybrid intelligence. *Morals & Machines* 2021;1:24-39. doi:10.5771/2747-5174-2021-2-24.
81 Selvan R, Bhagwat N, Anthony LFW, Kanding B, Dam EB. Carbon footprint of selecting and training deep learning models for medical image analysis. *arXiv* . 2022. http://arxiv.org/abs/2203.02202
82 Concannon TW, Grant S, Welch V, et al, Multi Stakeholder Engagement (MuSE) Consortium. Practical guidance for involving stakeholders in health research. *J Gen Intern Med* 2019;34:458-63. doi:10.1007/s11606-018-4738-6
83 Schiller C, Winters M, Hanson HM, Ashe MC. A framework for stakeholder identification in concept mapping and health research: a novel process and its application to older adult mobility and the built environment. *BMC Public Health* 2013;13:428. doi:10.1186/1471-2458-13-428
84 Bogina V, Hartman A, Kuflik T, Shulner-Tal A. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *Int J Artif Intell Educ* 2022;32:808-33. doi:10.1007/s40593-021-00248-0.
85 Woudstra K, Reuzel R, Rovers M, Tummers M. An overview of stakeholders, methods, topics, and challenges in participatory approaches used in the development of medical devices: a scoping review. *Int J Health Policy Manag* 2023;12:6839. doi:10.34172/ijhpm.2022.6839
86 Halvorsrud K, Kucharska J, Adlington K, et al. Identifying evidence of effectiveness in the co-creation of research: a systematic review and meta-analysis of the international healthcare literature. *J Public Health (Oxf)* 2021;43:197-208. doi:10.1093/pubmed/fdz126.
87 Edwards HA, Huang J, Jansky L, Mullins CD. What works when: mapping patient and stakeholder engagement methods along the ten-step continuum framework. *J Comp Eff Res* 2021;10:999-1017. doi:10.2217/cer-2021-0043
88 Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J Innov Knowl.* 2023;8:100333. doi:10.1016/j.jik.2023.100333.
89 Koulu R. Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastrich J Eur Comp Law* 2020;27:720-35. doi:10.1177/1023263X20978649.
90 Daniel S, Luz A. Human oversight and control in AI-driven healthcare systems. 2024
91 Van Velsen L, Wentzel J, Van Gemert-Pijnen JE. Designing eHealth that matters via a multidisciplinary requirements development approach. *JMIR Res Protoc* 2013;2:e21. doi:10.2196/resprot.2547
92 Arnold TC, Freeman CW, Litt B, Stein JM. Low-field MRI: Clinical promise and challenges. *J Magn Reson Imaging* 2023;57:25-44. doi:10.1002/jmri.28408
93 Tran TT, Hlaing M, Krause M. Point-of-care ultrasound: applications in low- and middle-income countries. *Curr Anesthesiol Rep* 2021;11:69-75. doi:10.1007/s40140-020-00429-y

94 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195. doi:10.1186/s12916-019-1426-2
95 de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med* 2022;5:2. doi:10.1038/s41746-021-00549-7.
96 Huber FA, Chaitanya K, Gross N, et al. Whole-body composition profiling using a deep learning algorithm: influence of different acquisition parameters on algorithm performance and robustness. *Invest Radiol* 2022;57:33-43. doi:10.1097/RLI.0000000000000799
97 Solyst J, Xie S, Ogan A, Hammer J, Yang E, Eslami M. The potential of diverse youth as stakeholders in identifying and mitigating algorithmic bias for a future of fairer AI. *Proc ACM Hum-Comput Interact* 2023;7. doi:10.1145/3610213
98 Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publication. 2022. doi:10.6028/NIST.SP.1270
99 Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol* 2024;42:3-15. doi:10.1007/s11604-023-01474-3.
100 Zicari RV, Ahmed S, Amann J, et al. Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier. *Front Hum Dyn* 2021;3:688152. doi:10.3389/fhumd.2021.688152.
101 Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—a scoping review. *J Am Acad Dermatol* 2022;87:157-9. doi:10.1016/j.jaad.2021.06.884
102 Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SIM. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. *Mayo Clin Proc Digit Health* 2023;1:120-38. doi:10.1016/j.mcpdig.2023.02.004.
103 Kollerup NK, Johansen SS, Tolsgaard MG, et al. Clinical needs and preferences for AI-based explanations in clinical simulation training. *Behav Inform Technol* 2024:1-21. doi:10.1080/0144929X.2024.2334852.
104 Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)* 2020;23:18. doi:10.3390/e23010018.
105 Jin W, Fan J, Pasquier P. EUCA: the End-User-Centered Explainable AI Framework. 2021. http://weinajin.github.io/end-user-xai/
106 Haque AB, Islam AKMN, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. *Technol Forecast Soc Change* 2023;186:122120. doi:10.1016/j.techfore.2022.122120.
107 Murphy K, Di Ruggiero E, Upshur R, et al. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics* 2021;22:14. doi:10.1186/s12910-021-00577-8
108 Coghlan S, Gyngell C, Vears DF. Ethics of artificial intelligence in prenatal and pediatric genomic medicine. *J Community Genet* 2024;15:13-24. doi:10.1007/s12687-023-00678-4
109 Huang S, Lai X, Ke L, et al. AI Technology panic—is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. *Psychol Res Behav Manag* 2024;17:1087-102. doi:10.2147/PRBM.S440889
110 UK's Approach to Regulating the Use of Artificial Intelligence. Mayer Brown. 2023. https://www.mayerbrown.com/en/insights/publications/2023/07/uks-approach-to-regulating-the-use-of-artificial-intelligence
111 A pro-innovation approach to AI regulation: government response - GOV.UK. 2024. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response
112 Health Information Privacy. HIPAA Home. HHS.gov. https://www.hhs.gov/hipaa/index.html
113 Ethical Principles for Artificial Intelligence in Medicine. 2024. https://www.cpsbc.ca/files/pdf/IG-Artificial-Intelligence-in-Medicine.pdf
114 Hong Kong Government. Ethical Artificial Intelligence Framework. 2021. https://www.digitalpolicy.gov.hk/en/our_work/data_governance/policies_standards/ethical_ai_framework/doc/Ethical_AI_Framework.pdf
115 Indian Council of Medical Research, Government of India. Ethical guidelines for application of Artificial Intelligence in Biomedical Research and Healthcare. 2023. https://main.icmr.nic.in/content/ethical-guidelines-application-artificial-intelligence-biomedical-research-and-healthcare
116 Social Principles of Human-Centric AI. https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf
117 Government of Japan. The Hiroshima AI Process: Leading the Global Challenge to Shape Inclusive Governance for Generative AI. 2024. https://www.japan.go.jp/kizuna/2024/02/hiroshima_ai_process.html
118 Australia's AI Ethics Principles. Australia's Artificial Intelligence Ethics Framework. Department of Industry Science and Resources 2024. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles
119 Anthony LFW, Kanding B, Selvan R. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. *arXiv* 2020. http://arxiv.org/abs/2007.03051
120 Jia Z, Chen J, Xu X, et al. The importance of resource awareness in artificial intelligence for healthcare. *Nat Mach Intell* 2023;5:687-98. doi:10.1038/s42256-023-00670-0
121 Heidenreich PA, Bozkurt B, Aguilar D, et al, ACC/AHA Joint Committee Members. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* 2022;145:e895-1032. doi:10.1161/CIR.0000000000001063
122 Liberman L, Menell JH. Breast imaging reporting and data system (BI-RADS). *Radiol Clin North Am* 2002;40:409-30, v. doi:10.1016/S0033-8389(01)00017-3
123 DICOM. https://www.dicomstandard.org/
124 Petkovic D. A survey of artificial intelligence risk assessment methodologies. https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf
125 Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Rev Biomed Eng* 2021;14:156-80. doi:10.1109/RBME.2020.3013489
126 Cheatham B, Javanmardian K, Samandari H. Confronting the risks of artificial intelligence. 2019. https://www.semanticscholar.org/paper/Confronting-the-risks-of-artificial-intelligence-Cheatham-Javanmardian/2022830bf9f896d99c1d7ff228e3adca08efe352
127 Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. *Array (N Y)* 2022;16:100258. doi:10.1016/j.array.2022.100258.
128 Mehmood A, Yang S, Feng Z, et al. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience* 2021;460:43-52. doi:10.1016/j.neuroscience.2021.01.002
129 Pianykh OS, Langs G, Dewey M, et al. Continuous learning AI in radiology: Implementation principles and early applications. *Radiology* 2020;297:6-14. doi:10.1148/radiol.2020200038
130 Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med* 2023;6:186. doi:10.1038/s41746-023-00927-3.
131 Nan Y, Ser JD, Walsh S, et al. Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions. *Inf Fusion* 2022;82:99-122. doi:10.1016/j.inffus.2022.01.001
132 Fortin JP, Parker D, Tunç B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 2017;161:149-70. doi:10.1016/j.neuroimage.2017.08.047
133 Kilintzis V, Kalokyri V, Kondylakis H, et al. Public data homogenization for AI model development in breast cancer. *Eur Radiol Exp* 2024;8:42. doi:10.1186/s41747-024-00442-4
134 How should my consent be requested? European Commission. https://commission.europa.eu/law/law-topic/data-protection/reform/rights-citizens/how-my-personal-data-protected/how-should-my-consent-be-requested_en
135 Vellido A. Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis (Basel)* 2019;5:11-7. doi:10.1159/000492428.
136 Alsaleh MM, Allery F, Choi JW, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: a systematic review. *Int J Med Inform* 2023;175:105088. doi:10.1016/j.ijmedinf.2023.105088
137 Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020;3:119. doi:10.1038/s41746-020-00323-1.
138 Linardos A, Kushibar K, Walsh S, Gkontra P, Lekadir K. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Sci Rep* 2022;12:3551. doi:10.1038/s41598-022-07186-4.
139 Yigzaw KY, Olabarriaga SD, Michalas A, et al. Health data security and privacy: challenges and solutions for the future. Roadmap to successful digital health ecosystems: a global perspective.2022;1;335-62. doi:10.1016/B978-0-12-823413-6.00014-8
140 General Data Protection Regulation (GDPR) Compliance Guidelines. 2023. https://gdpr.eu/

141 Centers for Disease Control and Prevention. Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html"https://www.cdc.gov/phlp/publications/topic/hipaa.html"
142 Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *J Am Med Inform Assoc* 2023;30:1561-6. doi:10.1093/jamia/ocad115
143 Li Y, Renqiang Min M, Lee T, et al. Towards robustness of deep neural networks via regularization. https://openaccess.thecvf.com/content/ICCV2021/html/Li_Towards_Robustness_of_Deep_Neural_Networks_via_Regularization_ICCV_2021_paper.html
144 Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. *BMC Med Imaging* 2022;22:69. doi:10.1186/s12880-022-00793-7.
145 Meng H, Lin Z, Yang F, et al. Knowledge distillation in medical data mining: a survey. ACM International Conference Proceeding Series. 2021;175-82. https://dl.acm.org/doi/10.1145/3503181.3503211
146 Soltan A, Washington P. Challenges in reducing bias using post-processing fairness for breast cancer stage classification with deep learning. *Algorithms* 2024;17:141. doi:10.3390/a17040141.
147 Seoni S, Jahmunah V, Salvi M, Barua PD, Molinari F, Acharya UR. Application of uncertainty quantification to artificial intelligence in healthcare: a review of last decade (2013-2023). *Comput Biol Med* 2023;165:107441. doi:10.1016/j.compbiomed.2023.107441
148 Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022;4:e384-97. doi:10.1016/S2589-7500(22)00003-6
149 Zhai S, Wang H, Sun L, et al. Artificial intelligence (AI) versus expert: a comparison of left ventricular outflow tract velocity time integral (LVOT-VTI) assessment between ICU doctors and an AI tool. *J Appl Clin Med Phys* 2022;23:e13724. doi:10.1002/acm2.13724
150 Shen J, Zhang CJP, Jiang B, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Med Inform* 2019;7:e10010. doi:10.2196/10010
151 Watson OpenScale fairness metrics – Docs. IBM Cloud Pak for Data as a Service. https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-fairness-metrics-ovr.html?context=cpdaas
152 Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013;26:1045-57. doi:10.1007/s10278-013-9622-7
153 Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12:e1001779. doi:10.1371/journal.pmed.1001779
154 Campello VM, Gkontra P, Izquierdo C, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms Challenge. *IEEE Trans Med Imaging* 2021;40:3543-54. doi:10.1109/TMI.2021.3090082
155 Garrucho L, Reidel CA, Kushibar K, et al. MAMA-MIA: a large-scale multi-center breast cancer DCE-MRI benchmark dataset with expert segmentations.*arXiv* 2024. http://arxiv.org/abs/2406.13844
156 Bakas S, Reyes M, Jakab A, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. Sandra Gonzlez-Vill. 2018. https://arxiv.org/abs/1811.02629v3
157 Kuo MD, Chiu KWH, Wang DS, et al. Multi-center validation of an artificial intelligence system for detection of COVID-19 on chest radiographs in symptomatic patients. *Eur Radiol* 2023;33:23-33. doi:10.1007/s00330-022-08969-z
158 Singhal L, Garg Y, Yang P, et al. eARDS: A multi-center validation of an interpretable machine learning algorithm of early onset acute respiratory distress syndrome (ARDS) among critically ill adults with COVID-19. *PLoS One* 2021;16:e0257056. doi:10.1371/journal.pone.0257056
159 Sun L, Bull SB. Reduction of selection bias in genomewide studies by resampling. *Genet Epidemiol* 2005;28:352-67. doi:10.1002/gepi.20068
160 Dang VN, Cascarano A, Mulder RH, et al. Fairness and bias correction in machine learning for depression prediction across four study populations. *Sci Rep* 2024;14:7848. doi:10.1038/s41598-024-58427-7.
161 Kondylakis H, Catalan R, Alabart SM, et al. Documenting the de-identification process of clinical and imaging data for AI for health imaging projects. *Insights Imaging* doi:10.1186/s13244-024-01711-x
162 Paz F, Pow-Sang JA. Current trends in usability evaluation methods: a systematic review. Proceedings - 7th International Conference on Advanced Software Engineering and Its Applications, ASEA 2014;11-15.
163 Shackel B. Usability – Context, framework, definition, design and evaluation. *Interact Comput* 2009;21:339-46. doi:10.1016/j.intcom.2009.04.007.
164 Hajesmaeel-Gohari S, Khordastan F, Fatehi F, Samzadeh H, Bahaadinbeigy K. The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health. *BMC Med Inform Decis Mak* 2022;22:22. doi:10.1186/s12911-022-01764-2
165 Topff L, Ranschaert ER, Bartels-Rutten A, et al. Artificial intelligence tool for detection and worklist prioritization reduces time to diagnosis of incidental pulmonary embolism at CT. *Radiol Cardiothorac Imaging* 2023;5:e220163. doi:10.1148/ryct.220163
166 Nelson R, Whitener E, Philcox H. The assessment of end-user training needs. *Commun ACM* 1995;38:27-39. doi:10.1145/213859.214793.
167 Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health* 2024;6:e367-73. doi:10.1016/S2589-7500(24)00047-5
168 Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. *Brief Bioinform* 2019;20:1699-708. doi:10.1093/bib/bby043.
169 van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes?*Pediatr Radiol* 2022;52:2087-93. doi:10.1007/s00247-021-05114-8
170 Medina-Lara A, Grigore B, Lewis R, et al. Cancer diagnostic tools to aid decision-making in primary care: mixed-methods systematic reviews and cost-effectiveness analysis. *Health Technol Assess* 2020;24:1-332. doi:10.3310/hta24660
171 Schwendicke F, Rossi JG, Göstemeyer G, et al. Cost-effectiveness of artificial intelligence for proximal caries detection. *J Dent Res* 2021;100:369-76. doi:10.1177/0022034520972335
172 Khanna NN, Maindarkar MA, Viswanathan V, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. *Healthcare (Basel)* 2022;10:2493. doi:10.3390/healthcare10122493
173 Han R, Acosta JN, Shakeri Z, et al. Randomized controlled trials evaluating AI in clinical practice: a scoping evaluation.*medRxiv* 2023. doi:10.1101/2023.09.12.23295381
174 Park SH, Choi JI, Fournier L, Vasey B. Randomized clinical trials of artificial intelligence in medicine: why, when, and how? *Korean J Radiol* 2022;23:1119-25. doi:10.3348/kjr.2022.0834
175 Yale Test-Retest Dataset. https://fcon_1000.projects.nitrc.org/indi/retro/yale_trt.html
176 Raisi-Estabragh Z, Gkontra P, Jaggi A, et al. Repeatability of cardiac magnetic resonance radiomics: a multi-centre multi-vendor test-retest study. *Front Cardiovasc Med* 2020;7:586236. doi:10.3389/fcvm.2020.586236
177 M&Ms challenge. 2020. https://www.ub.edu/mnms/
178 Tsamos A, Evsevleev S, Bruno G. Noise and blur removal from corrupted X-ray computed tomography scans: a multilevel and multiscale deep convolutional framework approach with synthetic training data (BAM SynthCOND). *Tomogr Mat Struct* 2023;2:100011. doi:10.1016/j.tmater.2023.100011 .
179 Campello VM, Martín-Isla C, Izquierdo C, et al. Minimising multi-centre radiomics variability through image normalisation: a pilot study. *Sci Rep* 2022;12:12532. doi:10.1038/s41598-022-16375-0.
180 Sovrano F, Vitali F. Highlights. An objective metric for explainable AI: how and why to estimate the degree of explainability. 2023. https://github.com/Francesco-Sovrano/DoXpy
181 Ahmed N, Alpkoçak A. A quantitative evaluation of explainable AI methods using the depth of decision tree. 2022. https://journals.tubitak.gov.tr/cgi/viewcontent.cgi?article=3924&context=elektrik
182 Nauta M, Trienes J, Pathak S, et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *arXiv* 2022;55. http://arxiv.org/abs/2201.08164
183 Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion* 2021;76:89-106. doi:10.1016/j.inffus.2021.05.009.
184 Mertes S, Huber T, Weitz K, Heimerl A, André E. GANterfactual-Counterfactual explanations for medical non-experts using generative adversarial learning. *Front Artif Intell* 2022;5:825565. doi:10.3389/frai.2022.825565
185 Vasconcelos H, Jörke M, Grunde-Mclaughlin M, Bernstein MS, Krishna R, Gerstenberg T. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction 2023;7. doi:10.1145/3579605.
186 Riveiro M, Thill S. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artif Intell* 2021;298:103507. doi:10.1016/j.artint.2021.103507.
187 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689. doi:10.1136/bmj.m689.
188 Fel T, Vigouroux D, Cadène R, Serre T. How good is your explanation? Algorithmic stability measures to assess the quality of explanations for deep neural networks. *arXiv* 2021. doi:10.48550/arXiv.2009.04521

189 Voulgaridis K, Lagkas T, Angelopoulos CM, Boulogeorgos AAA, Argyriou V, Sarigiannidis P. Digital product passports as enablers of digital circular economy: a framework based on technological perspective. *Telecommun Syst* 2024;85:699-715. doi:10.1007/s11235-024-01104-x
190 Omoteso K. The application of artificial intelligence in auditing: looking back to the future. *Expert Syst Appl* 2012;39:8490-5. doi:10.1016/j.eswa.2012.01.098.
191 Spolaôr N, Lee HD, Mendes AI, et al. Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets. *Multimed Tools Appl* 2024;83:27305-29. doi:10.1007/s11042-023-16529-w
192 Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun* 2020;11:5131. doi:10.1038/s41467-020-18918-3.
193 Malik H, Farooq MS, Khelifi A, Abid A, Nasir Qureshi J, Hussain M. A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging. *IEEE Access* 2020;8:139367-86. doi:10.1109/ACCESS.2020.3004766.
194 Wang Y, Nazir S, Shafiq M. An overview on analyzing deep learning and transfer learning approaches for health monitoring. *Comput Math Methods Med* 2021. doi:10.1155/2021/5552743.
195 Zhang O, Delbrouck JB, Rubin DL. Out of distribution detection for medical images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021:102-11. https://link.springer.com/chapter/10.1007/978-3-030-87735-4_10
196 Nikiforaki K, Karatzanis I, Dovrou A, et al. Image quality assessment tool for conventional and dynamic magnetic resonance imaging acquisitions. *J Imaging* 2024;10:115. doi:10.3390/jimaging10050115.
197 Cipollari S, Guarrasi V, Pecoraro M, et al. Convolutional neural networks for automated classification of prostate multiparametric magnetic resonance imaging based on image quality. *J Magn Reson Imaging* 2022;55:480-90. doi:10.1002/jmri.27879
198 Xia Y, Zhang Y, Liu F, Shen W, Yuille AL. Synthesize then compare: detecting failures and anomalies for semantic segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020:145-61. https://link.springer.com/chapter/10.1007/978-3-030-58452-8_9
199 Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ. Calibration of machine learning models. https://dmip.webs.upv.es/papers/BFHRHandbook2010.pdf
200 Gupta SK, Singh H, Joshi MC, Sharma A. Digital dashboards with paradata can improve data quality where disease surveillance relies on real-time data collection. *Digit Health* 2023;9:20552076231164098. doi:10.1177/20552076231164098
201 Li J, Jin L, Wang Z, et al. Towards precision medicine based on a continuous deep learning optimization and ensemble approach. *NPJ Digit Med* 2023;6:18. doi:10.1038/s41746-023-00759-1.
202 Ao SI, Fayek H. Continual deep learning for time series modeling. *Sensors (Basel)* 2023;23:7167. doi:10.3390/s23167167.
203 Quarta A, Bruno P, Calimeri F. Continual learning for medical image classification. 2022. http://ceur-ws.org
204 Lee CS, Lee AY. Applications of continual learning machine learning in clinical practice. *Lancet Digit Health* 2020;2:e279. doi:10.1016/S2589-7500(20)30102-3
205 Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Comput Biol Med* 2023;158:106848. doi:10.1016/j.compbiomed.2023.106848
206 Allen B. The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices. *J Am Coll Radiol* 2019;16:208-10. doi:10.1016/j.jacr.2018.09.007
207 Gilbert S. The EU passes the AI Act and its implications for digital medicine are unclear. *NPJ Digit Med* 2024;7:135. doi:10.1038/s41746-024-01116-6.
208 Center for Devices and Radiological Health. De Novo Classification Process (Evaluation of Automatic Class III Designation) Guidance for Industry and Food and Drug Administration Staff Preface Public Comment. 2021.https://www.fda.gov/regulatory-information/search-fda-guidance-documents/de-novo-classification-process-evaluation-automatic-class-iii-designation

Web appendix 1: Appendix
Web appendix 2: FUTURE-AI Consortium