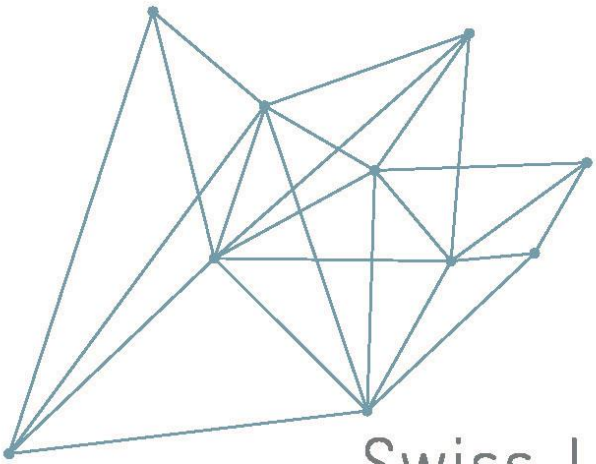

Swiss Learning Health System

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

Written by: Eliane Maalouf; Marie-Annick Le Pogam, MD, PhD; Paul Cotofrei, PhD

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

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

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 progress, 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, such as the EU, are responding with frameworks like the "AI Act," which establishes different risk levels for AI systems. Specifically, the AI Act categorizes medical devices using AI as high-risk but exempts 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 the ethical use of AI 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, ethics 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 variety 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. Challenges in applying and updating guidelines, along with cultural differences between health research and digital innovation, hinder structured quality assessment and delay the implementation of AI.

Options 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 approach. Key options for action include:

- **Embed ethical AI considerations in the project life cycle**. Integrate ethical assessments throughout the entire process, from initial idea generation to design, development, evaluation, and clinical translation. This involves forming a **project advisory committee** including representatives of clinicians, AI developers, patients, ethicists and other 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 management in hosting organizations**. Organizations involved in health-AI development should maintain a portfolio of their projects. This includes using stage-based evaluations aligned with each project's lifecycle. Building "**ethics as a service**" capacities within organizations can provide expertise, optimize resource use, and ensure compliance. Risk and quality management offices should expand their oversight to include ethical AI assessments and reporting.
- **Funding for responsible and transparent health-AI innovation**. Funding bodies and scientific publishers should encourage and incentivize projects that demonstrate rigorous ethical assessments of AI 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 and 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 simultaneously enhance health on a large scale while also posing risks of undermining this fundamental right and infringing on others.

The development of Artificial Intelligence/Machine Learning (AI/ML or simply AI) accelerated considerably in the last decades. 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 increasingly prevalent (see Box 1).

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, better control over the health system's costs, and other benefits.

Box 1 Submitted papers to the main conferences on the topic have shown a year-on-year increase anywhere between 1000 and 3000 papers [2]. Since the advent of ChatGPT in November 2022, public attention oriented towards AI grew immensely (see a comparison of Google search terms 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 seem to be also fueling a "fear of missing out" anxiety among technology leaders [7] and an eagerness to implement AI solutions despite a lack of return on investments guarantees [8]. Specifically in health research in Switzerland, the Swiss Association of Research Ethics Committees (swissethics) approvals of AI-related projects increased (from 0.065% of all approvals in 2016 to 1.463% in 2023; as of September 2024, this proportion was already at 2.163% for 2024).

In the public and scientific debate, an important focus is put on AI trustworthiness and responsible development and deployment. It is now widely acknowledged that risks to privacy, safety, security, autonomy, equality, and other concerns accompany the positive promises of AI. On the OECD.AI policy observatory's "AI Incidents Monitor"[1] we see a steady increase of reported incidents in the category "Healthcare, drugs and biotechnology". The economic incentives and the competition structure in the market do not favor setting AI safety as a priority for industrial AI leaders [10]. Regulatory bodies are counterbalancing this "neglect" by putting AI risks and their mitigation as political priorities. The EU "AI Act" [11], which came into force on 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 risk assessment and mitigation measures, logging of activities for traceability, human oversight and monitoring. High-risk systems are required to undergo a conformity assessment, self-led or via a third-party [12]. The AI Act is a cross-domain initiative that does not adopt specific dispositions for the health sector or health research, but rather explicitly categorizes medical devices using AI as high-risk. The act fully

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

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]. 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. In this context, ethical obligations should be considered in a 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[2] and does not provide more guidance in this regard. In 2020, the Swiss confederation adopted guidelines[3] that will orient its AI regulations efforts, including requirements for "Transparency, traceability and explainability" which 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 into Swiss law was announced[4]. Specifically in health research involving human subjects, the currently revised ordinances[5] 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]. In April 2025, swissethics adopted a guiding document for research projects involving the development or the use of an AI algorithm. The document is a list of questions to answer in the protocols on the themes of justification for AI use, justification for the AI model used, data confidentiality and security, and responsibility [17]. The document is not concerned with processes for quality control, transparency, security and other dimensions that require reporting.

Multiple health-AI research findings suffer from quality issues (see Box 2), increasing the risks of undermining trust in the scientific endeavor in general and requiring a renewed attention

Box 2

Quality issues frequently found in health-AI research: low computational reproducibility [18][19][20]; 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].

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

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

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

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

to methodological rigor. 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]. Currently, there are no broad, established and generally accepted standards governing the trustworthy lifecycle of health-AI, including its design, development, evaluation, and deployment regarding its robustness, safety, ethics, and legality [23].

Researcher stakeholders have a responsibility to address these challenges proactively by establishing quality and applicability assessments for AI systems [24]. However, their task is complex for four main reasons, which we develop further in the next section:

(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), based on a narrative review of scientific publications and relevant grey literature, we aim to provide initial options for action to address the following questions:

- **How can health researchers continuously monitor the efficacy and the risks of AI systems used in research projects?**
- **What reporting is necessary to prepare for and adhere to evaluation requirements on AI systems to ethics 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 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[6] is a set of principles that defines how AI systems align with moral principles and societal values. An extensive review of AI ethics guidelines [25] shows an emerging set of eleven principles, such as transparency, justice and fairness, non-maleficence, responsibility, or beneficence, for example. However, those principles do not share common definitions between the guidelines. For the purpose of the following discussion, we focus on an aggregate subset of core principles defined by the Coalition for Health AI (CHAI)[7] 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, such as 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 outcomes.
Safety & Reliability: Thorough testing and risk management are conducted before implementation to prevent 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 promptly address incidents and effectively safeguard data.

Another classification around six principles of "**Fairness, Universality, Traceability, Usability, Robustness, Explainability**" can be found in [23]. However, both classifications [26][23] are

[6] The terms "responsible AI" and "trustworthy AI" are also often used as stand-ins 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.

[7] https://chai.org/

very recent and returns on their usefulness are not yet present; both are also consensus-based, involving large groups of stakeholders, which might be favorable for their adoption in practice.

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 (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 organizations	Structure hosting health AI research, typically where the health researchers are based. These can 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 hospitals, pharmaceutical companies, university departments
AI developers/providers	Main experts on AI systems. They come 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 AI tools for whom the system is designed to generate a benefit (e.g., improved diagnostics, optimized clinical workflow, etc.).	Patients, clinical practitioners, healthcare professionals
Data contributors	Can 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
Ethics committees	Reviewers and authorizers of projects' protocols and their follow up.	Cantonal ethics committees

Stakeholder	Description	Examples
Regulators and policy-makers	Authorizers of healthcare solutions, creators of laws that provide 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 project protocols and providers of resources for project development.	SNSF, foundations, swissuniversities
Scientific publishers	Reviewers and curators of project 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 concerned, venture capitalists are funding stakeholders' to be included.

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

AI system characteristics vs health research questions vs data

Health research involving an AI system is often criticized for methodological and reporting issues. Depending on the study type (e.g., clinical trial, diagnostic accuracy study, etc.), it is recommended to follow guidelines for protocol writing and reporting to ensure the quality of the study design, its execution, and transparency. Study design and transparency remain central in guaranteeing the quality of such studies and efficient evaluation of their outcomes. Researchers are advised to assess whether specific guidelines are commonly used for their study type and if 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, and usability assessments. There might be a need to complement those guidelines with other measures on non-covered dimensions. It is also 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.

The research question will define the purpose of the AI-system. This purpose orients multiple technical choices, such as the AI model family, the required data 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
AI system maturity 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]; mis-use 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 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

Dimension	Example risks to ethical AI
	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].
Data manipulation	**Data preparations:** "data leakage [19] by a wrong separation of training and test datasets; sampling biases. **Features selection:** the use of features that are proxies of the outcome; interdependence between test and training observations.

Data Alert

Researchers should be particularly aware of the risks associated with data collection, sampling, or pre-processing strategies and their impact on generalizability, as well as the sensitivity of AI model performance to slight changes in data manipulation. Transparency in reporting on assessments of "data leakage", mitigation actions, and data manipulations is crucial for the reproducibility/replicability of the results. Furthermore, the use of synthetic data is becoming increasingly common as a means to preserve privacy. Synthetic data requires its own assessments [54].
Finally, the issue of consent for collecting, storing, using, and re-using or re-purposing patients' or health professionals' data presents new legal and technical challenges, especially when large datasets are used for training or fine-tuning and when there are requirements to implement the right to withdraw consent. [24][47][49].

Diversity of frameworks and guidelines

Many ethical AI guidelines already exist. Jobin et al. identified 84 in total [25], and we found 30 specifically related to health research (methodology p. 44). Another recent review listed 26 reporting guidelines for medical AI [55], yet their practical use remains limited. Despite the availability of guidelines, researchers often struggle to apply them due to mismatches with study designs (e.g., fewer tools for observational studies), lack of awareness, and weak enforcement, quality evaluation and usefulness [20].

Health research routinely compares and combines study results (e.g., systematic reviews, meta-analyses). 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 lack of structured quality assessment during AI development, evaluation and implementation may be slowing adoption in clinical 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]. These cultural differences can affect stakeholder expectations and communication.

It is not believed that another framework is needed. Instead, emphasis should be placed on processes and practices to support the practical application of ethical AI practices throughout research projects and beyond. An interdisciplinary and flexible approach should be established and maintained to guide the assessment and reporting of ethical AI use. Such a process would support adaptation to the evolving nature of health-AI research, ensure stakeholder involvement, and aid in meeting regulatory and ethical requirements.

Options for action

The diversity of challenges and their continuous evolution hinder the development of a "gold standard" and a one-size-fits-all framework that encompasses all aspects of developing and reporting on health-AI [20][23].
It is recommended to focus on operationalizing the relevant frameworks for the project. This assumes quality management processes at project and organizational levels to select appropriate frameworks, define implementation strategies, and allocate resources for implementation and maintenance. Even though each organization follows its own project and quality management philosophy, general process steps often include **Define, Measure, Analyze, Improve** and **Control** [56]. Ethical design, development, evaluation and implementation of an AI-system should follow clinical norms grounded in biomedical ethics and research ethics [29]. In the following, recommendations are offered at project, organizational, and regulatory/policy levels. Those recommendations do not form a complete set, they only aim to designate initial areas of action.

Option 1: Embed ethical AI considerations in the project life-cycle

It is recommended to integrate ethical AI considerations into all phases of health-AI research projects. Researchers can then fit their ethical assessments into wider organizational AI governance and quality management processes (see O.1).

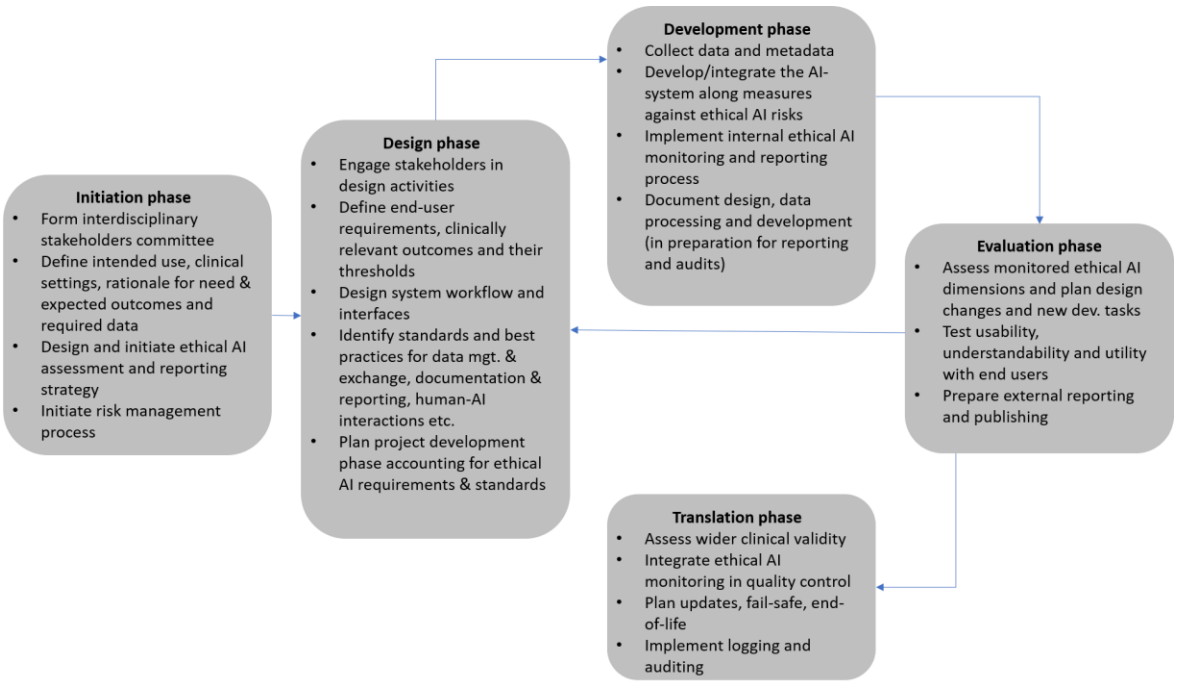

Figure 1 - Agile health-AI project management process embedding ethical aspect. Extended from Lekadir et al. [23].

Figure 1 is an example of such integration. An exhaustive list of tasks at each phase was not included. However, a few priority elements are highlighted.

Initiation and stakeholder involvement

The initiation phase is separated from the design phase to highlight the importance of early preparation in ethical health-AI assessment. This phase may start before the official start of the research project, namely, at idea generation, evidence review (e.g., systematic reviews) and protocol writing. In this phase, a strong rationale for the use of AI in the project needs to be established, including its intended use, workflows it would impact, expected clinical outcomes, and associated ethical risks. An understanding of the breadth of ethical AI risks and what could cause them is crucial during this phase (including design and conceptualization risks). Early stakeholder engagement is strongly encouraged [23][29].

Once the project moves toward implementation, a structured approach of involving stakeholders should be established. This includes forming a project management committee representing key stakeholder groups (see p. 39 for example on composition). The committee should be involved throughout the project's life cycle in key operational decisions, from design to clinical translation. This helps to 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. AI-driven treatment decisions must be supported by strong scientific evidence, and health system stakeholders should approve the development and use of AI to ensure its clinical relevance [24].

Ethical AI assessment and reporting strategy

An ethical AI assessment and reporting strategy should be designed early. AI based technologies and algorithms are very diverse and evolving. Each research project involving AI components will have specificities and challenges that are likely to be new. A specific assessment of the relevance and priorities of ethical AI issues for the project needs to be conducted 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 measuring and how it can be impacted by the dataset composition (a typical example is the accuracy's sensitivity to class imbalance in the dataset).

Standardized outcomes should be used when possible. When not, researchers should explain clearly how outcomes were derived. Along model performances metrics, clinical outcomes that enable comparison across studies, with or without AI, should also be reported.

Due to the early stage of regulating and reporting on health AI and the proliferation of frameworks, a project-specific quality and risk assessment strategy should be developed to evaluate relevant guidance and plan updates as needed. This ensures comprehensive ethical integration, facilitates publication and audit-readiness, and improves transparency and evidence quality. The advisory committee should contribute to identifying key risks, mitigation actions (e.g., procedural or technical), and monitoring strategies. Examples of strategy elements organized by ethical principles are provided on p. 40.

Framework selection and transparency

Commonly used assessment and reporting frameworks for a given dimension of interest, should be selected whenever possible. A catalogue of main health-AI frameworks identified

during the literature review has been compiled[8]. However, not all frameworks are rigorously validated. While it is common to select frameworks popular within a domain, validated frameworks with supporting publications should be prioritized. The advisory committee should review validation evidence to assess framework quality and identify gaps. The AGREE-II tool [20] may be used to asses guidelines based on scientific rigor, stakeholder involvement, applicability, and procedures for revision and monitoring.

To enhance transparency, details about the ethical AI assessment and reporting strategy should be published as supplementary materials. This can help other researchers undertake similar efforts and improve the appraisal of study outcomes.

Option 2: Professionalize health-AI projects portfolio management in hosting organizations

There are many ethical AI assessments relevant to health research and clinical translation. Their successful implementation depends on supporting components within the research organization, like a university hospital or R&D department, and broader industry elements such as scientific publishers and funding agencies.

Financial investment in developing or acquiring a new AI system is substantial [57] and these decisions are not made by health-AI researchers themselves. A wider assessment of required resources, real-world clinical impact, and societal, environmental, and reputational responsibilities can only be done by the organizations hosting the research. To maintain this big-picture view, it is recommended that organizations maintain and professionally manage health-AI project portfolios. Each project in this portfolio undergoes a set of evaluations by the end of which a decision must be made on whether to proceed to the next phase or stop/pivot the project (stage-gate process) [58]. Those evaluations are synchronized with the project life-cycle phases (see Option 1). Health-AI portfolio management ensures that the highest value initiatives are prioritized and that these initiatives are aligned with the organization's AI governance and data/cyber-security strategies, which usually extend to other types of initiatives within the organization. The FURM assessment framework is an example implementation of such a stage-gate process [57]. Figure 2 illustrates how this organizational process coordinates with the specific project life cycle. The project's ethical AI assessment strategy and its implementation are coordinated with the organizational processes. The presence of infrastructure experts, clinical implementors and quality and risks managers representatives in the project advisory committee guarantees such coordination.

Furthermore, given the increasing number of research projects involving AI in clinical settings, it is recommended that organizations build capacities for “ethics as a service”[9] [51]. Some re-

[8] https://github.com/elianemaalouf/ethical_AI_evaluation

[9] "Ethics as a Service" is an AI ethical governance model, inspired by Platform as a Service, that distributes responsibility across an independent multi-disciplinary ethics board and an organization’s internal AI practitioners. The board provides the core "infraethics"—developing an ethical code and defining the processes for pro-ethical design and audits, while technical practitioners are responsible for the contextual application and transparent documentation of these principles in specific algorithmic systems.

search projects require direct involvement by the organization, for example, projects implementing decentralized federated learning based on collaborations between multiple research organizations [59]. Organizations can be held responsible for data, cyber, and privacy violations, affecting their reputation due to scientific misconduct. A proactive ethical health-AI approach offers better control and positions organizations as pioneers. Regulations like the EU AI Act advocate risk-informed processes [23] and risk management. It is recommended that risk and quality management oversee ethical AI assessments and reporting.

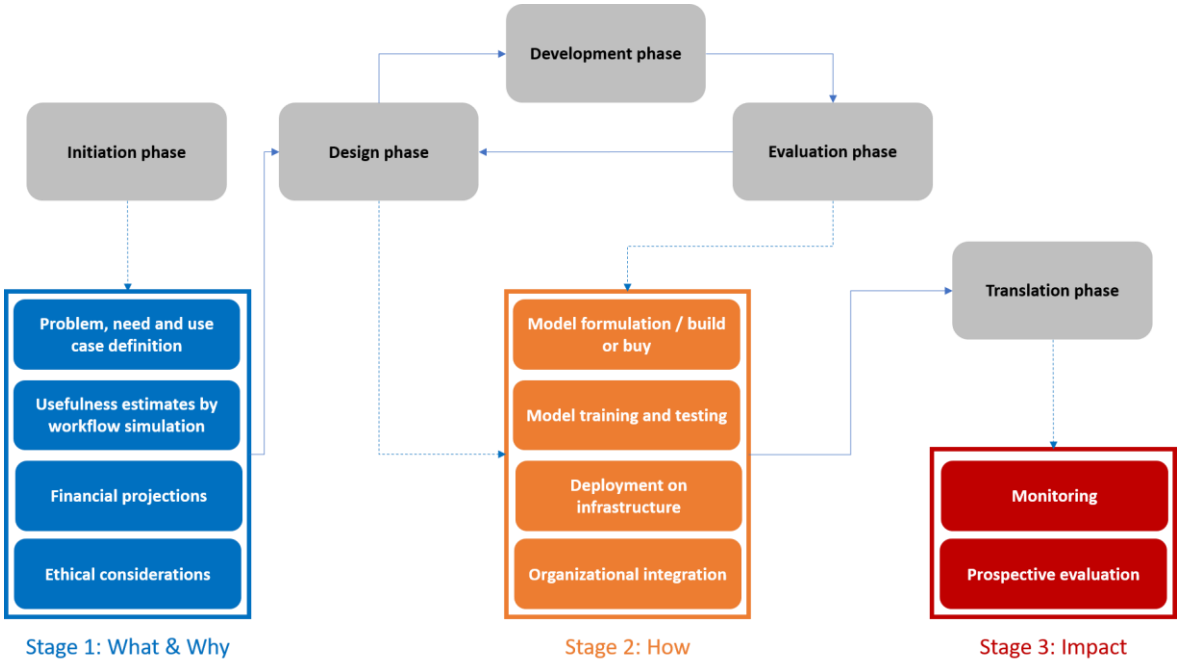

Figure 2 – Health-AI project life-cycle integrated in FURM stage-gate process [59]

Option 3: Ethical health-AI motivated funding in research and private investments

Standardizing ethical AI assessment and reporting is an emerging field. Scientific funding bodies should 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 develop and update guidelines and to implement well-established datasets and code-sharing concepts [20]. However, it is not advisable to judge model performance only on "benchmark" contests because such performances do not necessarily translate to AI-based science [19]. External validity does not guarantee clinical usefulness [62].

Furthermore, research funders[10] and scientific publishers have a strong influence in enabling

[10] 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)

ethical health-AI. They should exercise this influence by supporting and incentivizing projects that show rigorous and structured ethical AI assessments 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 should also support their reviewers in assigning a reproducibility label for the reviewed studies, which could be made 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 [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 in this. When possible, researchers on a project should also invite external researchers and ask them to attempt to reproduce the work [19][24][31]. Reproducibility researchers should be recognized in the community by making their contributions more visible, along the classical contributions they helped assess. It is important for researchers to define reproducibility clearly and establish standards that encompass not only computational reproducibility but also the accuracy of the findings [19]. This might require a cultural transformation in research, refocusing on transparency, quality and exhaustive documentation over the rush of publishing [20].

To control overoptimism with regards to AI capabilities, a research agenda to investigate efficacy of AI methods across scientific fields could also help to support better the debate on what can be realistically expected from the predictions of those models [19].

Option 4: Provide guidance on regulation applicability & agile regulatory processes

Establishing roles and responsibilities for monitoring AI performance and ensuring accountability is crucial to prevent and address any negative impacts [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 were generally ill-equipped to evaluate AI-centred research studies [77][16] adequately. Recent revisions to the Swiss Human Research Ordinance[11] require strengthening IT expertise in ethics committees with at least one member with specialized knowledge in health-related information technology, however, there are still no studies verifying the benefits of this addition. The nature of AI-systems requires periodic updates to their weights and the software/libraries. This adaptive nature adds complexities to the regulation process, namely for previously authorized health-AI technology. Regulators would need to adapt to the possibility of updates without requiring new certifications/authorizations [24] but considering

[11] https://www.fedlex.admin.ch/filestore/fedlex.data.admin.ch/eli/oe/2024/60/fr/pdf/fedlex-data-admin-ch-eli-oe-2024-60-fr-pdf-1.pdf#page=43.08

accelerated authorizations for previously authorized tools, for example based on a previously authorized “Change Control Plan” similar to the FDA’s[12].

Recommendations for increased agility in the regulatory process are reiterated [65]. This agility should be enabled by a set of actions such as: reinforcement of regulatory science [27] supported by knowledge and capacity building [65] and a nation-wide regulatory competence centre [65]; wider stakeholder involvement in the co-creation of regulations [65]; encouragement of "self-regulation" [27] by shifting the focus from innovation alone and placing more emphasis on the credentials of innovators and incentives to take responsibility; and the adoption of a proactive stance toward innovation, through continuous monitoring of new technologies [27].

Public-private partnerships should be initiated by regulators and policy-makers to establish "AI assurance laboratories" [66] for independent audits of critical applications and health research projects involving AI. These networks can also provide technical consultation during regulatory development, helping translate guidelines into practical actions. In coordination with research organizations, those laboratories could support "ethics as a service" expertise, pooling financial and technical resources to enable regular assessments and reporting. This coordination should also guide the establishment of a criticality classification for health-AI and define appropriate audit frequency and modality (e.g., third-party vs. internal).

[12] https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles

Implementation considerations

The Swiss regulatory context on AI is advancing. On the one hand, this is encouraging because the laws are evolving towards more explicit requirements for the ethical use of AI. On the other hand, however, this transitory phase can also be confusing to navigate for researchers until clear standards and specific guidelines are consolidated and recommended.

A reactive approach is strongly discouraged, and a proactive attitude toward ethical issues is encouraged. The current transition period should be used as a learning phase to build expertise in these areas. This proactivity is viewed not only as the responsibility of individual researchers but also as a shared obligation across all health research stakeholders to ensure the quality of scientific evidence and the safety and benefit of patients and health professionals.

In this section, the main challenges to implementing the overarching options for actions (see previous section) are presented, and some initial initiatives to facilitate their implementation and maintenance are proposed. These are summarized 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 conduct early consultations with stakeholders during the project design phase, especially for projects likely to involve complex changes to clinical workflows and/or patient healthcare modalities. Health research organizations could create permanent forums to recruit stakeholders for discussions and consultations on critical health initiatives, gathering beliefs, perceptions, expectations, and refining understanding of end-users' needs. In the long term, academic institutions should integrate education about AI into the curricula of clinicians and health researchers to increase understanding of the technology use, its limitations, and its development cycles [67]. AI developers working in healthcare should be mandated and certified in health research ethics (e.g., GCP certifications). Interdisciplinary master curricula combining technical, scientific, and ethical training would benefit future research project managers. Such education promotes a unified language and improves collaboration.
Regulatory transitory period limiting availability of clear assessment requirements	Ethics committees should propose a set of recommended documents to provide, in addition to the usual documentation required, when submitting a health-AI research protocol. This documentation could include the following: a clear rationale 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). Recently updated swissethics guidelines cover many of those aspects.

Challenge	Facilitating initiatives
	Health research organizations could provide support by operationalizing and internally standardizing this documentation by allocating data scientists, data engineers, and quality analysts to support researchers on projects, enabling them to efficiently and thoroughly identify and respond to those requirements. 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 led. Researchers on the project themselves should remain critical of AI's potential and uphold their ethical responsibility and scientific integrity as a guiding principle. 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. 40, 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 into the same organization, then during the development phase, the responsibility should be 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 standardizing interfaces with other systems (e.g., data exchange standards implementation, cybersecurity risks assessments). After development, the organization integrating an AI system into their workflows is primarily responsible for its maintenance, similar to other digital systems. AI systems require more frequent updates, both to the weights and the coding libraries, which involve specialized activities and knowledge of potential new risks (e.g., performance degradation, emerging security threats). A team of data scientists and data engineers should be available to carry out these updates and continuously assess their impact. Furthermore, given the critical nature of the AI system and its level of integration, the organization should be responsible for conducting at least internal audits (to prepare for potential external audits), with the frequency depending on the risk level.

Stakeholders' consultations

We performed individual consultations with 7 participants belonging to the stakeholders' groups of: regulators & policy makers, ethics committees, health-AI researchers, funding agencies. Those interviews brought updates and specifications to elements discussed in the previous sections.

Evolution of the regulatory context

At the time of writing this brief, there was no overarching regulation specifically dealing with AI. In February 2025, the federal council announced[13] [14] its intention to ratify the Council of Europe's AI Convention[15] and amend the Swiss law by defining the necessary legal measures, particularly related to transparency, data protection, non-discrimination and supervision. The council mandated federal offices to submit a bill on AI regulation for consultation as well as an implementation plan for legally non-binding measures by end of 2026[16] [17] [18].

A federal sectoral survey identified that the Federal Office of Public Health is currently developing a comprehensive overview of the potential use of AI applications, their general thematization and their regulation, considering the development of technical standards [72]. The survey mentioned the impact on Swissmedic's monitoring activities and potential requirement for an adaptation of the law due to the use of AI in medical devices [72], although it was not clear whether the impact mentioned is a general perception of AI's impact on the authorization process or whether it was referring to the context of adaptation to the EU AI convention. In fact, 12 product categories with AI components, including medical devices, will require a double conformity assessment before entering the European market starting August 2027 [73]. This fact might force Switzerland to review its product regulations for these products sectors, to conform to the EU AI Act and to consider expanding the Mutual Recognition Agreements with the EU[19] [73].

The upcoming development of the Swiss AI regulation will honour a set of regulatory principles, namely, conductivity to innovation for business and research, upholding fundamental rights as basis, and pursuing a principle-based technology-neutral and competition-neutral

[13] https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-104110.html

[14] https://www.bakom.admin.ch/en/artificial-intelligence

[15] https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[16] https://www.bk.admin.ch/bk/en/home/digitale-transformation-ikt-lenkung/bundesarchitektur/kuenstliche_intelligenz.html

[17] https://www.bj.admin.ch/bj/fr/home/staat/gesetzgebung/kuenstliche-intelligenz.html

[18] The relevant documentation summarizing these investigations are listed in : https://www.bakom.admin.ch/en/artificial-intelligence

[19] https://www.seco.admin.ch/seco/en/home/Aussenwirtschaftspolitik_Wirtschaftliche_Zusammenarbeit/Wirtschaftsbeziehungen/Technische_Handelshemmnisse/Mutual_Recognition_Agreement_MRA0/MRA_Schweiz_EU.html

legislative approach [73]. The intended objectives of such regulation are to guarantee continued access to the EU market for Swiss products integrating AI components, maintaining Switzerland as an innovation location, safeguarding fundamental rights and strengthening public trust in AI [73]. In supporting innovation under this upcoming legislation, the overview in [73] proposes different measures such as legislative sandboxes, industry-specific codes of conduct, AI strategies by the economic sectors or scientific endeavours. However, it remains to see which of these measures will be adopted in the AI bill or the implementation plan. It is important to mention as well that each canton will define the implementation of such AI bill in their jurisdiction; however, we did not extend our analysis to the cantonal actions.

In the meantime, and until it becomes clear which laws will be impacted[20], the current federal legal framework responds to potential issues caused by misuse of AI. In the context of health, health professionals and their employers bear responsibility of the outcomes of using an AI-integrated instrument in the clinical practice, in the same vein of using other instruments. It falls on the health professional to assess the adequacy of the use of the instrument in the clinical context. This assumes the ability of the professional to make such assessments based on communicated information declared by the producer/developer of the AI-integrated instrument, namely information about training and validation processes and demographics covered by the data. It is not clear how such assessments are currently happening in practice. In general, the Swiss civil law and liability law are the basis for managing litigations that might arise in the context of using AI in clinical practice. Extending responsibility to manufacturers, producers or importers is governed by the Product Liability Act in case a defective product causes harm.

In health research, recent reviews to the Federal Act on Research involving Human Beings and its ordinances[21], ICH Good Clinical Practice[22] and the declaration of Helsinki[23] introduce or reinforce dispositions related to consent for data reuse, data security and participants privacy and safety, risk-based approach to research conduct, increased use of digital tools and increased data analysis requirements. Those reviews are naturally relevant for oversight of health-AI in research. We recall that health-AI research should conform to the same ethical guidelines and laws as other health research projects involving humans[24]. The ethics committees' approval for health-AI projects follows the same approval process as other research projects. In their submitted protocols the researchers are now required to answer a list of questions pertaining to "issues to consider"[25], such as, justification of the use of the AI method and data sources, AI's model risk assessment, impact of errors, bias assessment and mitigation measures, data security, accountability and liability in case of harm, etc. Health researchers

[20] The legal analysis in [74] details the specific laws and their readiness or shortcomings in preparation for the adoption of the EU AI convention and the EU AI act.

[21] https://www.bag.admin.ch/en/regulation-of-human-research-in-switzerland

[22] https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf

[23] https://www.wma.net/policies-post/wma-declaration-of-helsinki/

[24] For a complete list of legal and ethical guidelines to abide by: https://www.easy-gcs.ch/grid/yecs/development/ethics-and-laws.html

[25] https://swissethics.ch/assets/pos_papiere_leitfaden/web_swissethics-position-statement_issues-to-consider-in-research-involving-ai_v1.0.pdf

should also abide by the Federal Act on Data Protection given the sensitive nature of health data they collect or reuse.

For completeness, Swissmedic emitted guidelines for the use of AI in medicinal product development[26]. Depending on the declared "intended purpose" and risk class of the AI-integrated instrument, the instrument might be categorized as medical device[27] and would have to comply with a specific set of regulations[28]. Although the authorization process differs between medicinal products and medical devices, both are subject to market surveillance. We were not able to assess in this work the impact of the different forms and levels of AI use on the authorization process. In research, the situation differs whether the project is intending to use a medical device, including software types of devices, or whether the project is investigating/developing the device. In the approval process, ethics committees and Swissmedic will assess whether the device is already authorized (e.g., CE marked), whether it will directly impact participants, etc.[29]. This process remains valid with AI-integrated software/medical devices today. We could not assess whether this process might be impacted by the upcoming federal regulations.

With this short update on the regulatory situation, we complement the situation presented in the policy brief text. However, this review is not exhaustive and we encourage the reader to further consult the provided pointers or to consult with legal experts.

Importance of reinforcing scientific integrity principles and participants protection

Regulatory evolutions are an explicit acknowledgment and response to the challenges brought by AI use and its pervasiveness, in health or otherwise. Although the technology itself, given its accelerated pace of development, hype, complexity and potential systemic impact, warrants particular attention and possible specific measures, it is important to remember that the context, health research for example, in which the AI technology is being used or developed, is not necessarily new and many of its aspects are already regulated. In health research, respect of the dignity and safety of participants as well as scientific integrity[30] are key basic principles honoured by ethics committees and funders and anchor the fulfilment of their duties in periods of technological innovations. Although new rules might not be needed, funders such as SNSF

[26] https://www.swissmedic.ch/swissmedic/en/home/humanarzneimittel/authorisations/artificiel-intelligence.html

[27] https://www.swissmedic.ch/dam/swissmedic/en/dokumente/medizinprodukte/mep_urr/bw630_30_007d_mbmedizinprodukte-software.pdf.download.pdf/BW630_30_007e_MB%20Medical%20Device%20Software.pdf

[28] https://www.swissmedic.ch/swissmedic/en/home/medical-devices/regulation-of-medical-devices/neue-eu-verordnungen-mdr-ivdr.html

[29] See «Decision tree for approval applications» in: https://www.swissmedic.ch/dam/swissmedic/en/dokumente/medizinprodukte/mep_urr/bw600_00_015d_mb_klinische_versuche_mep.pdf.download.pdf/BW600_00_015e_MB_Clinical%20investigations%20with%20medical%20devices.pdf

[30] https://www.snf.ch/media/en/LIogiKBPPwpCrszc/Kodex-wissenschaftliche-Integritaet-en.pdf

and their European partners[31], are considering updating their communications about the principles of scientific integrity, namely transparency and reproducibility, in relation to research developing or integrating AI-tools. There exists a need to re-explain and agree on feasible and actionable guidance to researchers on those aspects.

Researchers are not only developing AI-tools in the context of their research projects. They are also using them in their daily writing tasks. Funders, such as SNSF, and publishers are perceiving a positive impact on the quality of submissions 'redaction from non-native English speakers reducing their barriers to entry to funds and increasing equity among researchers. Although a general increase in the number of submissions is observed in recent years, this increase did not spike with the use of generative AI in the writing process. A shift started more than 10 years ago in funders 'evaluation processes, with the Declaration on Research Assessment[32] and other initiatives, to shift from indexes and impact factors-based evaluations into narrative CVs with a limited number of publications that can be verified for relevance to the project by reviewers, favouring quality over quantity and helping to manage this increase in submissions. Emphasising originality instead of novelty as a characteristic of excellent research[33] is also a way to encourage replication works and alternative tools to solve research questions.

Overall, the researchers remain ultimately responsible for following the code of conduct of scientific research and of honouring all regulations relevant in their context. Although we encourage proactive self-regulation, the report in [73] deems it to be insufficient to meet the obligations imposed by the EU convention on the incoming AI bill and government measures would still be needed.

Importance of building modern data infrastructures to face implementation challenges

Data infrastructures ensuring secure and interoperable health data exchange in the Swiss health system are still work in progress. The Digisanté[34] program, led by Federal Office of Public Health and the Federal Office of Statistics and started in January 2025, contains 50 projects to promote digitalization in the Swiss health system. The program will run for 10 years with the core objective to build a Swiss Health Data Space[35], a technical architecture supporting a unified data model with standardized interfaces for interoperability, unique identifiers, semantic structuring and support for plugging-in digital services. The program also intends to build the necessary legal frameworks for such infrastructure viability. One stated overarching objective of Digisanté is enabling secondary data use for scientific research. It is widely recognized today that training or updating AI algorithms requires large amounts of data, many of which would fall under secondary data use because they were collected for other purposes during routine clinical and administrative work or during previous research projects. Research projects

[31] https://www.scienceeurope.org/our-priorities/artificial-intelligence/working-group/

[32] https://sfdora.org/

[33] https://www.snf.ch/en/tf8nnJBdUJPCYODl/topic/the-snsfs-model-of-excellence

[34] https://www.digisante.admin.ch/fr/programme

[35] https://www.digisante.admin.ch/fr/espace-donnees-sante

are required to guarantee informed consent[36] for primary and secondary data use. The Digisanté architecture includes a module for "consent management", however, it is not clear yet what type of functions this module would serve and whether its implementation and use would require regulatory modifications.

Although seen as an important enabler for health system innovations, secondary health data use still faces important challenges in Switzerland. A recent issue paper [75] extensively analysed this topic. Unclear and fragmented regulations across cantons, increased risks and unclear benefits for sharing in data holders' institutions, blurring between risk and responsibility, lack of common technical standards and alignment on governance, among other challenges, cause risk aversion to data sharing in institutions. The issues expressed in [75] are interrelated with health-AI and might carry over during the future implementation of the Swiss AI bill, and specifically to responsible AI use and assessments both in terms of risks and quality. In essence, we borrow the recommendations by the author in [75] and see them equally relevant to responsible health-AI: support to institutions willing to spend resources on guaranteeing responsible-AI implementation during research or in production, encourage communication about the importance and the benefits to society from those efforts, provide a clear coherent regulatory framework consistent between federal and cantonal implementation but also coherent among cantons, propose technical guidelines to support faster operationalization effort, provide forums for experience sharing and technical training. We further agree with the author in [75] on the importance of a holistic and synergistic view of regulating data privacy, data use and, we add, AI oversight. These areas have pervasive impacts on populations' safety and public resources management, but also potential benefits and their challenges are not independent from each other. Data is an asset and it gives a competitive advantage when analysed and acted upon. We speculate that incentivizing for data sharing will become harder - another reason to encourage synchronizing efforts to solve for, what we see, as systemic challenges. Maybe health research, with or without AI, would suffer to a lesser extent from data reuse hurdles by their bounded scopes and their specific provisions (although three use cases in public health research [76] illustrate difficulties that are probably very common). Nevertheless, we expect long term consequences in hindering implementation and integration in production which would require significant additional infrastructure work.

Translating AI-based research into the clinical setting could face practical regulatory and infrastructure challenges. AI-based clinical tools might fall in the software as medical device category, based on their «intended use». Research organizations like university hospitals might not be equipped to manage such categorization with its legal, quality and technical requirements. Integration might face limited IT readiness to automate an AI-algorithm in production which might require developing specific operationalization infrastructure and integration interfaces with existing warehouses and databases. Those developments might be done with open-source code that are not necessarily managed by the institution's IT department increasing information security risks. Institutions' existent software might be lacking tools (e.g., APIs) to support integration and automation in the first place. Lack of resources (e.g., budget, qualified personnel, tools) for sustainable exploitation and maintenance of the AI system once in production limits the achievement of long-term benefits of the integrated tool and increases infrastructure and information security risks.

[36] https://www.edoeb.admin.ch/en/human-research

Although no specific technical guidance exists today for implementing responsible-AI in health, researchers need to be aware of and favour the use of methods and tools with technical and formal guarantees, responsible by design. For example, in the context of privacy, the current standard approach in multi-centric research is to anonymize the data and centralise it in a Trusted Research Environment[37]. Federated learning removes the need for transferring data by transferring model weights trained in a decentralized manner at the data source. The generation of synthetic data to train models instead of the real data avoids membership or attributes inference attacks or data reconstruction risks that can be gleaned from model weights trained on real data. Although federated learning and the use of synthetic data are the state of the art in privacy preserving technologies, their clinical utility compared to the standard approach[38] still needs to be proven in real-world comparative studies. Furthermore, interpretability by design is also more favourable than ad-hoc explainability tools since those tools depend themselves on modelling assumptions that cannot necessarily be verified. It also matters to account for the actual clinicians' need for explainability when interpreting and using AI-based recommendations. There is clear need by clinicians to trust the reliability of the predictions or the recommendations, by gauging an information about the uncertainty of the model, without necessarily a need, neither interest, in understanding the inner workings of a model. These inner workings might still be important for developers to apply diagnostics assessing the model behaviour, but such diagnostics are not necessarily relevant for clinicians. Traceability in a multi-agentic approach would still be important to clinicians to understand the chain of tools and intermediate system calls that lead to a given recommendation. Testing for fairness by comparing model predictions in demographic subgroups to the overall model predictions metrics could also be made commonplace methods to assess risks of bias. Improving fairness could also benefit from synthetic data to rebalance ill-represented groups in the dataset.

Responsible technology by design is still largely in the research domain and would benefit from assessments in realistic clinical settings to inform recommendations and standardization efforts. Although regulations have to remain technically neutral, we believe that regulators and policymakers can encourage the creation of technical forums, consortia or conferences to tackle technical guarantees and emit such guidelines that are adequate to the Swiss context[39]. Standards might also already exist; those forums can provide guidance on how to implement them or how to fill their potential gaps.

An important part for controlling risks in transitioning to secure infrastructures is post-market surveillance. A categorization as a software as medical device would warrant surveillance by Swissmedic, although, we have currently no information on how this process would be adapted, if any, to AI-use. However, when the AI-tool does not categorize as a device, there is

[37] For example BioMedIT(https://www.biomedit.ch/), which still requires setting up data transfer and processing agreements and providing security guarantees at the central depot

[38] For example, generating privacy preserving synthetic data involves applying differential privacy methods that add calibrated noise to the real data. This addition could bias the model outcomes trained with synthetic data and structured validation is required to assess this bias.

[39] For example in an implementation act, similar to what the EU does by citing a standard in the Official Journal of the EU after completing a standards assessment for its conformity with the regulation. See: https://artificialintelligenceact.eu/standard-setting-overview/

a risk of it not being assessed for bugs, behaviour shifts, and updated adequately if no specific resources are assigned for such tasks. Furthermore, publicly available tools, without any form of oversight, could still be used in clinical settings although without authorization. Anecdotal evidence refers to doctors/nurses sharing patients' data via WhatsApp because it was easier and faster during remote visits or chatting with chatGPT about a patient 's case during consultations. The extent to which patients' data is shared cannot be controlled with accessibility to such public tools from a multitude of platforms. Those forms of AI, that do not fall in the usual categorization might still pause risks to patients 'safety and privacy. We believe that the health ecosystem should consider these situations as sources of risk and apply measures of education and adequate technical restrictions to detect and avoid them.

The state of technical standards and certifications

The EU requested the creation of standards that are conform with the AI-act, and for the high-risk applications in which conformity assessments would be required. Many documents are being drafted and seem to be behind schedule[40]. Regardless, coordinated efforts between CEN and ISO/IEC[41] are the most advanced in this regard to date. SNV[42] is the National Standards Body representing Switzerland in those committees and would be responsible to adapt those standards to the Swiss context. These standards form a framework for the responsible development and deployment of Artificial Intelligence, with a strong focus on aligning with the EU AI Act. The drafts establish terminology and concepts, AI risk management, functional safety, and a detailed trustworthiness framework encompassing logging, transparency, accuracy, and robustness. Significant emphasis is placed on ethical and societal concerns, with dedicated guidelines for managing bias, conducting human rights impact assessments, and defining competence requirements for AI ethicists. Finally, the standards solidify the infrastructure for commercial and regulatory compliance through documents on quality management systems, conformity assessment, and specific evaluation methods for systems like computer vision and natural language processing. Although AI standards would drive harmonization, their development process is criticized[43] as gameable by big technology companies and driven by politics with little influence of independent experts and civil society. Organization hosting health-AI research should follow on these efforts, especially, in relation to translational projects that have a high chance of being implemented in clinical contexts.

Standards bodies rely on certification labs to perform necessary evaluations assessing an AI-product conformity with a given standard. At the current stage of standards development, it is not clear what those evaluations would exactly entail for health-related applications. In general terms, an evaluation process will first define the dimensions on which the AI-based tool needs to be evaluated, define metrics for each of those dimensions and their target values, perform a power analysis based on a chosen margin of tolerance for each metric's value to

---

40 https://standards.cencenelec.eu/dyn/www/f?p=205:22:0::::FSP_ORG_ID,FSP_LANG_ID:2916257,25&cs=1827B89DA69577BF3631EE2B6070F207D

41 https://www.iso.org/sectors/it-technologies/ai

42 https://www.snv.ch/en/information-on-standards/which-subjects-can-i-contribute-to/details-of-committees/nk/nk-149.html

43 https://corporateeurope.org/en/2025/01/bias-baked

define the amount of data to collect for metrics‘ estimation, collect the necessary data, estimate the metrics and compare estimates to their target values. Important to note here that evaluation data is different and separate from the AI-model training data. In fact, certifications assessment might not require access to the model's data nor model weights, unless a specific assessment is required on the data that was used in the training. Naturally, assessing more dimensions with varied categories increases the requirements on evaluations ‘data collection (e.g., assessing fairness might require collection variables of gender, age, social group, etc.). A certification lab would emit a certificate that would be valid for a given period of time and the assessment is renewed to renew the certification. Given the diversity of health-AI applications and their specificities in their context, third-party certifiers, such as the certification labs, could only assess and certify certain aspects. Self-assessments would still be needed to prove conformity with local regulations and specific applications ‘claims and risks. Mirroring EU AI act requirements, self-assessments require maintaining a quality management system and a detailed technical documentation on data governance, risk management, transparency, human oversight, etc. Health organization hosting AI-based tools might not have qualified data scientist who are also versed in quality management in order to perform those internal controls on-premises. We believe that those human capacities would be important to build in health organizations not only to assess newly developed health-AI tools used in clinical practice but also to practically assess external manufacturers' claims on the performances of algorithms integrated in their devices. We expect, in the long run, that such AI quality assessment capacities would improve patients ‘care and limit costs.

Limitations of the current work

The scope of the brief was large, which allowed us to have a wide overview on the dimensions of the problem. However, this wide scope was impossible to fully cover by our consultations. For example, we were not able to reach Swissmedic in time but their insight on AI-based medical devices would have been crucial, cantonal representatives would have been informative on practical aspect of upcoming legislation adaptations at cantonal level and their related enforcement challenges. We also were not able to reach hospital governance and other health organization implementing and working with health-AI projects to assess their current challenges and the expected change they foresee in the future on their operations. We also did not reach to health professionals interacting with AI-based recommendations to assess the views and challenges. We suggest that future works take a complementary approach to ours and target specific dimensions of the overarching issues we distilled in this brief in order to improve on our suggestions and validate their operationalization and actionable potential.

The conversation about ethical AI quality and risk assessments seemed too early in some circles and many stakeholders are in the process of building experience and learning from early developments. Nevertheless, we still believe that a proactive approach is useful here and early assessments and communication about this topic are important. Extending involvement to health professionals and patients to clarify expectations, education on risks of over reliance on AI-based tools before secure guardrails are put in place might help minimize risks. Reminding AI developers of the importance of making ethical design choices and health organizations of their role in enabling such choices in practice help make structured assessments common place. Other risks are also important to introduce early in the conversation, such as environmental impact, impact on equity and access to health technology, etc. An approach to continuously monitor those issues and their evolution, clarifying them and initiating solutions to act on them would be beneficial for health-AI science and for society in general.

Acknowledgments

We would like to express our sincere gratitude to all the stakeholders that participated in the consultations for their contributions. We thank:

- Prof. Ph.D. Jean-Louis Raisaro, CHUV/University of Lausanne, for clarifying implementation challenges in production systems and the use of responsible by design technologies.
- Ph.D. André Anjos, IDIAP Research Institute, for clarifying the current state of technical standards and the functioning of certification labs.
- Ph.D. Paola Daniore, Centre for Digital Trust, for clarifying challenges facing secondary health data reuse.
- Ph.D. Tobias Philipp, Swiss National Science Foundation, for clarifying the scientific funders perspective and challenges.
- Prof. Ph.D. Dominique Sprumont, CER-VD, for clarifying ethics committees' work.
- Anahita Gervais de Lafond, University of Neuchâtel, for clarifying health law and for her feedback on the policy brief text.

Finally, a sincere note of appreciation to the anonymous stakeholders whose support was instrumental in reaching this version of the policy brief.

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] "Artificial Intelligence (AI) and Research Involving Human Beings:Issues to Consider When Submitting a Projectto a Research Ethics Committee", swissethics, Accessed: Oct. 03, 2025. [Online]. Available: https://swissethics.ch/assets/pos_papiere_leitfaden/web_swissethics-position-statement_issues-to-consider-in-research-involving-ai_v1.0.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".

[72] "Rapport OFCOM, Aperçu des activités de réglementation sectorielles actuelles en lien avec l'intelligence artificielle," Accessed: Sept. 20, 2025. [Online]. Available: https://www.bakom.admin.ch/dam/en/sd-web/JlCv0SA42Os4/ueberblick_regulierungsvorhaben.pdf

[73] "Overview of artificial intelligence regulation, Report to the Federal Council," Accessed: Sept. 20, 2025. [Online]. Available: https://www.bakom.admin.ch/dam/en/sd-web/MKUEgci9oDRD/Auslegeordnung%20zur%20Regulierung%20von%20%20k%C3%BCnstlicher%20Intelligenz_def.pdf

[74] "Analyse juridique de base dans le cadre de l'état des lieux sur les approches de régulation en matière d'intelligence artificielle," Accessed: Sept. 20, 2025. [Online]. Available: https://www.bakom.admin.ch/dam/fr/sd-web/4HwAbRI-HeC1/analyse_juritisch.pdf

[75] P. Daniori, "Closing the benefit-risk loop: Realizing the value of secondary use of health data in Switzerland," C4DT, September 2025. Accessed: Sept. 20, 2025. [Online]. Available: https://drive.switch.ch/index.php/s/KMsuEJB1To1NFNA

[76] M. Bochud, M.-A. Le Pogam, "Recommendations for a more integrated Swiss health data ecosystem based on lessons learned from selected use cases for data re-use in academic public health research," June 2025. Accessed: Sept. 20, 2025. [Online]. Available: https://www.digisante.admin.ch/dam/fr/sd-web/tgDANxUSHu7O/20250625_Overall%20report%20uses%20cases%20academic%20research_UniSante%20V1.0_final.pdf

[67] Ferretti A, Ienca M, Velarde MR, Hurst S, Vayena E. “The Challenges of Big Data for Research Ethics Committees: A Qualitative Swiss Study”. Journal of Empirical Research on Human Research Ethics. 2021;17(1-2):129-143. doi:10.1177/15562646211053538

Project management committee

The committee could include the following members, along examples of their 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 Modelling 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 labelling process (if any) and assessment of labels quality and variability/reproducibility between labellers [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 modelling 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[44], nFADP[45]) such as the principle of "privacy by design" [24]; ensure informed consent for newly collected data or for secondary uses of historical ones [24]; fulfil 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].

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

[45] 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, reviewing and stakeholders consultations.
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).

Comment: 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.

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.