CLARITY AI: A Comprehensive Checklist Integrating Established Frameworks for Enhanced Research Quality in Medical AI Studies

Luca Marconi[1,*,†], Efrem Pirovano[1,†] and Federico Cabitza[1,2]

[1]University of Milano-Bicocca, 20126 Milan, Italy
[2]IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy

Abstract

The medical field is constantly evolving, integrating the latest technologies to enhance patient care and treatment efficacy. While various methodologies are available to evaluate the quality of research studies, checklists are often favored for their efficiency and ease of use. In this study, we contribute to this area of research by 1) analyzing the components of the most widely used checklists, and 2) proposing a more comprehensive checklist, CLARITY AI, which synthesizes the strengths of existing tools.

This study analyzed several established checklists—CLAIM, CONSORT, DECIDE, FUTURE, IJMEDI, PRISMA, SPIRIT, STARD, STARE-HI, and TRIPOD—with the goal of developing a comprehensive checklist for evaluating research studies. Each item in these checklists was carefully cataloged, labeled, and assessed. The analysis aimed to identify the most critical items for inclusion in a definitive checklist for research study evaluation.

The final version of the checklist is a coherent integration of structural elements—such as Title, Abstract, and Introduction—and essential parameters like Study Identification and Data Handling. This synthesis results in a comprehensive tool for thorough study and research evaluation.

By integrating the strengths of multiple established checklists, CLARITY offers a robust, systematic, and user-friendly framework for assessing research quality. This tool not only elevates research standards but also enhances transparency, reproducibility, and overall credibility in the field of medical AI studies. Its application has the potential to produce more reliable and effective healthcare solutions, ultimately improving patient outcomes and advancing medical research.

Keywords

AI in Healthcare, Research Evaluation, CLARITY Framework, Medical AI Studies, Reproducibility

1. Introduction

The integration of artificial intelligence (AI) into healthcare is transforming medical research and practice, offering significant opportunities to improve diagnostic accuracy, treatment personalization, and enhance patient outcomes. Despite these advancements, the rapid proliferation of AI technologies also brings critical challenges, particularly in ensuring that AI research adheres to rigorous standards of quality, transparency, and reproducibility.

Existing frameworks, like CLAIM[1][2][3], CONSORT[4][5][6], PRISMA[7][8][9] and OPTICA[10] address specific aspects of AI research but often lack the breadth required to evaluate the full complexity of AI in healthcare. The absence of a unified, adaptable framework hinders the reliable integration of AI technologies in clinical practice.

A critical gap in current evaluation methodologies is their limited scope, often neglecting key dimensions such as ethical considerations, data management, and usability—factors that are essential for the safe and effective deployment of AI in healthcare environments. Existing checklists such as TRIPOD[11][12] and STARE-HI[13], though valuable, do not adequately account for the iterative nature

3rd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2024), 25-28 November 2024, Bolzano, Italy
*Corresponding author.
†These authors contributed equally.
luca.marconi@unimib.it (L. Marconi); e.pirovano8@campus.unimib.it (E. Pirovano); federico.cabitza@unimib.it (F. Cabitza)
0000-0002-0236-6159 (L. Marconi); 0000-0002-4065-3415 (F. Cabitza)

of AI models and their reliance on dynamic datasets. This shortcoming is further exacerbated by the absence of standardized approaches to addressing ethical challenges, such as algorithmic bias and patient privacy, which are increasingly recognized as fundamental concerns in AI research. Thus, a comprehensive tool that integrates the strengths of existing frameworks while broadening their scope is urgently needed to address these gaps and ensure the responsible development and deployment of AI in healthcare.

In response to these challenges, we propose CLARITY AI, a synthesized and adaptable checklist designed for the comprehensive evaluation of AI-driven medical studies . CLARITY AI combines critical elements from ten established checklists into a unified framework. This tool addresses both technical and methodological rigor while also emphasizing data handling, ethical governance, and usability, ensuring that AI studies are scientifically robust, ethically sound, and practically relevant. With its structured yet flexible approach, CLARITY AI offers a more complete evaluation system that enhances research quality, reproducibility, and ultimately supports the safe integration of AI technologies into clinical practice.

By providing a holistic solution for evaluating AI research, this paper presents CLARITY AI as a key contribution to the field, aiming to establish a new standard for research quality in AI-driven medical studies. The implications of its adoption extend beyond improved transparency and rigor, offering the potential to accelerate the responsible deployment of AI tools in healthcare, ultimately advancing patient care.

2. Methods

The development of CLARITY proceeded through a structured, multifaceted process aimed at creating a comprehensive checklist to address gaps in existing frameworks for evaluating AI-driven medical research, particularly within healthcare. Our approach synthesized elements from existing checklists while expanding their scope to meet the unique challenges posed by AI technologies, including data handling, ethical considerations, and the dynamic nature of AI models. The methods used ensured that CLARITY captured all critical aspects of AI-driven research, from technical rigor to ethical deployment in real-world settings.

The methodological approach underpinning CLARITY's development was rooted in a detailed analysis of key checklists, each contributing specific strengths to create a robust, flexible framework adaptable to the evolving demands of AI in healthcare.We conducted a systematic review of widely-used AI checklists and our guiding research question was how to design a checklist that addresses the technical evaluation of AI models while integrating essential aspects of transparency, reproducibility, ethical governance, and usability. This approach aligns directly with the goal of ensuring reliable and ethical implementation of AI technologies in clinical practice.

2.1. Identification and contribution of Key Checklists

Several established checklists were selected, analyzed, and categorized into macro topics and structural items to shape the CLARITY framework, as detailed in [Tab. 1] and [Tab. 2]. Specifically, we selected these checklists based on their widespread adoption and relevance to AI in healthcare.

The CLAIM (Checklist for Artificial Intelligence in Medical Imaging) played a pivotal role in developing the Model Details and Data Handling sections of CLARITY. CLAIM's rigorous emphasis on data validation and performance metrics ensured that CLARITY effectively captured key aspects of model transparency and reproducibility, which are particularly crucial in medical AI, where model reliability must be demonstrated through robust data management practices.

Similarly, CONSORT (Consolidated Standards of Reporting Trials) provided a structured approach to study design, particularly for randomized controlled trials. CONSORT's focus on transparency and participant flow informed CLARITY's study design and methods categories, ensuring that AI studies adhere to the highest standards of rigor and reproducibility. This framework was vital in shaping how AI studies are documented and reported, creating a foundation for reliable implementation.

Table 1
Checklist Table Macro Topic Items

	STUDY IDENTIFICATION	STRUCTURED SUMMARY	BACKGROUND OBJECTIVES	STUDY DESIGN METHODS	DATA HANDLING	MODEL DETAILS
CLAIM AI	X	X	X	X	X	X
CONSORT AI	X	X	X	X		
DECIDE AI	X	X	X	X	X	X
FUTURE AI					X	X
IJMEDI AI					X	X
PRISMA AI	X	X	X	X	X	
SPIRIT AI	X	X		X		
STARD AI	X	X	X	X		
STARE HI	X	X	X	X		X
TRIPOD AI	X	X	X	X	X	X
	PERFORMANCE METRICS	RESULTS FINDINGS	DISCUSSION IMPLICATIONS	ETHICS GOVERNANCE	HUMAN FACTORS USABILITY	TRANSPARENCY REPRODUCIBILITY
CLAIM AI	X	X	X			
CONSORT AI		X	X		X	X
DECIDE AI		X	X	X	X	X
FUTURE AI	X	X	X	X	X	X
IJMEDI AI	X			X	X	X
PRISMA AI		X	X	X		X
SPIRIT AI						
STARD AI		X				
STARE HI		X	X			
TRIPOD AI	X	X	X	X		X

Table 2
Checklist Table Structural Items

	TITLE	ABSTRACT	INTRODUCTION	METHODS
CLAIM AI	X	X	X	X
CONSORT AI	X	X	X	X
DECIDE AI	X	X	X	X
FUTURE AI	?	?	?	?
IJMEDI AI	X		X	
PRISMA AI	X	X	X	X
SPIRIT AI	X		X	X
STARD AI	X	X	X	X
STARE HI	X	X	X	X
TRIPOD AI	X	X	X	X
	RESULTS	DISCUSSION	CONCLUSIONS	OTHER ELEMENTS
CLAIM AI	X	X		X
CONSORT AI	X	X		X
DECIDE AI	X	X		
FUTURE AI	?	?	?	?
IJMEDI AI	X	X	X	X
PRISMA AI	X	X		X
SPIRIT AI				X
STARD AI	X	X		X
STARE HI	X	X	X	X
TRIPOD AI	X	X		

DECIDE (Developmental and Exploratory Clinical Investigations of Decision Support Systems)[14] underscored the importance of clear communication in study reporting, influencing CLARITY's Structured Summary and Background sections. These ensure that AI research is methodologically sound and accessible to both clinicians and researchers.

FUTURE (Fairness, Universality, Traceability, Usability, Robustness, Explainability) guided the integration of ethical considerations into CLARITY. While FUTURE's focus on fairness and usability was central to shaping CLARITY's Ethics and Governance sections, its contributions to the structural organization of the framework were limited, as indicated by the use of the symbol "?" to denote its minimal role in these areas [Tab. 2]. Thus, FUTURE's input is more visible in the macro-level topics rather than in the core structural elements [Tab. 1].

Additional contributions came from IJMEDI [15], PRISMA (Preferred Reporting Items for Systematic

Reviews and Meta-Analyses), SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials)[16], STARD (Standards for Reporting Diagnostic Accuracy Studies)[17][18], STARE-HI (Standards for Reporting of Health Informatics), and TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis). Each of these checklists provided insights, particularly in data management and predictive model evaluation, enriching CLARITY's scope.

2.2. Comprehensive Analysis and Integration of Checklists

The integration of these checklists was a complex and iterative process aimed at developing a unified framework without redundancy. Each checklist's structure and focus areas were carefully analyzed to ensure that CLARITY incorporated the most valuable elements while eliminating duplicative or overly narrow criteria. We used a comparative approach to analyze the strengths of each checklist, identifying overlaps and gaps across them. For instance, CLAIM's emphasis on model validation was harmonized with CONSORT's focus on participant flow and study design. This synthesis ensured that CLARITY addressed both technical and methodological rigor, while also encompassing ethical governance and usability.

During the integration, common themes such as transparency, reproducibility, and ethical standards were identified across multiple checklists and synthesized into CLARITY. This ensured that the framework was not constrained by the limitations of any single checklist but rather offered a more versatile, adaptable tool for AI research, particularly given the dynamic nature of AI models and evolving datasets.

CLARITY's flexibility was a key consideration in its development. Unlike traditional checklists, which may be rigid, CLARITY was designed to evolve alongside advancements in AI technology and emerging ethical challenges. This adaptability ensures that the framework remains relevant and useful in a rapidly changing field. Furthermore, it was designed to be user-friendly, providing clear guidance for applying the checklist in diverse research contexts, from diagnostic imaging to predictive modeling.

One of the primary challenges in developing CLARITY was ensuring that the checklist remained comprehensive without becoming overly burdensome. To address this, we consolidated overlapping criteria while ensuring that no critical aspects were omitted. For example, although CLAIM and PRISMA both emphasize transparency, their approaches differ significantly. We integrated the most relevant elements from each, creating a unified guideline applicable to a broad range of AI research, ensuring a thorough evaluation without unnecessary complexity.

Another significant challenge was accommodating the iterative nature of AI models, which are often refined in real-time as new data becomes available. Traditional checklists, designed for static research studies, do not account for this iterative development. CLARITY, however, includes specific guidelines for evaluating data integrity, scalability, and security throughout the model lifecycle, ensuring that AI models are rigorously assessed over time.

Lastly, usability was a key consideration in CLARITY's development. The framework includes guidelines for evaluating the human factors and usability of AI tools, ensuring that they are practical and accessible to clinicians and researchers in real-world healthcare environments. This is particularly important for the success of medical AI, where seamless integration into clinical workflows is essential.

To further streamline the evaluation process, we developed tables summarizing both structural [Tab. 3] and macro topic items [Tab. 4] across the integrated checklists. These tables help researchers and practitioners quickly identify overlaps and gaps, ensuring that all critical aspects of AI research are addressed and facilitating the application of CLARITY in diverse research contexts.

In conclusion, CLARITY is a comprehensive, adaptable framework that integrates the strengths of multiple established checklists into a single tool. By addressing gaps in existing methodologies and expanding the scope of evaluation to include ethical and practical considerations, CLARITY establishes a new standard for evaluating AI-driven medical research. This framework ensures that AI technologies are not only scientifically robust but also ethically sound and practically relevant, promoting their responsible integration into healthcare.

Table 3
Structural Items List

TITLE	ABSTRACT	INTRODUCTION	METHODS
Identification and Scope	Structured Summary	Scientific Background	Study Design
Study Type and Objectives	Objectives and Implications	Rationale for the Study	Data Sources and Collection
Evaluation Metrics		Study Objectives	Participant Selection
			Interventions and Comparisons
			Outcome Measures
			Statistical Analysis
			Blinding and Randomization (if applicable)
			Ethical Considerations
RESULTS	DISCUSSION	CONCLUSIONS	OTHER ELEMENTS
Participant Flow and Demographics	Interpretation of Results	Summary of Main Findings	Funding
Primary and Secondary Outcomes	Comparison with Existing Literature	Clinical and Practical Implications	Conflicts of Interest
Adverse Events	Study Limitations		Ethical Approval
Statistical Analysis and Significance	Implications for Practice		Data Sharing and Availability
Subgroup Analyses (if applicable)	Recommendations for Future Research		Protocol Registration
Model Performance (for predictive models)	Generalizability		Author Contributions
	Strengths and Weaknesses		Acknowledgements

Table 4
Macro Topic Items List

STUDY IDENTIFICATION	STRUCTURED SUMMARY	BACKGROUND & OBJECTIVES	STUDY DESIGN AND METHODS	DATA HANDLING	MODEL DETAILS
Study Type and Purpose	Study Design and Methods	Background	Study Design 1	Data Management Plans	Model Architecture
Target Population and Clinical Context	Key Outcomes and Results	Objectives	Study Design 2	Data Security and Storage	Training Data
Objectives and Hypotheses	Conclusions		Data Collection	Data Quality Assurance	Evaluation Metrics
			Test Methods		
			Data Preparation		
			Outcome and Predictors		
			Model Development		
			Model Validation		
			Analysis		
PERFORMANCE METRICS	RESULTS & FINDINGS	DISCUSSION & IMPLICATIONS	ETHICS & GOVERNANCE	HUMAN FACTORS & USABILITY	TRANSPARENCY & REPRODUCIBILITY
Estimation Methods	Participant Flow	Interpretation of Findings	Research Ethics Approval	User Engagement	Data Availability
Handling Indeterminate Results	Baseline Characteristics	Comparison with Existing Studies	Consent or Assent	Usability Testing	Data Sharing
Handling Missing Data	Results of the AI Intervention	Implications for Practice	Confidentiality		
Sample Size Determination		Future Research Directions	Declaration of Interests		
			Access to Data		

3. Gap Analysis

The development of CLARITY began with a thorough examination of the limitations in existing AI-focused medical research checklists. The ten checklists analyzed—CLAIM, CONSORT, DECIDE, FUTURE, IJMEDI, PRISMA, SPIRIT, STARD, STARE-HI, and TRIPOD—each serve distinct purposes within the medical research landscape. However, when assessed against the comprehensive needs required to assess AI-driven studies in healthcare, several critical gaps became evident. These gaps reveal significant shortcomings in existing methodologies, which must be addressed to effectively evaluate the unique challenges posed by AI technologies in healthcare.

3.1. Analysis of Gaps in Existing Checklists

One of the most prominent issues identified in the analysis was the limited scope and narrow focus of many existing checklists. For instance , CLAIM[19] and STARE-HI, while valuable for specific areas such as diagnostic accuracy and health informatics, do not offer comprehensive guidance on broader aspects, particularly study design and ethical considerations. Due to its complexity, AI research requires a multi-dimensional evaluation approach that considers both the technical robustness and the ethical implications of these technologies in real-world healthcare settings. However, checklists like CONSORT and SPIRIT, which are primarily designed for clinical trials, do not sufficiently address the transparency and reproducibility challenges that are essential for the reliable deployment of AI systems. These checklists may overlook the iterative nature of AI models, especially in managing dynamic datasets that continuously evolve—a key characteristic that distinguishes AI from conventional technologies.
Additionally, while TRIPOD is comprehensive in reporting predictive models, it lacks sufficient coverage in areas such as usability and human factors, which are critical for ensuring that AI technologies are not only technically sound but also practical for end-users like clinicians and patients. Usability is essential in determining whether an AI tool will be successfully integrated into clinical practice, yet many checklists lack comprehensive guidance on this aspect. While DECIDE and FUTURE address ethical and fairness considerations, they fall short of providing a systematic framework for the ongoing governance of AI systems. The governance of AI in healthcare must address critical concerns such as ensuring fairness in AI-driven decisions, managing biases, and safeguarding patient data. However, many checklists lack structured approaches for evaluating these dimensions, particularly as AI technologies are implemented in diverse healthcare environments with varying regulatory and ethical standards.
Additionally , existing checklists lack flexibility, making them less adaptable to the rapidly evolving AI landscape in healthcare. Tools like CLAIM and PRISMA exemplify this rigidity, as they are often focused on specific research methods or applications, leaving little room for integration with emerging methodologies or newer AI technologies. This rigidity can limit the utility of such checklists when researchers work with innovative AI systems that do not align with traditional evaluation frameworks. For example, as AI models evolve toward more complex architectures, such as neural networks and unsupervised learning systems, the ability to adapt evaluation frameworks becomes essential. Checklists like DECIDE and FUTURE offer some decision-making structures, but even these tools lack the adaptability required to keep pace with AI advancements, particularly in complex clinical applications where AI systems are continuously evolving.
Another significant gap involves the handling of ethical issues specific to AI. While some checklists, such as CONSORT, SPIRIT[20], and PRISMA, provide minimal guidance on AI-specific ethical concerns—such as bias mitigation, data privacy, and the long-term impact of AI on patient care—these aspects are becoming increasingly critical as AI technologies become more pervasive in healthcare. Without comprehensive ethical guidance, AI-driven medical research risks introducing biases, compromising patient privacy, or deploying models that have unintended negative consequences on patient outcomes. Although FUTURE and DECIDE emphasize fairness and ethical considerations, they lack a structured framework that fully addresses the ethical governance of AI systems, particularly as it relates to continuous monitoring and accountability for AI decisions. AI systems are not static, and governance structures must be established to ensure their ongoing ethical performance.

3.2. Addressing the Gaps and Developing CLARITY

To address the gaps identified in the existing checklists, CLARITY was developed as a comprehensive and adaptable framework. It not only integrates the strengths of established checklists but also expands their scope, particularly by including detailed guidelines on ethical governance and data privacy. CLARITY provides a thorough assessment of critical concerns such as bias mitigation, data handling, and ongoing AI governance, ensuring that AI technologies are deployed responsibly and equitably across diverse healthcare environments. By incorporating these elements, the framework facilitates continuous monitoring of ethical risks, such as bias, which can directly impact patient care. Additionally, the

framework remains adaptable, evolving with new AI technologies and methodologies. Its adaptability ensures it remains relevant in the evolving landscape of healthcare AI, accommodating new challenges and technologies. As AI models advance, its modular structure allows updates while preserving core principles. As AI models advance, CLARITY's modular structure allows updates while preserving its core principles.

While addressing gaps in the ethical and technical evaluation of AI systems, the framework also leverages the strengths of existing checklists. It incorporates key principles from these checklists, such as transparency, reproducibility, and structured reporting. By requiring detailed documentation of AI model development, data handling, and performance metrics, it ensures that studies can be independently verified, thereby contributing to the integrity and transparency of AI research in healthcare.

In addition to ethical and technical aspects, CLARITY emphasizes practical usability, ensuring that AI tools are accessible and functional for end-users. By incorporating human factors and usability assessments , CLARITY promotes the development of AI systems that are not only effective but also user-friendly. In healthcare environments, where clinicians may have limited time or technical expertise, the usability of AI tools can determine whether they are adopted into routine practice. CLARITY's inclusion of usability testing bridges the gap between technical innovation and practical application, ensuring that AI technologies are truly beneficial in real-world clinical settings.

By addressing these gaps, CLARITY provides a comprehensive, flexible, and user-friendly framework for evaluating AI-driven research in healthcare. It not only fills the critical gaps identified in existing checklists but also builds on their strengths, offering a unified tool that adapts to the evolving demands of AI research in medical settings.

4. Results

After establishing the aforementioned categories, each was assigned a specific evaluation method and scoring system. This design allows researchers to easily and intuitively assess the quality of the study and determine whether it meets the necessary requirements.[Tab. 5][Tab. 6]

Table 5
Structural Score Ranges and Completeness Levels

Score Ranges and Completeness Levels [Structural]	
Excellent Completeness	**Score Range: 171 - 190** **The study is extremely comprehensive,** **covering nearly all required structural items with high quality.** **Few, if any, aspects are missing or insufficiently detailed.**
High Completeness	**Score Range: 152 - 170** **The study is very thorough, addressing most structural items adequately.** **There may be minor gaps or areas that could use additional detail,** **but overall the study is robust.**
Moderate Completeness	**Score Range: 133 - 151** **The study addresses many of the structural items,** **but there are some significant gaps or areas that need improvement.** **Additional detail and refinement are necessary to ensure a comprehensive evaluation.**
Basic Completeness	**Score Range: 114 - 132** **The study meets the minimum requirements for several structural items** **but lacks depth and detail in many areas.** **Significant improvements and additions are needed for a thorough evaluation.**
Low Completeness	**Score Range: 95 - 113** **The study addresses only a few structural items adequately.** **Most aspects are either missing or poorly described.** **Major revisions are required to meet the basic standards of completeness.**
Very Low Completeness	**Score Range: 76 - 94** **The study lacks comprehensiveness and detail in most structural items.** **It fails to provide sufficient information for a meaningful evaluation.** **Extensive work is needed to reach even basic completeness.**
Incomplete	**Score Range: 0 - 75** **The study is severely lacking in addressing structural items.** **It provides minimal to no useful information for evaluation.** **A complete overhaul is necessary to achieve any meaningful level of completeness.**

Table 6
Macro Topic Score Ranges and Completeness Levels

Score Ranges and Completeness Levels [Macro Topic]	
Excellent Completeness	**Score Range: 193 - 215** **The study is exceptionally comprehensive,** **addressing nearly all required macro topics with high quality.** **Very few, if any, aspects are missing or inadequately detailed.**
High Completeness	**Score Range: 171 - 192** **The study is very thorough, covering most macro topics adequately.** **There may be minor gaps or areas that could use additional detail,** **but overall the study is robust and well-rounded.**
Moderate Completeness	**Score Range: 150 - 170** **The study covers many of the macro topics,** **but there are some significant gaps or areas needing improvement.** **Additional detail and refinement are necessary to ensure a more comprehensive evaluation.**
Basic Completeness	**Score Range: 128 - 149** **The study meets the minimum requirements for several macro topics** **but lacks depth and detail in numerous areas.** **Significant improvements and additions are needed for a thorough evaluation.**
Low Completeness	**Score Range: 107 - 127** **The study addresses only a few macro topics adequately.** **Most aspects are either missing or poorly described.** **Major revisions are required to meet basic standards of completeness.**
Very Low Completeness	**Score Range: 85 - 106** **The study lacks comprehensiveness and detail in most macro topics.** **It fails to provide sufficient information for a meaningful evaluation.** **Extensive work is needed to reach even basic completeness.**
Incomplete	**Score Range: 0 - 84** **The study is severely lacking in addressing macro topics.** **It provides minimal to no useful information for evaluation.** **A complete overhaul is necessary to achieve any meaningful level of completeness.**

4.1. Scoring and Evaluation

Each category in CLARITY was analyzed to establish criteria for evaluation. The scoring system, ranging from 0 to 5 for each item, provides a straightforward mechanism for assessing study completeness and quality. This approach identifies strengths and weaknesses, helping researchers focus on areas for improvement.

Study Identification: Assesses how clearly the study is identified and defined, including the use of appropriate AI methodologies.

Structured Summary: Evaluates the completeness and clarity of the study's abstract and summary, ensuring they provide a comprehensive overview.

Background and Objectives: Examines the context and rationale behind the study, ensuring the objectives are clear and justified.

Study Design and Methods: Focuses on the robustness and appropriateness of the study design and methodologies used.

Data Handling: Evaluates the procedures for data collection, processing, and management to ensure data integrity and reliability.

Model Details: Analyzes the specifics of the AI model used, including its development, validation, and any comparative analyses performed.

Performance Metrics: Assesses the measures used to evaluate the AI model's performance, including accuracy, precision, and other metrics.

Results and Findings: Reviews the presentation and interpretation of the study's results, ensuring they are clear and well-supported by data.

Discussion and Implications: Evaluates the depth and breadth of the discussion, including the implications of the findings for clinical practice and future research.

Ethics and Governance: Ensures that ethical considerations and governance issues are thoroughly addressed and documented.

Human Factors and Usability: Assesses the involvement of end-users in the design and usability testing of the AI tool, ensuring it meets user needs and expectations.

Transparency and Reproducibility: Evaluates the study's transparency in reporting and its potential for reproducibility by other researchers.

4.2. Visualization and Data Distribution

To complement the scoring system, CLARITY employs bar charts to provide intuitive visualizations of data distribution and study results. These visualizations enable researchers to quickly assess study quality across various categories and compare overall scores.

The bar charts [Fig. 1] [Fig. 2] offer a multi-dimensional view of each study's performance. Each axis in the chart corresponds to a specific category (e.g., transparency, ethical governance, data handling), allowing for a quick and comprehensive assessment of a study's strengths and weaknesses. Structural items are mapped in [Fig. 1], while macro topic items are depicted in [Fig. 2].

In addition to these bar charts, a third chart [Fig. 3] represents the total score, combining both macro topic items and structural items. This bar chart highlights the "Total Score," providing a clear comparison of the completeness and quality of each evaluated study. The bar chart simplifies the comparative analysis by offering a straightforward visual that juxtaposes the final scores, making it easy to identify studies with stronger or weaker evaluations.

The integration of these categories, scoring techniques, and visualization tools has resulted in the creation of CLARITY, a comprehensive framework for the scientific evaluation of AI research in the medical field. By incorporating the strengths of various established checklists and refining them into a unified tool, CLARITY aims to set a new standard for assessing the quality and completeness of AI-related medical research. This holistic approach ensures that researchers have a reliable and user-friendly method for evaluating their studies, ultimately advancing the field of medical AI research.

5. Discussion

The CLARITY AI checklist provides a much-needed, comprehensive framework for evaluating the quality of AI-driven medical research, addressing gaps identified in previous checklists such as CLAIM, PRISMA, and CONSORT. CLARITY was developed in response to the rapidly evolving field of AI in healthcare, which requires rigorous and adaptable tools to capture the complexities inherent in AI systems, including issues of transparency, data handling, and ethical governance.

Key findings from the study indicate that CLARITY effectively bridges critical gaps by integrating diverse evaluation elements, including study design, ethical considerations, human factors, and usability. Its multi-dimensional approach ensures that AI studies are scientifically robust, practical, and ethical—crucial in the clinical deployment of AI systems.

The integration of established frameworks like CLAIM and STARE-HI underpins CLARITY's emphasis on transparent reporting and performance metrics, which are essential for reproducibility and effective validation of AI models. However, CLARITY goes beyond these existing tools by offering a more holistic approach that includes evaluating the usability of AI tools—an aspect often overlooked in traditional checklists. This focus on usability ensures that AI models are tested for real-world practicality, increasing their chances of successful adoption in clinical settings.

An unexpected finding during the checklist's development was the lack of comprehensive guidelines addressing the ongoing governance and ethical oversight of AI tools, particularly for mitigating algorithmic bias and ensuring patient safety. While frameworks like FUTURE have advanced fairness and ethics, CLARITY's structured focus on these elements sets it apart by embedding ethical considerations into every stage of the research process. This approach helps prevent AI models that may perform well technically but pose ethical risks from advancing to clinical implementation.

Comparison with previous research reveals that CLARITY not only integrates best practices from existing tools but also innovates by expanding the scope of evaluation. For instance, while TRIPOD provides detailed guidance on predictive models, it does not address usability or the iterative nature

Figure 1: Structural Items Average Score Bar Chart

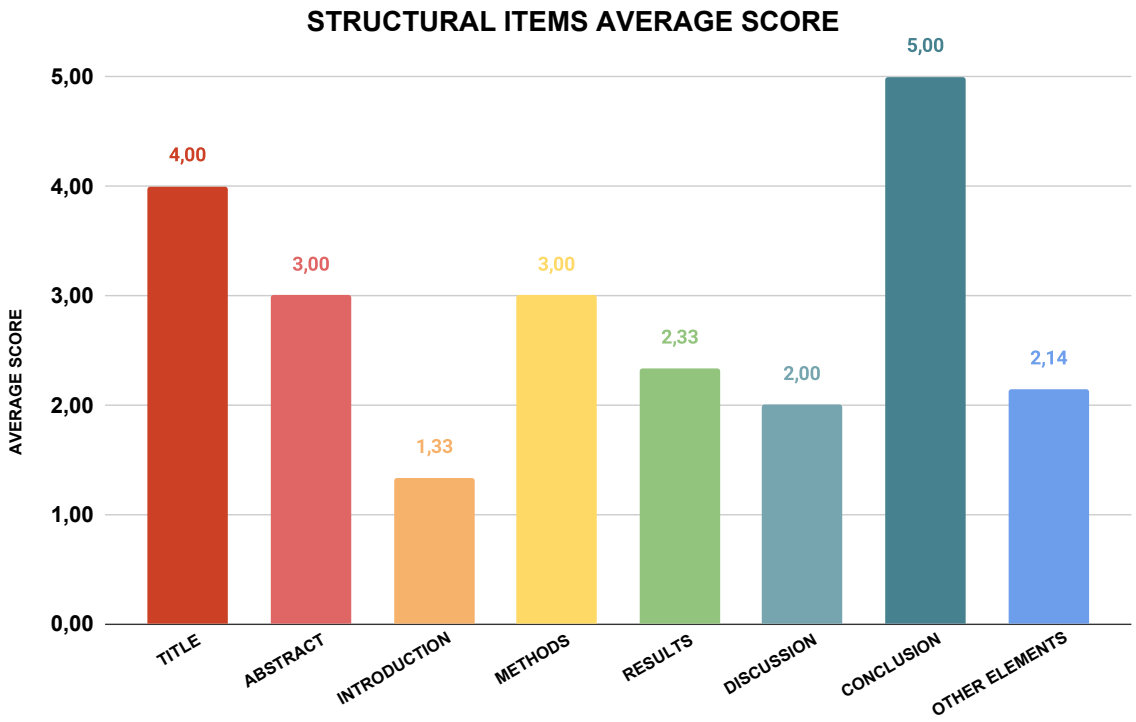


Figure 2: Macro Topic Items Average Score Bar Chart

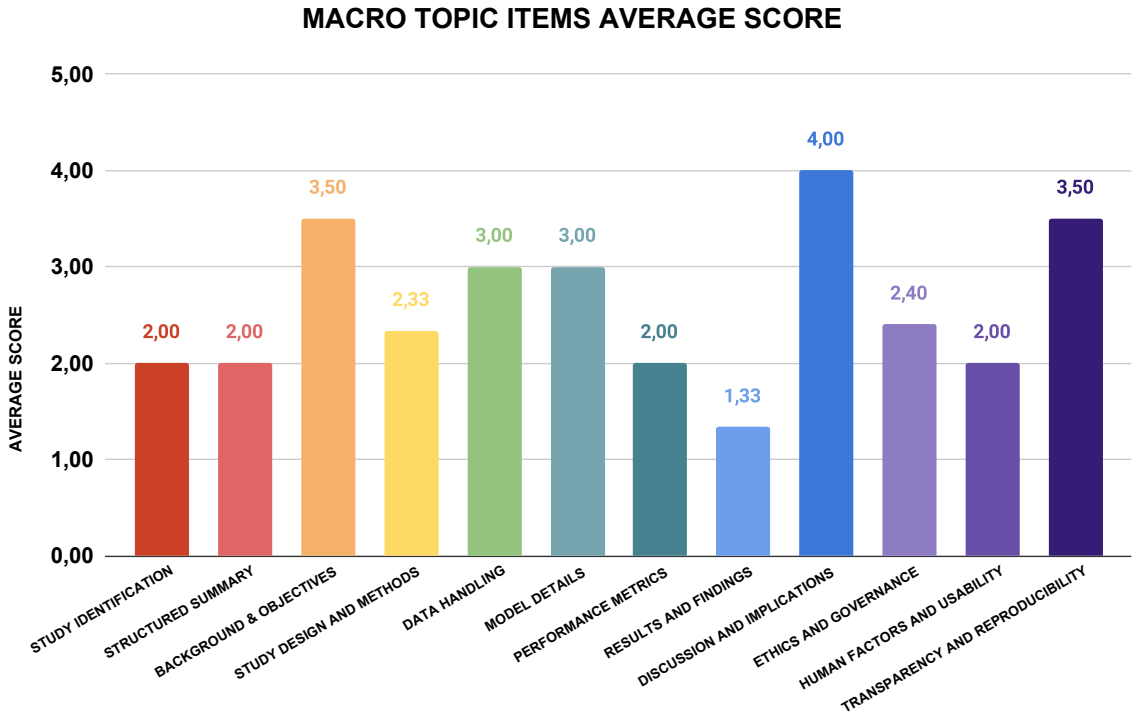
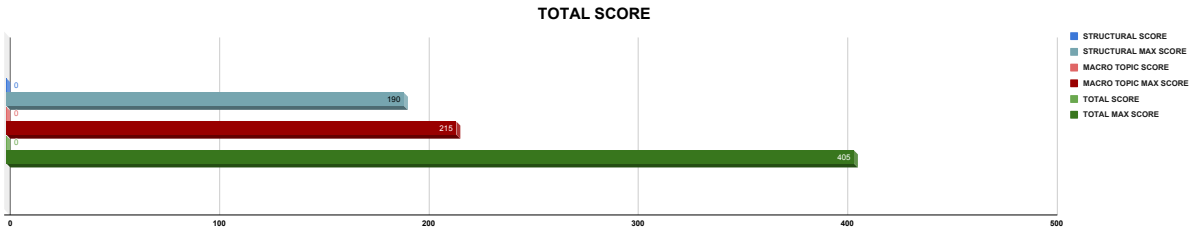


Figure 3: Total Score Bar Chart



of AI tool development, which CLARITY incorporates. Similarly, CONSORT's strength in clinical trials is complemented by CLARITY's broader applicability to AI studies, particularly in terms of data management and ethical governance—areas where CONSORT is less focused.

With its structured and flexible design, CLARITY is poised to make a significant impact on future AI research by setting a new standard for comprehensive and systematic evaluation. The integration of user feedback mechanisms allows CLARITY to evolve alongside advances in AI technology, ensuring its continued relevance and effectiveness as AI becomes more embedded in healthcare. Additionally, CLARITY's emphasis on transparent reporting and reproducibility will likely enhance the overall reliability of AI studies, facilitating more accurate meta-analyses and systematic reviews—essential for the responsible scaling of AI technologies in healthcare.

However, CLARITY's current iteration has some limitations. The framework has not yet undergone extensive empirical validation, and its real-world application across diverse research environments remains untested. Future research should aim to address this by conducting empirical studies in various clinical settings. Additionally, training and usability challenges pose potential barriers to widespread adoption, particularly in resource-constrained environments. Future work should focus on simplifying the checklist for ease of use and developing comprehensive training modules to support adoption by researchers of varying levels of expertise. Limitations and recommendations for further studies will be detailed in Section 6 and Section 7.

In conclusion, CLARITY represents a significant advancement in evaluating AI-driven medical research. By offering a rigorous, user-friendly, and adaptable tool, it addresses critical gaps in existing frameworks and promotes high standards of research quality. As AI continues to transform healthcare, tools like CLARITY will be essential to ensure that the supporting research is robust, reproducible, and ethically sound. Future studies should focus on refining and validating CLARITY in practiceto ensure it can adapt to the evolving landscape of AI in medicine.

6. Limitations

While CLARITY represents a significant advancement in evaluating AI-driven medical research, its current form has certain limitations. These can be categorized into three main areas: lack of empirical validation, complexity and usability challenges, and barriers to training and adoption. First, although CLARITY is built on a solid theoretical foundation, it has yet to undergo extensive real-world testing, leaving its practical utility unproven. The lack of empirical validation raises the possibility that some elements may not function as intended across diverse research environments, underscoring the need for studies that test the checklist in various contexts to ensure its broader applicability, as detailed in Section 7.

Secondly, while CLARITY's comprehensive nature aims to ensure thorough evaluations, it may present challenges for users unfamiliar with AI research or those in resource-limited settings. The detailed criteria can be overwhelming, indicating a need for simplification or tiered complexity to accommodate varying levels of expertise and resources. Finally, the specialized knowledge required for effective use of CLARITY may hinder its adoption. The steep learning curve, particularly in teams lacking AI expertise, could impede widespread implementation.

Moreover, the current version does not incorporate user feedback, meaning practical challenges in its application have not yet been addressed. To overcome these barriers, it would be beneficial to develop training programs to facilitate the checklist's use and gather feedback from early adopters to refine the framework. In summary, while CLARITY shows promise as a robust tool for AI research evaluation, further empirical validation, simplification, and user-focused improvements will be essential to fully realize its potential in enhancing research quality in healthcare.

7. Recommendations

Although CLARITY was meticulously developed by synthesizing existing frameworks, its adoption and long-term utility in evaluating AI-driven medical research will benefit from rigorous empirical validation. We recommend conducting pilot studies and case studies in diverse healthcare settings to assess the checklist's effectiveness in real-world applications. These studies should test CLARITY's comprehensiveness, usability, and adaptability across various research contexts and clinical environments.
A potential framework for empirical validation could involve the following steps:

1. *Pilot Implementation:* Researchers could retrospectively apply CLARITY to a range of AI-driven medical studies, including diagnostic imaging, predictive modeling, and treatment planning. These applications should cover diverse AI methodologies (e.g., supervised and unsupervised learning) to ensure broad relevance.
2. *Comparative Analysis:* To assess CLARITY's utility, comparisons with other established checklists (e.g., CLAIM, CONSORT) could be conducted. Evaluating studies with both CLARITY and these checklistswill help determine if the integrated approach provides additional insights or uncovers previously overlooked issues..
3. *User Feedback:* Engaging researchers and clinicians is fundamental to understanding the checklist's usability. Structured interviews or surveys with users applying CLARITY can provide feedback on ease of use, clarity, and the ability to inform research improvements.
4. *Iterative Refinement:* Based on pilot findings and feedback, the checklist can be adjusted. Continuous validationacross various clinical settings and AI applications will help ensure CLARITY remains relevant and adaptable to evolving AI technologies.
5. *Outcome Assessment:* Ultimately, the checklist's success should be measured by improvements in research quality, transparency, and reproducibility. Meta-analyses of studies evaluated by CLARITY could demonstrate whether its adoption correlates with higher standards in AI-driven research.

By following these steps, future studies can provide critical insights into improving CLARITY and ensure that it remains a valuable tool for evaluating AI research in healthcare..

8. Conclusions

CLARITY represents a significant advancement in evaluating medical research involving artificial intelligence. Its meticulously designed framework integrates best practices from various established checklists. By extracting and refining the most valuable elements from these sources, CLARITY provides a comprehensive tool that addresses both structural items and macro topic items, ensuring a thorough assessment of AI studies in healthcare.
The scoring system, along with visualization tools like spider charts and bar charts, enhances the checklist's usability and clarity, making it easier for researchers to evaluate their work. This systematic approach highlights a study's strengths and weaknesses while guiding researchers toward areas for improvement, fostering a culture of continuous enhancement in AI medical research.
CLARITY's structured evaluation method ensures that all critical aspects of a study are thoroughly assessed. This includes study identification clarity, completeness of structured summaries, relevance

and justification of background and objectives, robustness of study design and methods, integrity of data handling, model details, accuracy of performance metrics, clarity of results and findings, depth of discussion and implications, adherence to ethical standards, consideration of human factors and usability, and transparency and reproducibility .

By providing a detailed, standardized evaluation framework, CLARITY helps researchers produce high-quality, transparent, and reproducible AI studies. This enhances the credibility of individual studies and contributes to the overall integrity and advancement of AI research in the medical field.

CLARITY aims to set a new standard for evaluating AI studies in healthcare, promoting best practices and high standards across the research community. Its comprehensive approach ensures all relevant aspects of a study are considered, reducing the risk of oversight and enhancing the robustness of research findings. By fostering rigorous evaluations, CLARITY helps ensure that AI technologies developed and tested in healthcare settings are reliable, effective, and safe.

Moreover, using CLARITY can lead to more consistent and comparable evaluations of AI studies, facilitating meta-analyses and systematic reviews. This, in turn, can accelerate the adoption of effective AI technologies in clinical practice, ultimately improving patient outcomes and advancing medicine.

As AI evolves and its healthcare applications expand, CLARITY must adapt to incorporate new insights and developments. Ongoing feedback from the research community will be crucial for refining and updating the checklist to keep it relevant and effective. Future iterations of CLARITY may include additional categories or refined scoring criteria to better capture emerging trends and technologies in AI research.

Overall, CLARITY provides a comprehensive, systematic, and user-friendly framework for evaluating AI studies in the medical field.

The success and evolution of CLARITY rely heavily on interdisciplinary collaboration between researchers in AI, healthcare, ethics, and policy-making. As AI becomes more integrated into healthcare, input from diverse fields—including medical professionals, data scientists, legal experts, and ethicists—will be essential. CLARITY provides a robust starting point for evaluating AI research, but its continued relevance will depend on contributions from these diverse fields to address emerging challenges, such as patient safety, data privacy, and algorithmic fairness. By fostering interdisciplinary partnerships, CLARITY can evolve into a universally accepted standard, guiding AI innovation to be both scientifically rigorous and socially responsible. This collaborative approach will ensure CLARITY remains adaptive to the complex and evolving landscape of AI in healthcare.

References

[1] J. Mongan, et al., Checklist for artificial intelligence in medical imaging(claim): A guide for authors and reviewers, Radiology: Artificial Intelligence 2 (2020) 2638–6100.

[2] A. S. Tejani, M. E. Klontzas, A. A. Gatti, J. Mongan, L. Moy, S. H. Park, C. E. Kahn, Updating the checklist for artificial intelligence in medical imaging (claim) for reporting ai research, Nature Machine Intelligence 5 (2023) 950–951. URL: https://doi.org/10.1038/s42256-023-00717-2. doi:10.1038/s42256-023-00717-2.

[3] A. Bhandari, L. Scott, M. Weilbach, R. Marwah, A. Lasocki, Assessment of artificial intelligence (ai) reporting methodology in glioma mri studies using the checklist for ai in medical imaging (claim), Neuroradiology 65 (2023) 907–913. URL: https://doi.org/10.1007/s00234-023-03126-9. doi:10.1007/s00234-023-03126-9.

[4] M. D. Schulz KF, Altman DG, Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials., Journal of Pharmacology and Pharmacotherapeutics. 1(2) (2010) 100–107.

[5] A. P. L. Martindale, C. D. Llewellyn, R. O. de Visser, B. Ng, V. Ngai, A. U. Kale, L. F. di Ruffano, R. M. Golub, G. S. Collins, D. Moher, M. D. McCradden, L. Oakden-Rayner, S. C. Rivera, M. Calvert, C. J. Kelly, C. S. Lee, C. Yau, A.-W. Chan, P. A. Keane, A. L. Beam, A. K. Denniston, X. Liu, Concordance of randomised controlled trials for artificial intelligence interventions with the

consort-ai reporting guidelines, Nature Communications 15 (2024) 1619. URL: https://doi.org/10.1038/s41467-024-45355-3. doi:10.1038/s41467-024-45355-3.

[6] R. Shahzad, B. Ayub, M. A. R. Siddiqui, Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review, BMJ Open 12 (2022). URL: https://bmjopen.bmj.com/content/12/9/e061519. doi:10.1136/bmjopen-2022-061519. arXiv:https://bmjopen.bmj.com/content/12/9/e061519.full.pdf.

[7] M. J. Page, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021).

[8] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, D. Moher, The prisma 2020 statement: an updated guideline for reporting systematic reviews, Systematic Reviews 10 (2021) 89. URL: https://doi.org/10.1186/s13643-021-01626-4. doi:10.1186/s13643-021-01626-4.

[9] W. Z. Andrea C. Tricco, Erin Lillie, et al., Prisma extension for scoping reviews (prisma-scr): Checklist and explanation, Annals of Internal Medicine 169 (2018) 467–473. URL: https://doi.org/10.7326/M18-0850. doi:10.7326/M18-0850. arXiv:https://doi.org/10.7326/M18-0850, pMID: 30178033.

[10] N. Dagan, et al., Evaluation of ai solutions in health care organizations — the optica tool, NEJM AI 0 (2024) AIcs2300269.

[11] Collins, et al., Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024). doi:10.1136/bmj-2023-078378.

[12] G. S. Collins, K. G. M. Moons, Dhiman, et al., Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024). doi:10.1136/bmj-2023-078378.

[13] J. Talmon, E. Ammenwerth, J. Brender, N. de Keizer, P. Nykänen, M. Rigby, Stare-hi—statement on reporting of evaluation studies in health informatics, International Journal of Medical Informatics 78 (2009) 1–9.

[14] B. Vasey, et al., Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai, BMJ 377 (2022) e070904.

[15] A. Iancu, I. Leb, H.-U. Prokosch, W. Rödle, Machine learning in medication prescription: A systematic review, International Journal of Medical Informatics 180 (2023) 105241.

[16] C. Rivera, et al., Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension, The Lancet Digital Health 2 (2020) e549–e560.

[17] Sounderajah, et al., Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the stard-ai protocol, BMJ Open 11 (2021).

[18] D. A. Korevaar, J. F. Cohen, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, D. Moher, H. C. W. de Vet, D. G. Altman, L. Hooft, P. M. M. Bossuyt, Updating standards for reporting diagnostic accuracy: the development of stard 2015, Research Integrity and Peer Review 1 (2016) 7. URL: https://doi.org/10.1186/s41073-016-0014-7. doi:10.1186/s41073-016-0014-7.

[19] M. E. Klontzas, A. A. Gatti, A. S. Tejani, C. E. Kahn, Ai reporting guidelines: How to select the best one for your research, Radiology: Artificial Intelligence 5 (2023) e230055. URL: https://doi.org/10.1148/ryai.230055. doi:10.1148/ryai.230055. arXiv:https://doi.org/10.1148/ryai.230055.

[20] H. Ibrahim, X. Liu, S. C. Rivera, D. Moher, A.-W. Chan, M. R. Sydes, M. J. Calvert, A. K. Denniston, Reporting guidelines for clinical trials of artificial intelligence interventions: the spirit-ai and consort-ai guidelines, Trials 22 (2021) 11. URL: https://doi.org/10.1186/s13063-020-04951-6. doi:10.1186/s13063-020-04951-6.