Swiss Learning Health System

Quality & risk management of ethical AI use in human health research

Written by: Eliane Maalouf

Supervised by: MD-PhD Marie-Annick Le Pogam, PhD Paul Cotofrei

Keywords

Ethical AI in Health Research

Ethical-AI assessment

AI Governance

Authors

Eliane Maalouf, Scientific Collaborator – Information Management Institute, University of Neuchâtel, Neuchâtel, Switzerland

MD Marie-Annick Le Pogam, Associate Physician & Lecturer – Department of Epidemiology and Health Systems, Unisanté, Lausanne, Switzerland

Paul Cotofrei, Senior Lecturer –Information Management Institute, University of Neuchâtel, Neuchâtel, Switzerland

Address for correspondence

Eliane Maalouf
Information Management Institute
University of Neuchâtel
A.L. Breguet 2, 2000 Neuchâtel
E-Mail: eliane.maalouf@unine.ch

MD Marie-Annick Le Pogam
Department of Epidemiology and Health Systems
Unisanté
Route de la Cornich 10, 1010 Lausanne
E-Mail: marie-annick.le-pogam@unisante.ch

Paul Cotofrei
Information Management Institute
University of Neuchâtel
A. L. Breguet 2, 2000 Neuchâtel
E-Mail: paul.cotofrei@unine.ch

Suggested citation

The text of this policy brief may be freely quoted and printed, provided proper acknowledgement is given.

Maalouf E., Le Pogam M.-A., Cotofrei P. (2025). Quality & risk management of ethical AI use in human health research. Swiss Learning Health System.

Table of Contents

Policy Briefs and Stakeholder Dialogues of the Swiss Learning Health System

The Swiss Learning Health System (SLHS) was established as a nationwide project in 2017, involving academic partners across Switzerland. One of its overarching objectives is to bridge research, policy, and practice by providing an infrastructure that supports learning cycles.

Learning cycles enable the continuous integration of evidence into policy and practice by:

- continuously identifying issues relevant to the health system,
- systemizing relevant evidence,
- presenting potential courses of action, and
- if necessary, revising and reshaping responses.

Key features of learning cycles in the SLHS include the development of **Policy Briefs** that serve as a basis for **Stakeholder Dialogues**.

A **Policy Brief** describes the issue at stake by explaining the relevant contextual factors. It formulates a number of recommendations to address the issue (evidence-informed recommendations, when available), and for each possible recommendation, it explains relevant aspects and potential barriers and facilitators to their implementation. Policy Briefs serve as standalone products to inform interested audiences on potential courses of actions to address the issue, as well as input for Stakeholder Dialogues.

A **Stakeholder Dialogue** is a structured interaction where a variety of key stakeholders are brought together for the purpose of defining a common ground and to identify areas of agreement and disagreement on how to solve issues in the Swiss health system. Based on a Policy Brief, stakeholders discuss the issue, recommendations, and barriers and facilitators, and work collaboratively towards a common understanding of the issue and the best course of action. The dialogue takes the form of a deliberation to ensure that stakeholders work together to develop an understanding and solutions that are acceptable to all parties.

Key Messages

Background and Context

Everyone has a fundamental right to benefit from scientific advancement, especially in healthcare. The development of Artificial Intelligence/Machine Learning (AI/ML) has accelerated significantly, attracting public attention and investment. While AI promises to revolutionize healthcare and health research with more accurate diagnostics, personalized treatments, and streamlined processes, it also poses risks to fundamental rights and raises concerns about trustworthiness. Regulatory bodies like the EU are responding with frameworks such as the "AI Act," which establishes different risk levels for AI systems, categorizing medical devices using AI as high-risk but exempting scientific research. Despite increasing use of AI in health research, as evidenced by rising approvals for AI-related projects, there is a lack of specific guidance on how health research stakeholders should practically assess and monitor AI systems within their ethical obligations. This absence of clear guidance can lead to quality issues in health-AI research.

The Issue

Evaluating and reporting on ethical AI use in health research is a multidimensional process. This complexity arises from several interacting factors:

- **Multiple principles underlying ethical AI**, such as usefulness, fairness, safety, transparency, security, and privacy, each requiring specific metrics and monitoring.
- **Interdisciplinary stakeholders**, including health researchers, AI developers, end-users, ethical committees, regulators, funding agencies, and scientific publishers, who may have different perspectives and incentives. Ineffective collaboration among these stakeholders can hinder user-centric innovation and data sharing.
- **Characteristics of the AI system** (e.g., stage of development, algorithm type, integration level) combined with wide-ranging clinical research questions and data considerations (bias, quality, consent), all of which can introduce specific ethical risks.
- **A diversity of guidelines and frameworks** for quality assessments and reporting on AI use, making it difficult for researchers to choose the most appropriate ones for their specific context. The applicability and continuous evaluation of these guidelines remain challenging. Furthermore, a cultural conflict exists between the guideline-reliant health research culture and the more risk-tolerant digital innovation culture. The lack of structured quality assessment can slow down the implementation of AI models in practice.

Recommendations for Action

Instead of proposing new frameworks, the focus should be on operationalizing ethical AI practices during research projects and beyond through an interdisciplinary and agile process. Key recommendations include:

- **Embed ethical AI considerations in the project life cycle**: Integrate ethical assessments from the initial idea generation through design, development, evaluation, and clinical translation. This involves forming a **project advisory committee** with diverse stakeholders for shared decision-making and accountability. Design a specific ethical **AI assessment and reporting strategy** tailored to the project's unique challenges and select validated assessment and reporting frameworks whenever possible. Publish information about this strategy to benefit the wider scientific community.

- **Professionalize health-AI projects portfolio in hosting organizations**: Organizations should maintain a portfolio of health-AI projects with stage-gate evaluations aligned with the project lifecycle. Building "**ethics as a service**" capacities within organizations can provide expertise, optimize resource use, and ensure compliance. Risk and quality management offices should extend their oversight to ethical-AI assessments and reporting.
- **Ethical health-AI motivated funding in research and private investments**: Funding bodies and scientific publishers should encourage and incentivize projects demonstrating rigorous ethical AI assessments from the outset and enforce transparency and reproducibility requirements. Support for evaluation science, operationalization efforts, and data/code sharing is crucial. Private funders also have a responsibility to incentivize attention to societal responsibilities in AI products they finance.
- **Provide guidance on regulation applicability & agile regulatory processes**: Establish clear oversight and accountability mechanisms for AI performance. Regulators need to increase agility in their processes by reinforcing regulatory science, building capacity, involving stakeholders, encouraging self-regulation, proactively screening new technologies, and initiating public-private partnerships for "**AI assurance laboratories**".

Implementation Considerations

The current transitory phase of AI regulation in Switzerland can be challenging. A proactive approach to ethical considerations is strongly advised, with a shared responsibility among all stakeholders. Main barriers to break include **inefficient communication/coordination between stakeholders, the regulatory transitory period, and a lack of clarity on who is responsible for AI system validation**. To counter these challenges, initiatives could include early stakeholder consultations in permanent forums, integrating AI education into relevant curricula, ethics committees proposing recommended documentation for ethical-AI in research protocols, organizations providing support through data scientists and quality analysts, and a shared responsibility model for system validation involving development and operational teams with ongoing monitoring and audits.

Background and Context

“Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.” Article 27 of the universal declaration of human rights [1]. In no other human endeavor is the right of all humans to benefit from scientific advancement as fundamental as in healthcare. In no other time in human history have we developed a technology that can reinforce the betterment of health on large scales while simultaneously posing risks of undermining this same fundamental right and infringing on others.

The development of Artificial Intelligence/Machine Learning (AI/ML or simply AI) accelerated considerably in the last decades. Submitted papers in the main conferences on the topic shows a year on year increase anywhere between 1000 and 3000 papers [2]. Since the advent of ChatGPT[1] in November 2022, the public attention oriented towards AI grew immensely (see a comparison of Google search terms[2] trends) and investment in AI peaked [3]. “AI” is commonly used as a lump sum term for all “human like intelligent” machine behavior, while the landscape of AI methods is very diverse. With chat interfaces, AI became easily accessible and interactive [4] not only hidden in the background of services in recommender systems, internet of things or image processing tools for example. AI is an ubiquitous topic on social media and the news with predictions for the future ranging from reassessment of expectations [5] to extinction level threat to humanity [6]. The hype and the constant stream of new AI-powered products seems to be also fueling a “fear of missing out” anxiety among technology leaders [7] and an eagerness to implement AI solutions despite lack of return on investments guarantees [8].

AI is revolutionizing almost all sectors of the economy and healthcare, respectively health research is no exception [9]. The use of AI in health research is becoming more and more frequent. From 2016 to 2023, Swiss Association of Research Ethics Committees (swissethics) approvals of AI related projects[3] are increasing (from 0.065% of all approvals in 2016, to 1.463% in 2023; in September 2024, this proportion was already at 2.163% for 2024). AI/ML algorithms promise to accelerate the pace at which knowledge can be gleaned from complex health data in the hopes of more accurate diagnostics, personalized treatments, predictive health insights, streamlining care processes, a better control over the health systems costs and other benefits.

In the public and the scientific debate, an important focus is put on AI trustworthiness and responsible development and deployment. It is now common knowledge that the positive promises of AI are associated with risks to privacy, safety, security, autonomy, equality and other issues. On the OECD.AI policy observatory's “AI Incidents Monitor”[4] we see a steady increase of reported incidents in the category “Healthcare, drugs and biotechnology”. The economic incentives and the competition structure on the market does not favor setting AI safety

[1] https://chatgpt.com/

[2] https://tinyurl.com/GoogleTrendsAI

[3] https://raps.swissethics.ch/runningProjects_list.php?goto=920&orderby=dBASECID

[4] https://www.oecd.org/en/topics/ai-risks-and-incidents.html

as a priority for industrial AI leaders [10]. Regulatory bodies are counter balancing this “neglect” by putting AI risks and their mitigation as political priorities. The EU “AI Act” [11], into force since August 1st 2024, is the first legal framework on AI. The act establishes four levels of risks of AI systems: unacceptable risk, high-risk, limited risk and minimal risk [12]. Most of the act focuses on high-risk AI systems and requires their developers/providers to report, among others, on the data used for training, proofs of risks assessment and mitigation measures, logging of activities for traceability, human oversight and monitoring. High-risk systems are required to undergo a conformity assessment, self-lead or via a third-party [12]. The AI Act is a cross-domains initiative and it does not adopt specific dispositions for the health sector or health research, only explicitly categorizing medical devices using AI as high-risk. The act fully exempts scientific research from any obligations, hence allowing medical research and clinical trials to “escape” the regulation without necessarily guaranteeing lower risks to patients [13]. At this early stage of the regulations development, and given the act’s focus on AI developers/providers obligations, an important void is left unfilled on how health research stakeholders could practically, and in a standardized way, assess and monitor AI systems integrated in their research as part of their ethical obligations towards human participants. We consider the ethical obligations in the broader sense, not only in terms of harm minimization, but also in terms of scientifically proven effectiveness and usefulness of any newly developed tool or process (see swissethics mission statement [14]). AI regulation in Switzerland is still at the drafting stage[5] and does not provide more guidance in this regard. In 2020, the Swiss confederation adopted the guidelines[6] that will orient its AI regulations efforts, the third one of those being “Transparency, traceability and explainability” which we speculate might lead to requirements in terms of oversight and monitoring; and the fourth being “Responsibility” which is expected to define accountabilities in cases of damage, accidents or rights’ infringement. More recently a clear direction towards incorporating the European AI convention in Swiss law was publicized[7]. Specifically in health research involving human subjects, the currently revised ordinances[8] do not mention the topic of the use of AI in research – emphasis was put on data protection and data security as well as the need to include health information technologies experts in the ethics committees [15]. While privacy protection and data governance remain essential, it is equally important to investigate AI ethics issues [16]. The CER-VD ethics committee has a one page memo with a list of questions to answer on the themes of: justification of the AI use, justification of the AI model used, data confidentiality and security, and responsibility [17]. The memo does not provide practical guidance on how to go about putting in place processes for quality control, transparency, security and other dimensions that it requires reporting on.

The absence of clear guidance in health research is leading to multiple health-AI research findings suffering from quality issues such as: low computational reproducibility [18][19][20];

[5] https://digital.swiss/en/strategy/focus-topics/swiss-approach-to-regulating-ai-systems

[6] https://www.sbfi.admin.ch/sbfi/fr/home/actualite/communiques-de-presse/test-bit.msg-id-81319.html

[7] https://www.bakom.admin.ch/bakom/en/homepage/digital-switzerland-and-internet/strategie-digitale-schweiz/ai.html

[8] https://www.bag.admin.ch/bag/en/home/medizin-und-forschung/forschung-am-menschen/revision-verordnungen-hfg.html

low data quality ("data leakage" risks) [19] or biased data [21]; clinically irrelevant or wrong metric choices [19]; lack of standard modeling and evaluation procedures [19][20]; limited use of reporting standards [18][20][22] impacting transparency, lack of strong evidence of clinical effectiveness [22] and low robustness of results [18][19]; methodological conduct issues (inadequate sample size [18], lack of external validation [18], limited assessment of calibration [18], limited assessment of bias [18], limited error analysis [18][19]). Unfortunately, the academic environment is still encouraging for publishing "best-performing" AI prototypes instead of rigorously verified systems [20]. This list is not exhaustive, but we can already see a systematic relationship and dependence between the issues. Those issues increase the risks of undermining trust in the scientific endeavor in general and require a renewed attention to the hypothesis made, methodologies used and healthy skepticism about the results. AI-based science (i.e., making scientific claims using the performance of AI models [19]) should still adhere to scientific principles despite its complex challenges (i.e., explainability, reproducibility, governance, ethical implications) [18][20][22], as the burden of proof for ensuring the validity of the claims made falls on the scientists making them [19]. Unlike existing health technologies, health-AI is still missing universally accepted measures for quality assurance [23].

Health research stakeholders have a responsibility to address those challenges proactively by establishing AI-systems quality and applicability assessments [24]. However, we believe that their task is complex for four main reasons which we develop further in section “The Issue”: (1) assessments underlain by multiple dimensions of ethical AI; (2) the interdependence between the stakeholders involved; (3) the characteristics of the AI system itself and the research question being investigated; (4) a plethora of guidelines and frameworks for ethical AI available in the scientific literature. Unfortunately, health researchers might be left alone to make the difficult choice of the right guideline for the specific use case they have [20]. In this Policy Brief (PB), in a narrative review of scientific publications and relevant grey literature, we aim to provide initial recommendations on how to go about answering the following questions:

- **How could health research stakeholders continuously monitor the efficacy and the risks of AI systems used in research projects?**
- **What reporting would be necessary to prepare for and adhere to, potentially upcoming, evaluation requirements on the AI system to ethical committees, regulators or scientific reviewers?**
- **What is needed to transition an AI system monitoring from the research phase to the operational phase in real-life usage settings?**

The Issue

Evaluating and reporting on ethical AI use in health research is a multidimensional process reflecting interactions between: (1) multiple principles underlying ethical AI; (2) interdisciplinary stakeholders; (3) characteristics of the AI system combined with wide-ranging clinical research questions; and (4) diversity of guidelines and frameworks for quality assessments and reporting on AI use.

Ethical AI principles

Ethical AI is a set of principles that defines how AI systems align with moral principles and societal values. An extensive review of AI ethics guidelines in [25] shows an emerging set of eleven principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, solidarity. The same research identified that those principles do not share common definitions between the guidelines. For the purposes of our discussion, we focus on an aggregate subset of core principles defined by the Coalition for Health AI (CHAI)[9] in their "Assurance Standards Guide" [26]:

Usefulness, Usability & Efficacy: AI should address specific challenges and provide clear advantages for patients and healthcare providers, such as improved clinical outcomes and enhanced patient satisfaction. It should be user-friendly and seamlessly integrate into existing workflows. AI achieves its goals and maintains good performance over time via continuous testing and monitoring.
Fairness, Equity & Bias Management: AI should deliver consistent performance across diverse groups, ensuring it does not rely on protected attributes like race or gender. It should contribute to reducing health disparities. Bias should be managed by regular monitoring and correction of data and the AI system.
Safety & Reliability: Ensure thorough testing and risk management before implementation to avoid harm. Clear accountability and governance structures must be in place to guarantee safety and reliability over time.
Transparency, Intelligibility & Accountability: It is essential to provide clear and accessible information about how the AI system operates, its outputs, and its limitations. The decision-making process of the AI should be transparent and understandable to all stakeholders. Responsibilities for minimizing harm and addressing adverse effects must be clearly outlined and defined.
Security & Privacy: The AI system must incorporate robust security measures to ensure data is handled in compliance with privacy regulations. Ongoing security monitoring protocols should be implemented to address incidents promptly and safeguard data effectively.

Another classification around six principles of "**Fairness, Universality, Traceability, Usability, Robustness, Explainability**" can be found in [23]. We note here that we are not favoring any classification over the other. Both works in [26][23] are very recent and returns on their usefulness is not yet present; however, both are consensus based involving large groups of stakeholders which might be favorable for their adoption in practice.

[9] https://chai.org/

The terms "responsible AI" and "trustworthy AI" are also often used as stand-in for "ethical AI", encompassing similar principles without additional specifications, while the terms "trust" and "responsibility" appear as building principles of ethical AI in many guidelines. We will restrict ourselves to the term "ethical AI" to refer to the umbrella term encompassing all the principles mentioned before. Each of these principles requires an adapted set of metrics and/or checklists to be defined, measured, monitored and reported on. Not to mention that each dimension brings its technical implementation challenges and would influence the AI-system characteristics (e.g., choice of model as tradeoff between explainability and accuracy, use of federated learning instead of central model training for privacy, etc.).

Stakeholders' diversity

The landscape of health-AI research stakeholders is diverse and interdisciplinary. We list them in Table 1.

Table 1 – Main Health AI research stakeholders

Stakeholder	Description	Examples
Health researchers	Principal investigators, usually with limited AI experience. Strong methodological conduct.	Clinical, biomedical, public health, health services researchers
Research organization	Structure hosting the health AI research, typically where the health researchers are issued from. These could have varying degrees of expertise and resources for AI experimentation and deployment at scale (e.g., cloud infrastructure, data engineering and data science teams, etc.)	university hospital, pharmaceutical company, university department
AI developers/providers	Main experts on the AI systems. They can be issued from the hosting research organization or from tech companies entering the health AI space whether big-tech (e.g., Alphabet, Amazon, Meta, etc.) or start-ups.	Data scientist, AI researchers, programmers
AI-system end users	User of the AI tool for whom the system is designed to generate a benefit (e.g., improved diagnostics, optimized clinical workflow, etc.)	Patients, clinical practitioners, healthcare professionals

Stakeholder	Description	Examples
Data contributors	These could be the same as the end-users, but not necessarily. Providers of datasets to train and fine-tune AI models.	Patients, healthcare professionals, clinicians, data brokers
Ethical committees	Reviewers and authorizers of projects' protocols and projects conducts follow up.	swissethics
Regulators and policy makers	Authorizers of healthcare solutions, creators of laws that dictate more or less stringent oversight and enforcers of the civil code in case of harm.	FOPH, swissmedic, local governments, local health authorities
Funding agencies	Reviewers of projects' protocols and providers of resources for projects development.	SNSF, foundations, swiss universities
Scientific publishers	Reviewers and curators of projects' results.	Journals, peer-reviewers

Other stakeholders that could also be involved/impacted: the scientific community at large, clinical implementers and the wider population of end-users. When health tech startups are involved, we should also include venture capitalists as funding stakeholders as well.

Landers et al. [27] identified three impediments clusters to responsible digital health in Switzerland, of which health-AI is a component: (1) Ineffective stakeholder collaboration due to their complex interdependence and lack of incentives for joint innovation result in distraction from user-centricity (seen as central to responsible digital health) in innovation and slower data-sharing; (2) lack, for digital health innovators, of ethical awareness and resources in implementing responsible innovation, yet their early-stage design choices have a long-lasting impact; (3) inadequate reaction of regulators to the digital health innovation pace, generating a regulatory uncertainty that hinders innovation and drives out compliance focused innovators.

AI system characteristics vs health research questions vs data

Health research involving an AI system is criticized for methodological and reporting issues. Given the study type (e.g., clinical trial, diagnostic accuracy study, etc.) it is advised to follow guidelines at protocol writing and reporting to guarantee the quality of study design, its delivery and its transparency, with or without AI. Nevertheless, when AI is involved, study design and transparency remain central in guaranteeing quality of such studies and efficient evaluation of their outcomes. In that regard, researchers are advised to assess whether specific guidelines are commonly used for their study type and whether extensions for AI specific reporting are provided [28]. These extensions are not necessarily oriented towards covering ethical AI use per se. However, they contain multiple elements of ethical AI assessment such as bias assessment, reproducibility, early clinical validation, usability assessments. There might be a need to complement those guidelines with other measures on non-covered dimensions. Furthermore, it is important to consider how the study design itself might negatively influence ethical dimensions (e.g., Randomized Controlled Trials suffer from limited inclusivity [29]) and how this influence might be reinforced in AI-systems. An understanding of underlying health inequalities [29] and social determinants of health [27] could help to proactively account for risk of bias and unfair access to the technology being developed.

The research question will define the purpose of the AI-system in the research project and beyond. This purpose orients multiple technical choices, such as the model family, the data to be used and how the new AI system would interact with existing systems and workflows. The assessment and reporting on the AI-system needs to take into account those choices and the risks they might incur. Table 2 presents a non-exhaustive list of AI-system characteristics and how they could induce ethical risks.

Table 2 - AI system characteristics and potential induced ethical AI risks

Dimension	Example risks to ethical AI
Stage	**Development** : Inherited biases and reinforcement of inequities [27][29]; lack of generalizability[19][30]; low clinical validity and superiority/non-inferiority proofs [31]; imbedding misaligned norms [32]. **Implementation**: security and privacy infringement [30]; low usability/understandability and inefficient integration in existing workflows [30]; unclear accountabilities for errors [30].
Developers/AI solution provenance	**Academia vs industry (e.g., big tech, startups):** low incentives/knowledge/resources for transparency and responsible development [27]; imbalanced accountabilities due to power dynamics among stakeholders [32][33]; loss of agency and concentration of normative and technical mastery with developers due to imposed/non-negotiated designs [32]. **Country (Europe vs US vs China) :** compliance with local and regional regulations due to different approaches to AI governance [33][34].
Type of algorithm	**Discriminative or generative**: inheriting and perpetuating biases from (big) data used in training [29]; introducing erroneous data/omitting important data [35]. **Model complexity**: lack of robustness (e.g., due to overfitting) or generalizability (e.g., due to underfitting) [36][37]; limited traceability and identification of erroneous outputs [36][38][39]; tradeoff between accuracy and transparency [37][40]; decreased trust in model output due to opacity and inability to explain decisions [37][41]; overconfidence in predictions [40]; legal compliance issues due to lack of traceability and opaque decision making [24][42]. **Deterministic or stochastic**: limited reproducibility of outcomes [31][43]; misuse and misunderstanding of outcomes impacting decision-making [16][44].
Integration & autonomy level	**Background processes or end-user facing**: introducing erroneous data/omitting important data [35]; compromising quality of care due to low integration

Dimension	Example risks to ethical AI
	in clinical workflows, overwhelming/complex interfaces, or inability to explain outcomes to patients/health professionals [40][45]; underlying data/patient privacy [30][40]; overconfidence (automation bias) in model outputs due to inability to judge clinical reliability and model limitations [24][29][40]; biased decisions when presentation of model outcomes lack contextual information on underlying data and its representativeness [46]. **Fully autonomous or human oversight**: Misplaced/unclear accountabilities for errors and adverse events [47][48]; low traceability and explainability [48]; perpetuating biases without human corrective measures [43][47]; harm to patients/users when not accounting for their current context [43][49]; ethics dumping [32]; inefficient human oversight due to complex interfaces, opaque outcomes or lack of knowledge [43][49]; impact to patients/users self-determination and autonomy when autonomous decisions are not guaranteed to be aligned with relevant users values [50].
Resources requirements	**Infrastructure and code management**: inefficient integration in existing clinical systems and workflows increasing risks of harm [45]; closed-source software limits ability to audit [45]; inability to continuously assess, evaluate and react to AI-system quality and risks due to onerous costs [34][51]; outdated AI-systems and data management infrastructures increase risks of harms to end-users and cybersecurity risks [34]; **One-time model tuning or periodic**: adversarial attacks and malicious data injection [30]; one-time model training would not account for context shifts (usually reflected in the data) over time while periodic updates may introduce new biases in newly collected data [52]; model efficacy and performance might change considerably confusing end-users and/or causing new harms [16]; loss of compliance with regulations and pre-acquired authorizations/validation [23][53].

The data itself deserves its own assessment, whether freshly collected, reused or acquired from brokers. Data manipulations risk to introduce spurious correlations between the outcome variable and the input features. "Data leakage", a leading cause of errors in AI/ML applications [19], is caused by a wrong separation of training and test datasets (e.g., by preprocessing on both training and test sets, presence of duplicates), the use of features that are proxies of the outcome, an interdependence between test and training observations (e.g., both sets contain measurements about the same units) and sampling biases (e.g., sampling the test set from a geography/ethnicity/population group and making claims about model performance in other geographies/ethnicities/groups). Researchers should be aware of the risks related to data collection, sampling, or pre-processing strategies and their influence on generalizability, irreproducibility as well as how sensitive model performance is to slight changes in data manipulations. Transparency in reporting on assessments of "data leakage", mitigations actions and data-manipulations are crucial for reproducibility/replicability of the results and for an efficient critical appraisal of the study design and conduct. Furthermore, the use of synthetic data is becoming more frequent as a way to preserve privacy. Synthetic data requires its own assessments (e.g., synthetic distribution similar to the original data distribution, usability in downstream tasks, re-identification risks) [54]. Finally, the question of consent to collection, storage, use and re-use/re-purposing of patients or health professionals data poses new legal and technical challenges, especially in cases where very large datasets are used for training or fine-tuning and requirements to implement the right to withdraw consent are imposed [24][47][49].

Diversity of frameworks and guidelines

There is currently a plethora of guidelines and frameworks related to ethical AI. Jobin et al. identified 84 [25]. Our limited, yet structured, search (methodology p. 37) revealed the existence of no less than 30 such frameworks/guidelines in the context of health and health research. Very recently, a review identified 26 reporting guidelines in medical artificial intelligence [55]. The authors in [20] compared the quality of 11 of such reporting frameworks. Despite high-quality guidelines, applicability remains a prominent challenge. Applicability seems to be limited by study design (e.g., guidelines focused on clinical trials and a gap exists for observational studies), by awareness about the guidelines and their enforcement within the research community [20]. Furthermore, despite the proliferation of guidelines and frameworks, an extensive and continuous evaluation of their quality and usefulness is still missing [20]. In health research, it is a necessary practice to compare and combine outcomes from research works addressing the same clinical question (e.g., systematic reviews, meta-analysis). Such reviews are crucial in informing clinical practice guidelines and health policies. Evidence appraisal is a critical process to evaluate the quality, relevance, and reliability of research findings. Standardized reporting, transparency and reproducibility are crucial [22]. Nevertheless, the absence of a structured quality assessment across the AI-system development, evaluation and implementation may be contributing to the slow implementation of those models in practice [24]. We also highlight here a conflict between the cultures of health research and digital innovation, where the first is reliant on adherence to guidelines, stability and accuracy while the latter is more tolerant of ambiguity and an open attitude to risk [27]. This cultural difference between stakeholders could also influence their intercommunication channels and expectations from the project. Inefficient communication might lead to the inefficient choices of guidelines (or specific subsets) and their implementation, if any.

We doubt that yet another framework is actually needed. Instead, we advocate to focus on processes and practices to operationalize the ethical AI practices during research projects and beyond. We propose to build and maintain an interdisciplinary and agile process for assessment and reporting on ethical AI use. Such a process will help account for the complexities illustrated previously by pulling on all stakeholders' resources to adapt to the inevitable changing nature and dynamic process of health-AI research and to prepare for regulatory and ethical compliance requirements.

Recommendations

The diversity of the challenges and their continuous evolution impedes the crafting of a "gold standard" and a one-size-fits-all framework that covers all aspects of developing and reporting on health-AI [20][23].
Our general recommendation is to focus on operationalizing the relevant frameworks for the project. This assumes both project level and organizational level quality management processes that allow for the selection of the correct frameworks, the definition of their implementation strategy and the allocation of necessary resources for such implementation and further maintenance. Even though each organization follows its own project and quality management philosophy, the general steps of those processes would include **Define, Measure, Analyze, Improve** and **Control** steps [56]. And there is no harm in reminding that ethical design, development, evaluation and implementation of an AI-system should follow the clinical environment norms coded into the principles of biomedical ethics and clinical research ethics [29]. In the following we make recommendations tailored at project level, organizational level and regulatory/policy level. Those recommendations do not form a complete set, they only aim to designate initial areas of action.

R.1 Embed ethical AI considerations in the project life-cycle

We recommend integrating ethical AI considerations in all the phases of health-AI research projects. Researchers could then fit their ethical assessments in a wider organizational AI governance and quality management processes (see O.1).

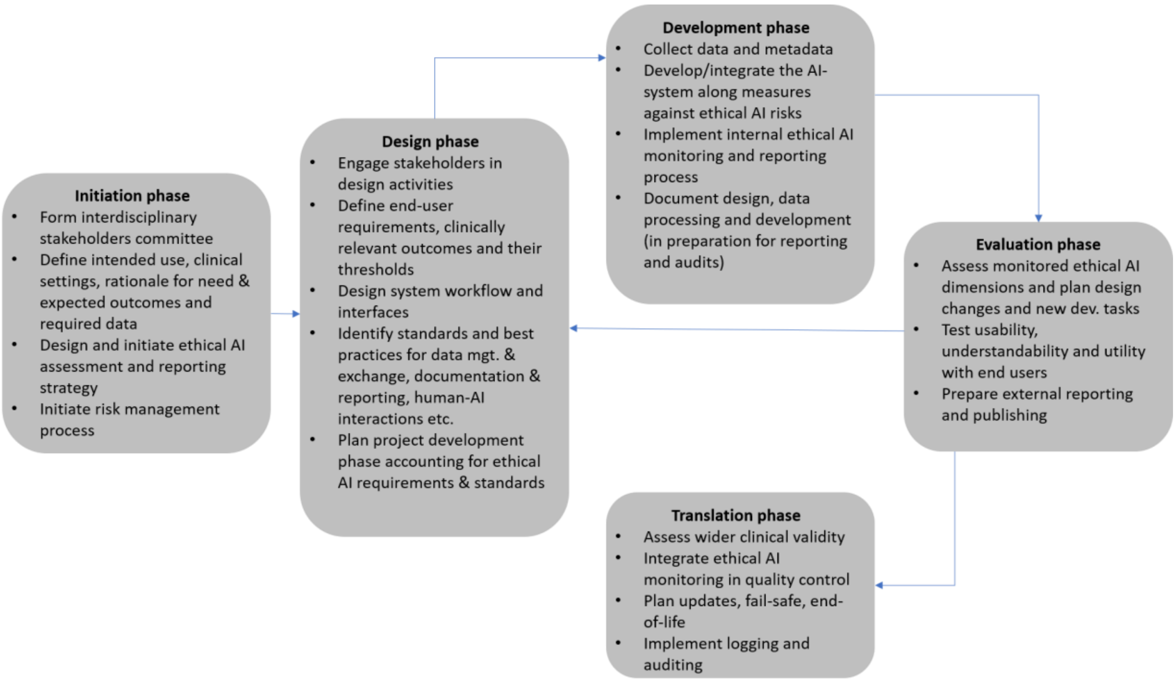
Initiation phase
Form interdisciplinary stakeholders committee
Define intended use, clinical settings, rationale for need & expected outcomes and required data
Design and initiate ethical AI assessment and reporting strategy
Initiate risk management process
Design phase
Engage stakeholders in design activities
Define end-user requirements, clinically relevant outcomes and their thresholds
Design system workflow and interfaces
Identify standards and best practices for data mgt. & exchange, documentation & reporting, human-AI interactions etc.
Plan project development phase accounting for ethical AI requirements & standards
Development phase
Collect data and metadata
Develop/integrate the AI-system along measures against ethical AI risks
Implement internal ethical AI monitoring and reporting process
Document design, data processing and development (in preparation for reporting and audits)
Evaluation phase
Assess monitored ethical AI dimensions and plan design changes and new dev. tasks
Test usability, understandability and utility with end users
Prepare external reporting and publishing
Translation phase
Assess wider clinical validity
Integrate ethical AI monitoring in quality control
Plan updates, fail-safe, end-of-life
Implement logging and auditing

Figure 1 - Agile health-AI project management process embedding ethical aspect. Extended from Lekadir et al. [23].

Figure 1 is an example of such integration. We did not include an exhaustive list of tasks at each phase. However, we bring the attention to a few priority elements.

First, we separate the initiation phase from the design phase to highlight the importance of this "pre-project" period in the whole ethical health-AI assessments procedure. In fact, the initiation phase could start much earlier than the research project official start date, namely, at idea generation and its related understanding (e.g., systematic reviews) and protocol writing. In this phase, a strong rational for the use of AI in the project needs to be supported with a clarification of intended use, the workflows it would impact, its expected clinical outcomes but also a clarification of ethical AI risks. An understanding of the breadth of ethical AI risks and what could cause them is crucial during this preparation (including design and conceptualization risks). Principal investigators are invited to engage with stakeholders early on to gather their perspectives [23][29]. Once the project is in fact planned to start, we advise for a structured approach of involving stakeholders by forming a project advisory committee representing stakeholders' groups (see composition example in p. 32). This committee would participate to the whole life-cycle of the project from design, development, evaluation and clinical translation. A shared decision-making process can be established with such committee to share responsibilities and accountabilities and avoid "ethics dumping" [32], where ethical responsibilities are offloaded from AI developers, regulators and ethics guidelines authors and dumped on end-users and local environments ill-equipped to handle them. Patient treatment strategies based on AI decisions should have a strong scientific foundation. Health system stakeholders should approve the AI's development and application to ensure its clinical utility and acceptance [24].

Second, we recommend to design an ethical AI assessment and reporting strategy. AI based technologies and algorithms are very diverse and evolving. Each research project involving AI components will have specificities and challenges that are most likely new. A specific assessment of the relevance and priorities of ethical-AI issues for the project needs to be done in order to guide the selection of definitions, metrics, and techniques. [22]. The choice of metrics is not trivial, as it is important to understand what each metric is actually measuring and how it can be impacted by the dataset composition (a very typical example is accuracy's sensitivity to class imbalance in the dataset). We remind of the importance of measuring standardized outcomes as much as possible and/or provide clear explanations how these outcomes were generated. Alongside model assessment outcomes, researchers should measure clinical outcomes that can be combined and compared across research works (using or not using AI).
Furthermore, given the early stage of regulating and reporting on health AI and the proliferation of frameworks, it is important that a specific quality and risks assessment strategy be designed for the project in order to assess novel recommendations and best practices at the outset of the project and plan necessary changes throughout. Such a strategy will prepare the project team to cover all relevant aspects of ethical AI integration and proactively prepare for reporting, publishing and external auditing when necessary. It also guarantees that the output of the research project is transparent to the community and their evidence is of high quality and actionable in future initiatives. The project stakeholders advisory committee contributes to crafting such a strategy by helping surface priority risks, their mitigation actions (e.g., procedural changes or via technical choices) and their monitoring metrics. The committee helps in planning the implementation throughout the project phases, change management and evaluation of monitored outcomes. We provide in p. 33 examples of recommendations to include in a strategy organized by ethical AI principles.

Finally, we recommend selecting commonly used assessment and reporting frameworks for a given dimension of interest, whenever possible. To support in this task, we started building a catalogue[10] of existing frameworks in health-AI, identified during our literature search. Importantly, not all frameworks are rigorously validated and it was out of our scope to assess the quality of the listed frameworks. It is common to select a framework that is often chosen in one domain of application or developed by members of that domain. However, we recommend that the project advisory committee to look for validated frameworks first and publications on the validation process to appraise quality, relevance and completeness of the frameworks and to provide guidance on filling possible gaps. The AGREE-II tool used in [20] could help to asses guidelines based on the following dimensions: scientific rigor via a systematic evaluation of evidence synthesis and an explicit link between the guideline and the body of evidence; external expert revisions; a procedure to update or modify the guideline; stakeholders involvement in the process of development; applicability via a detailed presentation of suggested tools and instructions for using the guidelines effectively; resource implications of applying the guidelines and whether monitoring or auditing criteria are presented.

We also recommend publishing, as supplementary materials to the research project, information about the ethical AI assessment and reporting strategy and its development. We believe this supplement could be very useful for the scientific community to start similar rigorous efforts and to support appraisal of the resulting outcomes.

R.2 Professionalize health-AI projects portfolio in hosting organizations

Relevant ethical AI assessments in the context of a health research project and clinical translation are numerous. The successful implementation of the assessment and reporting strategy depends on supporting components in the context of the research work, namely in the organization hosting the research (e.g., university hospital, research academic department or R&D company department, etc.) and the research industry at large, namely the scientific publishers and funding agencies.

Financial investment in developing a new AI-system or acquiring one are very large [57]. Decisions for such investments are not in the hands of the health-AI researchers themselves. A wider assessment of required resources (e.g., skilled professionals, computing infrastructure, interfacing with existing systems, etc.), real-world clinical impact and societal, environmental and reputational responsibilities, can only be done by the organizations hosting the research (or the ones planning to integrate the project's outcome). To maintain this big-picture view, we recommend organizations to maintain and professionally manage a health-AI projects portfolio. Each project in this portfolio goes through a set of evaluations by the end of which a decision needs to be made on whether to move on to the next phase or stop/pivot the project (stage-gate process) [58]. Those evaluations are synchronized with the project life-cycle phases presented in recommendation P.1. Health-AI portfolio management guarantees that the highest value initiatives are prioritized and that those initiatives are aligned with the organization's AI governance and data/cyber-security strategies which usually extend to other types of initiatives in the organization. FURM assessment framework is an example implementation of such a stage-gate process [57]. Figure 2 illustrates how such an organizational process coordinates with the specific project life-cycle. The project ethical AI assessment strategy and its implementation are coordinated with the organizational processes. Such coordination

[10] Accessible at https://github.com/elianemaalouf/ethical_AI_evaluation.

is guaranteed by the presence of infrastructure experts, clinical implementors and quality & risks managers representatives in the project advisory committee.

Furthermore, given the increasing number of research projects involving AI in clinical settings, we recommend organizations to build capacities for ethics as a service [51] in order to: build expertise and processes for ethical AI assessment and reporting, optimize the use of resources involved in such assessments across projects, avoid ethics shopping/washing on projects [51], perform internal audits and guarantee compliance with existing/upcoming regulations. Some research projects require a direct involvement by the organization, for example projects implementing decentralized federated learning based on collaborations between multiple research organizations [59]. The organization could be made responsible for data, cyber and privacy infringement and its reputation would be impacted by scientific misconduct. We believe that enforcing a proactive attitude towards ethical health-AI will bring hosting organizations a better control over the outcomes and a pioneering/reference spot in this domain. Regulations (such the EU-act) and framework/guideline developers follow or recommend to follow a risk informed process [23] and setup a risk management process. We recommend that the internal oversight of the ethical-AI assessments and reporting actions be an extended responsibility of risk and quality management offices.

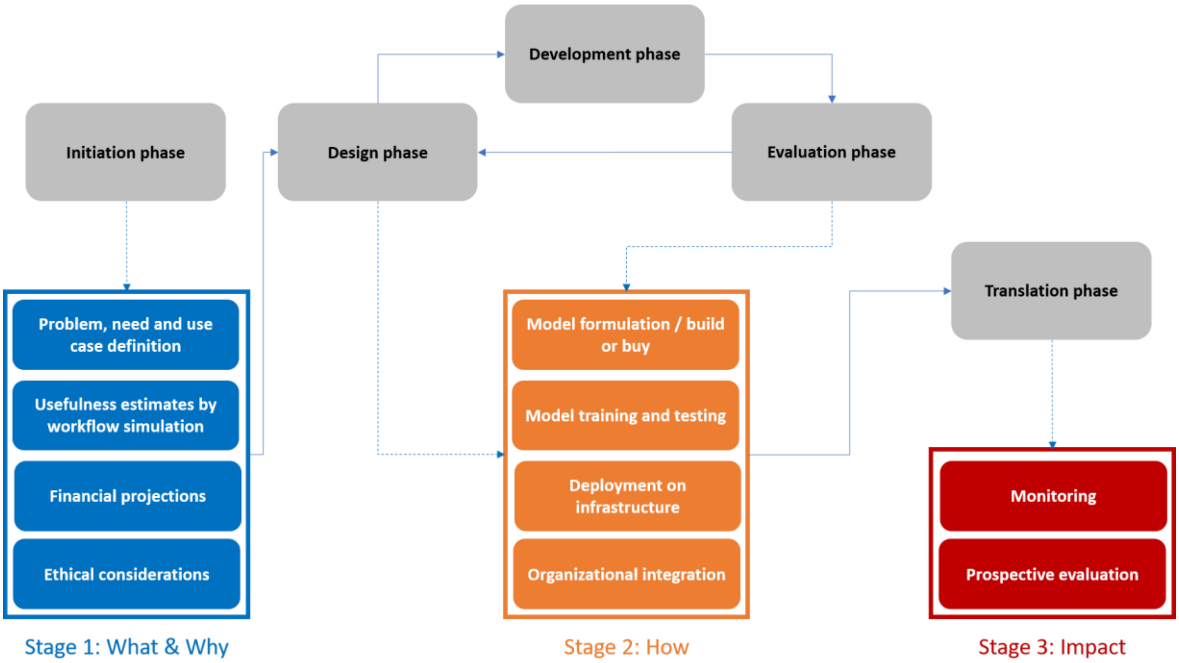

Figure 2 – Health-AI project life-cycle integrated in FURM stage-gate process [59]

R.3 Ethical health-AI motivated funding in research and private investments

Standardizing ethical AI assessment and reporting is an emerging field. We recommend scientific funding bodies to encourage programs that are focused on evaluation science and standards/guidelines appraisal, to encourage operationalization efforts (e.g., by incentivising sharing assessment and evaluation strategies and developed tools), the creation of reference catalogues of implementable tools and methods (e.g., catalogue built by Xia et al. [60], book and python codes by Japkowicz et al. [61]). It is also essential to not only develop and update guidelines but also to implement well-established datasets and code sharing concepts [20]. However, we do not advise to judge model performance only on "benchmark" contests because such performances do not necessarily translate to AI based science [19]. External validity (generalization) does not guarantee clinical usefulness (i.e., ability of a model to recommend decisions that enhance outcomes or reduce costs) [62].

Furthermore, research funders[11] and scientific publishers have a strong influence in enabling ethical health-AI and operationalizing it in research and beyond. They could exercise this influence by supporting and incentivizing projects that show rigorous and structured ethical AI assessments (from the project initiation phase/study protocol) and by enforcing transparency and reproducibility requirements and enabling them (e.g., provide guidance on recommended reporting frameworks, on data and code sharing tools, or by creating reproducibility challenges such as https://reproml.org/). Heil et al. [63] proposed a reproducibility standard with 3 degrees of "bronze," "silver," and "gold" with requirements to meet for each ("bronze" represents the bare minimum and "gold" represents full automation via containerized applications) [20][63]. Publishers could support their reviewers in assigning a reproducibility label for the reviewed studies and such label can be made clearly visible on the published article page. When creating and managing such labels it is important to keep in mind that it might be difficult for peer-reviewers to actually assess reproducibility (e.g., by rerunning the analysis themselves) [19] or there might be legitimate reasons for not sharing data and codes. Providing a forum to share reproducibility efforts performed by the community could be a step toward engaging the research community as a whole. When possible, researchers on a project could also recruit/invite a researcher/colleague external to the project and ask them to try to reproduce the work [19][24][31]. Reproducibility researchers could be rewarded as co-authors and/or recognized in the community by making their contributions more visible along traditional contributions. It is important for researchers to define reproducibility clearly and to establish standards that encompass not just computational reproducibility but also the accuracy of the findings [19]. Making health-AI sustainably ethical might require a cultural transformation in research, refocusing on transparency, quality and exhaustive documentation over the rush of publishing [20]. Addressing reproducibility failures pre-emptively can correct a lot of scientific research that would otherwise be flawed [19]. To control overoptimism with regards to AI capabilities, a research agenda to investigate efficacy of AI methods across scientific fields could help better support the debate on what can be realistically expected from the predictions of those models [19]. Encouragement to ethics-centred research reviews would also increase knowledge of the applicability and relevance of ethical dimensions assessments [39].

Innovation in health-AI is being largely developed by the private sector, big tech companies or startups. Those products could then be purchased and integrated by health organizations. During this product oriented development process very little incentives are present to consider ethical AI risks [27]. In this discussion, it is important to remind about the private funders responsibility and their power to incentivize attention to societal responsibilities [27] in the products they finance.

[11] The current disposition of SNSF regarding research projects integrating AI is summarized in the disposition "Researchers must assume responsibility". (https://www.snf.ch/en/LE2hc62fQoNDMoFb/topic/towards-a-responsible-use-of-ai-in-research)

R.4 Provide guidance on regulation applicability & agile regulatory processes

A lack of clear oversight and accountability mechanisms can lead to ethical risks. Establishing roles and responsibilities for monitoring AI performance and ensuring accountability is crucial to prevent and address any negative impacts on patients [22][29]. Ethical risks can also arise from the failure to provide necessary frameworks, usage agreements, and best practices for the implementation and evaluation of AI tools [32]. However, technical expertise is required to translate legal requirements into actionable measures [64]. Regulators and research ethics boards are ill-equipped to adequately evaluate AI-centered research studies [16]. The nature of AI-systems requires periodic updates to their weights and the software/libraries. This adaptive nature adds complexities into the regulation process, namely for previously authorized health-AI technology. Regulators would need to adapt to the possibility of updates without needing to request new certifications/authorizations [24].

We reiterate recommendations for increased agility in the regulatory process [65], enabled by a set of actions such as: reinforcing regulatory science [27] supported by knowledge and capacity building [65] and a nation-wide regulatory competence centre [65]; wider involvement of stakeholders in regulations co-creation [65]; encouraging "self-regulation" [27] by moving the focus from the innovation per se and putting more emphasis on innovators credentials and incentives to take responsibility; taking a proactive stance on innovation by maintaining continuous screening of new technologies [27].

Regulators and policy makers could initiate public-private partnerships to establish "AI assurance laboratories" [66] to perform independent audits for critical applications and health research project involving AI. They can also pull on those networks for technical support/consultation during regulation definition, to consolidate best practices and translating guidelines into practical implementable actions. In coordination with research organizations, those laboratories could be coordinating the "ethics as a service" expertise and pooling on necessary resources, financial and technical, to enable periodic assessments and reporting. Such coordination should also lead to the establishment of criticality classification for health-AI and audit frequency and modality (e.g., whether third-party or internal would be enough).

Implementation considerations

The Swiss regulatory context on AI is advancing. On one hand, this is encouraging because the laws are evolving towards more explicit requirements in regards to ethical AI use. On the other hand, however, this transitory phase could also be debilitating and confusing to navigate for researchers until clear standards and specific guidelines are consolidated and enforced. It could be tempting for researchers and research organizations to leave those ethical-AI considerations to a future time, and continue "business as usual" for the time being. We strongly advise against this approach and encourage a proactive attitude towards ethical questions and to use this transition period as a learning phase to acquire expertise in those areas. We also see this proactivity not only as the responsibility of the individual researchers on research projects, but a systematic involvement from all the health research stakeholders and a shared responsibility in guaranteeing scientific evidence quality and patients/health professionals' safety and benefits. In this section we discuss the main challenges that could face the implementation of our overarching recommendations and propose some initial initiatives to facilitate their implementation and maintenance. We summarize those in Table 3.

Table 3 - Main challenges and countering initiatives examples

Challenge	Facilitating initiatives
Inefficient communication/coordination between stakeholders' expectations and needs	In the short term, researchers could perform early consultations with the stakeholders at the design phase of the projects, especially for projects that could be expected to incur complex changes to clinical workflows and/or patients healthcare modalities. Health research organizations (supported by funders and ethical boards) could establish permanent forums to recruit stakeholders for these discussions/dialogues and open consultations on critical health initiatives to collect information about beliefs, perceptions, expectations and to refine understanding of actual needs of the end-users. In the long term, academics institutions should integrate education about AI in the clinicians and health researchers curriculums to increase understanding of the technology use, its limitations and its development cycles [67]. From the AI developers' side, working in healthcare and health research, training on health research and its ethical requirements could be made mandatory and certified. Ideally, new interdisciplinary master curriculums focusing on ethical health-AI combining technical and scientific conduct expertise would be a very useful capacity to have as research project managers. Interdisciplinary education would help establish a unified language and ease interactions.

Challenge	Facilitating initiatives
Regulatory transitory period limiting availability of clear assessment requirements	Ethics committees could propose a set of recommended documents to provide, along with the usual documentation required, when submitting a health-AI research protocol. This documentation could include the following: a clear rational for AI use; existing baselines (why they fall short and what are clinically relevant improvements on those baselines are sought after); initial assessment (and impact on existing workflows) of the clinical context and cost-benefit simulations; a proposal of ethical AI assessment plan including initial quality and risk assessments and mitigations (on a subset of priority dimensions, for example, bias sources, privacy risks, harms/errors sources, end-users involvement throughout the project). Health research organizations could provide support by operationalizing and internally standardizing this documentation by allocating (internally or externally contracted) data scientists, data engineers and quality analysts to support researchers on the projects in identifying and responding to those requirements efficiently and thoroughly. Such involvement from the organization, as early as the design phase, is necessary to inform and budget project and operationalization phase needs, while ensuring that impactful initiatives are adequately supported and lead. Researchers on the project themselves should remain critical of AI potential and maintain their ethical responsibility and scientific integrity as a guide. If they are not equipped with the necessary tools, they should not shy away from asking for support from their organization and request specific measures from the AI developers to support the ethical assessment they need to implement. We advise to look at the diverse set of possible assessments on p. 33, and references therein.
Lack of clarity on who is responsible for AI system validation and the modalities of this validation	If the system is being developed and integrated in the same organization, then during the development phase the responsibility is shared between the development team and the operational team. The development team naturally focuses on the performance and usability of the system while the operational team would be focused on integration questions and standardiz-

Challenge	Facilitating initiatives
	ing interfaces with other systems (e.g., data exchange standards implementation, cybersecurity risks assessments). After development, the organization integrating an AI system in their workflows is the main responsible for its maintenance, as is the case with other digital systems. AI systems require more periodic updates (both to the weights and the coding libraries), activities requiring expertise and knowledge of potentially incurring new risks (e.g., deteriorating performances, new security risks). A team of data scientists and data engineers should be available to perform those updates and the continuous assessment of their impact. Furthermore, given the criticality of the AI-system and its integration level, the organization should be responsible for performing at least internal audits (as preparation of potential requirements for external audits), the frequency of which would depend on the risk level.

Discussion

At the conclusion of this work, it is pertinent to explicit the inherent methodological limitations. First, we limited our search of guidelines and frameworks related to ethical AI to only two databases. Our search string used the terms "trustworthy" and "responsible" to describe ethical AI. Based on the initial screening, we identified that the term "ethical AI" serves our purposes better, but we could not repeat the search and the screening phases due to lack of resources. We acknowledge that we might have missed some relevant references. To compensate for this shortcoming, we strived to follow whenever possible a snowballing approach to retrieve further references that we deemed relevant. Second, the broader domain of digital health innovation encompasses AI, and there are possible relevant references there we missed by our chosen focus. However, whenever such useful references were identified, we tried to integrate them as well.

Despite those limitations, we believe this report brings a better insight to the complexity of ethical-AI assessments in health. In fact, our recommendations are not only relevant for health research; they can be extended further beyond, to healthcare in general and operationalizations after research. We mainly advocate for an attitude and cultural change of considering health-AI in research: it should be considered as relevant as it would be in clinical care, it should be treated with the same scrutiny, and it should require the same levels of proofs of efficacy and validity in practice. It is important along promising transformative potential to be able to assess sustainability of use and the expected needs in resources and adaptations (skills, infrastructures, etc.) in order to bring about this transformation. We believe that thinking about quality and risks assessment of health-AI should be done on the full path from research to implementation, be considered as a continuum process starting from the design phase, kept and updated all along, and passing it around by clear definition of responsibilities at the different phases. For an efficient and smooth process, a high maturity in AI-governance is required in research organizations hosting health-AI. An awareness of the importance of ethical AI principles for the healthcare domain, a knowledge about the complexities and interdependences, as well as knowledge and resources to put the necessary processes in place requires also agility, openness to change and to integrate new assessment tools, and evaluation and reporting techniques. In parallel, regulators policy makers and ethical committees are also expected to follow a more agile validation and assessment process, to acquire the needed skills or help building competence centres on which they can lean to make the necessary audits. Accountabilities should also involve regulators' responsibilities in allowing for context and regulatory changes that reinforce this testing and evaluation attitude, of importance of quality assessments and of transparency.

AI tools also have societal and environmental impacts that are important to take into account in a wider assessment of impact and ethical responsibility, dimensions that we did not cover in our review. Existing socio-economical factors might reinforce inequalities of access to the technology, at individuals' level (e.g., access to the internet/computer/smartphone) but also organizational level (e.g., some health organizations do not have the resources to acquire or develop such systems). Furthermore, when AI-based tools will prove to be efficient and clinically valid and are required to improve quality of care, resources limitations need to be overcome to avoid reinforcing those inequalities. It is currently hard to identify who is responsible to start this system reflection, projecting required changes, planning and imagining solutions for the upcoming challenges. As often, we can expect that a multi-stakeholder's approach would be the best, but someone should initiate the effort to engage the health-system stakeholders.

Nevertheless, we expect that developing ethical health-AI is expected to reduce costs on the long term by proving efficiency and usefulness, which helps inform investment decisions and avoid wasted costs due to corrective measures caused by inefficient systems. Questions of labour and change in health professions due to AI are also often brought up. In those systemic discussions, we need to think how healthcare professions might need to change, the skills that would be needed, how would responsibility/accountability concepts will change and what would patients expect from their relationship with a health professional in the not so far future.

References

[1] U. Nations, "Universal Declaration of Human Rights," United Nations. Accessed: Nov. 22, 2024. [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights

[2] "NeurIPS Statistics - Paper Copilot." Accessed: Sep. 13, 2024. [Online]. Available: https://papercopilot.com/statistics/neurips-statistics/

[3] T. Madiega and R. Ilnicki, "AI investment: EU and global indicators," www.europarl.europa.eu. Accessed: Sep. 13, 2024. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/760392/EPRS_ATA(2024)760392_EN.pdf

[4] C. Kozyrkov, "What's Different About Today's AI?," Medium. Accessed: Sep. 13, 2024. [Online]. Available: https://kozyrkov.medium.com/whats-different-about-today-s-ai-380569e3b0cd

[5] "The Top Artificial Intelligence Trends | IBM." Accessed: Sep. 13, 2024. [Online]. Available: https://www.ibm.com/think/insights/artificial-intelligence-trends

[6] B. Perrigo, "U.S. Must Act Quickly to Avoid Risks From AI, Report Says," TIME. Accessed: Nov. 19, 2024. [Online]. Available: https://time.com/6898967/ai-extinction-national-security-risks-report/

[7] "Tech leaders suffering from GenAI 'FOMO' with 75% believing they are behind competitors – Wavestone." Accessed: Sep. 13, 2024. [Online]. Available: https://www.wavestone.com/en/news/tech-leaders-suffering-from-genai-fomo-with-75-believing-they-are-behind-competitors/

[8] "CIOs eager to scale AI despite difficulty demonstrating ROI, survey finds," CIO. Accessed: Sep. 13, 2024. [Online]. Available: https://www.cio.com/article/2095301/cios-eager-to-scale-ai-despite-difficulty-demonstrating-roi-survey-finds.html

[9] Groupe de travail interdépartemental Intelligence artificielle and Secrétariat d'Etat à la formation, à la recherche et à l'innovation SEFRI, "Défis de l'intelligence artificielle", Accessed: Nov. 19, 2024. [Online]. Available: https://www.sbfi.admin.ch/dam/sbfi/fr/dokumente/2019/12/bericht_idag_ki.pdf.download.pdf/bericht_idag_ki_f.pdf

[10] B. Perrigo, "Employees at Top AI Labs Fear Safety Is an Afterthought," TIME. Accessed: Nov. 19, 2024. [Online]. Available: https://time.com/6898961/ai-labs-safety-concerns-report/

[11] "AI Act | Shaping Europe's digital future." Accessed: Oct. 05, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[12] "High-level summary of the AI Act | EU Artificial Intelligence Act." Accessed: Nov. 19, 2024. [Online]. Available: https://artificialintelligenceact.eu/high-level-summary/

[13] H. van Kolfschooten and J. van Oirschot, "The EU Artificial Intelligence Act (2024): Implications for healthcare," *Health Policy*, vol. 149, p. 105152, Nov. 2024, doi: 10.1016/j.healthpol.2024.105152.

[14] "Mission statement of swissethics." Accessed: Feb. 25, 2025. [Online]. Available: https://swissethics.ch/en

[15] swissethics and swissmedic, "Partial revision of the HRA and StRA ordinances." Accessed: Oct. 05, 2024. [Online]. Available: https://kofam.ch/upload/downloads/anschauungsmaterial/HFG-Verordnungsrevision_EN_klein.pdf

[16] S. Bouhouita-Guermech, P. Gogognon, and J.-C. Bélisle-Pipon, "Specific challenges posed by artificial intelligence in research ethics," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1149082.

[17] CER-VD, "Intelligence Artificielle (IA), algorithmes et machine learning; Pense-bête de la CER-VD (v1.0, 26.10.2023)." Accessed: Oct. 05, 2024. [Online]. Available: https://static1.squarespace.com/static/60b94bed393f8064950b2821/t/653a3ca300ef590e3933ae43/1698315427799/CER-VD_IA_pense-bete_231026.pdf

[18] J. C. C. Kwong *et al.*, "APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support," *JAMA Netw. Open*, vol. 6, no. 9, p. e2335377, Sep. 2023, doi: 10.1001/jamanetworkopen.2023.35377.

[19] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[20] K. B. Shiferaw, M. Roloff, I. Balaur, D. Welter, D. Waltemath, and A. A. Zeleke, "Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review," *JAMIA Open*, vol. 8, no. 1, p. ooae155, Feb. 2025, doi: 10.1093/jamiaopen/ooae155.

[21] D. Schwabe, K. Becker, M. Seyferth, A. Klass, and T. Schaeffter, "The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review.," *NPJ Digit. Med.*, vol. 7, no. 1, p. 203, Aug. 2024, doi: 10.1038/s41746-024-01196-4.

[22] G. E. Cacciamani *et al.*, "PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare," *Nat. Med.*, vol. 29, no. 1, pp. 14–15, Jan. 2023, doi: 10.1038/s41591-022-02139-w.

[23] K. Lekadir *et al.*, "FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, p. e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[24] de H. AAH *et al.*, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review.," *NPJ Digit. Med.*, vol. 5, no. 1, p. 2, Jan. 2022, doi: 10.1038/s41746-021-00549-7.

[25] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[26] Coalition for health AI, "Assurance Standards Guide Coalition for Health AI (CHAI)," CHAI - Coalition for Health AI. Accessed: Nov. 22, 2024. [Online]. Available: https://chai.org/assurance-standards-guide/

[27] C. Landers, E. Vayena, J. Amann, and A. Blasimme, "Stuck in translation: Stakeholder perspectives on impediments to responsible digital health," *Front. Digit. Health*, vol. 5, 2023, doi: 10.3389/fdgth.2023.1069410.

[28] H. Ibrahim, X. Liu, and A. K. Denniston, "Reporting guidelines for artificial intelligence in healthcare research," *Clin. Experiment. Ophthalmol.*, vol. 49, no. 5, pp. 470–476, 2021, doi: 10.1111/ceo.13943.

[29] M. Mccradden *et al.*, "What's fair is ⋯ fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning," *ACM Int. Conf. Proceeding Ser.*, pp. 1505–1519, 2023, doi: 10.1145/3593013.3594096.

[30] L. Szabo *et al.*, "Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging," *Front. Cardiovasc. Med.*, vol. 9, 2022, doi: 10.3389/fcvm.2022.1016032.

[31] N. L. Crossnohere, M. Elsaid, J. Paskett, S. Bose-Brill, and J. F. P. Bridges, "Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks," *J. Med. Internet Res.*, vol. 24, no. 8, 2022, doi: 10.2196/36823.

[32] J.-C. Bélisle-Pipon and G. Victor, "Ethics dumping in artificial intelligence," *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1426761.

[33] T. Birkstedt, M. Minkkinen, A. Tandon, and M. Mäntymäki, "AI governance: themes, knowledge gaps and future agendas," *Internet Res.*, vol. 33, no. 7, pp. 133–167, 2023, doi: 10.1108/INTR-01-2022-0042.

[34] J.-C. Bélisle-Pipon, V. Couture, M.-C. Roy, I. Ganache, M. Goetghebeur, and I. G. Cohen, "What Makes Artificial Intelligence Exceptional in Health Technology Assessment?," *Front. Artif. Intell.*, vol. 4, 2021, doi: 10.3389/frai.2021.736697.

[35] J. Gallifant *et al.*, "The TRIPOD-LLM reporting guideline for studies using large language models," *Nat. Med.*, vol. 31, no. 1, pp. 60–69, Jan. 2025, doi: 10.1038/s41591-024-03425-5.

[36] M. B. A. McDermott, B. Nestor, and P. Szolovits, "Clinical Artificial Intelligence: Design Principles and Fallacies," *Clin. Lab. Med.*, vol. 43, no. 1, pp. 29–46, 2023, doi: 10.1016/j.cll.2022.09.004.

[37] J. Hassan, S. M. Saeed, L. Deka, M. J. Uddin, and D. B. Das, "Applications of Machine Learning (ML) and Mathematical Modeling (MM) in Healthcare with Special Focus on Cancer Prognosis and Anticancer Therapy: Current Status and Challenges," *Pharmaceutics*, vol. 16, no. 2, 2024, doi: 10.3390/pharmaceutics16020260.

[38] M. Kritharidou *et al.*, "Ethicara for Responsible AI in Healthcare: A System for Bias Detection and AI Risk Management.," *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2023, pp. 2023–2032, 2023.

[39] M. Drira, S. Ben Hassine, M. Zhang, and S. Smith, "Machine Learning Methods in Student Mental Health Research: An Ethics-Centered Systematic Literature Review," *Appl. Sci. Switz.*, vol. 14, no. 24, 2024, doi: 10.3390/app142411738.

[40] E. Nasarian, R. Alizadehsani, U. R. Acharya, and K.-L. Tsui, "Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework," *Inf. Fusion*, vol. 108, 2024, doi: 10.1016/j.inffus.2024.102412.

[41] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration.," *Healthc. Basel Switz.*, vol. 11, no. 20, Oct. 2023, doi: 10.3390/healthcare11202776.

[42] Z. A. Nazi and W. Peng, "Large Language Models in Healthcare and Medical Domain: A Review," *Informatics*, vol. 11, no. 3, Art. no. 3, Sep. 2024, doi: 10.3390/informatics11030057.

[43] The Royal Society, *Science in the age of AI*. 2024. Accessed: Mar. 08, 2025. [Online]. Available: https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

[44] A. Jain, M. Salas, O. Aimer, and Z. Adenwala, "Safeguarding Patients in the AI Era: Ethics at the Forefront of Pharmacovigilance.," *Drug Saf.*, Sep. 2024, doi: 10.1007/s40264-024-01483-9.

[45] P. Theriault-Lauzier *et al.*, "A Responsible Framework for Applying Artificial Intelligence on Medical Images and Signals at the Point of Care: The PACS-AI Platform.," *Can. J. Cardiol.*, vol. 40, no. 10, pp. 1828–1840, Oct. 2024, doi: 10.1016/j.cjca.2024.05.025.

[46] A. Felländer, J. Rebane, S. Larsson, M. Wiggberg, and F. Heintz, "Achieving a Data-Driven Risk Assessment Methodology for Ethical AI," *Digit. Soc.*, vol. 1, no. 2, p. 13, Aug. 2022, doi: 10.1007/s44206-022-00016-0.

[47] S. A. Hassan, A. I. Omar, and N. R. Ahmed, “Exploring the ethical implications of ai in public health research: A comprehensive analysis,” *South East. Eur. J. Public Health*, vol. 25, pp. 108–115, 2024, doi: 10.70135/seejph.vi.1309.

[48] B. Murdoch, “Privacy and artificial intelligence: challenges for protecting health information in a new era,” *BMC Med. Ethics*, vol. 22, no. 1, 2021, doi: 10.1186/s12910-021-00687-3.

[49] J. Fehr, B. Citro, R. Malpani, C. Lippert, and VI. Madai, “A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare.,” *Front. Digit. Health*, vol. 6, p. 1267290, 2024, doi: 10.3389/fdgth.2024.1267290.

[50] N. Aguilar, A. Y. Landau, S. Mathiyazhagan, A. Auyeung, S. Dillard, and D. U. Patton, “Applying Reflexivity to Artificial Intelligence for Researching Marginalized Communities and Real-World Problems,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2023, pp. 712–721, 2023.

[51] J. Morley, A. Elhalal, F. Garcia, L. Kinsey, J. Mökander, and L. Floridi, “Ethics as a Service: A Pragmatic Operationalisation of AI Ethics,” *Minds Mach.*, vol. 31, no. 2, pp. 239–256, 2021, doi: 10.1007/s11023-021-09563-w.

[52] F. McKay, B. J. Williams, G. Prestwich, D. Bansal, D. Treanor, and N. Hallowell, “Artificial intelligence and medical research databases: ethical review by data access committees,” *BMC Med. Ethics*, vol. 24, no. 1, 2023, doi: 10.1186/s12910-023-00927-8.

[53] E. Wellenhofer, “Real-World and Regulatory Perspectives of Artificial Intelligence in Cardiovascular Imaging.,” *Front. Cardiovasc. Med.*, vol. 9, p. 890809, 2022, doi: 10.3389/fcvm.2022.890809.

[54] VB. Vallevik *et al.*, “Can I trust my fake data - A comprehensive quality assessment framework for synthetic tabular data in healthcare.,” *Int. J. Med. Inf.*, vol. 185, p. 105413, May 2024, doi: 10.1016/j.ijmedinf.2024.105413.

[55] FR. Kolbinger, GP. Veldhuizen, J. Zhu, D. Truhn, and JN. Kather, “Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis.,” *Commun. Med.*, vol. 4, no. 1, p. 71, Apr. 2024, doi: 10.1038/s43856-024-00492-0.

[56] L. M. Monday, “Define, Measure, Analyze, Improve, Control (DMAIC) Methodology as a Roadmap in Quality Improvement,” *Glob. J. Qual. Saf. Healthc.*, vol. 5, no. 2, pp. 44–46, Jun. 2022, doi: 10.36401/JQSH-22-X2.

[57] A. Callahan *et al.*, “Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems,” *NEJM Catal.*, vol. 5, no. 10, p. CAT.24.0131, Sep. 2024, doi: 10.1056/CAT.24.0131.

[58] J. Y. Kim *et al.*, “Organizational Governance of Emerging Technologies: AI Adoption in Healthcare,” in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT ’23. New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 1396–1417. doi: 10.1145/3593013.3594089.

[59] D. Upreti, E. Yang, H. Kim, and C. Seo, “A Comprehensive Survey on Federated Learning in the Healthcare Area: Concept and Applications,” *CMES - Comput. Model. Eng. Sci.*, vol. 140, no. 3, pp. 2239–2274, 2024, doi: 10.32604/cmes.2024.048932.

[60] B. Xia, Q. Lu, L. Zhu, S. U. Lee, Y. Liu, and Z. Xing, “Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability,” in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, in CAIN ’24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 100–111. doi: 10.1145/3644815.3644959.

[61] N. Japkowicz and Z. Boukouvalas, *Machine Learning Evaluation: Towards Reliable and Responsible AI*. Cambridge: Cambridge University Press, 2024. doi: 10.1017/9781009003872.

[62] A. Youssef, M. Pencina, A. Thakur, T. Zhu, D. Clifton, and N. H. Shah, “External validation of AI models in health should be replaced with recurring local validation,” *Nat. Med.*, vol. 29, no. 11, pp. 2686–2687, Nov. 2023, doi: 10.1038/s41591-023-02540-z.

[63] B. J. Heil, M. M. Hoffman, F. Markowetz, S.-I. Lee, C. S. Greene, and S. C. Hicks, “Reproducibility standards for machine learning in the life sciences,” *Nat. Methods*, vol. 18, no. 10, pp. 1132–1135, Oct. 2021, doi: 10.1038/s41592-021-01256-7.

[64] L. Lucaj, P. Van Der Smagt, and D. Benbouzid, “AI Regulation Is (not) All You Need,” *ACM Int. Conf. Proceeding Ser.*, pp. 1267–1279, 2023, doi: 10.1145/3593013.3594079.

[65] “Responsible digital health innovation roadmap,” DigitalHealthRoadmap. Accessed: Feb. 04, 2025. [Online]. Available: https://digitalhealthroadmap.ethz.ch

[66] N. H. Shah *et al.*, “A Nationwide Network of Health AI Assurance Laboratories,” *JAMA*, vol. 331, no. 3, pp. 245–249, Jan. 2024, doi: 10.1001/jama.2023.26930.

[67] T. Schubert, T. Oosterlinck, R. D. Stevens, P. H. Maxwell, and M. van der Schaar, “AI education for clinicians,” *eClinicalMedicine*, vol. 79, Jan. 2025, doi: 10.1016/j.eclinm.2024.102968.

[68] A. J. Vickers and E. B. Elkin, “Decision curve analysis: a novel method for evaluating prediction models,” *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.

[69] M. D. McCradden *et al.*, “A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning,” *Am. J. Bioeth.*, vol. 22, no. 5, pp. 8–22, May 2022, doi: 10.1080/15265161.2021.2013977.

[70] “Model Cards for Model Reporting | Proceedings of the Conference on Fairness, Accountability, and Transparency.” Accessed: Feb. 23, 2025. [Online]. Available: https://dl.acm.org/doi/10.1145/3287560.3287596

[71] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Lariviere, and A. Beygelzimer, “Improving Reproducibility in Machine Learning Research”.

Project advisory committee

The committee could include the following members, along examples of there potential involvement:

- **Scientific principal investigator** (typically a health researcher): effectively the project manager and the responsible for all aspects of scientific and methodological conduct related to the project.
- **AI developers/integrators/experts lead**: depending whether the AI-system is being developed from scratch for the project or an existing tool is being assessed, these representatives would be responsible for the technical aspects of the AI-system implementation, the automation of tasks (e.g. model training/updating) and data assessments. They also support in selecting and implementing the correct ethical-AI assessments metrics for the project.
- **Legal advisor and ethicist**: to validate that the legal requirements and legal risks are being tracked and managed.
- **Representative of AI system end users**: to challenge the project conduct from the point of view of the final users, to confirm that end-users' perspectives are being heard and their needs and expectations are considered in a timely manner throughout the project. They participate into the requirements definition/consultations as well as the interfaces designs and validations.
- **IT/information system/infrastructure experts**: to help in assessing infrastructure requirements for implementation/integration phases (even if those phases are not central to the research project itself, these members can help in early cost-benefit assessments). They would also help in the integration aspects by supporting with data engineering efforts, data structuring/management and general management and provision of computational resources.
- **Clinical implementors representative**: they also need to be involved early in the design of the project in order to assess clinical assessment and integration requirements from the clinical workflows' points of view, required change management and the practical limitations the project will be faced when routinely collecting and updating the model. Their input informs the scope of the project, design choices, costs estimation, etc.
- **Quality management representative**: to inform and guide on the overall risk management and correct reporting practices. These are the representatives of the research organization hosting the research project, which support in standardizing assessment strategies, reporting on them internally and externally. They also support in sustaining those assessments beyond the lifespan of the research project and into the operationalization of the AI-system when needed.

Ethical AI assessments

Ethical AI principle	Examples of assessment and reporting topics
Usefulness, Usability & Efficacy	Clear definition of the health/clinical problem and its context [24]; involvement of end-users and relevant stakeholders to assess needs and why current approaches are not enough [24], definition of the AI-system goal and clinical success criteria [24]. Cost-benefit feasibility check accounting for costs of developing, maintaining and managing the risks and changes induced by the AI system integration [24]. Perform live clinical test to ensure safe and effective use focusing on understanding the AI-system functionality and workflow impact [24]. Sample size computation when available for the planned AI model, or evaluate the learning curve of the planned AI model on a similar publicly available dataset to make the sample size estimation [24]. It is important to survey existing literature on the topic and back claims on sample size selection from existing literature and specific simulation studies. Comparison with simpler models might also be helpful here. Data preprocessing steps documentation for data splits (e.g., into training, tuning and test), data augmentation, outliers management, variables/features transformations, standardization, missing data mechanisms and handling [24]; document software/libraries used. Take particular measure to avoid "data leakage" [19] which leads to overoptimistic model performance. Collaboration with experts is advised [24]. Internal validation to assess how the AI-system performs on new data from the same distribution as the training data [24]. Assessments of discrimination (e.g., with area under the curve), calibration (e.g., calibration plots). Rational for metrics choices and their significance thresholds [24]. It is recommended to select those thresholds with clinical experts. Document variability and uncertainty quantification (e.g., confidence intervals). Decision curve analysis to assess clinical utility and expected improvements to patient care [18][24][70]. Site-specific validation before deployment and recurring local validation after deployment (MLOps) [62]. Assessments of risk of overfitting on training data and document implemented measures to avoid this risk, knowing that the dataset might be fixed in size (e.g., dimension reduction, regularization, feature selection) [24]. It is recommended that feature selection be guided by domain expertise [19]. External performance evaluation and generalizability assessment to check how well the AI-system works in settings different from where it was created [24]. Compare the new model/system performance against reference/simpler/classical/explainable models' performances or common practices [18].

Ethical AI principle	Examples of assessment and reporting topics
	Document the AI-system interface design ensuring end-users understand the AI purpose, its outcomes and are provided with facts labels with technical details and limitations explanations, could send feedback and request reviews [24]. Perform and document usability testing [24]. Impact studies to compare outcomes between groups exposed to the AI-system and those following the current standard/tool [24]. The outcomes of these studies are diverse: clinical and patient-reported outcomes, cost-effectiveness, decision-making changes, and patient experience [24]. Decision Analytical Modeling can provide early estimates of clinical utility before a full impact study. Multiple reader multiple case study designs can measure AI's effect on decision-making which will be necessary knowledge for any future clinical implementation [24]. Communicate the results of impact studies to the wider stakeholders' community (i.e., healthcare professionals, administrators, policymakers). Silent evaluation [29][71]; prospective clinical evaluation [29][71] in the form of randomized controlled trials, stepped-wedge, before-after, and observational studies [24]. Human judgement is often used as baseline for comparison. Yet, human decision making could be flawed. Noise audits help assess noise in human judgment [21]. Preview necessary education for end-users to use the AI-system correctly. It also concerns regular trainings to all stakeholders to cover general AI knowledge, assumptions, limitations, legal aspects, risks, benefits, decisions understanding and assessment, automation bias, security breaches, etc. [24]
Fairness, Equity & Bias Management	Data representativeness of the target population and healthcare setting [24]; detailed information on collection time, location, population traits and inclusion/exclusion criteria used and document any differences from the target population [24]. Assessment of data quality [21] via checks for missing data, measurement errors and their causes; definition of measurement tools; reports on data quality risks and their impact on predictions and validations [24]; details about the labeling process (if any) and assessment of labels quality and variability/reproducibility between labelers [24]. Document choice of fairness definition given the AI-system purpose and the corresponding metrics for this definition [24]. Provide a "Fairness Statement" to project stakeholders for review and acceptance [24]. Document bias assessments and mitigation strategies [24]. Diverse stakeholders involvement in the design and regular assessment of fairness throughout the project [24].

Ethical AI principle	Examples of assessment and reporting topics
Safety & Reliability	Regular data quality checks and error correction processes [24]. Make sure unit tests are implemented and repeated regularly to ensure software reliability [24]. Use of data coding standards (e.g., SNOMED CT, ICD-10) and data exchange protocols [24]. Ensure interoperability of the new AI-system with existing digital infrastructure [24]. It is recommended to use open-source libraries in the development of the AI-system [24]. Plan software updates and clear documentation and notification about changes [24]. Ensure logging and traceability of changes and ways of rolling/reverting back [24]. Plan automatic deployment, shadow deployment and rollback processes for continuous monitoring and streamlined updating [24]. From the outset of the research project and after deployment, plan processes and systems to continuously monitor and document on model performance, data quality, error types, user feedback, (clinical) outcomes, fairness, dataset shifts. These areas influence the AI system accuracy over time and the frequency of monitoring should match the risk level of the system [24]. Develop a risk management plan to identify and manage risks, extreme situations, adverse events and system failures [24]. It includes setting safety levels, quality checks, and reporting errors or near misses, with a plan detailing role, risk assessment, reporting, monitoring, and addressing issues [24].
Transparency, Intelligibility & Accountability	Rational for modeling technique accounting for prediction accuracy, ease of understanding, end-user familiarity, computational needs, costs, maintenance, privacy, bias assessment possibilities, data size, and data structure, etc. [24] Assessment of the impact of each feature/group of features on predictions, identify errors, biases, and potential vulnerabilities in the models [24]. AI-system interface should help end-users understand how inputs lead to model outputs [24]. Assessment for requirements of explainability and understandability given the integration level of the AI-system and decision automation level; documentation and rational for explainability methods used [24]. Report details of training and hyperparameters tuning process, including final values, number of models trained, performance evaluation, final model structure, model inputs, model outputs [24]. Provide code and data for reproducibility [24]. Model cards [72], model info sheets [19], reproducibility checklists [73] [74]. Ensure standard reporting frameworks and guidelines are used whenever available for the application/study type or the specific assessment being performed. Maintain up to date documentation and auditing framework [24] on all the aspects of the development of the AI system, data, evaluation, monitoring, updating, risks, failures, integration with other systems in preparation for external audits (i.e., conformity assessments, internal audits). Plans should be in place to handle incidents.

Ethical AI principle	Examples of assessment and reporting topics
	This includes reporting failures, discussing them, and possibly changing the model's design or usage to prevent future issues [24]. Developers, implementers and research organizations are encouraged to disclose their innovation pathways and commercialization routes. It is important to consider the risks, investments, roles, and responsibilities of the different parties involved in the development of an AI system. This can help in the allocation of benefits and the economic impact analysis [24].
Security & Privacy	Ensure compliance with relevant privacy legislation and their requirements (e.g., GDPR[12], nFADP[13]) such as the principle of "privacy by design" [24]; ensure informed consent for newly collected data or for secondary uses of historical ones [24]; fulfill any legal requirement to appoint a data protection officer for data protection oversight and to assess how the right to withdraw consent (right to forget, right to object) would influence design and maintenance of the AI-system [24]. Ensure AI-system interface preserves data privacy when providing outcomes to end users [24]. Perform risk assessments for data vulnerabilities and adversarial attacks [24]. Establish an incidents response plan before deployment detailing how to handle security breaches and who is responsible [24]. Communicate about the timeframes for security updates and ensure that any new software vulnerabilities are addressed promptly and thoroughly tested before implementation [24]. Perform necessary load and penetration testing [24]. New vulnerabilities and changes to the AI system should be documented and reported [24].

[12] https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[13] https://www.kmu.admin.ch/kmu/en/home/facts-and-trends/digitization/data-protection/new-federal-act-on-data-protection-nfadp.html

Authors contributions, literature search methodology and AI use disclaimer

1- **Contributions:** EM contributed to topic definition, methodology, databases search, screening, writing and reviewing.
M-A LP contributed to topic definition, methodology and reviewing.
PC contributed to topic definition and reviewing.

2- **Methodology**: The following databases were used to search for literature regarding frameworks and guidelines for evaluating and reporting on ethical AI use in health research:

- In PubMed :

  Search string : ("artificial intelligence"[MeSH] OR "Machine learning" [tiab] OR "AI" [tiab] OR "artificial intelligence" [tiab]) AND ("risk"[Mesh] OR "risk"[tiab] OR "quality"[tiab] OR "Efficacy" [tiab] OR "Effectiveness" [tiab]) AND ("Trustworthy" [tiab] OR "responsible" [tiab]) AND ("control" [tiab] OR "assess*" [tiab] OR "report*" [tiab] OR "evaluat*" [tiab]) AND ("framework"[tiab] OR "governance" [tiab] OR "guidelines" [tiab])

  Date: 30.01.2025

  Results: retrieved 138, of which 38 were retained after title and abstract screening

- In Scopus:

  Search string: (ALL ("artificial intelligence" OR "Machine learning" OR ai) AND ALL (risk OR quality OR efficacy OR effectiveness) AND ALL (trustworthy OR responsible) AND ALL (control OR assess* OR report* OR evaluat*) AND ALL (framework OR governance OR guidelines) AND ALL ({health research})) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2022) OR LIMIT-TO (PUBYEAR , 2023) OR LIMIT-TO (PUBYEAR , 2024) OR LIMIT-TO (PUBYEAR , 2025))

  Date: 04.01.2025

  Results: retrieved 2168, of which 87 were retained after title and abstract screening

The main inclusion criteria were: (primary) elaborating about frameworks/guidelines; (secondary) applications/works focusing on aspects of governance, policy making, interpretability/explainability, fairness, privacy, etc. and their evaluation criteria and/or methods/processes; (tertiary) general applications and recommendations for those applications in relation to responsible/ethical/trustworthy AI use.

We excluded references that are not related to health or human health (e.g. veterinary), not involving AI or an AI application (e.g., general clinical research) and not related to ethical/responsible/ trustworthy AI or perceptions of those (e.g., general application of DL or ML).

3- **AI use:** During the writing of this document, we used the following tools to support in the screening, summarization and data extraction from the retained papers:

- Rayyan.ai: used for screening retrieved papers based on title and abstracts.
- Scispace (typeset.io): used for summarizing the content of specific articles and for data extraction.
- NotebookLM (notebooklm.google.com): used to write the “key Messages” section based on the body of the document.