NextGen

Genome-Centric Multimodal Data Integration in Personalised Cardiovascular Medicine.

Deliverable 6.1

Overview of the Ethics of Human-in-the-Loop AI in Medicine

Grant Agreement Number: 101136962

Funded by the European Union

UK Research and Innovation

The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No. 10098097, No. 10104323]

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

NextGen	
Project full title	Next Generation Tools for Genome-Centric Multimodal Data Integration In Personalised Cardiovascular Medicine
Call identifier	HORIZON-HLTH-2023-TOOL-05-04
Type of action	RIA
Start date	01/ 01/ 2024
End date	31/12/2027
Grant agreement no	101136962

Funding of associated partners
The Swiss associated partners of the NextGen project were funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No 10098097, No 10104323]

D6.1 - Overview of the Ethics of Human-in-the-Loop AI in Medicine	
Author(s)	Jojanneke Drogt, Karin Jongsma, Saskia Haitjema
Editor	Jojanneke Drogt
Participating partners	UMCU
Version	1.0
Status	Final
Deliverable date	M18
Dissemination Level	PU - Public
Official date	June 30th 2025
Actual date	June 30th 2025

Disclaimer

The NEXTGEN consortium consists of the following partners:

No	PARTNER ORGANISATION NAME	ABBREVIATION	COUNTRY
1	UNIVERSITAIR MEDISCH CENTRUM UTRECHT	UMCU	NL
2	HIRO MICRODATACENTERS B.V.	HIRO	NL
3	EURECOM GIE	EURE	FR
4	JOHANN WOLFGANG GOETHE-UNIVERSITAET FRANKFURT AM MAIN	GUF	DE
5	KAROLINSKA INSTITUTET	KI	SE
6	HUS-YHTYMA	HUS	FI
7	UNIVERSITY OF VIRGINIA	UVA	US
8	KLINIKUM RECHTS DER ISAR DER TECHNISCHEN UNIVERSITAT MUNCHEN	TUM-Med	DE
9	HL7 INTERNATIONAL FOUNDATION	HL7	BE
10	MYDATA GLOBAL RY	MYDTA	FI
11	DATAPOWER SRL	DPOW	IT
12	SOCIETE EUROPEENNE DE CARDIOLOGIE	ESC	FR
13	WELLSPAN HEALTH	WSPAN	US
14	LIKE HEALTHCARE RESEARCH GMBH	LIKE	DE
15	NEBS SRL	NEBS	BE
16	THE HUMAN COLOSSUS FOUNDATION	HCF	CH
17	SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA	SUPSI	CH
18	DRUG INFORMATION ASSOCIATION	DIA	CH
19	DPO ASSOCIATES SARL	DPOA	CH
20	QUEEN MARY UNIVERSITY OF LONDON	QMUL	UK
21	EARLHAM INSTITUTE	ERLH	UK

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
30/06/2025	1.0	Submitted version of the deliverable	UMCU

Authors

AUTHOR/EDITOR	ORGANISATION
Jojanneke Drogt	UMCU
Karin Jongsma	UMCU
Saskia Haitjema	UMCU

Reviewers

REVIEWER	ORGANISATION

List of terms and abbreviations

ABBREVIATION	DESCRIPTION
AI	Artificial intelligence
HITL	Human-In-The-Loop
HLEG	High-Level Expert Group (HLEG) EU
FDA	U.S. Food and Drug Administration
CE	Conformité Européeenne (European Conformity)
ML	Machine learning
GAI	Generative AI
LLM	Large Language Model

Table of contents

1 Executive summary

This document offers a comprehensive overview of the ethics of Human-in-the-Loop (HITL) AI in medicine, highlighting that the concept is based on a range of diverging assumptions about who the 'human' is or should be, what roles they have, and what they will be able to accomplish.

Application of HITL in medical contexts

Studies highlight the risks associated with using AI in medical settings and illustrate the widely held belief that HITL is a necessary condition for managing risks when using AI to support medical decision-making. The literature suggests that HITL is particularly critical in high-stakes medical decisions, such as cancer treatment planning and pediatric care, where errors can have severe consequences. Stakeholders, including patients and the general public, confirmed that HITL is viewed as an essential approach towards implementing AI responsibly.

Human roles and responsibilities

In various articles, HITL is presented as an alternative to fully autonomous AI systems in medicine, where medical professionals are kept in the loop instead of being replaced by AI. Medical oversight in HITL frameworks often involves validating or refining AI-generated outcomes; yet, medical oversight does not need to be limited to validating AI-generated outcomes, as they can be involved as medical officers guiding institutions and companies to target the most relevant and topical clinical decision-making tools, act as consultants providing continuous supervision to AI bodies, and supervise AI tools in practice to prevent medical errors. Many authors have high expectations and assumptions about what such human oversight by medical professionals could offer. It was stated that medical professionals could improve AI's safety and quality, ensure that AI-driven insights are clinically meaningful, ethically sound, and contextually appropriate, and identify biases and shortcomings of AI, such as hallucinations. Additionally, there is a call for expanding multidisciplinary involvement to include ethicists, AI developers, and patients to ensure oversight.

Conceptual ambiguity

When referring to HITL, authors may mean different things: some describe humans who simply verify AI outputs in clinical contexts, others refer to their participation in various stages of AI development and implementation, while some refer to maintaining control at all times. This indicates that HITL lacks a universally agreed-upon definition in the context of medical AI. A problem that challenges reaching a consensus on HITL is that it is not only difficult to determine the appropriate level of oversight, but also that there is a gap between how HITL is interpreted in the technical literature and discussions on AI governance and ethics. Technically, it is often framed as human oversight aimed at improving system performance, training, and error mitigation. Ethically, HITL emphasizes the preservation of human autonomy and control in clinical decision-making. Since such a discrepancy can significantly affect the meaning of human requirements and control in HITL approaches, it is crucial to clarify the term, also in relation to other related concepts such as Human-Centered AI (HAI).

Further ethical and practical considerations

Several ethical and practical considerations are mentioned when discussing effective HITL implementation:

- **Explainability:** For humans to effectively oversee AI, it may be necessary to provide transparent and interpretable outputs, enabling human involvement in decision-making processes.
- **New human competencies:** The evolving nature of AI can demand new competencies from healthcare professionals to interact effectively with AI tools.
- **Limits of human oversight:** Continuous monitoring of AI outputs can lead to human fatigue or an increased risk of errors. Moreover, human oversight does not always lead to better outcomes, as humans can have an automation bias, confirmation bias, or override a correct AI system.
- **Removing humans from the loop or humans being 'near' the loop:** Whether there are medical settings in which humans can be removed from the loop or be ‘‘near’’ it instead of being fully in the loop, still needs to be determined and may be considered when medical expertise or control is less valuable or not needed for differentiating AI errors.

In summary, while HITL is widely recognized as essential for safe, ethical, and effective AI integration in healthcare, its practical realization requires clearer conceptualization, a robust practical framework, and multidisciplinary collaboration to overcome current limitations and ensure responsible deployment.

2 Introduction

Artificial Intelligence (AI) systems are increasingly developed for medical purposes and are gradually implemented in various clinical contexts, achieving notable success in image classification. (Rajpurkar et al., 2022) Given the accelerating rate at which FDA (Benjamens et al., 2020) and CE (Muehlematter et al., 2021) approvals are provided for AI-based medical systems, it seems likely that AI will be adopted more widely in medicine in the coming years. At the same time, the use of medical AI systems in medical practice remains challenging, particularly given the severe risks and ethical concerns associated with their use. (Rajpurkar et al., 2022) Recently, a new case of previously undetected harmful bias was, for instance, found in one of the most cited AI models used to scan chest x-rays for diseases, signalling that it doesn't accurately detect potentially life-threatening diseases in marginalized groups, including women and Black people (Ortega, 2025; Yang et al., 2025)

Considering the risks associated with medical AI systems, several authors and guidelines have argued that human oversight and monitoring play a critical role in their use. (Commission, 2024; Guidance, 2021; Morandín-Ahuerma, 2023) In *the Ethics Guidelines for Trustworthy AI,* the EU High-Level Expert Group (HLEG) on Artificial Intelligence has stated that ''human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects.'' (HLEG, 2019) One prominent approach to human oversight is to employ human-in-the-loop (HITL). HITL is a term used to describe two different processes: 1) it refers to a semi-automatic annotation approach (Chen et al., 2023), 2) it is more broadly applied in medical contexts for ''crucial decisions [that] must be subject to human control instead of being processed fully automatically'' (Salloch & Eriksen, 2024) or, as the HLEG formulates, ''the capability for human intervention in every decision cycle of the system.'' (HLEG, 2019)

Even though HITL is often described as a desirable and likely way forward when implementing medical AI (Jotterand & Bosco, 2020; Rajpurkar et al., 2022), the approach has received criticism for being overly broad and providing insufficient grounding for determining what kind of oversight is required. Haselager et al. (2024), for example, maintain that ''a human in the loop does not ensure that effective human oversight will be exerted to the extent required for moral and legal responsibility. Rather, humans might end up being 'under' the loop, merely playing a symbolic role by providing formal 'stamps of approval' without genuine reflection.'' (Haselager et al., 2024) Similarly, Santoni de Sio and Van den Hoven argue that 'being in the loop' is insufficient for maintaining control over an activity, proposing that Meaningful Human Control (MHC) is a more suitable approach to ensuring human oversight. (Santoni de Sio & Van den Hoven, 2018) Additionally, Salloch and Eriksen (2024) flag that the 'human factor' remains too vague in HITL and that it remains to be determined which humans need to be in control in medical contexts. (Salloch & Eriksen, 2024) Even the HLEG states that HITL, as having the ability to intervene in every decision cycle, is ''in many cases (…) neither possible nor desirable.'' (HLEG, 2019)

Despite this ongoing debate, to our knowledge, no comprehensive overview of the ways in which HITL is used in the academic discourse in medicine exists. Such an overview is a valuable contribution to the academic literature, as it provides insights into assumptions about HITL, the humans and AI systems involved, which is helpful for formulating more precise guidance and provisions for human oversight over AI in medical decision-making.

3 Methods

For this overview, we employed scoping review methodology to investigate how the term 'Human-in-the-Loop' is applied in the literature on AI in medical decision-making and what kind of human oversight is considered desirable. Scoping reviews are particularly suited for emerging topics where there is still ambiguity about how the topic is conceptualized and addressed.(Ienca et al., 2018; Munn et al., 2018) Furthermore, unlike a systematic review which typically focus on evaluating specific interventions or clinical practices, scoping reviews are more appropriate for exploring how particular concepts are employed across the literature. This approach allows for a comprehensive mapping of concept usage rather than a narrow focus on outcomes or efficacy of medical treatments or practices (Munn et al., 2018)

On May 20, 2025, we performed a literature search using four databases (Embase, Scopus, Web of Science, and PubMed) to retrieve eligible publications. We searched title, abstract, keywords and complete publications for the terms: ("Human-in-the-Loop" OR “HITL” OR “in-the-loop” OR “in the loop”) AND (“AI” OR “Artificial Intelligence” OR “machine learning” OR “ML” OR “deep learning” OR ‘’DL’’ OR “decision support” OR “CDSS” OR “DSS” OR “autonomous system” OR “automated system”) AND (“healthcare" OR "health care" OR ‘’medicine’’ OR "medical care" OR ‘’medical practice’’ OR "clinical care" OR ‘’clinical practice’’ OR "health services" OR "health system" OR "medical system" OR "patient care" OR “cardiology” OR “pathology” OR “radiology” OR “ophthalmology” OR “dermatology” OR “general care” OR “general practice” OR “intensive care” OR ‘’gastroenterology’’ OR ‘’obstetrics’’ OR ‘’gynaecology’’ OR ‘’immunology’’ OR ‘’anesthesiology’’ OR ‘’oncology’’ OR ‘’surgery’’ OR ‘’neurology’’ OR ‘’critical care’’ OR ‘’nephrology’’ OR ‘’pediatrics’’ OR ‘’hematology’’ OR ‘’neonatology’’). Queries were adapted to accommodate the language used by each engine or database; see Appendix 1. Screening identified 823 entries. All entries were imported into the RIS format for the Rayyan screening tool, whereby the software identified 374 duplicates. Guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework (Moher et al., 2009), we conducted the screening process according to the four steps of identification, screening, eligibility and inclusion (**Figure 1**).

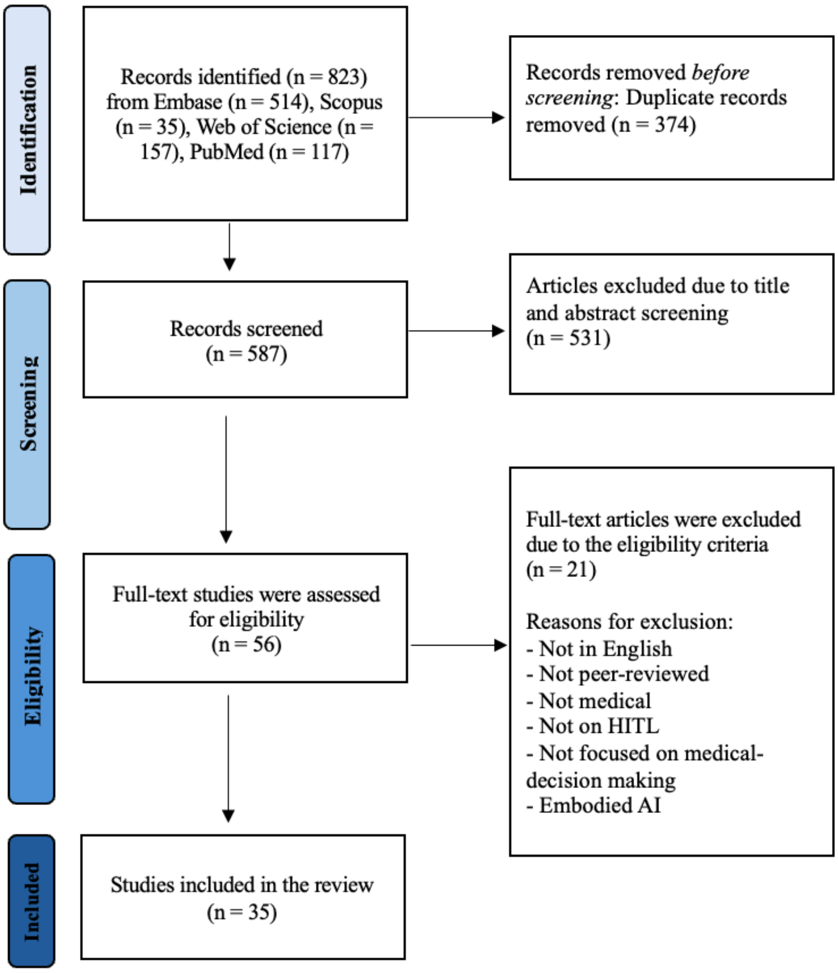


Figure 1. Scoping literature review flow cart (PRISMA)

After removing duplicates, the eligibility assessment was conducted independently by two of the co-authors (JD and KJ) on 587 articles through title-abstract screening; JD performed full-text screening for the remaining 56 studies. Diverging inclusion choices between reviewers were discussed in the research group with documented reasons. Studies included in the overview had the following features: (1) written in English; (2) peer-reviewed article; (3) full-text available; (4) the term Human-in-the-Loop or a similar variation of the term is mentioned (for example, clinician-in-the-loop); (5) focused on medical decision-making, and (6) focused on non-embodied AI. Articles mentioning a term containing 'loop' that refer to different concepts, such as Human-on-the-Loop or Machine-in-the-Loop, were excluded. Additionally, articles focusing on HITL as a technical annotation method or development approach were also excluded.

Full-text screening was conducted systematically using a data charting table (Appendix 2). Based on the recommendations to improve scoping study methodology from Levac et al. (Levac et al., 2010) and the approach taken by Ienca et al. (Ienca et al., 2018) in their scoping review, our research team worked together to create a data charting form that identified the variables to extract from the review data. We decided that the data charting form should focus on (1) the human component in HITL and how it is specified, (2) the roles and tasks envisioned for AI in a HITL model, (3) how HITL is connected to other AI ethics discussions on, for example human-AI collaboration, (4) which explicit ethical aims and considerations are mentioned in relation to HITL. During the process, we worked iteratively and tailored the data charting form to align with the research data. (Levac et al., 2010)

4 Findings

In our search, we found 36 articles that discussed 'human-in-the-loop' or a variation of the term, such as 'clinician-' or 'doctor-in-the-loop,' for AI in medical decision-making. All records were published between 2018 and 2025, with a surge in articles in 2024 and 2025. We present our results in accordance with the data charting form used for analysing the articles and discuss (1) the ethical and public rationale for needing a human in the loop, (2) the proposed roles and responsibilities of medical professionals in the context of HITL, (3) the desirable ways in which AI is developed and used in HITL and (4) open questions and criticism regarding HITL in health.

4.1 Ethical and public rationale for needing a human in the loop

Many authors believe that HITL is a reliable way to mitigate the risks associated with medical AI systems; they strongly advocate for human-in-the-loop approaches, convinced that human expertise and oversight are crucial in balancing the risks of medical AI systems. (Haemmerli et al., 2023; Holmes et al., 2025; Jotterand & Bosco, 2020; Ku, 2019; Livingston et al., 2025; Malaguti et al., 2025; McCaffrey et al., 2025; Montomoli et al., 2024; Robinson et al., 2024) In a review article on the challenges of AI in medicine, Aldosari et al. (Aldosari et al., 2025) demonstrate that various authors are, for example, concerned about algorithmic errors and patient harm resulting from AI usage, and that incorporating a HITL can offer a solution to reduce the likelihood of mistakes; here, HITL is referred to as ‘‘having someone in control at all times.’’ McCaffrey et al. also discuss how the shortcomings of AI, such as biases, confabulations, and practical challenges of implementing AI in existing workflows, as well as end-user acceptance, can be mitigated by incorporating a human-in-the-loop.(McCaffrey et al., 2025) Due to the severity of the risks associated with utilizing AI, Gunes et al. argue that only by adopting a HITL approach can the full potential benefits of AI, such as increased efficiency, be realized. (Gunes et al., 2025)

Stakeholders confirmed that HITL is an essential approach towards implementing AI responsibly. In their interview study, Lee et al. describe how patients emphasized the ongoing need for human involvement in providing therapy for mental health issues. In order to foster safe AI use, participants noted that humans should at least be able to respond in emergencies (i.e., self-harm, suicidal ideation, or thoughts of harming others), and many also believed that humans should remain the primary providers of care. (Lee et al., 2025) Another empirical study found that children and young persons also deemed medical professional oversight essential when AI is used for imaging purposes, expressing a preference for medical professionals in the loop. (Lee et al., 2024) Moreover, a mixed-methods study found that the general public viewed human supervision in the form of HITL as an essential condition for the use of AI in pathology, to verify the validity of diagnoses, reduce errors, and maintain human control. (Lewis et al., 2025) Lastly, a scoping review of stakeholders' perspectives on the future of AI in radiology found that radiologists themselves believed they should be kept in the loop in terms of responsibility. (Yang et al., 2022)

4.2 HITL in medical practice: proposed roles and responsibilities of medical professionals

In various articles, HITL is presented as an alternative to fully autonomous AI systems in medicine, where medical professionals are the ones who are kept in the loop instead of being replaced by

AI.(Ali et al., 2024; Lee et al., 2025; Liew, 2018; Malaguti et al., 2025; Mudgal & Das, 2020; Phongpreecha et al., 2025; Salih et al., 2024; Sezgin, 2023; Wadden, 2024; Woo et al., 2024; Zhang et al., 2025) Keeping a physician or nurse in the loop is presented as a desirable approach to the integration of AI in different medical contexts, such as the use of AI in multidisciplinary team meetings (Ali et al., 2024) assessing discharge needs (Duckworth et al., 2024), MRI artifact detection and correction (Gunes et al., 2025), and recognizing and predicting health changes in smart home environments (Fritz et al., 2022). Malaguti et al. (Malaguti et al., 2025) argue that the increasing use of the term 'clinician-in-the-loop' points to the recognition that clinician involvement ensures that AI-driven insights are clinically meaningful, ethically sound and contextually adapted, which they recognized in recent studies that state that clinicians can enhance the reliability of predictions, prevent algorithmic errors, and foster patient trust.

Medical oversight in HITL frameworks often involves validating or refining AI-generated outcomes. (Gunes et al., 2025; Liew, 2018; McCaffrey et al., 2025; Patel et al., 2019) In radiology, for example, the importance of radiologists providing the final verification of diagnosis is discussed (Liew, 2018); where it might be particularly important if radiologists would check lower confidence outputs from AI. (Patel et al., 2019) Yet, medical oversight does not need to be limited to validating AI-generated outcomes, as is also exemplified in the radiology context. Mudgal and Das (Mudgal & Das, 2020) describe how radiologists can be involved as medical officers guiding institutions and companies to target the most relevant and topical clinical decision-making tools, act as consultants providing continuous supervision to AI bodies, and supervise AI tools in practice to prevent medical errors. Mudgal and Das emphasize that involving radiologists in multiple roles and maintaining the final verdict is especially crucial for utilizing AI in the context of high-stakes decisions, such as choosing between oncological treatments and initiating palliative care.

4.2.1. Physician-in-the-loop oversight for the effective use of generative AI

Notably, several authors view physician oversight through a human-in-the-loop approach as a desirable method for applying generative AI (GAI). (Haemmerli et al., 2023; Holmes et al., 2025; Livingston et al., 2025; Lu et al., 2024; McCaffrey et al., 2025; Roberts et al., 2024; Woo et al., 2024) Haemmerli et al., for example, argue that ChatGPT can be useful for decision-making in brain glioma adjuvant therapy, but that shortcomings, such as bias or errors in interpreting medical information, mean a HITL approach should be used. (Haemmerli et al., 2023) Lu et al. (Lu et al., 2024) and McCaffrey et al. (McCaffrey et al., 2025) also describe the potential of GAI through human-in-the-loop clinical decision-making in pathology. McCaffrey et al. go into the potential GAI systems have to automate certain tasks in pathology, allowing human expertise to focus on higher-level report review and evaluating generated outputs. Meanwhile, Lu et al. believe generalist AI models can function as consultants to pathologists, providing an initial AI-assisted assessment that can be further contextualized with input from the pathologist. They consider HITL AI assistance especially valuable in situations involving lengthy and complex evaluations, such as cancers of unknown primary, and in low-resource settings where access to experienced pathologists may be restricted. In their article, they describe a specific example of what a HITL process would entail, included here as Appendix 3.

Livingston et al. have also developed a clinician-in-the-loop evaluation framework that enables clinicians to systematically assess Generative AI (GAI) outputs. (Livingston et al., 2025) They assert that such a framework is essential for quantifying the potential risks of GAI in clinical practice, particularly when these types of AI systems are prone to producing hallucinated outputs. Furthermore, they believe that the wide variation in evaluation tools calls for a standard GAI evaluation framework and that metrics often fall short: ‘‘text-comparison metrics alone cannot adequately assess whether an LLM’s response is clinically appropriate, nor can it assess usefulness in a healthcare context.’’(Livingston et al., 2025)

Based on the most pressing clinical concerns in the literature, they identify five dimensions by which clinicians should judge GAI: (1) *helpfulness*, the overall value of the response for clinical practice; (2) *comprehension*, the system’s understanding of the clinical query, from basic text processing to deeper clinical interpretation; (3) *correctness*, the factual accuracy of each line against the provided peer-reviewed literature and clinical resource references; (4) *completeness*, whether the response addresses all clinically relevant aspects of the query; (5) *clinical harmfulness*, potential patient safety risks if the information in the response were applied without clinical judgement and followed through on without safety systems in clinical care. In their framework, clinicians can score these dimensions on a 3- or 5-point scale to quantify the correctness and harm posed by GAI systems in a specific context.

4.3 Desirable ways in which AI is developed and used in HITL approaches

Authors saw a broad range of ways in which AI could be used in healthcare, yet there was a general focus on AI systems that would function as a clinical decision support tool; examples include a CDSS for supporting clinical decision-making during MDT meetings on cancer (Ali et al., 2024), LLMs as decision-support in pathology (McCaffrey et al., 2025) or CDSS to support the early and accurate identification of onward care needs. (Duckworth et al., 2024) Yet, in HITL approaches to AI, the degree to which medical AI can be truly helpful and effectively utilized often depends on whether humans have been sufficiently involved in the deployment of AI. Montomoli et al. (Montomoli et al., 2024), for example, argue that the AI workflow necessitates evaluating the role that human expertise and judgment play in AI’s learning process, interpreting AI outcomes, and determining its usefulness and usability. HITL is thus said to strike a balance between automation and human expertise, where AI systems are guided, communicated, and supervised by human expertise, thereby enhancing both safety and quality. (Phongpreecha et al., 2025; Sezgin, 2023)

To enable human involvement and control, it is frequently mentioned that interpretability or explainability of AI is essential for HITL approaches in medicine. (Fuchs et al., 2024; Montomoli et al., 2024; Plass et al., 2023; Salih et al., 2024) Fuchs et al. state that HITL necessitates AI systems to offer easily understandable insights that align with medical guidelines in radiology and capture relevant information without incurring additional costs. (Fuchs et al., 2024) Liew also argues that, for legal liability to be assigned to a human authority and for radiologists to assume responsibility, there must be effective ways to clearly explain the role that AI has played in the decision-making process, and this explanation should be understandable to the human in the loop.(Liew, 2018) Moreover, Salih et al. (Salih et al., 2024) argue that the risks of using AI are more pronounced in certain medical domains, such as pediatrics, where misinterpreting AI outcomes can have significant consequences; this emphasizes the need for explainability to enable pediatricians to stay in the loop and evaluate outcomes effectively. Plass et al. agree and recommend explainable human-AI interfaces that are targeted at causal understanding and ''allow the domain expert to ask interactive what-if questions. (…) A human-in-the-loop can [thus] (sometimes–not always) bring human experience and conceptual knowledge to AI processes – something that the current best AI algorithms (still) lack.'' (Plass et al., 2023)

Since HITL often assumes a partnership, or a synergistic or symbiotic relationship between humans and AI, developing AI using a HITL approach is, in some cases, viewed as being closely related to the idea of human-AI collaboration.(Fuchs et al., 2024; Lee et al., 2024; Montomoli et al., 2024; Mudgal & Das, 2020; Patel et al., 2019; Phongpreecha et al., 2025; Roberts et al., 2024; Sezgin, 2023; Wadden, 2024) Fuchs et al. (Fuchs et al., 2024), for instance, describe a ''closer collaboration between AI systems and clinicians'' and view HITL as a part of two-way communication between radiologists and AI, where radiologists are kept in the loop by being able to understand AI outcomes, and radiologists, in turn, need to provide information in a machine-readable format. Moreover, Roberts et al. (Roberts et al., 2024) describe how an HITL approach can ensure that the surgeon's nuanced comprehension and personal touch can be maintained when collaborating with an LLM to elevate efficiency when drafting and refining clinical letters. Both examples show that there can be a synergy between HITL and human-AI collaboration.

4.4 Open questions and criticism regarding HITL in health

Although most authors agree that human involvement and oversight are essential for the application of AI in healthcare, several authors criticize HITL for being approached too narrowly. Sezgin, for instance, proposes expanding the HITL approach and involving multidisciplinary teams; this might include clinicians, IT experts, managers, administrators, patient and community advocates.(Sezgin, 2023) Furthermore, Bhatia et al. contend that HITL can be too limited in the context of AI in pediatric neuroradiology; ''AI (…) not only requires a 'human in the loop,' but humans in the center.'' (Bhatia et al., 2024) They argue that a human-in-the-loop approach can serve as a valuable starting point, but the development and implementation of AI should also more broadly focus on health outcomes for patients, the experiences of patients and their families, cost reduction, and the well-being of the healthcare team, with an emphasis on equity, diversity, and inclusivity.

A related question that Salloch and Eriksen (Salloch & Eriksen, 2024) pose and believe is in need of further discussion is what HITL refers to in moral terms. In other words, they state that ethical principles guiding the integration of medical AI in practice necessitate human interpretation or judgment, and that it remains to be determined how such judgments should be made. In their article, they propose the possibility of moral co-reasoning in the patient-physician encounter, where patients and physicians jointly reason to reach a shared understanding of what the clinical process requires. They believe that a partnership model of clinical reasoning is warranted in the context of AI, as patients can play a genuine epistemic role in avoiding dangers such as automation bias, providing feedback on how systems should be designed, and supporting the interpretation of AI principles in practice. Their argument suggests the possibility of also considering involving patients in HITL.

4.4.1 The lack of a (universally adopted) definition

Besides the possibility of expanding how HITL is approached, a recurring concern about the use of the term ‘human-in-the-loop’ in health is its lack of a clear, universally adopted definition, which allows for various interpretations of the concept in the literature. (Muyskens et al., 2025) Salloch and Eriksen (Salloch & Eriksen, 2024), for instance, maintain that ‘‘it remains fundamentally unclear what humans are supposed to be doing in the loop.’’ Malaguti et al. (Malaguti et al., 2025) also note in their review of AI use in Parkinson’s Disease (PD) care that there is no consensus on the definition or use of the term ‘clinician-in-the-loop’:

> While nearly all studies recognize the significant contributions of clinicians, the term ‘‘clinician-in-the-loop’’ lacks a universally accepted definition, and its operationalization differs markedly among studies. In some instances, clinicians are merely interpreters of AI-generated outputs, while in others, they play a pivotal role in data collection, variable selection, and even model development. This variability underscores the need for a more standardized framework to articulate the clinician’s involvement at different stages of the AI lifecycle.

Malaguti et al. argue that the lack of a unified and operationalizable conceptualization results in ambiguity regarding the role clinicians play in AI development and use. (Malaguti et al., 2025) In their article, they therefore propose differentiating between ‘clinician in the pre-processing loop’ and ‘clinician in the modeling (or post-processing) loop’ to clarify how and when clinicians are fundamental to AI development and use. They elaborate that, in the pre-processing loop, clinicians would play an essential role in aligning AI goals with clinical practice by defining clinical questions, selecting relevant variables, and ensuring data quality. In the modeling loop, clinicians collaborate with data scientists and AI specialists during model development, providing insights into clinical contexts and offering real-time feedback. Clinicians can also be involved in the post-processing loop, where they interpret AI outputs and focus on validation, contextualization, and communication of AI-derived insights to patients and other stakeholders.

4.4.2. Is HITL sufficient to maintain human control, and when is control required?

Additionally, some authors question whether HITL is sufficient to ensure actual control over AI. (Van Der Stigchel et al., 2023) Ku et al. (Ku, 2019), for example, agree with Zerilli et al. (2019) and worry that humans can lack the technical ability or have physical limitation to supervise AI, that the human becomes complacent, over-reliant, or unduly diffident when faced with outputs of reliable autonomous systems. Van der Stigchel et al. designed an experiment to test whether HITL would lead to better control and detection of biases by focusing on human-AI collaboration for patient triage during a pandemic crisis; in this experiment, human control was operationalized as being able to detect and correct for bias in the AI advice. (Van Der Stigchel et al., 2023) They conclude that participants were unable to exert oversight, and that it is essential to deliberately design AI to facilitate human control. Moreover, they emphasize that insufficient human control could lead to people being unable to detect biases in AI and thus unable to prevent machine biases from influencing decisions. In line with their findings, Holmes et al. (Holmes et al., 2025) argue that ‘’there is a clear need for further research into developing AI systems that can effectively integrate professionals into the loop.’’

In evaluating when it is appropriate to maintain human control and enable HITL, Muyskens et al. (Muyskens et al., 2025) argue that we should consider whether there are medical situations in which humans should be ‘’kicked out of the loop.’’ They propose that this might be ethically permissible in cases where: (1) the technology is as effective (or better) than a human at a given task in terms of error rates; (2) the risk to patients (or any humans involved) is low in the event of an error and (3) the wellbeing that is gained by the speed, accuracy and cost-efficiency of automation is high. By formulating these criteria, they aim to challenge the notion that HITL is synonymous with good ethical practice and that insisting on keeping humans in the loop can be unwarranted in certain contexts, particularly in instances where it comes with significant opportunity costs. They believe that in the future, it should be assessed ‘’when the training wheels can safely come off’ and certain AI systems meet a threshold to be used automatically.

5 Discussion

Studies in this overview highlight the risks associated with using AI in medical settings and illustrate the widely held belief that HITL is a necessary condition for managing risks when using AI to support medical decision-making. Several authors argue that HITL is especially crucial in high-risk medical situations, such as pediatrics, or when AI produces low-confidence results. At the same time, the results confirm that HITL is an ambiguous term referring to a variety of ways humans can be involved in medical AI development and use, and raise the question whether HITL will ensure that effective oversight is maintained. In part, this can be explained by the fact that HITL, as a term, has only been relatively recently introduced in the medical literature. That is, the term 'HITL' did not originate in the medical field; technology developers have been using it since the 1990s to describe the role of human expertise in, or human oversight of, machine learning (ML) technology and other advanced software in various domains. (Barmore et al., 2005; Fales et al., 2005; Lee & Hsu, 2003; Looney & Tacker, 1990) One of the earliest examples in the medical AI literature seems to be the article by Kieseberg et al. (Kieseberg et al., 2015) which was published in 2015; the overview shows that, since then, the term has been increasingly used to describe the role of the 'human' in the development and integration of advanced AI systems in medical decision-making. Yet, as the term is still relatively new in the medical context, as is the integration of AI in medical practice, it is not surprising that questions remain about how HITL should be applied to medical AI and which medical tasks can be delegated without undesirably losing control over clinical processes.

A close examination of the application of HITL in the literature reveals that numerous articles merely mention the importance of HITL, without specifying the particular requirements for human oversight over AI in medical decision-making. Nevertheless, some articles included in this overview have attempted to clarify the meaning and practical implications of HITL. Notable examples include Livingston et al. (Livingston et al., 2025) , who have developed a concrete evaluation framework for clinicians in the loop to assess GAI outputs, and Mudgal and Das (Mudgal & Das, 2020), who describe specific roles radiologists in the loop can play during AI development and use. Interestingly, these proposals seem to go beyond HITL approaches that involve ''simple human presence'' or authorizing

AI outputs (Santoni de Sio & Van den Hoven, 2018), and also include deciding how well AI systems fit into a medical setting and having control over assessing in which cases the benefits of AI outweigh the risks. Still, despite these more detailed proposals, the overview raises questions about how to move forward with the term HITL and how humans are adequately equipped to be a HITL in clinical settings. In the upcoming sections, we will address two questions: (1) What should HITL entail in the medical context? and (2) What skills and expertise are required for humans in the loop?

5.1 What should HITL entail in the medical context?

When referring to HITL, authors may mean different things: some describe humans who simply verify AI outputs in clinical contexts, others refer to their participation in various stages of AI development and implementation, while some refer to maintaining control at all times. This variation highlights the lack of consensus on what HITL should entail in the medical field, a point also noted by several authors in this overview. However, a problem that challenges reaching a consensus on HITL is that it is not only difficult to determine the appropriate level of oversight, but also that there is a gap between how HITL is interpreted in the technical literature and discussions on AI governance and ethics. In the technical literature on medical AI development, HITL is widely regarded as a methodology to ensure system performance and accurate outcomes, where human experts primarily provide feedback and maintain some level of control over the AI system's learning process. (Seeuws et al., 2024; Shu et al., 2024; Vásquez-Venegas et al., 2024) In the AI ethics and governance literature, HITL refers to the broader concept that humans, rather than autonomous systems, should remain accountable and in control of decision-making processes in medical settings. (HLEG, 2019; Salloch & Eriksen, 2024) The discrepancy between the two discussions raises an important question about the use of the term HITL: should HITL mainly refer to the value of human expertise in training and labeling AI outcomes, or should it also include human oversight that preserves human autonomy and control over decision-making processes? It is crucial to answer this question because the current discrepancy might indicate a mismatch between what ethical guidelines prescribe by means of HITL and what is technically developed. That is, developers might adopt a HITL methodology for annotation or segmentation purposes (excluded in this overview, n = 143), believing they have adequately applied a HITL approach, while ethical guidelines call for more thorough human involvement and broader control over medical AI.

In response to the discrepancy between technical and ethical interpretations of HITL, Zheng et al. (Zheng et al., 2024) caution against taking the technical interpretation of the HITL term at face value in the medical field:

> The term HITL originated in the machine learning (ML) community, where it assumes that the role of humans is to aid ML models' autonomous decisions when the ML models cannot fully handle edge cases or unexpected scenarios. In this role, instead of actively engaging in clinical decisions, humans are relegated to merely checking or

> validating AI models' decisions despite bearing the responsibility and accountability for those decisions. Therefore, the assumption of HITL downplays the role of humans as primary epistemic subjects. Such assumptions, prevalent in the technical sector, are inappropriate and unjust in medical settings. (…) The ethical debate [on the other hand] centers on balancing human capability with AI technologies that may erode autonomy, skills, and independence of humans.

While Zheng et al. rightly warn against reducing the role of domain experts to epistemic subjects who do not have a decisive role in clinical decisions, the technical sector may tend to use other terms to describe the essential role humans should play in overseeing AI use; terms such as Human-Centered AI (HAI) are, for example, used to ''[approach] AI from a human perspective by considering human conditions and contexts.'' (Mosqueira-Rey et al., 2023) Moreover, technical developers such as Kieseberg et al. (Kieseberg et al., 2016) are also aware that HITL, as a methodology, still has ethical consequences for physicians, especially in terms of maintaining responsibility for decision-making. It may therefore primarily be a matter of determining which terminology best matches specific human roles in medical AI development and implementation.

5.2 What skills and expertise are required for humans in the loop?

A primary goal of this overview was to identify which humans should participate in HITL and what roles they should assume. Most articles included in the overview focused on medical professionals as the human in the loop, with radiologists and pathologists being particularly frequently mentioned, which can be explained by the advanced state of AI development in image-based medicine. (Drogt et al., 2024) Many authors have high expectations and assumptions about what human oversight by medical professionals could offer. It was stated that medical professionals could improve AI's safety and quality, ensure that AI-driven insights are clinically meaningful, ethically sound, and contextually appropriate, and identify biases and shortcomings of AI, such as hallucinations. In reality, it might not be possible for medical practitioners to meet all these expectations. As Holzinger et al. (Holzinger et al., 2024) argue: ''the demands of interpreting complex AI outputs and maintaining attention during prolonged monitoring can result in oversight fatigue or errors.'' Additionally, a study by Rosbach et al. (Rosbach et al., 2025) found that pathologists may fail to recognize AI-induced errors and suffer from automation bias, particularly under time pressure. Similarly, physicians might have a confirmation bias – agreeing with a wrong AI system – or override a correct AI diagnosis. (Rosenbacke et al., 2024) Human oversight is thus not a safeguard against all errors and it should be considered whether medical professionals may also need to develop specific competencies to work responsibly with and meaningfully use AI outcomes. (Sand et al., 2022; Vos et al., 2025)

The skills and expertise required for a HITL approach will also likely vary depending on the type of AI system and the associated risks of its use. As several authors in this overview mention, GenAI poses specific challenges: requiring a thorough assessment of errors and clinical significance, as it tends to confabulate and is, as a system, more challenging to validate. As a result, assessing its outputs may demand more specialized medical expertise. At the same time, these validation challenges may make it even more relevant to include a HITL. Whether there are medical settings in which humans can be removed from the loop (Muyskens et al., 2025) or be ‘‘near’’ it instead of being fully in the loop (Jackson & Pinto, 2024), still needs to be determined and may be considered when medical expertise is less valuable or not needed for differentiating AI errors. As Jackson and Pinto state, ‘‘there are currently no guidelines or distinctions that clarify how far a human can drift away from the loop and for how long.’’ (Jackson & Pinto, 2024) Nevertheless, it should be considered that there are also calls for more thorough human involvement in HITL approaches, exemplified by Scheek et al. (Scheek et al., 2021) who argue that ‘‘over time, organizations should not limit their interactions with radiologists to only occasional consultants; rather they should actively and systematically bring radiologists on board, with responsibilities and authorities to become part of the development team.’’ This highlights the importance of evaluating the extent to which the expertise of medical professionals should be included in determining the clinical relevance of AI.

6 Conclusion and limitations

6.1 Conclusion

HITL is a term frequently used in the context of medical AI to refer to a type of human oversight or control. Despite an ongoing academic debate on this topic, this study presents the first comprehensive overview of how HITL is used in academic literature on medical AI. Our findings illuminate several key features of the academic debate thus far, namely that the concept of human-in-the-loop (HITL) in healthcare is underpinned by a range of diverging assumptions about who the 'human' is or should be, what roles they have, and what they will be able to accomplish. This shows that it is far from evident that HITL is sufficient for control, let alone for ethical medical AI. Without clear task allocation, adequate training, context-sensitive system design, and supportive organizational structures, the human role remains fragile—and may even become merely symbolic.

Furthermore, the differences in HITL approaches identified in our overview reveal a discrepancy and possible mismatch between technical and ethical interpretations of HITL. In technical literature, HITL primarily refers to the importance of human expertise for training AI models, whereas ethical guidelines suggest it should also include broad human oversight that preserves human autonomy and control over decision-making processes. Since such a discrepancy can significantly affect the meaning of human requirements and control in HITL approaches, it is crucial to clarify the term, also in relation to other related concepts such as Human-Centered AI (HAI). By drawing attention to these crucial points in the debate on HITL in medicine, this overview facilitates advancing the discussion and identifies key contributions that can inform future decisions on how the HITL concept is used.

6.2 Limitations

We have provided a comprehensive overview of the literature on HITL in medical AI. The articles presented were included after a thorough screening of the academic literature on the topic by two independent reviewers, based on a search strategy that was guided by experienced librarians. Nonetheless, this overview has several limitations. First, reviewing the literature in such a way always involves reporting bias; a different group of researchers could have selected or grouped the included ethical aspects differently. Second, we were unable to systematically perform a quality assessment of the included literature, as there is no established screening instrument to evaluate the quality of normative papers. Finally, we note that it was beyond the scope of this overview to assess the scientific validity of HITL applications and different forms of HITL discussed in the included articles.

References

Aldosari, B., Aldosari, H., & Alanazi, A. (2025). Challenges of Artificial Intelligence in Medicine. *Envisioning the Future of Health Informatics and Digital Health*, 16-20.

Ali, S. R., Dobbs, T. D., Tarafdar, A., Strafford, H., Fonferko-Shadrach, B., Lacey, A. S., Pickrell, W. O., Hutchings, H. A., & Whitaker, I. S. (2024). Natural language processing to automate a web-based model of care and modernize skin cancer multidisciplinary team meetings. *British Journal of Surgery*, *111*(1), znad347.

Barmore, B., Abbott, T., & Capron, W. (2005). Evaluation of airborne precision spacing in a human-in-the-loop experiment. AIAA 5th ATIO and16th lighter-than-air sys tech. and balloon systems conferences,

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, *3*(1), 118. https://doi.org/10.1038/s41746-020-00324-0

Bhatia, A., Khalvati, F., & Ertl-Wagner, B. B. (2024). Artificial Intelligence in the Future Landscape of Pediatric Neuroradiology: Opportunities and Challenges. *American Journal of Neuroradiology*, *45*(5), 549-553.

Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "Human-in-the-Loop" system. *Systems*, *11*(11), 548.

Commission, E. (2024). *AI Act*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Drogt, J., Milota, M., Veldhuis, W., Vos, S., & Jongsma, K. (2024). The Promise of AI for Image-Driven Medicine: Qualitative Interview Study of Radiologists' and Pathologists' Perspectives. *JMIR Human Factors*, *11*, e52514.

Duckworth, C., Burns, D., Lamas Fernandez, C., Wright, M., Leyland, R., Stammers, M., George, M., & Boniface, M. (2024). Predicting onward care needs at admission to reduce discharge delay using machine learning. *medRxiv*, 2024.2008. 2007.24311596.

Fales, R., Spencer, E., Chipperfield, K., Wagner, F., & Kelkar, A. (2005). Modeling and control of a wheel loader with a human-in-the-loop assessment using virtual reality.

Fritz, R., Wuestney, K., Dermody, G., & Cook, D. J. (2022). Nurse-in-the-loop smart home detection of health events associated with diagnosed chronic conditions: A case-event series. *International Journal of Nursing Studies Advances*, *4*, 100081.

Fuchs, M., Gonzalez, C., Frisch, Y., Hahn, P., Matthies, P., Gruening, M., Dos Santos, D. P., Dratsch, T., Kim, M., & Nensa, F. (2024). Closing the loop for AI-ready radiology. RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren,

Guidance, W. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization*.

Gunes, Y. C., Cesur, T., Camur, E., Cifci, B. E., Kaya, T., Colakoglu, M. N., Koc, U., & Okten, R. S. (2025). Textual Proficiency and Visual Deficiency: A Comparative Study of Large Language Models and Radiologists in MRI Artifact Detection and Correction. *Academic Radiology*.

Haemmerli, J., Sveikata, L., Nouri, A., May, A., Egervari, K., Freyschlag, C., Lobrinus, J. A., Migliorini, D., Momjian, S., & Sanda, N. (2023). ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? *BMJ Health & Care Informatics*, *30*(1), e100775.

Haselager, P., Schraffenberger, H., Thill, S., Fischer, S., Lanillos, P., van de Groes, S., & van Hooff, M. (2024). Reflection Machines: Supporting Effective Human Oversight Over Medical Decision Support Systems. *Cambridge Quarterly of Healthcare Ethics*, *33*(3), 380-389. https://doi.org/10.1017/S0963180122000718

HLEG, A. (2019). Ethics guidelines for trustworthy artificial intelligence. *High-Level Expert Group on Artificial Intelligence*, *8*.

Holmes, G., Tang, B., Gupta, S., Venkatesh, S., Christensen, H., & Whitton, A. (2025). Applications of Large Language Models in the Field of Suicide Prevention: Scoping Review. *Journal of Medical Internet Research*, *27*, e63126.

Holzinger, A., Zatloukal, K., & Müller, H. (2024). Is Human Oversight to AI Systems still possible? In: Elsevier.

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS one*, *13*(10), e0204937.

Jackson, J. M., & Pinto, M. D. (2024). Human Near the Loop: Implications for Artificial Intelligence in Healthcare. In (Vol. 33, pp. 135-137): SAGE Publications Sage CA: Los Angeles, CA.

Jotterand, F., & Bosco, C. (2020). Keeping the “human in the loop” in the age of artificial intelligence: accompanying commentary for “correcting the brain?” by Rainey and Erden. *Science and Engineering Ethics*, *26*(5), 2455-2460.

Kieseberg, P., Malle, B., Frühwirt, P., Weippl, E., & Holzinger, A. (2016). A tamper-proof audit and control system for the doctor in the loop. *Brain Informatics*, *3*, 269-279.

Kieseberg, P., Schantl, J., Frühwirt, P., Weippl, E., & Holzinger, A. (2015). Witnesses for the doctor in the loop. Brain Informatics and Health: 8th International Conference, BIH 2015, London, UK, August 30-September 2, 2015. Proceedings 8,

Ku, C.-Y. (2019). When AIs Say Yes and I Say No: On the Tension between AI's Decision and Human's Decision from the Epistemological Perspectives. *Information Society/Információs Társadalom (InfTars)*, *19*(4).

Lee, H. S., Wright, C., Ferranto, J., Buttimer, J., Palmer, C. E., Welchman, A., Mazor, K. M., Fisher, K. A., Smelson, D., & O’Connor, L. (2025). Artificial intelligence conversational agents in mental health: Patients see potential, but prefer humans in the loop. *Frontiers in Psychiatry*, *15*, 1505024.

Lee, J.-S., & Hsu, P.-L. (2003). Remote supervisory control of the human-in-the-loop system by using Petri nets and Java. *IEEE Transactions on industrial electronics*, *50*(3), 431-439.

Lee, L., Salami, R. K., Martin, H., Shantharam, L., Thomas, K., Ashworth, E., Allan, E., Yung, K.-W., Pauling, C., & Leyden, D. (2024). “How I would like AI used for my imaging”: children and young persons’ perspectives. *European Radiology*, *34*(12), 7751-7764.

Levac, D., Colquhoun, H., & O'brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation science*, *5*, 1-9.

Lewis, C., Groarke, J., Graham-Wisener, L., & James, J. (2025). Public Awareness of and Attitudes Toward the Use of AI in Pathology Research and Practice: Mixed Methods Study. *Journal of Medical Internet Research*, *27*, e59591.

Liew, C. (2018). The future of radiology augmented with artificial intelligence: a strategy for success. *European Journal of Radiology*, *102*, 152-156.

Livingston, L., Featherstone-Uwague, A., Barry, A., Barretto, K., Morey, T., Herrmannova, D., & Avula, V. (2025). Reproducible Generative AI Evaluation for Healthcare: A Clinician-in-the-Loop Approach. *medRxiv*, 2025.2003. 2004.25323131.

Looney, C. G., & Tacker, E. C. (1990). Human-in-the-loop control with majority vote neural networks. 1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings,

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., & Patel, A. (2024). A multimodal generative AI copilot for human pathology. *Nature*, *634*(8033), 466-473.

Malaguti, M. C., Gios, L., & Jurman, G. (2025). The third wheel or the game changer? How AI could team up with neurologists in Parkinson's care. *Parkinsonism & Related Disorders*, *134*, 107797.

McCaffrey, P., Jackups, R., Seheult, J., Zaydman, M. A., Balis, U., Thaker, H. M., Rashidi, H., & Gullapalli, R. R. (2025). Evaluating Use of Generative Artificial intelligence in clinical pathology practice: opportunities and the way forward. *Archives of Pathology & Laboratory Medicine*, *149*(2), 130-141.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, *339*.

Montomoli, J., Bitondo, M. M., Cascella, M., Rezoagli, E., Romeo, L., Bellini, V., Semeraro, F., Gamberini, E., Frontoni, E., & Agnoletti, V. (2024). Algor-ethics: charting the ethical path for AI in critical care. *Journal of clinical monitoring and computing*, *38*(4), 931-939.

Morandín-Ahuerma, F. (2023). Montreal declaration for responsible AI: 10 principles and 59 recommendations.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005-3054.

Mudgal, K. S., & Das, N. (2020). The ethical adoption of artificial intelligence in radiology. *BJR| Open*, *2*(1), 20190020.

Muehlematter, U. J., Daniore, P., & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *The Lancet Digital Health*, *3*(3), e195-e203.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18*, 1-7.

Muyskens, K., Ma, Y., Menikoff, J., Hallinan, J., & Savulescu, J. (2025). When can we kick (some) humans “out of the loop”? An examination of the use of AI in medical imaging for lumbar spinal stenosis. *Asian Bioethics Review*, *17*(1), 207-223.

Ortega, R. P. (2025). AI models miss disease in Black and female patients https://www.science.org/content/article/ai-models-miss-disease-black-female-patients#:~:text=A%20new%20study%2C%20published%20today,including%20women%20and%20Black%20people.

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., & Halabi, S. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine*, *2*(1), 111.

Phongpreecha, T., Ghanem, M., Reiss, J. D., Oskotsky, T., Mataraso, S. J., De Francesco, D., Reincke, S. M., Espinosa, C., Chung, P., & Ng, T. (2025). AI-guided precision parenteral nutrition for neonatal intensive care units. *Nature medicine*, 1-13.

Plass, M., Kargl, M., Kiehl, T. R., Regitnig, P., Geißler, C., Evans, T., Zerbe, N., Carvalho, R., Holzinger, A., & Müller, H. (2023). Explainability and causability in digital pathology. *The Journal of Pathology: Clinical Research*, *9*(4), 251-260.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature medicine*, *28*(1), 31-38.

Roberts, R. H., Ali, S. R., Dobbs, T. D., & Whitaker, I. S. (2024). Can large language models generate outpatient clinic letters at first consultation that incorporate complication profiles from UK and USA aesthetic plastic surgery associations? Aesthetic Surgery Journal Open Forum,

Robinson, A., Flom, M., Forman-Hoffman, V. L., Histon, T., Levy, M., Darcy, A., Ajayi, T., Mohr, D. C., Wicks, P., & Greene, C. (2024). Equity in Digital Mental Health Interventions in the United States: Where to Next? *Journal of Medical Internet Research*, *26*, e59939.

Rosbach, E., Ganz, J., Ammeling, J., Riener, A., & Aubreville, M. (2025). Automation Bias in AI-assisted Medical Decision-making under Time Pressure in Computational Pathology. BVM Workshop,

Rosenbacke, R., Melhus, Å., & Stuckler, D. (2024). False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. *Nature Communications*, *15*(1), 6896.

Salih, A. M., Menegaz, G., Pillay, T., & Boyle, E. M. (2024). Explainable Artificial Intelligence in Paediatric: Challenges for the Future. *Health Science Reports*, *7*(12), e70271.

Salloch, S., & Eriksen, A. (2024). What Are Humans Doing in the Loop? Co-Reasoning and Practical Judgment When Using Machine Learning-Driven Decision Aids. *The American Journal of Bioethics*, 1-12.

Sand, M., Durán, J. M., & Jongsma, K. R. (2022). Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics*, *36*(2), 162-169. https://doi.org/https://doi.org/10.1111/bioe.12887

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 323836.

Scheek, D., Rezazade Mehrizi, M. H., & Ranschaert, E. (2021). Radiologists in the loop: the roles of radiologists in the development of AI applications. *European Radiology*, *31*(10), 7960-7968. https://doi.org/10.1007/s00330-021-07879-w

Seeuws, N., De Vos, M., & Bertrand, A. (2024). A Human-in-the-Loop Method for Annotation of Events in Biomedical Signals. *IEEE Journal of Biomedical and Health Informatics*.

Sezgin, E. (2023). Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *Digital health*, *9*, 20552076231186520.

Shu, L., He, Q., Yan, B., Wu, D., Wang, M., Wang, C., & Zhang, L. (2024). Human-in-the-loop: Human involvement in enhancing medical inquiry performance in large language models. *Allergy*, *79*(5).

Van Der Stigchel, B., Van den Bosch, K., Van Diggelen, J., & Haselager, P. (2023). Intelligent decision support in medical triage: are people robust to biased advice? *Journal of Public Health*, *45*(3), 689-696.

Vásquez-Venegas, C., Sotomayor, C. G., Ramos, B., Castañeda, V., Pereira, G., Cabrera-Vives, G., & Härtel, S. (2024). Human-in-the-Loop—A Deep Learning Strategy in Combination with a Patient-Specific Gaussian Mixture Model Leads to the Fast Characterization of Volumetric Ground-Glass Opacity and Consolidation in the Computed Tomography Scans of COVID-19 Patients. *Journal of clinical medicine*, *13*(17), 5231.

Vos, S., Hebeda, K., Milota, M., Sand, M., Drogt, J., Grünberg, K., & Jongsma, K. (2025). Making Pathologists Ready for the New Artificial Intelligence Era: Changes in Required Competencies. *Modern Pathology*, *38*(2), 100657.

Wadden, J. J. (2024). Naming and Diffusing the Understanding Objection in Healthcare Artificial Intelligence. *Canadian Journal of Bioethics*, *7*(4), 57-63.

Woo, K.-m. C., Simon, G. W., Akindutire, O., Aphinyanaphongs, Y., Austrian, J. S., Kim, J. G., Genes, N., Goldenring, J. A., Major, V. J., & Pariente, C. S. (2024). Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings. *Journal of the American Medical Informatics Association*, *31*(9), 1983-1993.

Yang, L., Ene, I. C., Arabi Belaghi, R., Koff, D., Stein, N., & Santaguida, P. (2022). Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. *European Radiology*, *32*(3), 1477-1495.

Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E. J., Sahani, D., & Patel, S. (2025). Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, *11*(13), eadq0305.

Zhang, G., Gao, Z., Duan, C., Liu, J., Lizhu, Y., Liu, Y., Chen, Q., Wang, L., Fei, K., & Wang, T. (2025). A Multimodal Vision-text AI Copilot for Brain Disease Diagnosis and Medical Imaging. *medRxiv*, 2025.2001. 2009.25320293.

Zheng, E. L., Jin, W., Hamarneh, G., & Lee, S. S.-J. (2024). From human-in-the-loop to human-in-power. *The American Journal of Bioethics*, *24*(9), 84-86.

Appendix 1 – Search string

("Human-in-the-Loop" OR “HITL” OR “in-the-loop” OR “in the loop”) AND (“AI” OR “Artificial Intelligence” OR “machine learning” OR “ML” OR “deep learning” OR ‘‘DL’’ OR “decision support” OR “CDSS” OR “DSS” OR “autonomous system” OR “automated system”) AND (“healthcare" OR "health care" OR ‘‘medicine’’ OR "medical care" OR ‘‘medical practice’’ OR "clinical care" OR ‘‘clinical practice’’ OR "health services" OR "health system" OR "medical system" OR "patient care" OR “cardiology” OR “pathology” OR “radiology” OR “ophthalmology” OR “dermatology” OR “general care” OR “general practice” OR “intensive care” OR ‘‘gastroenterology’’ OR ‘‘obstetrics’’ OR ‘‘gynaecology’’ OR ‘‘immunology’’ OR ‘‘anesthesiology’’ OR ‘‘oncology’’ OR ‘‘surgery’’ OR ‘‘neurology’’ OR ‘‘critical care’’ OR ‘‘nephrology’’ OR ‘‘pediatrics’’ OR ‘‘hematology’’ OR ‘‘neonatology’’)

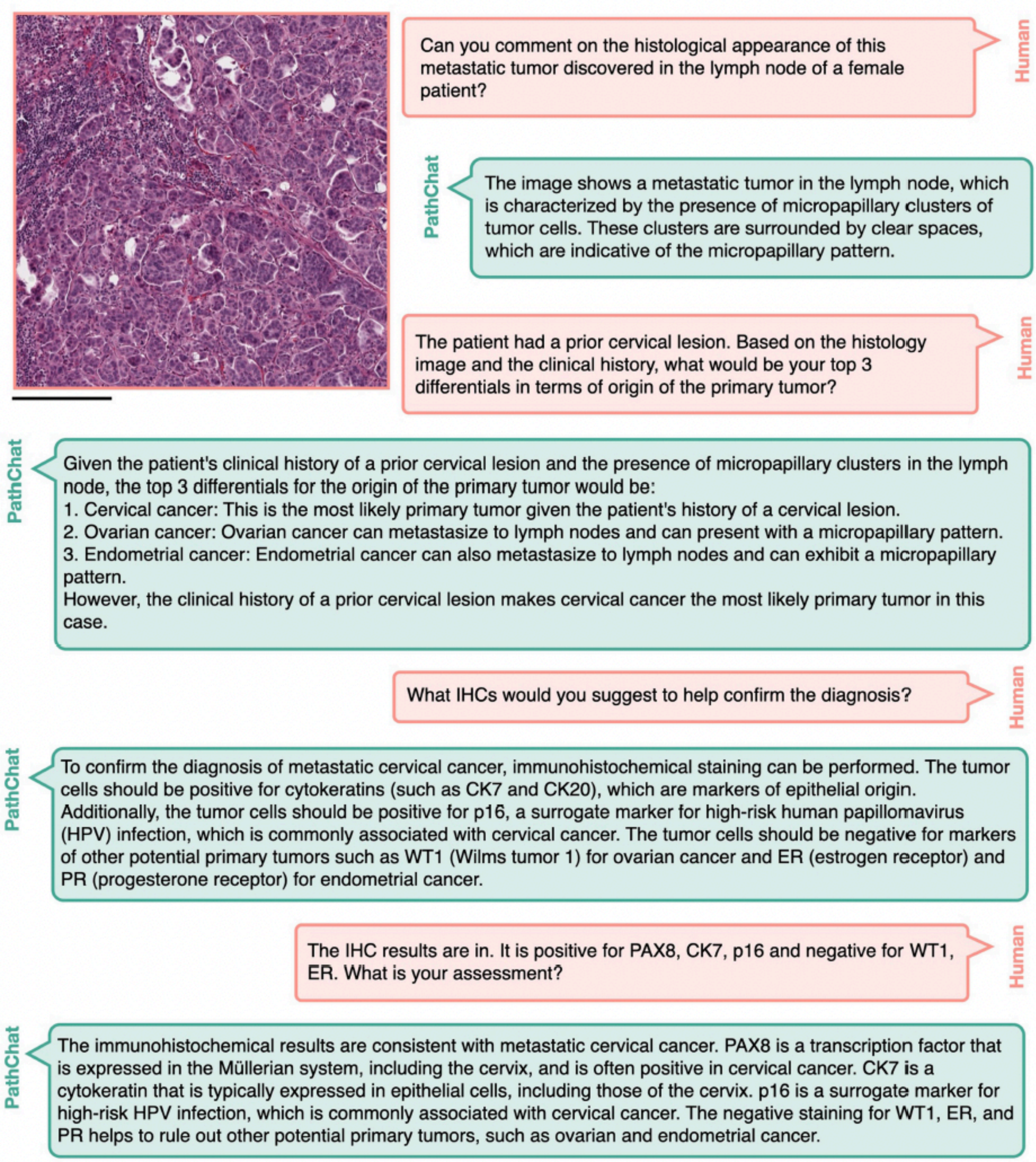
('Human-in-the-Loop' OR 'HITL' OR 'in-the-loop' OR 'in the loop') AND ('AI' OR 'Artificial Intelligence' OR 'machine learning' OR 'ML' OR 'deep learning' OR 'DL' OR 'decision support' OR 'CDSS' OR 'DSS' OR 'autonomous system' OR 'automated system') AND ('healthcare' OR 'health care' OR 'medicine' OR 'medical care' OR 'medical practice' OR 'clinical care' OR 'clinical practice' OR 'health services' OR 'health system' OR 'medical system' OR 'patient care' OR 'cardiology' OR 'pathology' OR 'radiology' OR 'ophthalmology' OR 'dermatology' OR 'general care' OR 'general practice' OR 'intensive care' OR 'gastroenterology' OR 'obstetrics' OR 'gynaecology' OR 'immunology' OR 'anesthesiology' OR 'oncology' OR 'surgery' OR 'neurology' OR 'critical care' OR 'nephrology' OR 'pediatrics' OR 'hematology' OR 'neonatology')

(healthcare OR health care OR medicine OR medical care OR medical practice OR clinical care OR clinical practice OR health services OR health system OR medical system OR patient care OR cardiology OR pathology OR radiology OR ophthalmology OR dermatology OR general care OR general practice OR intensive care OR gastroenterology OR obstetrics OR gynaecology OR immunology OR anesthesiology OR oncology OR surgery OR neurology OR critical care OR nephrology OR pediatrics OR hematology OR neonatology) AND (AI OR Artificial Intelligence OR machine learning OR ML OR deep learning OR DL OR decision support OR CDSS OR DSS OR autonomous system OR automated system) AND (Human-in-the-Loop OR HITL OR in-the-loop OR in the loop)

Appendix 2 – Data charting table

Data Charting Table	
Article	
Title, publication year, authors	
Medical context	
Type of AI	
Variation of HITL term used	
The human in HITL	
Who is the 'human'	
What role does the human have	
Roles and tasks of the AI	
Roles and tasks envisioned for AI in HITL	
General assumptions	
Assumptions on how HITL will be put to practice	
Relation of HITL to other AI ethics concepts (such as human-AI collaboration, HOTL)	
Ethical aims and considerations	
Ethical aims and considerations mentioned in relation to HITL	

Appendix 3 – Examples of Lu et al. illustrating human-in-the-loop differential diagnosis with GAI (PathChat) in a case of cancer of unknown primary


Human
Can you comment on the histological appearance of this metastatic tumor discovered in the lymph node of a female patient?
PathChat
The image shows a metastatic tumor in the lymph node, which is characterized by the presence of micropapillary clusters of tumor cells. These clusters are surrounded by clear spaces, which are indicative of the micropapillary pattern.
Human
The patient had a prior cervical lesion. Based on the histology image and the clinical history, what would be your top 3 differentials in terms of origin of the primary tumor?
PathChat
Given the patient's clinical history of a prior cervical lesion and the presence of micropapillary clusters in the lymph node, the top 3 differentials for the origin of the primary tumor would be:
1. Cervical cancer: This is the most likely primary tumor given the patient's history of a cervical lesion.
2. Ovarian cancer: Ovarian cancer can metastasize to lymph nodes and can present with a micropapillary pattern.
3. Endometrial cancer: Endometrial cancer can also metastasize to lymph nodes and can exhibit a micropapillary pattern.
However, the clinical history of a prior cervical lesion makes cervical cancer the most likely primary tumor in this case.
Human
What IHCs would you suggest to help confirm the diagnosis?
PathChat
To confirm the diagnosis of metastatic cervical cancer, immunohistochemical staining can be performed. The tumor cells should be positive for cytokeratins (such as CK7 and CK20), which are markers of epithelial origin. Additionally, the tumor cells should be positive for p16, a surrogate marker for high-risk human papillomavirus (HPV) infection, which is commonly associated with cervical cancer. The tumor cells should be negative for markers of other potential primary tumors such as WT1 (Wilms tumor 1) for ovarian cancer and ER (estrogen receptor) and PR (progesterone receptor) for endometrial cancer.
Human
The IHC results are in. It is positive for PAX8, CK7, p16 and negative for WT1, ER. What is your assessment?
PathChat
The immunohistochemical results are consistent with metastatic cervical cancer. PAX8 is a transcription factor that is expressed in the Müllerian system, including the cervix, and is often positive in cervical cancer. CK7 is a cytokeratin that is typically expressed in epithelial cells, including those of the cervix. p16 is a surrogate marker for high-risk HPV infection, which is commonly associated with cervical cancer. The negative staining for WT1, ER, and PR helps to rule out other potential primary tumors, such as ovarian and endometrial cancer.