AI for IMPACTS: A Comprehensive Framework for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

Christine Jacob, Noé Brasier, Emanuele Laurenzi, Sabina Heuss, Stavroula-Georgia Mougiakakou, Arzu Cöltekin, Marc K Peter

Submitted to: Journal of Medical Internet Research
on: October 14, 2024

Table of Contents

AI for IMPACTS: A Comprehensive Framework for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

Christine Jacob[1] PhD; Noé Brasier[2] MD; Emanuele Laurenzi[1] PhD; Sabina Heuss[1*] PhD, Prof Dr; Stavroula-Georgia Mougiakakou[3, 4*] PhD, Prof Dr; Arzu Cöltekin[5*] PhD, Prof Dr; Marc K Peter[1*] Prof Dr

[1]FHNW, University of Applied Sciences and Arts Northwestern Switzerland Olten CH
[2]Institute of Translational Medicine, Department of Health Science and Technology, ETH Zurich Zurich CH
[3]University of Nicosia Nicosia CY
[4]ARTORG Center for Biomedical Engineering Research, University of Bern Bern CH
[5]FHNW, University of Applied Sciences and Arts Northwestern Switzerland Windisch CH
[*]these authors contributed equally

Corresponding Author:
Christine Jacob PhD
FHNW, University of Applied Sciences and Arts Northwestern Switzerland
Riggenbachstrasse 16
Olten
CH

Abstract

Background: Artificial Intelligence (AI) has the potential to revolutionize healthcare by enhancing both clinical outcomes and operational efficiency. However, its clinical adoption has been slower than anticipated, largely due to the absence of comprehensive evaluation frameworks. Existing frameworks remain insufficient and tend to emphasize technical metrics like accuracy and validation, while overlooking critical real-world factors such as clinical impact, integration, and economic sustainability. This narrow focus prevents AI tools from being effectively implemented, limiting their broader impact and long-term viability in clinical practice.

Objective: This study aimed to create a comprehensive framework for assessing AI in healthcare, extending beyond technical metrics to incorporate social and organizational dimensions. The framework was developed by systematically reviewing, analyzing, and synthesizing the evaluation criteria necessary for successful implementation, focusing on the long-term real-world impact of AI in clinical practice.

Methods: A comprehensive search was performed in July 2024 across PubMed, Cochrane, Scopus, and IEEE Xplore databases to identify relevant studies published in English between January 2019 and mid-July 2024, yielding 3528 results, of which 44 studies met the inclusion criteria. The systematic review followed PRISMA guidelines and the Cochrane Handbook for Systematic Reviews to ensure a systematic approach. Data were analyzed using NVivo (QSR International) through thematic analysis and narrative synthesis to identify key emergent themes in the studies.

Results: By synthesizing the included studies, we developed a framework that goes beyond the traditional focus on technical metrics or study-level methodologies. It integrates clinical context and real-world implementation factors, offering a more comprehensive approach to evaluating AI tools. With our focus on assessing the long-term real-world impact of AI technologies in healthcare, we named the framework AI for IMPACTS. The criteria are organized into seven key clusters, each corresponding to a letter in the acronym: (I) integration, interoperability and workflow (M) monitoring, governance, and accountability (P) performance and quality metrics (A) acceptability, trust, and training (C) cost and economic evaluation (T) technological safety and transparency (S) scalability and impact. These are further broken down into 32 specific sub-criteria.

Conclusions: The AI for IMPACTS framework offers a holistic approach to evaluating the long-term real-world impact of AI tools in the heterogeneous and challenging healthcare context, but further validation through expert consensus and testing of the framework in real-world healthcare settings would strengthen the findings. It is important to emphasize that multidisciplinary expertise is essential for thorough assessment, yet many assessors lack the necessary training. Additionally, traditional evaluation methods struggle to keep pace with AI's rapid development. To ensure successful AI integration, flexible, fast-tracked assessment processes and proper assessor training are needed that maintain rigorous standards while adapting to AI's dynamic evolution. Clinical Trial: NA

(JMIR Preprints 14/10/2024:67485)
DOI: https://doi.org/10.2196/preprints.67485

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

AI for IMPACTS: A Comprehensive Framework for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

Christine Jacob[1], PhD; Noé Brasier[2], MD; Emanuele Laurenzi[1], PhD; Sabina Heuss[1]*, Prof, PhD; Stavroula-Georgia Mougiakakou[3,4]*, Prof, PhD; Arzu Cöltekin[1]*, Prof, PhD; Marc K Peter[1]*, Prof, DBA

[1] FHNW, University of Applied Sciences and Arts Northwestern Switzerland
[2] Institute of Translational Medicine, Department of Health Science and Technology, ETH Zurich, Zurich, Switzerland
[3] ARTORG Center for Biomedical Engineering Research, University of Bern, Switzerland
[4] University of Nicosia, Cyprus

** Equally contributing last authors*

Abstract

Background: Artificial Intelligence (AI) has the potential to revolutionize healthcare by enhancing both clinical outcomes and operational efficiency. However, its clinical adoption has been slower than anticipated, largely due to the absence of comprehensive evaluation frameworks. Existing frameworks remain insufficient and tend to emphasize technical metrics like accuracy and validation, while overlooking critical real-world factors such as clinical impact, integration, and economic sustainability. This narrow focus prevents AI tools from being effectively implemented, limiting their broader impact and long-term viability in clinical practice.

Objective: This study aimed to create a comprehensive framework for assessing AI in healthcare, extending beyond technical metrics to incorporate social and organizational dimensions. The framework was developed by systematically reviewing, analyzing, and synthesizing the evaluation criteria necessary for successful implementation, focusing on the long-term real-world impact of AI in clinical practice.

Methods: A comprehensive search was performed in July 2024 across PubMed, Cochrane, Scopus, and IEEE Xplore databases to identify relevant studies published in English between January 2019 and mid-July 2024, yielding 3528 results, of which 44 studies met the inclusion criteria. The systematic review followed PRISMA guidelines and the Cochrane Handbook for Systematic Reviews to ensure a systematic approach. Data were analyzed using NVivo (QSR International) through thematic analysis and narrative synthesis to identify key emergent themes in the studies.

Results: By synthesizing the included studies, we developed a framework that goes beyond the traditional focus on technical metrics or study-level methodologies. It integrates clinical context and real-world implementation factors, offering a more comprehensive approach to evaluating AI tools. With our focus on assessing the long-term real-world impact of AI technologies in healthcare, we named the framework AI for IMPACTS. The criteria are organized into seven key clusters, each corresponding to a letter in the acronym: (I) integration, interoperability and workflow (M) monitoring, governance, and accountability (P) performance and quality metrics (A) acceptability, trust, and training (C) cost and economic evaluation (T) technological safety and transparency (S) scalability and impact. These are further broken down into 32 specific sub-criteria.

Conclusions: The AI for IMPACTS framework offers a holistic approach to evaluating the long-term real-world impact of AI tools in the heterogeneous and challenging healthcare context, and lays the groundwork for further validation through expert consensus and testing of the framework in real-world healthcare settings. It is important to emphasize that multidisciplinary expertise is essential for thorough assessment, yet many assessors lack the necessary training. Additionally, traditional evaluation methods struggle to keep pace with AI's rapid development. To ensure successful AI integration, flexible, fast-tracked assessment processes and proper assessor training are needed that maintain rigorous standards while adapting to AI’s dynamic evolution.

Introduction

Background

Artificial intelligence (AI) is profoundly transforming healthcare across a range of applications, enhancing both clinical outcomes and operational efficiency. In medical imaging, AI algorithms improve diagnostic accuracy by analyzing complex imaging data, such as from MRI and CT scans, for highly precise and rapid clinical diagnostics [1]. Decision support systems (DSS) powered by AI assist clinicians in making evidence-based decisions by providing real-time, data-driven insights and predictive analytics [2]. Large Language Models (LLMs) are increasingly used for generating detailed medical reports and streamlining triage processes by analyzing and summarizing patient data quickly and accurately [3]. Additionally, innovative digital health technologies like electronic skins utilize wearable sensor technologies and AI to offer continuous, real-time monitoring of various health indicators, further enhancing personalized care [4]. These advancements have the potential to contribute to more efficient, accurate, responsive, and holistic healthcare, reshaping how patient care is delivered and managed.

Despite the growing body of literature on AI in healthcare, its implementation has lagged behind other industries [5,6]. A significant barrier identified by healthcare leaders worldwide is that, despite the emergence of various new frameworks for assessing AI in healthcare, most focus primarily on the quality of study methodologies or technical aspects [7,8]. There remains a lack of a comprehensive, systematic framework that assesses the real-world impact of AI and offers guidance on clinical implementation, monitoring, procurement, and evaluation [7,9]. Most research overlooks the complex, multi-step process required for successful AI integration, leaving critical gaps in understanding how to effectively implement and sustain AI tools in clinical practice [7,9]. As a result, the adoption of AI in clinical practice has fallen short of expectations, with only a few algorithms showing sustained clinical impact [10]. This gap is often due to inadequate or incomplete evaluation and the lack of universally recognized standards for AI assessment. The limited understanding of AI's true added value in healthcare highlights the need for a more comprehensive evaluation framework [11–13]. To ensure confidence in the added clinical value and successful integration of AI into healthcare workflows, a practical, comprehensive tool is needed so that the translational readiness of AI systems can be evaluated. Current approaches assessing AI in healthcare often focus on foundational technical metrics like sensitivity and specificity, which fail to capture the full clinical impact [11,14]. A robust valuation should encompass factors such as patient outcomes, effects on clinical decision-making, workflow efficiency, and the tangible benefits for patients to fully determine AI's true contribution to and impact on healthcare [8,15,16].

In the context outlined above, regulatory approval is an important milestone for demonstrating overall performance, though the scientific evidence supporting AI tools in healthcare remains limited compared to traditional medical standards [7,17]. Additionally, new regulations are being introduced to keep pace with rapidly evolving AI technologies, such as the EU AI Act, which aims to ensure the trustworthiness of high-risk AI tools including healthcare [18]. Despite the potentially positive impact of regulatory frameworks on AI-related developments, a recent study revealed that nearly half of FDA-authorized AI devices lacked clinical validation data, raising concerns about their safety and effectiveness [19]. Without robust clinical validation, these technologies could pose significant risks to patient care. Despite efforts to create reporting guidelines for AI in healthcare, such as Standard Protocol Items Recommendations for Interventional Trials – AI (SPIRIT-AI) [11], Consolidated Standards of Reporting Trials – AI (CONSORT-AI) [12], Standards for Reporting of Diagnostic Accuracy Studies – AI (STARD-AI) [20], Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [21], Prediction model Risk Of Bias ASsessment Tool – AI (PROBAST-AI) [22], and others, a unified international consensus on the evaluation of AI-based tools has yet to be established.

While these guidelines address key methodological issues and share significant overlap, indicating the importance of certain assessment criteria, the absence of a standardized, universally accepted framework remains a significant challenge [4]. This lack of consensus complicates the consistent evaluation and implementation of AI technologies in clinical practice.

Objectives

The goal of this study was to develop a comprehensive framework for assessing the impact of AI tools in healthcare. This involved synthesizing and consolidating the various evaluation criteria found in existing literature regarding the quality and impact of AI tools. Based on the outcomes of this study, we plan on validating the framework through expert consensus using the Delphi process. However, this validation effort will be addressed in the subsequent phase of the project and is beyond the scope of this foundational paper. This approach aims to create a rigorous, evidence-based structure for AI evaluation, ensuring its relevance and applicability in healthcare settings.

In doing so, we adopted the World Health Organization’s (WHO) perspective on AI in healthcare, defining it as "the ability of algorithms and software to analyze complex medical data and support healthcare providers by improving decision-making, predicting outcomes, and enhancing clinical efficiency” [23]. AI tools in healthcare span a broad spectrum of applications, such as, (i) diagnostic support, (ii) prognosis of diseases course, (iii) personalized treatment recommendations, (iv) patient monitoring, and (v) overall health management, driving innovation across the healthcare landscape [23].

To address this, a systematic review was conducted to offer a comprehensive and current analysis of the criteria used in existing research to evaluate the quality and impact of AI in healthcare, from technological, social, and organizational perspectives. The review also explores the potential implications of AI implementation for key stakeholders and offers recommendations on how to effectively assess AI-powered clinical tools under consideration for clinical impact. This study builds upon and extends the findings of a prior research project, which examined sociotechnical assessment criteria for patient-facing eHealth tools, already published [24,25].

We believe the results of this review will provide valuable insights for clinicians, pharmaceutical leaders, insurance professionals, technology providers, and policymakers by presenting an up-to-date, thorough overview of the criteria used to assess AI-powered clinical tools. These insights will help stakeholders make informed decisions about which tools to implement, recommend to patients, invest in, partner with, or provide reimbursement for, based on their assessed quality and potential impact.

Method

Overview

The methodology for this review was based on established best practices, specifically following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [26] and the Cochrane Handbook for Systematic Reviews of Interventions [27]. These frameworks were chosen to ensure a rigorous and methodologically sound approach to the systematic literature review process. All review methods were pre-determined and documented in advance, with the protocol being publicly registered in the Research Registry (reference: reviewregistry1859) to enhance transparency and accountability [28]. The study adhered closely to the initial protocol, with no significant deviations required throughout the process. The primary research question guiding this systematic review was: What technical, social, and organizational criteria should be considered when assessing the quality and impact of AI-powered clinical tools? This question served as the foundation for the analysis and exploration of the criteria relevant to AI's evaluation in clinical settings.

Search Strategy

A comprehensive search of the PubMed, Cochrane, Scopus, and IEEE Xplore databases was conducted in July 2024 to identify relevant studies. The review was limited to peer-reviewed papers published in English between January 2019 and mid-July 2024. We focused on this specific time frame and limited the search to the last five years to ensure the findings reflect the most recent advancements and challenges, particularly with the emergence of new generative AI technologies. Going back further would have added limited value, as older studies may not capture the rapid technological shifts and evolving complexities that are relevant today. Only fully published research articles were included, while other formats, such as editorials and study protocols, were excluded from the analysis. In accordance with the Cochrane Handbook for Systematic Reviews of Interventions, we chose not to include articles sourced through manual reference list searches, as "positive studies are more likely to be cited," which could introduce bias [27].

Textbox 1 illustrates the search string designed using the Participants, Intervention, Comparators, and Outcomes (PICO) framework. To ensure the relevance of the retrieved papers, the search was mostly restricted to manuscript titles, focusing on studies that addressed AI assessment criteria comprehensively rather than those evaluating specific tools or pilot studies. Since comparators were not relevant to this review, they were excluded from the search parameters.

Textbox 1. The search string according to the participants, intervention, comparators, and outcomes framework

Participants: Clinicians
Focus on AI-powered tools for clinicians, excluding those designed solely for patients or medical students.

Intervention: AI-powered clinician tools
Focus is on AI-powered clinician tools, the search targeted manuscript titles containing the terms (AI OR "Artificial Intelligence").

Comparator: Not Applicable
There were no restrictions on eligible conditions for inclusion.

Outcome: Assessment criteria
The search targeted manuscript titles also containing
AND (assessment OR assess OR evaluation OR evaluating OR effectiveness OR efficacy OR quality OR efficiency OR usability OR usefulness).
As well as manuscript tiles and abstracts containing
AND (criteria OR framework OR method OR methodology OR methodologies OR measurement OR toolkit OR tool OR tools OR approach OR scorecard).

Study Selection

A total of two researchers (CJ and EL) participated in the screening, eligibility, and inclusion phases of the study. Any discrepancies during these stages were resolved through discussion among the two. If consensus could not be reached, a third co-author was consulted to make the final decision. The team utilized the open-source Rayyan app (Qatar Computing Research Institute) to streamline collaborative screening efforts [29]. The screening process took place between July and August 2024. The inclusion and exclusion criteria, outlined in Textbox 2, were developed following the participants, intervention, comparators, and outcomes (PICO) framework.

Following the completion of the screening process and resolution of any conflicting views among the researchers, CJ and EL proceeded to assess the full texts of the selected studies for eligibility. Any remaining disagreements were addressed through consultation with a third co-author. CJ evaluated the risk of bias using the Critical Appraisal Skills Programme (CASP) checklist [30], which assesses key quality criteria in the included studies. These criteria include: the presence of a clear statement of the research aims, the appropriateness of the methodology for the research objectives, the suitability of the research design in addressing those aims, the relevance of the recruitment strategy, the adequacy of data collection methods in relation to the research question, the consideration given to the researchers' roles, the evaluation of ethical issues, the rigor of data analysis, the clarity of the study's findings, and whether the researchers discussed the study's contribution to existing knowledge, such as its implications for current practice, policy, or relevant literature. The results of this appraisal are available in Multimedia Appendix 1.

Textbox 2. Inclusion and exclusion criteria according to the participants, intervention, comparators, and outcomes framework

Inclusion criteria

- Participants
  - Focused on clinicians
- Intervention
  - Focused on AI-powered clinician tools
- Comparators
  - Does not apply
- Outcomes
  - Addresses the different criteria used to assess the quality and impact of AI-powered clinician tools regardless of the condition
- Publication type
  - Peer-reviewed and published papers
- Time frame
  - Studies published between January 2019 and mid-July 2024
- Language
  - Studies published in English

Exclusion criteria

- Participants
  - Focused solely on patients or medical students
- Intervention
  - Technologies utilized outside of clinical environments, such as chatbots employed by patients to obtain healthcare information
- Comparators
  - Does not apply
- Outcomes
  - Individual assessments of pilot studies singling out specific tools; and assessment frameworks that focus on the reporting and methodological quality of AI research and clinical studies, rather than

    - evaluating the AI tool itself
- Publication type
    - Editorials and study protocols
- Time frame
    - Studies published before January 2019 or after mid-July 2024
- Language
    - Studies published in languages other than English

Data Collection and Synthesis

The procedures and outcomes across the included studies were too diverse to support a quantitative analysis. As a result, a narrative synthesis was employed following the sociotechnical approach, organized around the social, organizational, and technical criteria used to evaluate the quality and impact of AI-powered tools for clinicians. The authors were influenced by the sociotechnical theory, which emphasizes that the design and performance of innovations can only be fully understood when both social and technical aspects are considered as interdependent components of a larger system [31]. This approach aligns with recommendations from several scholars who advocate for moving beyond purely technology-focused frameworks to incorporate the broader context, including societal and implementation factors [32–34]. To facilitate this process, NVivo (QSR International) version 1.7.2, a qualitative data analysis software, was utilized.

Data coding began with a preliminary extraction grid, which was structured around themes derived from previous research and established technology acceptance frameworks. The initial codebook was informed by our prior work on factors influencing eHealth evaluation and adoption[24,25,34–36], with additional codes being incorporated as new themes emerged during the review. Thematic analysis, as outlined by Braun and Clarke [37], was conducted to identify and extract themes based on the social, technical, and organizational assessment criteria relevant to the research question. This analysis followed seven key phases: familiarizing with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, linking themes to explanatory frameworks, and producing the final report. The first author CJ conducted the initial analysis and coding, and NB reviewed the coding, and any cases of disagreement were discussed and mutually agreed upon in conjunction with a third author. This process was carried out from August to October 2024.

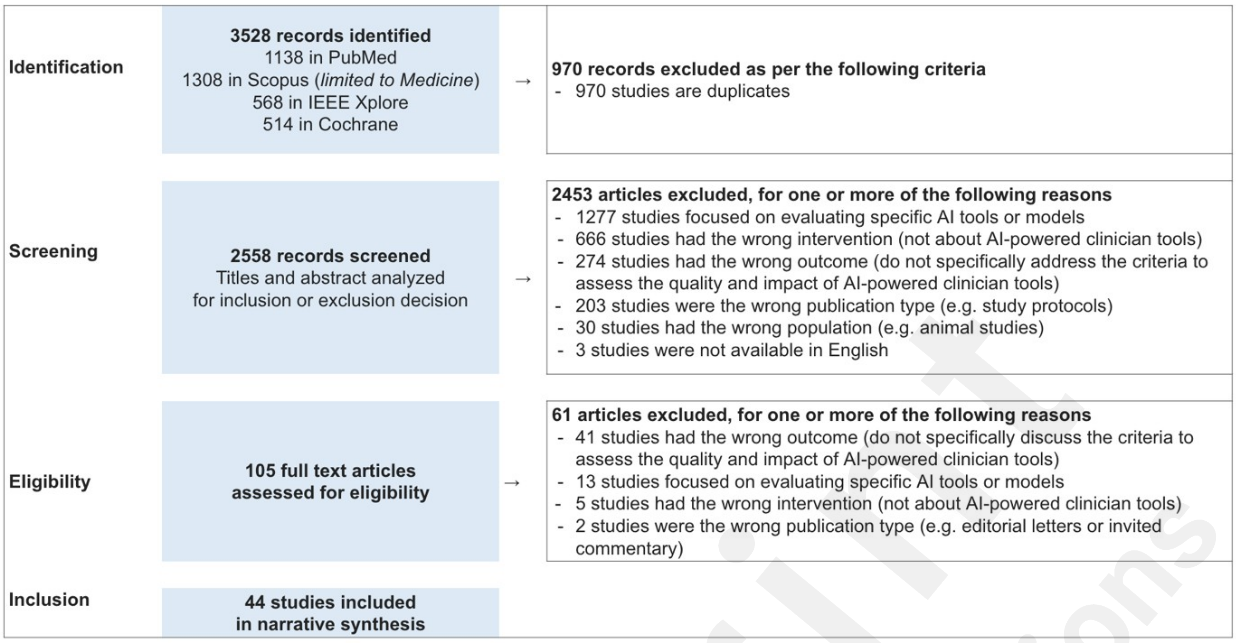
Results

Study Selection Flow and Characteristics of the Included Studies

Figure 1 presents the PRISMA flow diagram, illustrating the progression of study selection during the systematic review. It details the number of records identified, screened, included, and excluded, along with reasons for exclusion. After applying these criteria, 44 articles were selected for the qualitative synthesis.

Table 1 outlines the characteristics of these studies, offering insights into their research methodologies, geographic distributions, and clinical focuses. This comprehensive overview highlights the diversity of approaches and topics addressed within the included studies.

Figure 1. Study selection flow diagram based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines



Identification

3528 records identified
1138 in PubMed
1308 in Scopus (*limited to Medicine*)
568 in IEEE Xplore
514 in Cochrane

→ **970 records excluded as per the following criteria**

- 970 studies are duplicates

Screening

2558 records screened
Titles and abstract analyzed
for inclusion or exclusion decision

→ **2453 articles excluded, for one or more of the following reasons**

- 1277 studies focused on evaluating specific AI tools or models
- 666 studies had the wrong intervention (not about AI-powered clinician tools)
- 274 studies had the wrong outcome (do not specifically address the criteria to assess the quality and impact of AI-powered clinician tools)
- 203 studies were the wrong publication type (e.g. study protocols)
- 30 studies had the wrong population (e.g. animal studies)
- 3 studies were not available in English

Eligibility

105 full text articles assessed for eligibility

→ **61 articles excluded, for one or more of the following reasons**

- 41 studies had the wrong outcome (do not specifically discuss the criteria to assess the quality and impact of AI-powered clinician tools)
- 13 studies focused on evaluating specific AI tools or models
- 5 studies had the wrong intervention (not about AI-powered clinician tools)
- 2 studies were the wrong publication type (e.g. editorial letters or invited commentary)

Inclusion

44 studies included in narrative synthesis

Table 1: Characteristics of the included studies (N=44)

Study Characteristic	n (%)	References
Country of authors		
Multiple	21 (47.7%)	[38–58]
USA	5 (11.4%)	[59–63]
France	3 (6.8%)	[64–66]
Netherlands	3 (6.8%)	[67–69]
Australia	2 (4.5%)	[70,71]
Canada	2 (4.5%)	[72,73]

Others	8 (18.2%)	China [74];
		Denmark [75];
		Germany [76];
		Greece [77];
		India [78];
		Saudi Arabia [79];
		Sweden [80];
		UK [81]

Focus (some papers encompassed multiple areas of focus)		
No specific focus	9 (20.5%)	[46,51,53,60,62,64,65,67,72]

Clinical focus	20 (45.5%)	Cardiovascular [40,55,56];
		Dermatology [39,63];
		ENT (Ear, Nose, and Throat) [79];
		Medical Imaging [38,42,43,52,61,66,71,73,75,78,80,81];
		Nuclear Medicine [45];
		Radiation Oncology [44]

Technology focus	12 (27.3%)	Artificial Neural Network (ANN) [45];
		Clinical Decision Support Systems (CDSSs) [49,50,74];
		Diagnostic Quality Models (DQMs) [48];
		Large Language Models (LLMs) [54,59,70];
		Machine Learning (ML) [43,50];
		Prediction Models [55,68]

Thematic focus	10 (22.7%)	Economic Evaluations (EEs) [41,58,69,76];
		Ethics and Equity [57,60,63];
		Explainability [77];
		Regulatory and trust [47,67]

Paper type		

	Original research	5 (11.4%)	Delphi process [41,72,74]; Survey or questionnaire [77,79]
	Expert consensus	3 (6.8%)	[39,54,66]
	Expert perspective or comment	9 (20.5%)	[48,49,51,52,59,61–63,73]
	Guidelines or statements	6 (13.6%)	[38,44,45,55,57,60]
	Policy brief	1 (2.3%)	[67]
	Review	10 (22.7%)	[42,43,47,50,53,56,71,78,80,81]
	Scoping review	6 (13.6%)	[40,65,68,70,75,76]
	Systematic review	4 (9.1%)	[46,58,64,69]
Publication year			
	2019 (from January)	2 (4.5%)	[43,49]
	2020	2 (4.5%)	[50,57]
	2021	5 (11.4%)	[47,52,53,67,71,74]
	2022	10 (22.7%)	[39,42,45,56,68,69,75,76,80]
	2023	12 (27.3%)	[40,44,48,55,58,60,62,64,66,70,72,73,77,78]
	2024 (until mid-July)	13 (29.6%)	[38,41,46,51,54,59,61,63,65,79,81]
Frameworks resulting from the included studies			
	ABCDS (Algorithm-Based Clinical Decision Support)	1 (2.3%)	[60]
	CHEERS-AI (Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence)	1 (2.3%)	[41]
	CLEAR (Derm Consensus Guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group)	1 (2.3%)	[39]
	DQM (Diagnostic quality model)	1 (2.3%)	[48]
	DRIM France AI grid (French community grid for the evaluation of radiological artificial intelligence solutions)	1 (2.3%)	[66]
	ECLAIR (evaluating commercial AI solutions in radiology)	1 (2.3%)	[52]
	HEAL (Health equity assessment of machine learning performance)	1 (2.3%)	[63]
	MAS-AI (Model for ASsessing the value of AI in medical imaging)	1 (2.3%)	[75]
	RADAR (Radiology AI Deployment and Assessment Rubric)	1 (2.3%)	[38]
	RELAINCE Guidelines (Recommendations for EvaLuation of AI for NuClear medicinE)	1 (2.3%)	[45]
	R-AI-DIOLOGY checklist (a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology)	1 (2.3%)	[42]
	TEHAI (Translational Evaluation of Healthcare AI)	1 (2.3%)	[53]
	TREE (transparency, reproducibility, ethics, and effectiveness)	1 (2.3%)	[57]
Frameworks used in or referred to in the included studies			
	Consolidated Health Economic Evaluation Reporting Standards (CHEERS)	3 (6.8%)	[41,58,69]
	Checklist for Artificial Intelligence in Medical Imaging (CLAIM)	4 (9.1%)	[64,73,75,78]
	Consolidated Standards of Reporting Trials of AI (CONSORT-AI)	6 (13.6%)	[39,41,58,64,65,71]
	Reporting guideline for the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by AI (DECIDE-AI)	2 (4.5%)	[39,81]
	International consensus guideline for trustworthy and deployable artificial intelligence in healthcare (FUTURE-AI)	1 (2.3%)	[38]
	Good Evaluation Practice in Health	1 (2.3%)	[49]

	Informatics (GEP-HI)		
	Health Technology Assessment (HTA)	6 (13.6%)	[40,56,58,64,65,75]
	Model for Assessment of Telemedicine (MAST)	2 (4.5%)	[40,75]
	Prediction model Risk Of Bias ASsessment Tool-Artificial Intelligence (PROBAST-AI)	4 (9.1%)	[39,41,64,71]
	Quality Analysis of Medical Artificial Intelligence (QAMAI)	1 (2.3%)	[54]
	Quality Management System (QMS)	1 (2.3%)	[62]
	Radiomics Quality Score (RQS)	1 (2.3%)	[73]
	Standard Protocol Items: Recommendations for Interventional Trials - Artificial Intelligence (SPIRIT-AI)	6 (13.6%)	[39,41,58,64,65,71]
	Standards for Reporting of Diagnostic Accuracy Studies - Artificial Intelligence (STARD-AI)	3 (6.8%)	[39,64,71]
	Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI)	1 (2.3%)	[49]
	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis Artificial Intelligence (TRIPOD-AI)	7 (15.9%)	[39,41,55,57,60,65,71]

Critical Appraisal

We evaluated the quality of the included studies using the Critical Appraisal Skills Programme (CASP) checklist [30]. This tool was selected due to the variety of methodologies employed in the studies and the narrative approach of our synthesis, which differed from meta-analyses and other quantitative methods. The CASP is widely recognized as the most frequently used tool for appraising the quality of qualitative evidence in health research, with endorsement from the Cochrane Qualitative and Implementation Methods Group [82]. The studies included in our review utilized a range of methodologies (quantitative, qualitative, mixed methods, and systematic literature reviews) which meant that some questions on the checklist were not applicable to all study types. As per the checklist's recommendations, we did not assign scores to the studies.

Following the critical appraisal of the 44 studies, several issues were identified. While all studies clearly stated their aims, presented well-defined findings, and provided valuable insights for healthcare stakeholders, 21 (47.7%) studies lacked a dedicated methods section, making it difficult to assess the appropriateness and suitability of their approach. Similarly, the absence of clear methods in these studies hindered the evaluation of the research design and data collection techniques.

Additionally, 25 (56.8%) studies did not detail their analysis methods, making it challenging to gauge the rigor and reliability of their approach. Furthermore, 28 (63.6%) studies lacked validation of their findings, while 8 (18.2%) offered only partial validation (e.g., expert consensus), highlighting the need for empirical validation in real-world clinical applications to ensure the findings' robustness. The comprehensive quality assessment of the included studies can be found in Multimedia Appendix 1.

Studies were not excluded based on the results of the quality assessment, as this was unlikely to significantly impact the definition of the assessment criteria or the development of the aggregated framework. However, the quality assessment offered valuable insight into the overall robustness of the development processes behind the existing frameworks, helping to gauge the strength and reliability of the evidence presented [82]. A more in-depth exploration of this topic can be found in the Discussion section, where the challenges associated with current initiatives and frameworks are examined.

Synthesized Assessment Criteria

We synthesized comparable measures from various papers, frameworks, and initiatives, ultimately identifying a set of unique criteria that reflected all relevant assessment methods referenced in the included studies. Notably, several criteria are closely interrelated and could fit into multiple categories; however, they were placed in the most appropriate category based on their significance and impact. For instance, while "user trust" and "model explainability" are inherently linked, since trust often correlates with the level of explainability provided by an AI system, we categorized trust under the cluster "Acceptability, Trust, and Training," which focuses on user-centric aspects, whereas "explainability" was assigned to the cluster evaluating model performance metrics, given its technical focus. Additionally, we intentionally included assessment criteria applicable to high-risk tools, enabling us to compile a more comprehensive list. We recognized that not all criteria would apply to lower-risk AI-powered healthcare tools, such as patient safety assessments, which are more relevant to high-risk tools that pose potential safety concerns. We are guided by NICE's Evidence Standards Framework for Digital Health Technologies to assess and understand the risk levels of healthcare technologies [83].

Figure 2 provides a visual overview of the aggregated criteria, organized into clusters and sub-clusters, while Table 2 presents these criteria grouped into seven primary clusters and 32 sub-criteria, outlining their occurrences across the included studies, along with their definitions and corresponding references. A detailed exploration of each criteria cluster and its corresponding sub-criteria is provided in the Discussion section.

With our focus on assessing the long-term real-world impact of AI technologies in healthcare, we named the framework AI for IMPACTS. The criteria are organized into seven key clusters, each corresponding to a letter in the acronym: (I) integration, interoperability and workflow (M) monitoring, governance, and accountability (P) performance and quality metrics (A) acceptability, trust, and training (C) cost and economic evaluation (T) technological safety and transparency (S) scalability and impact.

Figure 2: Visual overview of the aggregated assessment criteria, organized into clusters and sub-criteria

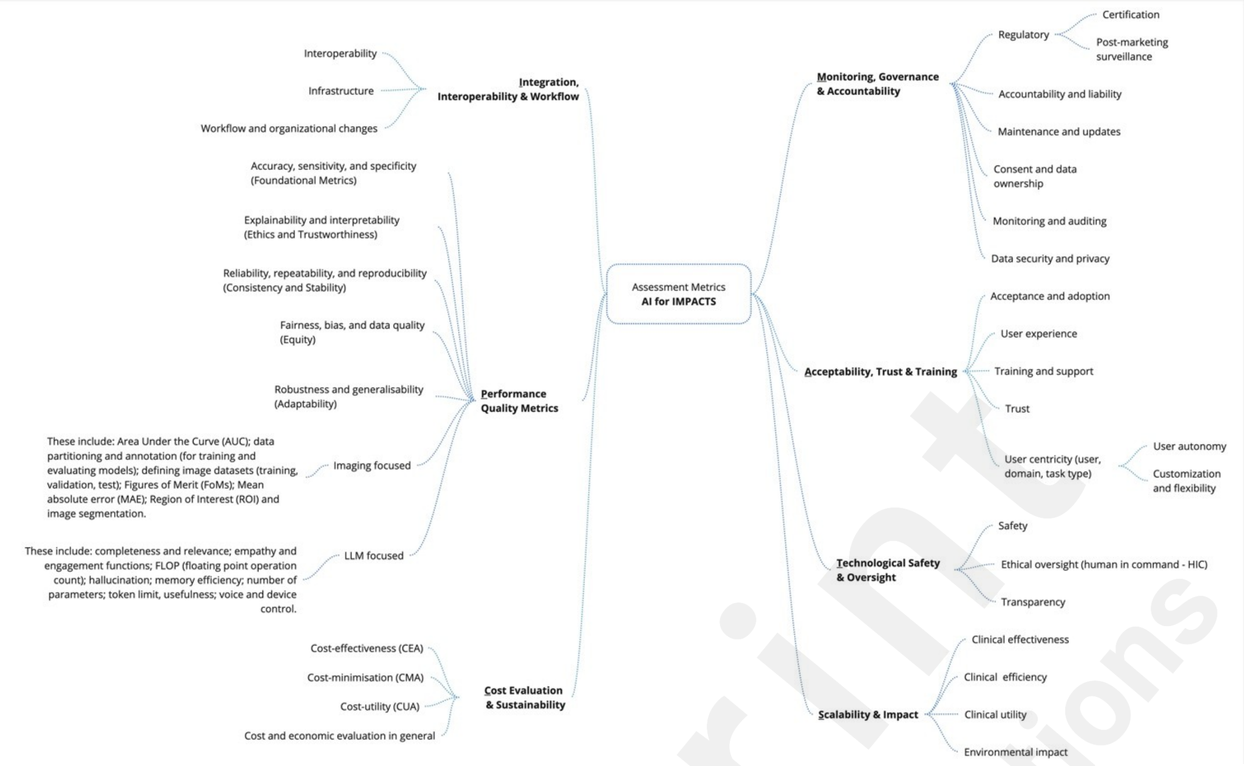

Assessment Metrics
AI for IMPACTS
Integration, Interoperability & Workflow
Interoperability
Infrastructure
Workflow and organizational changes
Performance Quality Metrics
Accuracy, sensitivity, and specificity (Foundational Metrics)
Explainability and interpretability (Ethics and Trustworthiness)
Reliability, repeatability, and reproducibility (Consistency and Stability)
Fairness, bias, and data quality (Equity)
Robustness and generalisability (Adaptability)
Imaging focused
These include: Area Under the Curve (AUC); data partitioning and annotation (for training and evaluating models); defining image datasets (training, validation, test); Figures of Merit (FoMs); Mean absolute error (MAE); Region of Interest (ROI) and image segmentation.
LLM focused
These include: completeness and relevance; empathy and engagement functions; FLOP (floating point operation count); hallucination; memory efficiency; number of parameters; token limit, usefulness; voice and device control.
Cost Evaluation & Sustainability
Cost-effectiveness (CEA)
Cost-minimisation (CMA)
Cost-utility (CUA)
Cost and economic evaluation in general
Monitoring, Governance & Accountability
Regulatory
Certification
Post-marketing surveillance
Accountability and liability
Maintenance and updates
Consent and data ownership
Monitoring and auditing
Data security and privacy
Acceptability, Trust & Training
Acceptance and adoption
User experience
Training and support
Trust
User centricity (user, domain, task type)
User autonomy
Customization and flexibility
Technological Safety & Oversight
Safety
Ethical oversight (human in command - HIC)
Transparency
Scalability & Impact
Clinical effectiveness
Clinical efficiency
Clinical utility
Environmental impact

Table 2: Assessment criteria, their definitions, occurrence, and respective references (N = 44)

	Criteria	Definition	n (%)	Papers where the criteria occurred
Integration				
	Infrastructure	The underlying technological, hardware, and software systems required to support the deployment and scalability of the AI tool.	15 (34.1%)	[38,41,46,50–52,57,62,65,66,68,72,73,75,81]
	Interoperability	The AI tool's ability to seamlessly exchange and integrate data with different healthcare platforms and devices.	19 (43.2%)	[38,42,48,50–53,56,61,64–66,68,71–75,79]
	Workflow and organizational changes	The degree to which the AI tool impacts existing clinical workflows and healthcare operations, ensuring minimal disruption while enhancing efficiency, communication, and overall care delivery.	22 (50%)	[38,42,44,46,48–53,55,61,62,64–66,68,73,75,76,79,81]
Monitoring, Governance & Accountability				
	Accountability and liability	The clear attribution of responsibility for errors or outcomes and the establishment of legal and ethical frameworks to address potential issues and ensure proper recourse.	13 (29.5%)	[46–48,51,52,60–62,67,72,75,78,81]
	Consent and data ownership	Evaluates the processes for obtaining informed consent from patients regarding the use of their data and ensuring clear policies on data ownership, privacy, and control.	5 (11.4%)	[46,51,65,75,78]
	Maintenance and updates	Evaluates the processes for ongoing support, including regular updates and bug fixes, to ensure the AI tool remains effective, secure, and aligned with evolving medical standards and practices.	13 (29.5%)	[41,42,46,47,49,50,52,56,62,66,68,73,74]
	Monitoring and governance	Evaluates the systems in place for overseeing the AI's performance, including regular assessments and audits to ensure ethical use and	22 (50%)	[41,45–53,56,57,61,62,65,66,68,71,73,78,80,81]

		effectiveness.		
	Regulatory compliance	Evaluates adherence to established regulations throughout the AI tool's lifecycle, including ongoing monitoring and reporting after deployment to ensure continued safety, efficacy, and adherence to legal requirements.	23 (52.3%)	[42,46–52,56,57,60–62,64–66,69,71,73,75,78,80,81]
	Security and privacy	Evaluates the measures implemented to protect sensitive patient data from unauthorized access and breaches while ensuring compliance with privacy regulations.	26 (59.1%)	[42,48–53,57,59,61–68,70–72,74,75,77–80]
Performance Quality Metrics				
	Accuracy, sensitivity, and specificity (foundational metrics)	Accuracy: The proportion of correct predictions (both true positives and true negatives) out of all predictions. It gives an overall measure of performance but may be misleading if the dataset is imbalanced (i.e., when one class dominates). Sensitivity (Recall): The ability of the model to correctly identify true positives (i.e., people with the condition). In healthcare, this often refers to how well the model detects cases like diseases. High sensitivity ensures that most cases of the disease are caught, reducing the chance of missing sick patients. Specificity: The ability to correctly identify true negatives (i.e., people without the condition). High specificity means the model avoids false positives, reducing unnecessary interventions for healthy people.	26 (59.1%)	[38,40,42–50,54,55,57,59,64,66,67,70,71,73,77–81]
	Explainability and interpretability (ethics and trustworthiness)	Explainability: Refers to the degree to which the model's predictions and decisions can be understood by humans. In healthcare, explainability is crucial because clinicians need to trust AI recommendations and understand why the AI made a particular decision. Interpretability: Closely related to explainability, it is about how easily a human can comprehend the internal workings of the model. For example, an interpretable model may allow clinicians to track how specific features (like patient age or lab results) influenced the AI's prediction.	19 (43.2%)	[42,44–46,48,49,52,57,59,64,65,68,70–72,75,77,78,80]
	Fairness (equity)	Fairness: Ensures that the AI model does not systematically discriminate against any specific group of people (e.g., based on race, gender, or socioeconomic status). Fairness in healthcare is key to avoid bias in diagnoses or treatments.	32 (72.7%)	[39–44,46,48–53,55,57,59–66,68,70–72,74,75,78,80,81]
	Reliability, repeatability, and reproducibility	Reliability: Refers to the consistency of the model over time. Can the AI be trusted to perform in the same	24 (54.5%)	[39,40,42,45,47–53,55–57,61,64,65,68,71,73,74,78,80,81]

	(consistency and stability)	way under similar conditions in the future? Repeatability: The ability of the model to provide consistent results when the same input is given multiple times in the same environment. In healthcare, this ensures that if a patient is re-evaluated using the same AI tool, it will give the same outcome. Reproducibility: Refers to how well the model performs when applied to different datasets or by different teams. This is critical in healthcare, where models trained on one population must still perform well when tested on different populations or data collected in different hospitals.		
	Robustness and generalizability (adaptability)	Robustness: The model's ability to maintain performance despite slight variations or noise in the input data. In a healthcare setting, this might mean the model works well even with slightly lower-quality images or lab results from different equipment. Generalizability: The ability of the model to perform well on new, unseen data that may differ from the training data. In healthcare, it's crucial that an AI model trained in one hospital or region can generalize to others.	23 (52.3%)	[38–43,45,46,49,51–53,55,57,61,64,65,67,68,72,73,80,81]
	Imaging-focused	These may include: Area Under the Curve (AUC); data partitioning and annotation (for training and evaluating models); defining image datasets (training, validation, test); Figures of Merit (FoMs); Mean absolute error (MAE); Region of Interest (ROI) and image segmentation.	10 (22.7%)	[39,42–45,52,66,73,78,80]
	LLM-focused	These may include: completeness and relevance; empathy and engagement functions; floating point operation count (FLOP); hallucination; memory efficiency; number of parameters; token limit, usefulness; voice and device control.	3 (6.8%)	[54,59,70]
Acceptability, trust, and training				
	Acceptance and adoption	Evaluates how well the AI tool is embraced by healthcare professionals and patients, including their willingness to integrate it into routine practice.	18 (40.9%)	[44,46,49–51,53,57,60,62,64–66,69,70,72,74,75,79]
	Training and support	Evaluates the effectiveness and availability of resources provided to users for learning and utilizing the AI tool, ensuring they have the necessary guidance and assistance for successful implementation and operation.	17 (38.6%)	[42,46,49–52,61,62,64,66,68,73,75,77–79,81]
	Trust	Evaluates the degree to which	11	[42,44,46,49,57,59,65,70,72,79,81]

		healthcare professionals and patients believe in the reliability, accuracy, and ethical considerations of the AI tool, influencing their willingness to use it.	(25%)	
	Usability	Evaluates how easily and effectively healthcare professionals and patients can interact with and utilize the AI tool, ensuring it enhances rather than hinders the user experience and clinical workflows.	18 (40.9%)	[44,46–50,52,53,59,60,65,68,70,72–74,79,81]
	User centricity (user, domain, task type)	Evaluates how well the AI tool is designed to meet the specific needs, preferences, and contexts of its users, domain-specific requirements, and task types it is intended to support.	19 (43.2%)	[39,42,46,49,50,52,53,55,57,59,61,62,68,70,72–74]
Cost and economic evaluation				
	Costs and economic evaluation in general	Evaluates the financial implications of implementing the AI tool, ensuring it provides value without imposing excessive financial burdens on healthcare systems or patients.	18 (40.9%)	[41,46,48,52,56–58,61,64–66,69,70,72,73,76,79,81]
	Cost-effectiveness analysis (CEA)	Compares the relative costs and outcomes of different interventions. The outcomes are typically measured in natural units like life years saved, cases prevented, or symptom-free days.	12 (27.3%)	[38,50,52,56–58,65,68–71,81]
	Cost-minimization analysis (CMA)	Used when two or more interventions or treatments are assumed to produce identical outcomes or equivalent effectiveness. Given that the outcomes are considered the same, the focus is entirely on minimizing costs.	5 (11.4%)	[38,58,69,75,79]
	Cost-utility analysis (CUA)	Measures outcomes in terms of both quantity (life expectancy) and quality of life. It uses a metric called quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs) to quantify health benefits.	3 (6.8%)	[38,58,69]
Technological safety, and transparency				
	Safety	Evaluation an AI tool's ability to avoid causing harm to patients by ensuring that it operates reliably, adheres to clinical standards, and mitigates potential risks.	26 (59.1%)	[39,44,46–53,57,59–65,68,70,72,75,77,78,80,81]
	Transparency	Refers to the extent to which an AI tool’s processes, decision-making logic, and data sources are made understandable and accessible to stakeholders.	27 (61.4%)	[39,41,43,46,47,49–51,53,55–58,60–62,64,65,67,68,70–73,75,77,78]
	Ethical oversight, human in command (HIC)	Assesses whether the AI tool is designed to support human decision-making, allowing clinicians to maintain control and override AI decisions when necessary, ensuring AI complements rather than replaces human judgment.	14 (31.8%)	[38,39,42,46,48,61,62,64,65,67,71,72,75,77]
Scalability and impact				
	Clinical effectiveness	Assesses how well the AI tool works in real-world practice, including its ability to achieve desired clinical outcomes across diverse populations and settings.	26 (59.1%)	[38–41,44–46,48–51,57,60–62,64–66,68–70,72,74,75,80,81]

	Clinical efficiency	Focuses on the optimal use of resources (time, staff, cost) to deliver care.	8 (18.2%)	[44,49,50,52,53,65,73,76]
	Clinical utility	Refers to the practical benefits of a treatment or intervention in improving patient care, such as guiding clinical decision-making or reducing risks.	14 (31.8%)	[38–40,44,45,47,49,53,55,57,59,66,74,79]
	Environmental impact	Evaluates how the development, deployment, and operation of AI tools affect environmental sustainability, such as energy consumption and carbon footprint.	1 (2.3%)	[72]

Discussion

A Comprehensive Framework for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

By synthesizing and aggregating the assessment criteria from all included studies, we developed the AI for IMPACTS framework. This framework goes beyond focusing solely on technical metrics or methodological guidance at the study level. It integrates the clinical context and real-world implementation factors to ensure AI tools are evaluated holistically. Most criteria in our proposed framework can be aligned with existing frameworks, but none covers all relevant categories without extensions. For successful AI implementation in healthcare, it is essential to integrate these tools within the broader organizational context. Frameworks should account for the complexities of the sociotechnical environment, recognizing the interplay between technical, social, and organizational dimensions. Our consolidated framework achieves this by synthesizing and expanding existing frameworks for AI assessment in healthcare. It uses a sociotechnical approach to consider all contextual factors, their interactions, and the long-term real-world impact of these technologies in clinical practice.

The AI for IMPACTS framework is organized around seven core criteria clusters: (a) integration, interoperability and workflow: focuses on evaluating how effectively the AI tool integrates into existing clinical workflows and healthcare systems (b) monitoring, governance, and accountability: focuses on evaluating how effectively the AI tool is monitored throughout its lifecycle, addressing critical aspects such as model drift, data governance, and adherence to ethical standards (c) performance and quality metrics: focuses on evaluating the performance and quality of the AI tool by assessing key metrics such as accuracy, fairness, explainability, reliability, and robustness (d) acceptability, trust, and training: evaluates user-centric aspects of the AI tool, focusing on its acceptance, trustworthiness, and the adequacy of user training and support (e) cost and economic evaluation: evaluates the economic implications of the AI tool to determine its financial viability and long-term sustainability (f) technological safety and transparency: assesses the safeguards in place to ensure safe and ethical operation (g) scalability and impact: focuses on evaluating scalability and impact by determining the AI tool’s clinical utility and effectiveness, and examining its broader impact.

Figure 3 depicts the seven assessment clusters of the AI for IMPACTS framework. Each cluster contains multiple sub-criteria, all of which are summarized in a comprehensive checklist presented in

Table 3. The framework provides a systematic approach for evaluating AI's holistic role and potential in healthcare applications. The following sub-sections provide a detailed analysis of each criteria cluster and their respective sub-criteria, offering a comprehensive breakdown of how each factor contributes to the overall assessment.

Figure 3: AI for IMPACTS: A Comprehensive Framework for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

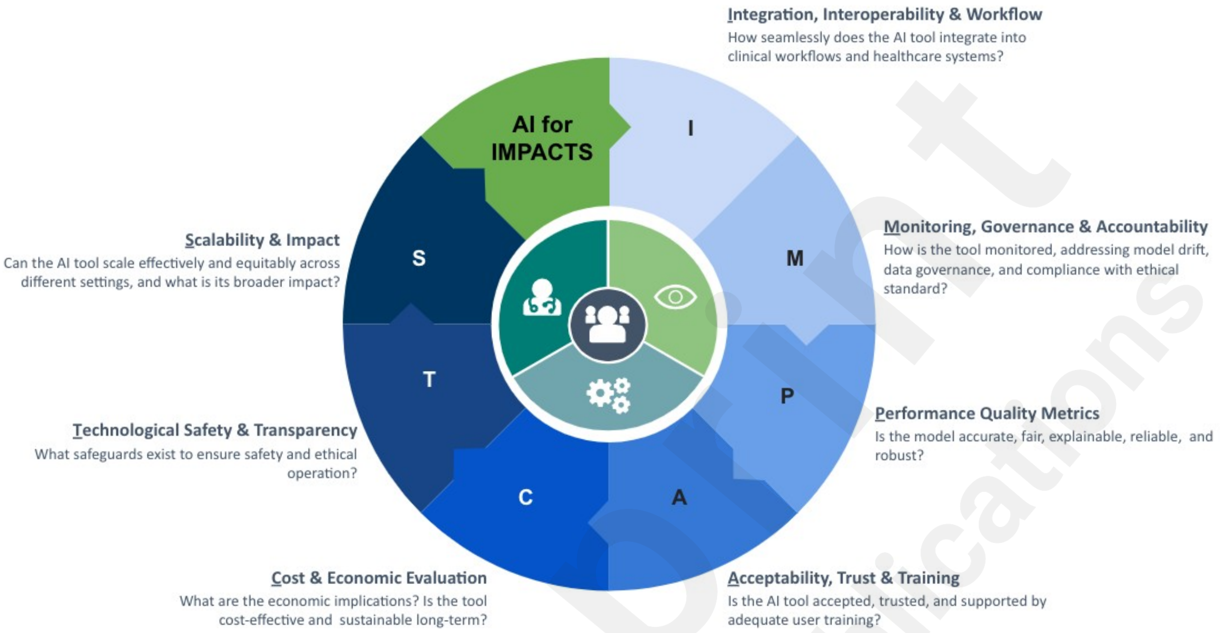

Integration, Interoperability & Workflow
How seamlessly does the AI tool integrate into clinical workflows and healthcare systems?
AI for IMPACTS
I
M
P
A
C
T
S
Monitoring, Governance & Accountability
How is the tool monitored, addressing model drift, data governance, and compliance with ethical standard?
Scalability & Impact
Can the AI tool scale effectively and equitably across different settings, and what is its broader impact?
Performance Quality Metrics
Is the model accurate, fair, explainable, reliable, and robust?
Technological Safety & Transparency
What safeguards exist to ensure safety and ethical operation?
Cost & Economic Evaluation
What are the economic implications? Is the tool cost-effective and sustainable long-term?
Acceptability, Trust & Training
Is the AI tool accepted, trusted, and supported by adequate user training?

Table 3: AI for IMPACTS Assessment Criteria for Evaluating the Long-term Real-world Impacts of AI-powered Clinician Tools

	Criteria	Assessment
Integration		
	Infrastructure	Does the deployment and scalability of the AI tool require additional technological, hardware, or software infrastructure beyond what is already available in the current clinical setting?
	Interoperability	Does the AI tool seamlessly integrate and exchange data with various healthcare platforms and devices, ensuring interoperability across different systems without requiring significant modifications?
	Workflow and organizational changes	Does the AI tool integrate smoothly into existing clinical workflows and healthcare operations, minimizing disruption while enhancing efficiency, communication, and the overall delivery of care?
Monitoring, Governance & Accountability		
	Accountability and liability	Is there clear attribution of responsibility for errors or outcomes, supported by well-defined legal and ethical frameworks that ensure accountability and proper recourse in the event of any issues?
	Consent and data ownership	Does the AI tool have clear and robust processes for obtaining informed consent from patients, including transparent policies on data ownership, privacy, and control, ensuring patients fully understand how their data will be used?
	Maintenance and updates	Does the AI tool have established processes for ongoing support, including regular updates and bug fixes, to ensure it remains effective, secure, and compliant with evolving medical standards and practices?
	Monitoring and governance	Does the AI tool have systems in place for ongoing oversight of its performance, including regular assessments and audits to ensure ethical use, effectiveness, and adherence to relevant standards?
	Regulatory compliance	Does the AI tool demonstrate adherence to established regulations throughout its entire lifecycle, with systems in place for ongoing monitoring and reporting post-deployment to ensure continued safety, efficacy, and compliance with legal requirements?

	Security and privacy	Does the AI tool have robust measures in place to protect sensitive patient data from unauthorized access and breaches, while ensuring full compliance with relevant privacy regulations?
Performance Quality Metrics		
	Foundational metrics	These are application-specific metrics to ensure each tool is assessed appropriately based on its function: Diagnosis & Prediction Applications: Use classification metrics (e.g. accuracy, sensitivity, specificity, Area Under the Curve (AUC)) for diagnosis tasks and regression metrics (e.g. Mean Absolute Error (MAE), Root Mean Square Error (RMSE)) for predicting continuous outcomes (classification, anomaly detection, recommendation systems). Image & Pattern Analysis: Focus on segmentation accuracy and reinforcement learning's long-term performance optimization (e.g. Dice Coefficient, Jaccard Index, Cumulative Reward). Text & Language Processing Applications: Evaluate the accuracy and quality of AI-extracted or generated text (e.g. completeness and relevance, empathy and engagement, Floating Point Operation Count (FLOP), hallucination).
	Explainability and interpretability (ethics and trustworthiness)	Does the AI tool offer a high degree of explainability and interpretability, allowing clinicians to understand its predictions and decisions, and enabling them to track how specific features influenced its outcomes?
	Fairness (equity)	Does the AI tool ensure fairness by avoiding systematic discrimination against any specific group, such as race, gender, or socioeconomic status, and promoting equitable outcomes in diagnoses and treatments?
	Reliability, repeatability, and reproducibility (consistency and stability)	Does the AI tool demonstrate reliability, repeatability, and reproducibility by consistently delivering the same results over time, under similar conditions, and when applied to different datasets or used by different teams?
	Robustness and generalizability (adaptability)	Does the AI tool demonstrate both robustness and generalizability by maintaining strong performance despite variations or noise in input data, and by performing well on new, unseen data from different hospitals or regions compared to its training data?
Acceptability, trust, and training		
	Acceptance and adoption	Does the AI tool demonstrate strong acceptance by healthcare professionals and patients, including their willingness to adopt and integrate it into routine clinical practice?
	Training and support	Does the AI tool provide comprehensive and readily available resources for users, ensuring they have the necessary guidance, training, and assistance to successfully implement and operate it in clinical practice?
	Trust	Does the AI tool inspire trust among healthcare professionals and patients in terms of its reliability, accuracy, and ethical considerations, thereby positively influencing their willingness to use it?
	Usability	Does the AI tool offer an intuitive and user-friendly interface that allows healthcare professionals and patients to interact with it easily and effectively, ensuring it enhances the user experience and integrates smoothly into clinical workflows?
	User centricity (user, domain, task type)	Does the AI tool effectively meet the specific needs, preferences, and contexts of its users, while addressing domain-specific requirements and supporting the relevant tasks for which it is intended?
Cost and economic evaluation		
	Costs and economic evaluation	Does the AI tool provide financial value by enhancing care without imposing excessive costs on healthcare systems or patients, ensuring that its implementation is economically sustainable? *This can be measured using one or more of the following methods:* Does the AI tool demonstrate cost-effectiveness by offering a favorable balance between its costs and the health outcomes it achieves, such as life years saved, cases prevented, or symptom-free days, when compared to alternative interventions? Does the AI tool demonstrate cost-utility by providing measurable improvements in both life expectancy and quality of life, quantified through metrics such as quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs)? Does the AI tool demonstrate cost-minimization by achieving equivalent outcomes or effectiveness compared to alternative interventions, while focusing on minimizing overall costs?
Technological safety, and transparency		
	Safety	Does the tool reliably adhere to clinical standards, consistently mitigate potential risks, and

		demonstrate the ability to avoid causing harm to patients through reliable operation and risk management?
	Transparency	Does the AI tool provider ensure transparency by making its processes, decision-making logic, and data sources understandable and accessible to all relevant stakeholders?
	Ethical oversight, human in command (HIC)	Does the AI tool incorporate ethical oversight by ensuring that it supports human decision-making, allowing clinicians to maintain control and override AI-generated decisions, when necessary, thereby complementing rather than replacing human judgment?
Scalability and impact		
	Clinical effectiveness	Does the AI tool demonstrate clinical effectiveness by consistently achieving the desired clinical outcomes in real-world practice, across diverse patient populations and healthcare settings?
	Clinical efficiency	Does the AI tool demonstrate clinical efficiency by optimizing the use of resources, including time, staff, and costs, to effectively deliver care without compromising quality?
	Clinical utility	Does the AI tool demonstrate clinical utility by offering practical benefits that improve patient care, such as guiding clinical decision-making or reducing risks during treatment?
	Environmental impact	Does the AI tool minimize its environmental impact by considering sustainability in its development, deployment, and operation, including factors such as energy consumption and carbon footprint?

Integration

This criteria cluster focuses on evaluating how effectively the AI tool integrates into existing clinical workflows and healthcare systems.

Infrastructure plays a crucial role in the successful implementation of AI tools in healthcare settings. Adequate computational power, specialized hardware, and robust IT infrastructure are often necessary to support the processing of large datasets and the operational demands of AI technologies [46,68]. This may include advanced components like Graphics Processing Units (GPUs), which are not always standard in healthcare systems [52]. Additionally, integrating these tools might require significant investment in new hardware or upgrades [57,66]. For cloud-based AI solutions, attention must be paid to network security and performance [52]. Ensuring infrastructure compatibility is essential for smooth deployment and optimal functionality of AI in healthcare [38,51].

Interoperability ensures seamless integration with existing systems, such as Electronic Health Records (EHRs) and imaging software, it allows AI tools to operate within current workflows without disrupting established clinical processes, enhancing data exchange across platforms [38,65]. It also ensures that AI tools adhere to industry standards, facilitating communication between different healthcare technologies and minimizing issues such as data misinterpretation or workflow inefficiencies [68]. Proper integration can reduce the resource burden on healthcare facilities and improve the overall usability and effectiveness of AI systems in diverse clinical settings [61].

Understanding the impact on **clinical workflows and organizational structures** is essential. AI tools must be seamlessly integrated into workflows to avoid disrupting clinical processes [46,79]. Evaluating how AI affects the redistribution of tasks among healthcare professionals and identifying necessary organizational changes are essential [61,64]. Poor integration or failure to align with clinical routines can negatively impact efficiency, increase cognitive burdens, and require significant resources to adapt systems [42,55].

Monitoring, Governance & Accountability

This criteria cluster focuses on evaluating how effectively the AI tool is monitored throughout its lifecycle, addressing critical aspects such as model drift, data governance, and adherence to ethical standards.

Clarity on **accountability and liability** is essential when assessing AI tools in healthcare due to the potential risks involved in their implementation [46,78]. AI systems can make errors or offer recommendations that may not be followed by clinicians, raising complex questions about who is responsible when mistakes occur [51,75]. The lack of clear guidelines on whether liability lies with the developer, the healthcare institution, or the clinician using the tool poses significant legal and ethical concerns [52,81]. Proper assessment frameworks must ensure that accountability is well-defined, including clear roles for all stakeholders involved (clinicians, developers, and institutions) particularly in cases of adverse events or errors [47,61,67].

Data **security, privacy, informed consent, and data ownership** are vital criteria for assessing AI tools in healthcare. These tools often require large amounts of sensitive patient data, which must be protected from unauthorized access, breaches, or misuse [72,80]. Ensuring compliance with relevant regulations, such as GDPR or HIPAA, is essential to safeguard patient privacy [52,57,68]. Additionally, clear processes for obtaining informed consent are critical, ensuring that patients understand how their data will be used [65,78]. Proper data ownership policies must also be in place, ensuring transparency around who controls the data and how it can be accessed or shared [46,75]. These measures are crucial for building trust and ensuring ethical AI deployment in healthcare settings [51,65].

Regulatory compliance and certification are essential but insufficient assessment criteria for AI tools in healthcare [19]. Although regulatory bodies like the FDA in the US and CE marking in the EU set minimum safety and efficacy standards, there are significant gaps between legal certification and real-world clinical validation, workflow integration, and ongoing use [19,42]. For instance, FDA clearance does not always assure users that an AI tool will meet their expectations for effective performance in all clinical settings, leading to skepticism among healthcare professionals [19,81]. Similarly, in the EU, AI tools with CE marking are often assumed to be clinically validated, but many lack sufficient validation for real-world clinical use, such as in dementia diagnosis via MRI [42,80]. These gaps highlight the need for stronger regulatory frameworks and post-market surveillance to ensure AI tools are not only certified but also thoroughly validated and integrated into healthcare workflows for effective and safe use [19,69,73].

Monitoring and governance mechanisms, including feedback loops, are critical for ensuring the continued safety, effectiveness, performance and reliability of AI tools in healthcare [68]. It is essential that the responsibility for monitoring these tools is shared between the developer, regulator, and the healthcare organization deploying the tool [81]. Developers are responsible for ongoing performance evaluations, including regular updates to address issues such as data drift or algorithmic failure [45,71]. Regulators must ensure compliance with post-market surveillance requirements and set clear guidelines for monitoring practices [57,81]. Healthcare organizations must implement local oversight systems, ensuring that the AI tool continues to meet clinical needs without causing disruption or harm [46,61,62,68,80]. By assigning responsibility to all three entities, healthcare systems can ensure comprehensive, multi-layered oversight that addresses technical, clinical, and regulatory concerns [81].

The **maintenance and updating** of AI tools are critical to ensuring their continued effectiveness and safety in healthcare [68]. Regular updates, including adjustments to algorithms and reference datasets, are essential to avoid performance degradation and ensure accurate results [50,68]. Without proper maintenance, different software versions could introduce biases or inconsistencies, which might affect clinical outcomes [42,73]. Establishing clear protocols for updates, including version control and procedures for managing software changes, ensures that AI tools remain reliable and aligned with current medical standards, safeguarding patient care [47].

Performance Quality Metrics

This criteria cluster focuses on evaluating the performance and quality of the AI tool by assessing key metrics such as foundational performance metrics, fairness, explainability, reliability, and robustness.

Foundational performance metrics play a crucial role in assessing the effectiveness of AI tools, and the systematic review revealed that 59.1% of studies primarily focused on accuracy, sensitivity, and specificity as key metrics. However, it's essential to consider application-specific metrics when evaluating AI performance, as different AI tools require tailored measures depending on their intended use. For example, diagnosis and prediction tools encompass applications like classification (e.g., disease diagnosis), regression (e.g., predicting disease progression), anomaly detection, and recommendation systems. These tools can be assessed through metrics such as accuracy, sensitivity, specificity, and the Area Under the Curve (AUC) for classification tasks [38,40,66], and mean absolute error (MAE) and root mean square error (RMSE) for regression tasks [43,78]. Image & Pattern Analysis covers tasks such as image segmentation and reinforcement learning, using metrics like the Dice coefficient and Jaccard index for segmentation accuracy [84,85], and cumulative reward for evaluating reinforcement learning performance [86]. On the other hand, Text & Language Processing applications, such as natural language processing (NLP) and large language models (LLMs), are assessed using metrics like relevance, engagement, empathy, token limits, hallucination rates, memory efficiency, and floating point operation count (FLOP) [54,59,70]. These metrics ensure the AI tool is properly evaluated based on its intended use and technology type.

Explainability and interpretability are essential for ensuring the ethical and trustworthy use of AI tools in healthcare. These criteria allow healthcare professionals to understand how AI models arrive at their conclusions, fostering trust in their recommendations [44,80]. Explainability helps to demystify the AI's decision-making process, making it transparent and accessible to users [64,65]. This, in turn, improves adoption, as clinicians are more likely to trust and rely on AI tools that are interpretable [46,57]. Ultimately, clear explainability supports ethical deployment, reducing risks associated with "black box" systems [70,75].

Fairness or equity ensures that AI models provide unbiased, consistent performance across diverse demographic groups, including those defined by race, gender, age, or socioeconomic status [59,63,80]. This criterion addresses the risk of bias in training data, including sample size and representativeness, which can lead to unequal treatment or outcomes for underrepresented populations [39,40,68]. By focusing on fairness, AI tools can avoid perpetuating disparities and contribute to more equitable healthcare delivery for all patients [52,60,68].

Reliability, repeatability, and reproducibility ensure that the AI tool can produce consistent outputs when presented with similar inputs, is repeatable under identical conditions, and is reproducible in diverse environments, including different institutions or patient populations [49,52,53,61]. Maintaining consistency and stability is essential for the tool's trustworthiness and its broader applicability in real-world healthcare scenarios [51,81].

Robustness and generalizability are essential criteria for assessing the adaptability of AI tools in healthcare [39,80]. Robustness ensures the tool can maintain high performance even when exposed to slight variations in input data or operational environments [67,80]. Generalizability, on the other hand, evaluates whether the AI tool can effectively perform across different populations, clinical settings, or geographic regions beyond the environment in which it was trained [45,80]. These criteria ensure that AI tools remain reliable and effective when scaled or applied to diverse healthcare contexts [46,51].

Acceptability, Trust, and Training

This criteria cluster evaluates user-centric aspects of the AI tool, focusing on its acceptance, trustworthiness, and the adequacy of user training and support.

User **acceptance and adoption** are crucial for the successful implementation and translation of AI-powered health tools in real-life settings [53,62,79]. Key challenges include fostering trust and confidence among healthcare professionals, ensuring ease of use, and integrating these tools seamlessly into clinical workflows [74]. User acceptance depends significantly on perceived benefits, transparency, and safety of the AI systems [44,46,49]. Moreover, ethical concerns, the potential for bias, and the need for comprehensive testing also impact adoption [64]. Clinicians are more likely to embrace these tools when they complement human expertise and are introduced with adequate training and support, ensuring they enhance patient outcomes without compromising safety [65].

Trust is built through factors such as validation, transparency, safety, privacy, and interpretability of the AI tool [59]. Both healthcare professionals and patients must trust that the AI tool is reliable, safe, and effective in clinical practice [57,65]. Validating AI performance using local data is essential to build clinician confidence, while demonstrating that the tool adheres to rigorous standards helps address concerns about its real-world application [46,81]. Trust also influences adoption, making it vital for the successful implementation of AI tools in healthcare [79].

User centricity emphasizes the need for a clear understanding of the intended users, domain, and specific tasks the AI tool is designed to support [39,55,68]. AI tools must be tailored to meet the unique requirements of their end-users, whether clinicians, nurses, or patients, and address the particular medical conditions they aim to diagnose, monitor, or treat [39,53]. Clarity in defining the tool's intended use, the healthcare domain it serves, and the tasks it performs ensures that it delivers meaningful value in its practical application [61,70].

Usability ensures that the tool is user-friendly and intuitive for both healthcare professionals and patients [46,50]. An AI tool's ease of use and minimal training requirements are essential for successful adoption [44,50]. Usability also impacts user satisfaction, influencing acceptance and trust in the system [49,74]. Proper design should minimize cognitive load, provide relevant information in context, and allow customization by users [68]. Evaluating usability ensures that AI tools can be effectively deployed in real-world clinical environments, enhancing rather than hindering care delivery [52,53].

Adequate **training** ensures that clinicians and other end-users can effectively utilize AI tools, minimizing user error and maximizing the tool's potential to improve patient outcomes [52,62,68]. Training programs should cover how to interact with the AI interface, interpret its outputs, and understand the tool's limitations [46,51]. Continuous education is also crucial, and end users should not only be trained on interpreting the algorithm's output but also made aware of the factors that can affect its performance [61]. Moreover, accessible and responsive technical **support** is necessary to address user concerns, provide ongoing assistance, and maintain confidence in the AI tool's reliability and safety over time [49]. Without proper **training and support**, the integration of AI tools into clinical practice may face significant barriers, limiting their overall effectiveness [42,62].

Cost and Economic Evaluation

This criteria cluster evaluates the economic implications of the AI tool to determine its financial viability and long-term sustainability.

Economic evaluation and cost considerations are crucial in assessing AI tools in healthcare. AI interventions must demonstrate not only clinical value but also health economic impact to ensure their long-term sustainability [73,81]. This includes evaluating both direct costs, such as acquisition, maintenance, and implementation, as well as indirect costs like staff training or workflow disruptions [46,58]. Transparent and comprehensive economic evaluations help healthcare organizations determine the financial viability of AI tools, guiding decision-making on investments, reimbursement, and long-term sustainability [56,61,64,76]. Incomplete or unclear cost assessments can hinder AI adoption and create financial risks [69,76].

The choice of economic evaluation method for an AI tool in healthcare depends on its intended use and desired outcomes. **Cost-effectiveness** analysis (CEA) is useful when comparing costs with health outcomes like life years saved [38,56,65,68]. **Cost-utility** analysis (CUA) is ideal when focusing on both life expectancy and quality of life improvements, measured in QALYs or DALYs [38,58,69]. **Cost-minimization** analysis (CMA) is appropriate when the AI tool achieves similar outcomes as alternatives but aims to reduce costs [58,69,75]. The method chosen should align with the tool's specific goals and intended healthcare impact.

Technological Safety and Transparency

This criteria cluster focuses on evaluating the technological safety and transparency of the AI tool by assessing the safeguards in place to ensure safe and ethical operation.

Safety ensures that AI systems operate reliably and securely in clinical environments beyond laboratory settings and clinical trials [70,75]. This includes compliance with safety regulations, minimizing the risks of harmful outcomes, and maintaining high standards for long-term safety and patient protection [50,64,65]. Safety also encompasses the reliability of the AI model after its implementation, ensuring it consistently avoids errors and unintended consequences [44,62,75]. Ongoing monitoring, risk management, and thorough clinical validation are necessary to ensure that AI tools remain safe and effective in diverse healthcare settings and the long-term safety of constant updates [46,64,65,80].

Transparency is a critical assessment criterion for AI tools in healthcare, ensuring clarity in data processing, coding standards, and the overall functioning of AI systems [68,75,77]. Transparent models allow healthcare professionals to understand how decisions are made, promoting trust and enabling accurate assessments of the AI's performance [60,64,65,70]. Clear documentation and disclosure of data processing methods, coding protocols, and the AI's decision-making processes ensures accountability and reproducibility [39,51,61]. A recent review of 692 FDA-approved AI enabled medical devices highlighted major gaps in transparency and safety reporting [88]. Key data such as ethnicity (reported in only 3.6% of approvals), socioeconomic information (absent in 99.1%), and study subjects' age (missing in 81.6%) were often underreported [88]. Additionally, only 46.1% of devices provided detailed performance results, and just 1.9% linked to scientific publications on safety and efficacy [88]. These findings underscore the urgent need for improved transparency and more comprehensive safety reporting to reduce algorithmic bias and ensure equitable healthcare outcomes.

Ethical oversight and Human-in-Command (HIC) ensure human control and responsibility in AI decision-making processes [61,67]. This criterion emphasizes that humans must retain ultimate authority over AI-generated decisions, particularly in critical healthcare scenarios [65,67]. HIC ensures that clinicians can review, intervene, or override AI decisions, maintaining ethical standards and safeguarding patient outcomes [39,67]. This oversight protects against over-reliance on automated systems and ensures that AI tools support, rather than replace, human judgment in clinical practice [42,65,77].

Scalability and Impact

This criteria cluster focuses on evaluating scalability and impact by determining the AI tool's clinical utility and effectiveness, and examining its broader impact.

Clinical effectiveness focuses on the tool's ability to positively impact patient outcomes [65,68,75]. This involves evaluating whether the AI tool contributes to better therapeutic results or patient-reported outcomes [40,68]. The assessment examines how well the AI integrates into real-world clinical settings and measures its tangible benefits in terms of patient health and healthcare quality [46,66]. Clinical effectiveness ensures that AI tools do more than function technically, they must provide meaningful improvements in patient care [44,48].

Clinical utility focuses on how effectively the tool supports clinical tasks and decision-making, including its ability to assist with diagnoses, treatment recommendations, and overall healthcare delivery [48,79]. Ensuring clinical utility means the AI tool must provide tangible benefits that align with clinical needs and enhance healthcare practices [40,76]. While **clinical efficiency** focuses on the tool's ability to optimize resource use while maintaining or improving care quality [65]. This includes evaluating how well it improves productivity, reduces time spent on routine tasks, and streamlines workflows for healthcare professionals [44,50,52].

Environmental impact is an important, yet often overlooked, criterion for assessing AI tools in healthcare; only one out of 44 studies addressed this criterion. The energy consumption and resource use associated with developing, deploying, and maintaining AI systems, such as data centers, computational power, and device infrastructures, can lead to significant environmental harm, including e-waste and greenhouse gas emissions [72]. Implementing eco-responsible practices, such as energy-efficient computing and sustainable data storage, is essential to minimizing the ecological footprint of AI tools [72].

Practical Implications and Persisting Challenges

The wide array of frameworks and initiatives focused on AI assessment in healthcare shown in this systematic review highlights the significant lack of standardization in this field, creating additional challenges for stakeholders [40,67,68]. Faced with a growing number of assessment tools, they often struggle to determine which approach is most appropriate or how to apply it effectively [60]. This diversity in assessment methods can lead to confusion and hinder comparability [40,65,66,76]. Variations in data collection and evaluation methods, ranging from self-reported to objective measures, and from qualitative to quantitative assessments, only add to the complexity, further complicating the establishment of clear, universal guidelines for AI evaluation in healthcare [59].

The majority of frameworks included in this analysis were driven by the recognition that many existing methods for assessing AI tools in healthcare were not specifically tailored to AI-based medical devices or healthcare applications [59,62,64]. Traditional technology assessments often lack a critical focus on the unique, dynamic challenges and opportunities AI presents [54]. This underscores the need for healthcare-specific frameworks that account for the evolving nature and complexities of AI systems in clinical environments [57]. Moreover, existing frameworks tend to prioritize technical metrics such as algorithm accuracy, precision, and validation [59,70]. While these factors are undeniably important, this narrow focus often overlooks broader considerations, including clinical relevance, practical application, and long-term impact on patient outcomes [38,53,79]. Consequently, these frameworks can fall short in delivering a holistic evaluation of AI tools, which is essential for ensuring their safe, effective, and seamless integration into real-world healthcare settings [44,80].

This study builds upon and advances the ongoing discussion in AI assessment in healthcare, aiming to address the recognized gaps by developing the AI for IMPACTS framework. This proposed framework integrates technical, social, and organizational dimensions, ensuring that the adaptive nature of AI and the complexity of the healthcare ecosystem are fully considered. By encompassing these critical aspects, the framework provides a more comprehensive and nuanced approach to evaluating AI tools, helping shape the field and offering a robust method for assessing AI's real-world impact in healthcare settings.

However, numerous challenges still remain. These challenges extend beyond just setting the assessment criteria, to include practical difficulties in implementing, validating, and standardizing these criteria across diverse healthcare environments. A key challenge in assessing AI tools in healthcare is the variation across different contexts and settings [61,68]. The majority of available evidence focuses on developed countries, limiting the generalizability of findings to diverse healthcare environments, particularly in low- and middle-income countries [40,46]. Recent studies underscore the importance of collaborative efforts and context-sensitive solutions to effectively address the unique healthcare challenges faced in these regions [89]. Another challenge is the need for a multidisciplinary team of assessors. Effective evaluation requires collaboration among professionals from various fields, such as medicine, information technology, and social sciences to ensure a comprehensive assessment [74,80]. This diversity of expertise is necessary to address the complexities of AI, from technical and ethical considerations to clinical relevance and real-world impact [52,62,81].

It is crucial to emphasize the importance of adequate training in assessment methods [66,71]. Many assessors may lack the specific expertise required to thoroughly evaluate AI-based tools [43]. Proper training in the complexities of AI technology and appropriate evaluation techniques is essential for conducting accurate and meaningful assessments [52]. Without this, the assessment process may be compromised, potentially leading to inaccurate or incomplete evaluations of an AI tool's safety and effectiveness, which could undermine its implementation in healthcare settings [65]. Furthermore, the rapid pace of AI development, with AI-based medical devices having shorter product lifecycles compared to traditional medical devices, underscores the need for more adaptive and fast-tracked Health Technology Assessment (HTA) processes [46,65]. Conventional HTAs are often too time-consuming, taking about a year to complete, which is incompatible with the fast-evolving nature of AI technologies [56]. Balancing the need for robust evidence with the dynamic nature of AI development is essential to ensure timely, informed decision-making, and avoid delays in implementation and potential reimbursement [56,65].

Limitations and Future Research

This study enhances the understanding of various criteria for assessing the quality and impact of AI tools in healthcare, but several limitations must be acknowledged. Relevant studies may have been missed due to language restrictions or limited database searches, and the exclusion of gray literature may have omitted valuable insights. Additionally, no follow-up was conducted with the study authors to validate the findings, and manual reference searches were avoided to minimize citation bias. As a result, some relevant frameworks or assessment criteria may not have been captured in this review. Future research could expand to include studies in other languages, offering a more comprehensive understanding of potential interregional or intercultural differences in the assessment of AI tools in healthcare.

The critical appraisal of the frameworks included in this review highlighted that many papers discussing AI tool assessment in healthcare lacked rigorous validation, with some omitting methods sections entirely. To address this gap, the authors propose rigorously validating the AI for IMPACTS framework proposed in this work through a Delphi process. This approach will involve key stakeholders to critically apply, reflect on, and refine the framework, ensuring it is relevant, comprehensive, and user-friendly. The goal is to co-create practical, accessible tools with industry experts that can support the effective evaluation of AI tools in real-world healthcare settings.

It is also important to highlight that new frameworks were published after the cutoff date of this systematic review, including the Organizational PerspecTIve Checklist for AI Adoption (OPTICA) [90], Stanford's Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems (FURM) [91], and the Transparent Reporting of Ethics for Generative AI (TREGAI) checklist [92]. While an initial review shows that their assessment dimensions align with this work, a deeper integration will be undertaken prior to the validation study. This will ensure that the foundation for the Delphi process is as comprehensive and up-to-date as possible.

Conclusions

AI has the potential to transform healthcare by improving clinical outcomes and operational efficiency. However, its adoption has been slower than expected due to the lack of comprehensive evaluation frameworks. Existing frameworks often focus too narrowly on technical metrics, such as accuracy and validation, neglecting real-world factors like clinical impact, workflow integration, and economic viability. Furthermore, the variety of frameworks and initiatives focused on AI assessment in healthcare, as highlighted in this systematic review, underscores a significant lack of standardization in the field, creating additional challenges for stakeholders, making it difficult to compare and implement AI tools effectively.

This study builds on and advances the ongoing discussion surrounding AI assessment in healthcare by developing the AI for IMPACTS framework. It aims to address key gaps identified in existing evaluation approaches, offering a comprehensive model that incorporates technical, social, and organizational dimensions. It is organized around seven key criteria clusters (a) integration, interoperability and workflow; (b) monitoring, governance, and accountability; (c) performance and quality metrics; (d) acceptability, trust, and training; (e) cost and economic evaluation; (f) technological safety and transparency; (g) scalability and impact.

While the framework provides a more holistic approach, significant challenges persist. The diverse contexts and settings in healthcare make it difficult to apply a one-size-fits-all framework. Multidisciplinary teams are necessary to evaluate AI tools thoroughly, as expertise from fields like medicine, IT, and social sciences is required to address the complexities of AI. Additionally, many assessors lack the specific training needed to evaluate these tools accurately. The rapid pace of AI development further complicates the assessment process, as conventional evaluation methods are often too slow to keep up with AI's short product lifecycles. To ensure successful AI integration in healthcare, adaptive and fast-tracked assessment processes are essential, allowing for timely decision-making and implementation while maintaining the necessary rigor.

References

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24–29. doi: 10.1038/s41591-018-0316-z
2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 6;3(1):17. doi: 10.1038/s41746-020-0221-y
3. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and Surgical

Applications of Large Language Models: A Systematic Review. J Clin Med 2024 May 22;13(11):3041. doi: 10.3390/jcm13113041

4. Xu C, Solomon SA, Gao W. Artificial intelligence-powered electronic skin. Nat Mach Intell 2023 Dec 18;5(12):1344–1355. doi: 10.1038/s42256-023-00760-z
5. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health 2021 Mar;3(3):e195–e203. doi: 10.1016/S2589-7500(20)30292-2
6. Zhang J, Whebell S, Gallifant J, Budhdeo S, Mattie H, Lertvittayakumjorn P, del Pilar Arias Lopez M, Tiangco BJ, Gichoya JW, Ashrafian H, Celi LA, Teo JT. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. Lancet Digit Health 2022 Apr;4(4):e212–e213. doi: 10.1016/S2589-7500(22)00032-2
7. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021 Apr 5;27(4):582–584. doi: 10.1038/s41591-021-01312-x
8. Brady AP, Visser J, Frija G, Bargalló N, Rockall A, Brkljacic B, Fuchsjäger M, Birch J, Becker M, Kröncke T. Value-based radiology: what is the ESR doing, and what should we do in the future? Insights Imaging 2021 Dec 27;12(1):108. doi: 10.1186/s13244-021-01056-9
9. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, Tyskbo D, Svedberg P. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res 2022 Dec 1;22(1):850. doi: 10.1186/s12913-022-08215-8
10. United States Government Accountability Office. Technology assess- ment: artificial intelligence in health care—benefits and challenges of technologies to augment patient care. 2023 Nov.
11. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, Ashrafian H, Beam AL, Collins GS, Darzi A, Deeks JJ, ElZarrad MK, Espinoza C, Esteva A, Faes L, Ferrante di Ruffano L, Fletcher J, Golub R, Harvey H, Haug C, Holmes C, Jonas A, Keane PA, Kelly CJ, Lee AY, Lee CS, Manna E, Matcham J, McCradden M, Moher D, Monteiro J, Mulrow C, Oakden-Rayner L, Paltoo D, Panico MB, Price G, Rowley S, Savage R, Sarkar R, Vollmer SJ, Yau C. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020 Oct;2(10):e549–e560. doi: 10.1016/S2589-7500(20)30219-3
12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Ashrafian H, Beam AL, Chan A-W, Collins GS, Deeks ADJ, ElZarrad MK, Espinoza C, Esteva A, Faes L, Ferrante di Ruffano L, Fletcher J, Golub R, Harvey H, Haug C, Holmes C, Jonas A, Keane PA, Kelly CJ, Lee AY, Lee CS, Manna E, Matcham J, McCradden M, Monteiro J, Mulrow C, Oakden-Rayner L, Paltoo D, Panico MB, Price G, Rowley S, Savage R, Sarkar R, Vollmer SJ, Yau C. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020 Oct;2(10):e537–e548. doi: 10.1016/S2589-7500(20)30218-1
13. Nsoesie EO. Evaluating Artificial Intelligence Applications in Clinical Settings. JAMA Netw Open 2018 Sep 28;1(5):e182658. doi: 10.1001/jamanetworkopen.2018.2658
14. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13(1):1. doi: 10.1186/s12916-014-0241-z
15. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R, Kotter E, Gichoya JW, Cook TS, Morgan MB, Tang A, Safdar NM, Kohli M. Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement. Canadian Association of Radiologists Journal 2019 Nov 29;70(4):329–334. doi: 10.1016/j.carj.2019.08.010
16. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, Halpern EF, Hess CP, Schiebler ML, Weiss CR. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the *Radiology* Editorial Board. Radiology 2020 Mar;294(3):487–489. doi: 10.1148/radiol.2019192515

17. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 2021 Jun 15;31(6):3797–3804. doi: 10.1007/s00330-021-07892-z
18. Stettinger G, Weissensteiner P, Khastgir S. Trustworthiness Assurance Assessment for High-Risk AI-Based Systems. IEEE Access 2024;12:22718–22745. doi: 10.1109/ACCESS.2024.3364387
19. Chouffani El Fassi S, Abdullah A, Fang Y, Natarajan S, Masroor A Bin, Kayali N, Prakash S, Henderson GE. Not all AI health tools with regulatory authorization are clinically validated. Nat Med 2024 Aug 26; doi: 10.1038/s41591-024-03203-3
20. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, Moons K, Collins G, Moher D, Bossuyt PM, Darzi A, Karthikesalingam A, Denniston AK, Mateen BA, Ting D, Treanor D, King D, Greaves F, Godwin J, Pearson-Stuttard J, Harling L, McInnes M, Rifai N, Tomasev N, Normahani P, Whiting P, Aggarwal R, Vollmer S, Markar SR, Panch T, Liu X. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 2021 Jun 28;11(6):e047709. doi: 10.1136/bmjopen-2020-047709
21. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020 Mar 1;2(2):e200029. doi: 10.1148/ryai.2020200029
22. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B, van Smeden M, Riley RD, Moons KG. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul;11(7):e048008. doi: 10.1136/bmjopen-2020-048008
23. WHO. Ethics and governance of artificial intelligence for health. Geneva; 2021 Jun. Available from: https://www.who.int/publications/i/item/9789240029200 [accessed Oct 2, 2024]
24. Jacob C, Lindeque J, Klein A, Ivory C, Heuss S, Peter MK. Assessing the Quality and Impact of eHealth Tools: Systematic Literature Review and Narrative Synthesis. JMIR Hum Factors 2023 Mar 23;10:e45143. doi: 10.2196/45143
25. Jacob C, Lindeque J, Müller R, Klein A, Metcalfe T, Connolly SL, Koerber F, Maguire R, Denis F, Heuss SC, Peter MK. A sociotechnical framework to assess patient-facing eHealth tools: results of a modified Delphi process. NPJ Digit Med 2023 Dec 15;6(1):232. doi: 10.1038/s41746-023-00982-w
26. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 2009 Jul 21;6(7):e1000097. doi: 10.1371/journal.pmed.1000097
27. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, Page V. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Wiley; 2019.
28. Jacob C. AI-powered Clinician Tools Assessment Criteria: Protocol of a systematic review of the literature. https://www.researchregistry.com/browse-the-registry#registryofsystematicreviewsmeta-analyses/registryofsystematicreviewsmeta-analysesdetails/669754107bdc5b002704adbe/. 2024.
29. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016 Dec 5;5(1):210. doi: 10.1186/s13643-016-0384-4
30. CASP. Critical Appraisal Skills Programme Checklists. https://casp-uk.net/casp-tools-checklists/. 2022.
31. Leonardi PM. Methodological Guidelines for the Study of Materiality and Affordances. Routledge Companion to Qualitative Research in Organization Studies 2017. p. 279–290.
32. Ammenwerth E. Technology Acceptance Models in Health Informatics: TAM and UTAUT. Stud Health Technol Inform 2019 Jul;
33. Shachak A, Kuziemsky C, Petersen C. Beyond TAM and UTAUT: Future directions for HIT implementation research. J Biomed Inform 2019 Dec;100:103315. doi: 10.1016/j.jbi.2019.103315
34. Jacob C, Sanchez-Vazquez A, Ivory C. Understanding Clinicians' Adoption of Mobile Health Tools: A Qualitative Review of the Most Used Frameworks. JMIR Mhealth Uhealth 2020 Jul

6;8(7):e18072. doi: 10.2196/18072

35. Jacob C, Sanchez-Vazquez A, Ivory C. Social, Organizational, and Technological Factors Impacting Clinicians' Adoption of Mobile Health Tools: Systematic Literature Review. JMIR Mhealth Uhealth 2020 Feb 20;8(2):e15935. doi: 10.2196/15935
36. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical Factors Affecting Patients' Adoption of Mobile Health Tools: Systematic Literature Review and Narrative Synthesis. JMIR Mhealth Uhealth 2022 May 5;10(5):e36284. doi: 10.2196/36284
37. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide For Beginners. SAGE Publications Ltd; 2013. doi: 10.1002/jmr.2361
38. Boverhof B-J, Redekop WK, Bos D, Starmans MPA, Birch J, Rockall A, Visser JJ. Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. Insights Imaging 2024 Feb 5;15(1):34. doi: 10.1186/s13244-023-01599-z
39. Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, Guitera P, Gutman D, Halpern A, Helba B, Kittler H, Kose K, Liopyris K, Malvehy J, Seog HS, Soyer HP, Tkaczyk ER, Tschandl P, Rotemberg V. Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology. JAMA Dermatol 2022 Jan 1;158(1):90. doi: 10.1001/jamadermatol.2021.4915
40. Di Bidino R, Piaggio D, Andellini M, Merino-Barbancho B, Lopez-Perez L, Zhu T, Raza Z, Ni M, Morrison A, Borsci S, Fico G, Pecchia L, Iadanza E. Scoping Meta-Review of Methods Used to Assess Artificial Intelligence-Based Medical Devices for Heart Failure. Bioengineering 2023 Sep 22;10(10):1109. doi: 10.3390/bioengineering10101109
41. Elvidge J, Hawksworth C, Avşar TS, Zemplenyi A, Chalkidou A, Petrou S, Petykó Z, Srivastava D, Chandra G, Delaye J, Denniston A, Gomes M, Knies S, Nousios P, Siirtola P, Wang J, Dawoud D, Arbour S, Asche C, Ashurst C, Balkanyi L, Bennett H, Boros G, Boyce R, Carswell C, Chaiyakunapruk N, Chhatwal J, Ciani O, Collins G, Dawson D, Vanness D, Di Bidino R, Faulding S, Felizzi F, Haig M, Hawkins J, Hiligsmann M, Holst-Kristensen AW, Isla J, Koffijberg E, Kostyuk A, Krief N, Lee D, Lee K, Lundin D, Markiewicz-Barreaux K, Mauskopf J, Moons K, Németh B, Petrova G, Pwu R-F (Jasmine), Rejon-Parrilla JC, Rogers G, Sampson C, Springborg AA, Steuten L, Sutherland E, Suutala J, Theisen D, Thompson A, van Gemert-Pijnen L, Walker T, Wilson E. Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI). Value in Health 2024 Sep;27(9):1196–1205. doi: 10.1016/j.jval.2024.05.006
42. Haller S, Van Cauter S, Federau C, Hedderich DM, Edjlali M. The R-AI-DIOLOGY checklist: a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology. Neuroradiology 2022 May 31;64(5):851–864. doi: 10.1007/s00234-021-02890-w
43. Handelman GS, Kok HK, Chandra R V., Razavi AH, Huang S, Brooks M, Lee MJ, Asadi H. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. American Journal of Roentgenology 2019 Jan;212(1):38–43. doi: 10.2214/AJR.18.20224
44. Jackson GP, Vergis R. Evaluation of Artificial Intelligence in Radiation Oncology. Artificial Intelligence in Radiation Oncology WORLD SCIENTIFIC; 2023. p. 359–368. doi: 10.1142/9789811263545_0016
45. Jha AK, Bradshaw TJ, Buvat I, Hatt M, KC P, Liu C, Obuchowski NF, Saboury B, Slomka PJ, Sunderland JJ, Wahl RL, Yu Z, Zuehlsdorff S, Rahmim A, Boellaard R. Nuclear Medicine and Artificial Intelligence: Best Practices for Evaluation (the RELAINCE Guidelines). Journal of Nuclear Medicine 2022 Sep;63(9):1288–1299. doi: 10.2967/jnumed.121.263239
46. Khan SD, Hoodbhoy Z, Raja MHR, Kim JY, Hogg HDJ, Manji AAA, Gulamali F, Hasan A, Shaikh A, Tajuddin S, Khan NS, Patel MR, Balu S, Samad Z, Sendak MP. Frameworks for procurement, integration, monitoring, and evaluation of artificial intelligence tools in clinical settings: A systematic review. PLOS Digital Health 2024 May 29;3(5):e0000514. doi: 10.1371/journal.pdig.0000514
47. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms:

Summary and Recommendations. Journal of the American College of Radiology 2021 Mar;18(3):413–424. doi: 10.1016/j.jacr.2020.09.060

48. Lennerz JK, Salgado R, Kim GE, Sirintrapun SJ, Thierauf JC, Singh A, Indave I, Bard A, Weissinger SE, Heher YK, de Baca ME, Cree IA, Bennett S, Carobene A, Ozben T, Ritterhouse LL. Diagnostic quality model (DQM): an integrated framework for the assessment of diagnostic quality when using AI/ML. Clinical Chemistry and Laboratory Medicine (CCLM) 2023 Mar 28;61(4):544–557. doi: 10.1515/cclm-2022-1151
49. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, Rigby M, Scott PJ, Vehko T, Wong ZS-Y, Georgiou A. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. Yearb Med Inform 2019 Aug 25;28(01):128–134. doi: 10.1055/s-0039-1677903
50. Mahadevaiah G, RV P, Bermejo I, Jaffray D, Dekker A, Wee L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. Med Phys 2020 May 17;47(5). doi: 10.1002/mp.13562
51. Mahmood U, Shukla-Dave A, Chan H-P, Drukker K, Samala RK, Chen Q, Vergara D, Greenspan H, Petrick N, Sahiner B, Huo Z, Summers RM, Cha KH, Tourassi G, Deserno TM, Grizzard KT, Näppi JJ, Yoshida H, Regge D, Mazurchuk R, Suzuki K, Morra L, Huisman H, Armato SG, Hadjiiski L. Artificial intelligence in medicine: mitigating risks and maximizing benefits via quality assurance, quality control, and acceptance testing. BJR|Artificial Intelligence 2024 Mar 4;1(1). doi: 10.1093/bjrai/ubae003
52. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE, Louvet-de Verchère F, Pinto Dos Santos D, Kober T, Richiardi J. To buy or not to buy—evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol 2021 Jun 5;31(6):3786–3796. doi: 10.1007/s00330-020-07684-x
53. Reddy S, Rogers W, Makinen V-P, Coiera E, Brown P, Wenzel M, Weicken E, Ansari S, Mathur P, Casey A, Kelly B. Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform 2021 Oct 12;28(1):e100444. doi: 10.1136/bmjhci-2021-100444
54. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, Bergonzani M, Boscolo-Rizzo P, Califano G, Cammaroto G, Chiesa-Estomba CM, Committeri U, Crimi S, Curran NR, di Bello F, di Stadio A, Frosolini A, Gabriele G, Gengler IM, Lonardi F, Maglitto F, Mayo-Yáñez M, Petrocelli M, Pucci R, Saibene AM, Saponaro G, Tel A, Trabalzini F, Trecca EMC, Vellone V, Salzano G, De Riu G. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. European Archives of Oto-Rhino-Laryngology 2024 May 4; doi: 10.1007/s00405-024-08710-0
55. van Royen FS, Asselbergs FW, Alfonso F, Vardas P, van Smeden M. Five critical quality criteria for artificial intelligence-based prediction models. Eur Heart J 2023 Dec 7;44(46):4831–4834. doi: 10.1093/eurheartj/ehad727
56. Vervoort D, Tam DY, Wijeysundera HC. Health Technology Assessment for Cardiovascular Digital Health Technologies and Artificial Intelligence: Why Is It Different? Canadian Journal of Cardiology 2022 Feb;38(2):259–266. doi: 10.1016/j.cjca.2021.08.015
57. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, Grainger D, Birse M, Branson R, Moons KGM, Collins GS, Ioannidis JPA, Holmes C, Hemingway H. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;l6927. doi: 10.1136/bmj.l6927
58. Vithlani J, Hawksworth C, Elvidge J, Ayiku L, Dawoud D. Economic evaluations of artificial intelligence-based healthcare interventions: a systematic literature review of best practices in their conduct and reporting. Front Pharmacol 2023 Aug 8;14. doi: 10.3389/fphar.2023.1220950
59. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, Sriram R, Yang Z, Wang Y, Lin B, Gevaert O, Li L-J, Jain R, Rahmani AM. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar

29;7(1):82. doi: 10.1038/s41746-024-01074-z
60. Economou-Zavlanos NJ, Bessias S, Cary MP, Bedoya AD, Goldstein BA, Jelovsek JE, O’Brien CL, Walden N, Elmore M, Parrish AB, Elengold S, Lytle KS, Balu S, Lipkin ME, Shariff AI, Gao M, Leverenz D, Henao R, Ming DY, Gallagher DM, Pencina MJ, Poon EG. Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. Journal of the American Medical Informatics Association 2024 Feb 16;31(3):705–713. doi: 10.1093/jamia/ocad221
61. Larson DB, Doo FX, Allen B, Mongan J, Flanders AE, Wald C. Proceedings From the 2022 ACR-RSNA Workshop on Safety, Effectiveness, Reliability, and Transparency in AI. Journal of the American College of Radiology 2024 Jul;21(7):1119–1129. doi: 10.1016/j.jacr.2024.01.024
62. Overgaard SM, Graham MG, Brereton T, Pencina MJ, Halamka JD, Vidal DE, Economou-Zavlanos NJ. Implementing quality management systems to close the AI translation gap and facilitate safe, ethical, and effective health AI solutions. NPJ Digit Med 2023 Nov 25;6(1):218. doi: 10.1038/s41746-023-00968-8
63. Schaekermann M, Spitz T, Pyles M, Cole-Lewis H, Wulczyn E, Pfohl SR, Martin D, Jaroensri R, Keeling G, Liu Y, Farquhar S, Xue Q, Lester J, Hughes C, Strachan P, Tan F, Bui P, Mermel CH, Peng LH, Matias Y, Corrado GS, Webster DR, Virmani S, Semturs C, Liu Y, Horn I, Cameron Chen P-H. Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study. EClinicalMedicine 2024 Apr;70:102479. doi: 10.1016/j.eclinm.2024.102479
64. Farah L, Davaze-Schneider J, Martin T, Nguyen P, Borget I, Martelli N. Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review. Artif Intell Med 2023 Jun;140:102547. doi: 10.1016/j.artmed.2023.102547
65. Farah L, Borget I, Martelli N, Vallee A. Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices: Scoping Literature Review. J Med Internet Res 2024 May 13;26:e51514. doi: 10.2196/51514
66. Guenoun D, Zins M, Champsaur P, Thomassin-Naggara I. French community grid for the evaluation of radiological artificial intelligence solutions (DRIM France Artificial Intelligence Initiative). Diagn Interv Imaging 2024 Feb;105(2):74–81. doi: 10.1016/j.diii.2023.09.002
67. Bimczok SP, Godynyuk EA, Pierey J, Roppel MS, Scholz ML. How are excellence and trust for using artificial intelligence ensured? Evaluation of its current use in EU healthcare. South East Eur J Public Health 2023;
68. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, Aardoom JJ, Debray TPA, Schuit E, van Smeden M, Reitsma JB, Steyerberg EW, Chavannes NH, Moons KGM. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2. doi: 10.1038/s41746-021-00549-7
69. Voets MM, Veltman J, Slump CH, Siesling S, Koffijberg H. Systematic Review of Health Economic Evaluations Focused on Artificial Intelligence in Healthcare: The Tortoise and the Cheetah. Value in Health 2022 Mar;25(3):340–349. doi: 10.1016/j.jval.2021.11.1362
70. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. Journal of the American Medical Informatics Association 2023 Dec 9; doi: 10.1093/jamia/ocad222
71. Goergen SK, Frazer HM, Reddy S. Quality use of artificial intelligence in medical imaging: What do radiologists need to know? J Med Imaging Radiat Oncol 2022 Mar 3;66(2):225–232. doi: 10.1111/1754-9485.13379
72. Lehoux P, Rocha de Oliveira R, Rivard L, Silva HP, Alami H, Mörch CM, Malas K. A Comprehensive, Valid, and Reliable Tool to Assess the Degree of Responsibility of Digital Health Solutions That Operate With or Without Artificial Intelligence: 3-Phase Mixed Methods Study. J Med Internet Res 2023 Aug 28;25:e48496. doi: 10.2196/48496

73. Tanguay W, Acar P, Fine B, Abdolell M, Gong B, Cadrin-Chênevert A, Chartrand-Lefebvre C, Chalaoui J, Gorgos A, Chin AS-L, Prénovault J, Guilbert F, Létourneau-Guillon L, Chong J, Tang A. Assessment of Radiology Artificial Intelligence Software: A Validation and Evaluation Framework. Canadian Association of Radiologists Journal 2023 May 6;74(2):326–333. doi: 10.1177/08465371221135760
74. Ji M, Genchev GZ, Huang H, Xu T, Lu H, Yu G. Evaluation Framework for Successful Artificial Intelligence–Enabled Clinical Decision Support Systems: Mixed Methods Study. J Med Internet Res 2021 Jun 2;23(6):e25929. doi: 10.2196/25929
75. Fasterholdt I, Naghavi-Behzad M, Rasmussen BSB, Kjølhede T, Skjøth MM, Hildebrandt MG, Kidholm K. Value assessment of artificial intelligence in medical imaging: a scoping review. BMC Med Imaging 2022 Oct 31;22(1):187. doi: 10.1186/s12880-022-00918-y
76. Gomez Rossi J, Feldberg B, Krois J, Schwendicke F. Evaluation of the Clinical, Technical, and Financial Aspects of Cost-Effectiveness Analysis of Artificial Intelligence in Medicine: Scoping Review and Framework of Analysis. JMIR Med Inform 2022 Aug 12;10(8):e33703. doi: 10.2196/33703
77. Panagoulias DP, Virvou M, Tsihrintzis GA. Applying DOI Theory to Assess the Required Level of Explainability in Artificial Intelligence-empowered Medical Applications. 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA) IEEE; 2023. p. 1–7. doi: 10.1109/IISA59645.2023.10345846
78. Bhatnagar S. Checklist for Medical Imaging using Artificial Intelligence by Evaluation of Machine Learning Models. 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) IEEE; 2023. p. 865–871. doi: 10.1109/ICIRCA57980.2023.10220939
79. Alshehri S, Alahmari KA, Alasiry A. A Comprehensive Evaluation of AI-Assisted Diagnostic Tools in ENT Medicine: Insights and Perspectives from Healthcare Professionals. J Pers Med 2024 Mar 28;14(4):354. doi: 10.3390/jpm14040354
80. Lundström C, Lindvall M. Mapping the Landscape of Care Providers' Quality Assurance Approaches for AI in Diagnostic Imaging. J Digit Imaging 2022 Nov 9;36(2):379–387. doi: 10.1007/s10278-022-00731-7
81. Ross J, Hammouche S, Chen Y, Rockall AG, Alabed S, Chen M, Dwivedi K, Fascia D, Greenhalgh R, Hall M, Halliday K, Harden S, Ramsden W, Shelmerdine S. Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment. Clin Radiol 2024 May;79(5):338–345. doi: 10.1016/j.crad.2024.01.026
82. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Research Methods in Medicine & Health Sciences 2020 Sep 6;1(1):31–42. doi: 10.1177/2632084320947559
83. Unsworth H, Dillon B, Collinson L, Powell H, Salmon M, Oladapo T, Ayiku L, Shield G, Holden J, Patel N, Campbell M, Greaves F, Joshi I, Powell J, Tonnel A. The NICE Evidence Standards Framework for digital health and care technologies – Developing and maintaining an innovative evidence framework with global impact. Digit Health 2021 Jan 24;7:205520762110186. doi: 10.1177/20552076211018617
84. Sarwar N, Irshad A, Naith QH, D.Alsufiani K, Almalki FA. Skin lesion segmentation using deep learning algorithm with ant colony optimization. BMC Med Inform Decis Mak 2024 Sep 27;24(1):265. doi: 10.1186/s12911-024-02686-x
85. Zubair M, Owais M, Mahmood T, Iqbal S, Usman SM, Hussain I. Enhanced gastric cancer classification and quantification interpretable framework using digital histopathology images. Sci Rep 2024 Sep 28;14(1):22533. doi: 10.1038/s41598-024-73823-9
86. Bourdillon AT, Garg A, Wang H, Woo YJ, Pavone M, Boyd J. Integration of Reinforcement Learning in a Virtual Robotic Surgical Simulation. Surg Innov 2023 Feb 3;30(1):94–102. doi: 10.1177/15533506221095298
87. deepchecks. DeepChecks Glossary. https://deepchecks.com/glossary/. 2024. Available from:

https://deepchecks.com/glossary/ [accessed Oct 2, 2024]

88. Muralidharan V, Adewale BA, Huang CJ, Nta MT, Ademiju PO, Pathmarajah P, Hang MK, Adesanya O, Abdullateef RO, Babatunde AO, Ajibade A, Onyeka S, Cai ZR, Daneshjou R, Olatunji T. A scoping review of reporting gaps in FDA-approved AI medical devices. NPJ Digit Med 2024 Oct 3;7(1):273. doi: 10.1038/s41746-024-01270-x
89. Yang J, Dung NT, Thach PN, Phong NT, Phu VD, Phu KD, Yen LM, Thy DBX, Soltan AAS, Thwaites L, Clifton DA. Generalizability assessment of AI models across hospitals in a low-middle and high income country. Nat Commun 2024 Sep 27;15(1):8270. doi: 10.1038/s41467-024-52618-6
90. Dagan N, Devons-Sberro S, Paz Z, Zoller L, Sommer A, Shaham G, Shahar N, Ohana R, Weinstein O, Netzer D, Kotler A, Balicer RD. Evaluation of AI Solutions in Health Care Organizations — The OPTICA Tool. NEJM AI 2024 Aug 22;1(9). doi: 10.1056/AIcs2300269
91. Callahan A, McElfresh D, Banda JM, Shah NH. Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. NEJM Catal 2024 Oct;5(10).
92. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, Mertens M, Ting DSW, Ong JCL, Liu M, Cao J, Dunn M, Vaughan R, Ong MEH, Sung JJ-Y, Topol EJ, Liu N. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. Lancet Digit Health 2024 Sep; doi: 10.1016/S2589-7500(24)00143-2

Supplementary Files

Multimedia Appendixes

CASP Appraisal of the Included Studies.
URL: http://asset.jmir.pub/assets/058c14988fba5a6fd781876275388719.xlsx