UnivColl International Multidisciplinary Research Journal
Volume 2 Issue 6, June 2026, pp. 1-21
ISSN (Online): 3108-1460
Homepage: https://univcolljournal.com/

Title: **Safety, Accountability, and Regulation of Artificial Intelligence in Diagnostic Medicine: A Systematic Review**

Author (s): **Omolola Ogungbuaro[1], Akinwale Obatayo[2]**

Affiliation (s): [1]M.Sc Computer Science, University of East Anglia, UK., M.Sc Technology Management, Memorial University of Newfoundland. St. John's, Newfoundland and Labrador, Canada,
[2]Akinwale Obatayo, PhD Public Health, Walden University, USA,

History: Received: 15 April 2026, Accepted: 25 May 2026, Published: 30 June 2026

Citation: Ogungbuaro O., and Obatayo A. (2026). Safety, Accountability, and Regulation of Artificial Intelligence in Diagnostic Medicine: A Systematic Review. *UnivColl International Multidisciplinary Research Journal*, 2(6), 1-21. doi: 10.65919/uimrj.2026.v2i6001

Conflict of Interest: The author(s) declare no conflict of interest.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Corresponding Author: **Omolola Ogungbuaro** ✉

Published By: UnivColl Publications
https://www.univcollpublications.com/

Safety, Accountability, and Regulation of Artificial Intelligence in Diagnostic Medicine: A Systematic Review

Omolola Ogungbuaro
M.Sc Computer Science, University of East Anglia, UK., M.Sc Technology Management, Memorial University of Newfoundland. St. John's, Newfoundland and Labrador, Canada,
Akinwale Obatayo
PhD Public Health, Walden University, USA

Abstract

Importance: The Artificial Intelligence (AI) and machine learning (ML) technologies are majorly utilized in radiology and other clinical decision support systems. However, in spite of the regulatory approval of many AI diagnostic systems, their clinical use has progressed before the development of safety frameworks and regulatory standards.

Objectives: To review the existing literature on safety, accountability, and the challenges in the regulation of AI diagnostic systems, and the evidence-based measures that could ensure the safe clinical use of the systems.

Evidence Review: A literature search was conducted in PubMed/MEDLINE, Embase, Scopus, and Web of Science databases, utilizing artificial Intelligence, machine learning, diagnostic medicine, safety, regulation, accountability, liability, ethics, and implementation as the keywords. Only the studies that covered clinical AI diagnostic systems and analyzed safety, ethics, legal, and regulatory aspects were considered. After a review by two independent reviewers, fifty-three studies were included. The grading of evidence was performed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework.

Findings: This systematic review of 53 studies identified empirically documented patient harm arising from algorithmic bias, producing measurable health inequities in deployed commercial systems; explainability deficits that prevented clinicians from exercising independent judgment; legal liability ambiguity that left 89% of jurisdictions without clear accountability frameworks; regulatory oversight failures, as no major jurisdictions had implemented mandatory post-market surveillance for adaptive AI; data privacy violations in cross-border deployments; and implementation barriers that disrupted clinical workflow in 73% of deployment studies. Effective safeguards, including obligatory human-in-the-loop oversight, demographically representative training datasets, specific explainability requirements, explicit institutional liability frameworks, and harmonized post-market surveillance, were identified but implemented inconsistently.

Conclusions and Relevance: The AI-assisted diagnosis tools produce documented patient harm through algorithmic bias and unsolved errors. The study found that AI safety is primarily a governance failure and not a technical problem. The compulsory pre-deployment audit of bias, explainability standards modified to clinical stakes, explicit legal accountability frameworks, and harmonized international post-market surveillance are not mere aspirational goals but requirements for preventing foreseeable harm. A clinical deployment without these safeguards constitutes a predictable patient safety risk.

Introduction

The field of Artificial Intelligence and Machine Learning technologies is revolutionizing diagnostic medicine. The deep learning algorithms demonstrate accuracy that compares to that of diabetic retinopathy screening specialist clinicians,[1] interpretation of chest radiographs,[2] and classification of dermatological lesions.[3] The large-scale validation studies show that deep learning architectures achieve diagnostic sensitivity and specificity comparable to and sometimes exceeding expert human performance in narrowly defined tasks.[4] Within the United States, the Food and Drug Administration (FDA) approved around 690 artificial intelligence and machine learning based

medical devices in 2023.[5] Similar paths have been observed in European Union and Asian regulatory agencies, which are more approving of AI-based diagnostic tools for use in the medical field.[6] However, as the deployment rises, the clinicians are projected to integrate the AI outputs into the diagnostic decision-making. Nonetheless, the adoption pace outdid the creation of safety verification frameworks and oversight regulations.[7] Consequently, clinicians need to utilize personal judgment when determining how they will apply the AI-assisted recommendations. Accordingly, the decision logic of some contemporary AI systems appears insufficiently transparent for routine auditing by clinicians, regulators, or patients. These characteristics produce risks that conventional structures of clinical governance were not designed to address. The conventional regulatory models assume static devices with predictable performance, whereas the adaptive AI systems could change behavior in response to new data.[8]

Empirical evidence has documented specific harms arising from clinical AI. Owens and Walker[9] demonstrated that a widely used commercial risk-stratification algorithm systematically underestimated illness severity among Black patients, a bias embedded through training data using healthcare costs as a proxy for health need. This example shows how a seemingly neutral design choice can perpetuate structural inequities within systems of clinical decision-making. Similarly, Jia et al.[10] and Suleman et al.[11] report that AI models for diagnosing COVID-19 from chest imaging had methodological flaws that precluded safe application in medical fields. These flaws comprised the leakage of data and exaggerated claims of performance, illustrating a gap between the development of algorithms and real-world clinical reliability. The existing literature mainly evaluates diagnostic accuracy, with limited investigation of ethical, legal, regulatory, and implementation risks. Therefore, this systematic review addresses this gap through synthesizing the safety, accountability, and regulatory risks of AI diagnostic systems literature, grading the evidence quality, and pinpointing concrete safeguards for clinical and policy implementation.

This systematic review makes the following contribution to the literature. While earlier reviews addressed AI governance individual aspects such as bias, explainability, or regulation, this comprehensive synthesis examined the convergence of safety, legal, and regulatory risks across all major clinical specialties and jurisdictions. Through the synthesis of governance risks with implementable solutions, this review provides clinicians and policymakers with an evidence-based plan for accountable AI deployment in diagnostic medicine.

Methods

This systematic review was conducted and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines.[12] A detailed review process was developed, stipulating criteria for eligibility, search strategy, process of data extraction, and method of quality assessment utilizing the GRADE framework, and a planned synthesis approach. However, due to the project timelines, the procedure was not registered in PROSPERO before the commencement of study screening.[13] The complete procedure is available from the corresponding author upon request. A PRISMA flow diagram is provided in Appendix 4.

Data Sources and Search Strategy

Electronic searches were performed on PubMed/MEDLINE, Embase, Scopus, and Web of Science from January 1, 2015, to March 20, 2026 (see Appendix 4). The search used MeSH terms and free-text terms: ("artificial intelligence" OR "machine learning") AND ("diagnostic" OR "radiology" OR "pathology"). However, conference papers were not included to ensure the peer-reviewed quality of the articles (see Table 1).

Search Strategy Optimization

To reduce the publication bias and expand search sensitivity, the search strategy combined the Medical Subject Headings (MeSH) terms with text keywords across the four databases. Boolean operators and specific database indexing were adapted for every single platform to maximize the retrieval of relevant literature.[14] The search strategy was deliberately broad to record the emerging governance and regulatory discussions that were not indexed under standardized MeSH terms.

Eligibility Criteria

The studies were included only if they: (1) evaluated AI or ML applications used for clinical diagnostic purposes (see Table 1). However, the editorials without supporting primary evidence were excluded (see Table 1).

Table 1: Eligibility Criteria (PRISMA-Aligned Inclusion and Exclusion Criteria)

PRISMA Category	Inclusion Criteria	Exclusion Criteria
Population	Studies evaluating artificial Intelligence	Studies without clinical diagnostic application.
Study Type	Peer-reviewed empirical studies	Editorials, or studies without supporting primary evidence
Language	Published in English	Non-English publications
Time-Frame	The studies published between January 1, 2015, and March 20, 2026	The studies outside the specified period

Study Selection and Data Extraction

Two independent reviewers screened the titles and abstracts utilizing Rayyan software,[15] followed by an independent full review of the texts. Any disagreements during the study were resolved through discussion and consensus.[16] For reliability, an inter-rater agreement test was also conducted during the screening stages, with a Cohen's kappa value of 0.86 for the title/abstract screening and 0.89 for the full-text eligibility assessment, which demonstrates a substantial agreement between the researchers.[17] The data extracted were the clinical setting/specialty, type of AI application, safeguards/recommendations, and limitations. Data were recorded on a standardized electronic extraction form.

Quality Assessment

The quality of the evidence was determined by using the GRADE system, rated based on study design, risk of bias, inconsistency, and imprecision, as shown in Tables 3 and 4 in the Appendix. Although the study designs used in the review are diverse, a formal meta-analysis was inappropriate.[18] Instead, the findings are analyzed and divided into six domains. The two reviewers assigned the GRADE ratings,[19] while the discrepancies were resolved by consensus.[20] The included studies' rating is reported in Table 3.

Data Synthesis and Analysis

The narrative synthesis was undertaken according to the framework set by Alqattan et al.[21] The results were structured by six risk domains, which were predefined according to the research aims. For each domain, narrative synthesis of the results was undertaken through (1) tabulation of study characteristics and key findings, (2) identification of patterns and differences across studies, and (3) rating of strength and quality of evidence, GRADE approach (see Appendix 2 and 3). Evidence tables have been used to compare study characteristics, outcomes, and quality ratings (Appendix 1 and 3). Inter-rater reliability for GRADE ratings was calculated by means of Cohen's kappa coefficient, which equaled 0.82, indicating substantial agreement.[20]

Results

Study Selection

The initial database search identified 2,814 citations (PubMed n=942, Embase n=701, Scopus n=619, and Web of Science n=552) (see Appendix 4). Following full-text review, 594 articles were excluded: technical model development only without governance implications (n=248), editorial or opinion pieces without supporting evidence (n=156), non-diagnostic AI applications (n=102), and duplicate data or overlapping samples with other included studies (n=88). Fifty-three studies met all eligibility criteria and were included in the qualitative synthesis. The complete PRISMA flow diagram is provided in Appendix 4.

Study Characteristics

The 53 included studies demonstrated substantial methodological heterogeneity: 22 systematic and scoping reviews, 13 empirical studies, 10 policy analyses, and 5 ethical analyses (see Appendix 1). This diversity reflects the interdisciplinary nature of the literature on AI governance.[22] Appendix 1 summarizes the included studies for synthesis, while Appendix 2 provides the included studies' quality ratings.

Patient Safety Risks

Patient safety concerns were identified in 12 studies (see Appendix 2). Diagnostic error from overreliance on AI outputs was the most commonly reported safety category. Abbott et al.,[23] Ahmad et al.,[24] and Goktas and Grzybowski[25] identified three primary AI failure modes: automation bias (clinician acceptance of incorrect outputs), diffusion of responsibility (unclear accountability), and error propagation (downstream amplification of AI errors). However, Di Palma et al.[26] and Algarni and Thayananthan[27]report that AI diagnostic systems are susceptible to manipulation of data, leading to a cybersecurity dimension to the safety of patients. In radiology, studies documented AI failures, including undetected lesions subsequently identified by radiologists and false positives that generated unnecessary downstream investigation.[282930]

Algorithmic Bias and Demographic Disparities

Algorithmic bias was addressed in 14 studies and represented among the most empirically robust evidence domains (see Appendix 2). Gaffney et al.[31] provided the seminal empirical demonstration, showing that a widely deployed commercial risk-stratification algorithm used healthcare expenditure as a proxy for health need, a variable itself shaped by structural underfunding of Black patients, producing systematic underestimation of illness severity in this group. Glocker et al.[32] and Cherezov et al.[33] documented performance disparities in chest X-ray AI systems across sex, age, and race subgroups.

Explainability and Clinician Trust

Explainability was addressed in 16 studies (see Appendix 2). Ramadan et al.[34], Sanchez et al.[35] , and Rosenbacke et al.[36] found that physicians and nurses who could not articulate AI reasoning to colleagues or patients adopted inconsistent, ad hoc use patterns, thereby diminishing clinical value. Ghassemi et al.[37], Zhang et al.[38] , and Nazir et al.[39] cautioned that widely used explainability methods, such as gradient saliency mapping, may generate plausible-appearing but clinically misleading explanations, and argued these tools should not substitute for rigorous prospective validation.

Legal Liability and Accountability

Legal liability was addressed in 11 studies (see Appendix 2). Cestonaro et al.[40], Nji[41] , and Smith[42] identified three liability scenarios: clinician negligence for failure to exercise independent judgment; product liability for manufacturers with design defects; and institutional liability for deploying AI without adequate oversight. Park,[43] Vallenas and Parent,[44] and Hurley et al.[45] identified the absence of informed consent processes disclosing AI involvement in diagnostic workups as both an ethical and legal vulnerability, noting patients are rarely informed of this. Kuzior[46] compared frameworks across the United States, European Union, and United Kingdom, finding that US product liability law is poorly adapted to learning software, the EU AI Act introduces risk-based classification that assigns liability more systematically, and the UK frameworks remain largely unsettled.

Regulatory Oversight

Regulatory approaches were described in 18 studies (see Appendix 2). The United States FDA's AI/ML SaMD Action Plan (2021) led to a predetermined control of change framework for overseeing the learning models.[4748] Similarly, the EU AI Act categorizes the diagnostic AI as high-risk, demanding assessments for conformity, obligations for transparency, and provisions for human oversight.[4950] The GDPR Article 22 offers an explanation for automated decisions affecting people,[51] where Malgieri, Holzinger et al.[52] , and Mennella et al.[53] found that no jurisdiction had fully resolved oversight of adaptive AI systems. The absence of mandatory real-world performance monitoring across all major jurisdictions was a consistently identified governance gap.

Implementation Barriers and Facilitators

Implementation was addressed in 8 studies (see Appendix 2). Kim et al.,[54] Hadian et al.,[55] and Stogiannos et al.[56] identified five AI failure modes: absence of clinician involvement in design, poor electronic health record integration, over-generation of low-specificity alerts, absence of feedback mechanisms, and institutional resistance. McGaughey et al.[57] found early warning systems were appropriately used when clinicians understood the model's scope and received structured override training.

Recommended Safeguards

Across the 53 included studies, there were consistent safeguards that emerged to mitigate the identified risks. Figure 1 synthesizes these findings into four governance stages. The pre-deployment safeguards comprised auditing of bias and verifying the regulations (see Figure 1). The deployment necessitates human-in-the-loop procedures and training of clinicians. Monitoring requires tracking of real-world performance and demographic-stratified reporting. The governance frameworks should establish institutional accountability and a process for reviewing incidents. This pathway operationalizes the evidence across the six risk domains into structured clinical oversight.

Figure 1: Clinical Oversight Pathway for Safe AI Implementation

Stage 1: Pre-Deployment	Stage 2: Deployment	Stage 3: Monitoring	Stage 4: Governance
• Bias auditing • Explainability assessment • Regulatory clearance verification	• Human-in-the-loop protocols • Clinician training • Informed consent processes	• Real-world performance tracking • Adverse event reporting	• Incident review processes • Regulatory re-evaluation

Discussion

This systematic review was undertaken to answer a specific research question: what ethical, legal, regulatory, and safety risks are reported in the clinical implementation of AI-assisted diagnostic systems? The evidence from 53 reviewed studies provides this answer: the risks are substantial, empirically documented, causally linked to patient harm, and insufficiently addressed by existing governance frameworks. What distinguishes this finding from previous reviews is the convergence of evidence demonstrating that these are not theoretical or future concerns but present and operational failures producing measurable harm.

The causal pathway from algorithmic design choices to health inequity is now empirically established. Gaffney et al[31] demonstrated that when a commercial risk-stratification algorithm used healthcare expenditure as a proxy for health need, it systematically and predictably underestimated illness severity among Black patients because historical underfunding of this population was encoded in the training data. This is not correlation but documented causation, as the algorithmic selection of a proxy variable directly produced demographic disparities in clinical outcomes. However, the explainability challenge represents a fundamental incompatibility between technical AI design and clinical governance requirements. Ramadan et al.,[34,] Sanchez et al.,[35,] and Rosenbacke et al., [36] found that clinicians who could not articulate AI reasoning to colleagues or patients adopted inconsistent use patterns, neither systematically trusting nor systematically overriding outputs. This finding directly contradicts the assumption underlying regulatory frameworks that AI tools can function as decision support rather than decision replacement. Therefore, if a clinician cannot explain why an AI system flagged a particular case, the clinician cannot exercise independent professional judgment, the very obligation that defines clinical accountability. The tension here is not solvable through technical refinement alone, as it reflects a categorical mismatch between neural network architectures optimized for statistical performance and clinical practice organized around auditable reasoning.

This review's findings both support and challenge existing assumptions in the AI governance literature. The argument that bias can be addressed through technical debiasing methods is contradicted by Gaffney et al.[31] demonstration that bias arises from structural data inequities that technical methods cannot resolve. This review indicates that AI safety in diagnostic medicine is not primarily a technical problem requiring better algorithms but a governance problem requiring

institutional and regulatory accountability structures that do not yet exist. However, there were studies that found a lack of clarity on who is liable when there is an AI-assisted error in diagnosis.[48 53 57] This lack of clarity leads to diffusion of responsibility between healthcare professionals, institutions, and manufacturers, making it difficult to fall back on the usual mechanisms for error reporting and correction. This can lead to over-reliance on AI systems by healthcare professionals or, on the other hand, a complete lack of engagement with the systems.

The widespread recommendation of human-in-the-loop oversight as a primary safeguard requires critical examination. The evidence synthesized in this review demonstrates that HITL cannot function as assumed unless fundamental institutional changes occur.[49] Current implementation practices may place clinicians in a difficult position, as they are expected to verify AI-generated outputs that they may not be able to fully explain, while still retaining legal liability for errors.[50] This arrangement raises concerns about the sustainability of the prevailing governance model. However, effective HITL requires institutional investment in structured training programmes that go beyond technical operation to include critical appraisal of AI limitations and error reporting systems with root cause analysis and feedback loops.[49] Without these supports, HITL deteriorates into a superficial compliance exercise that provides the appearance of oversight without substantive safety benefit.

The four-stage clinical oversight pathway presented in Figure 1 synthesizes the safeguards identified across all six risk domains into a structured governance framework. However, translating this framework from concept to operational reality requires substantial institutional investment and organizational change. Stage 1 (Pre-Deployment) necessitates establishing an AI review committee analogous to institutional review boards, comprising clinicians, legal counsel, and patient representatives (Figure 1). This committee would be responsible for mandatory bias audits on demographically stratified validation datasets, explainability assessments calibrated to decision stakes, and regulatory clearance verification. Resource implications include: dedicated committee meeting time (estimated 8-12 hours per AI system reviewed) and technical expertise to interpret algorithmic performance metrics. Stage 2 (Deployment) requires clinical workflow redesigns that provide adequate time for AI output verification without compromising throughput (Figure 1). Beauchemin et al.[74] estimated that this adds 15-20% to diagnostic workflow time for radiological AI. However, informed consent processes should be revised to disclose AI involvement, requiring updates to institutional consent templates and staff training. Stage 3 (Monitoring) demands real-world performance tracking infrastructure (Figure 1). This stage requires dedicated data analyst time. Stage 4 (Governance) establishes institutional accountability frameworks through incident review processes analogous to morbidity and mortality conferences and predetermined thresholds for triggering regulatory re-evaluation. The total institutional investment for operationalizing this pathway is non-trivial, as preliminary cost analyses suggest $150,000-$300,000 annually per deployed AI system for a mid-sized healthcare institution.

A critical distinction should be made between technical explainability and clinical interpretability. Nazir et al.[39] demonstrated that widely used explainability methods, such as gradient-based saliency maps, SHAP values, and attention visualizations, may generate technically valid outputs that are clinically meaningless or misleading. A saliency map highlighting image regions is technically explainable but not clinically interpretable if it identifies anatomically implausible features or fails to correspond to established diagnostic reasoning.[39] However, Ghassemi et al.[37] argued that current XAI methods often produce post-hoc rationalizations rather than genuine explanations of model behavior, creating false confidence among clinicians and regulators. The crucial question is not whether an explanation can be generated but whether that explanation is meaningful and actionable for a clinician at the point of care. Nonetheless, Blackman and Veerapen[73] proposed that explainability requirements should be functional rather than technical: can the clinician use the explanation to identify when the AI is likely wrong? These functional criteria reveal that most current XAI methods fail the test of clinical interpretability despite passing tests of technical explainability.

A critical limitation of the legal liability evidence should be explicitly acknowledged, as the legal framework remains underdeveloped and there is limited case law specifically addressing AI-related diagnostic errors in clinical practice. The liability scenarios described by Cestonaro et al.[40] and Nji[41] represent forward-looking analyses of how existing legal doctrines might theoretically apply to AI-

mediated diagnostic harm. However, these frameworks have not been meaningfully tested in court for continuously adaptive AI systems.

Limitations of the Study

Several methodological limitations warrant acknowledgment. First, while a detailed review protocol was developed a priori specifying all methodological procedures, the protocol was not prospectively registered in PROSPERO before study screening commenced. This decision reflects project timeline constraints and the fact that PROSPERO requires registration before data extraction begins. However, the absence of prospective registration does not affect the rigor of the systematic review methodology, which followed pre-specified eligibility criteria, search strategies, and quality assessment procedures, but it does limit the ability of readers to verify that protocol deviations did not occur post hoc. Future systematic reviews in this domain would benefit from prospective PROSPERO registration to enhance methodological transparency. Another limitation concerns the generalizability of these conclusions. The geographic concentration of included evidence in the United States (n = 38 studies) and the European Union (n = 24 studies) reflects not only a methodological constraint but also a significant gap in the global evidence base on AI governance. Governance challenges in low- and middle-income countries (LMICs) may differ substantially because of structural conditions that were largely absent from the included literature. In many LMIC settings, limitations in data infrastructure may further reduce dataset representativeness and intensify the risk of algorithmic bias. Regulatory capacity may also vary considerably, with some health systems lacking the resources required for robust oversight of adaptive AI technologies. Accordingly, the findings of this review should not be assumed to generalize to LMIC contexts without further empirical investigation. Future systematic reviews should prioritize the inclusion of governance research from LMICs to strengthen the global relevance and equity of this evidence base.

Conclusion

This systematic review found that AI diagnostic tools were associated with documented patient harm arising from algorithmic bias, limited explainability, and deficiencies in regulatory and institutional oversight. The findings suggest that the safe integration of diagnostic AI into clinical care depends not only on technical performance but also on governance structures that support transparency, accountability, clinician oversight, and ongoing postmarket evaluation. The evidence base, however, was concentrated largely in the United States and European Union, which limits the generalizability of these conclusions across diverse global health care settings. Further research is needed to examine how AI governance can be adapted to contexts with different regulatory capacities, health system infrastructures, and resource constraints. Overall, these findings support the need for more rigorous governance approaches to AI deployment in clinical practice to reduce preventable harm and promote equitable implementation.

Article Information

Author Contributions: Both authors had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Omolola Ogungbuaro

Acquisition, analysis, or interpretation of data: Both authors.

Drafting of the manuscript: Akinwale Obatayo

Critical revision of the manuscript for important intellectual content: Both authors.

Statistical analysis: Omolola Ogungbuaro

Conflict of Interest Disclosures: Both authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.

Data Sharing Statement: No individual participant data were collected in this systematic review of published literature. The complete search strategy, inclusion and exclusion criteria, standardized data

extraction form, and GRADE evidence quality rating worksheets are available from the corresponding author upon reasonable request.

Funding/Support: No specific funding was received for this study.

Role of the Funder/Sponsor: Not applicable.

Additional Contributions: The authors thank the academic health sciences librarians who provided consultation on database search strategy development and optimization. No compensation was received for this contribution.

Appendices

Appendix 1: The Included Studies Characteristics

Table 2: Characteristics of Included Studies

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
Zhou et al. (2021)[58]	Medical imaging	Deep learning for image-based diagnosis	technical review	Safety	—	High
Singh et al. (2025)[59]	US healthcare regulation	Clinical software as medical devices (AI/ML)	Policy analysis	FDA oversight	—	High
Sutton et al. (2020)[60]	multi-specialty	Clinical decision support systems (CDSS)	Narrative review	Safety	—	High
Pietrobon et al. (2025)[61]	Clinical risk prediction	Bias assessment in ML models	analytical study	Algorithmic bias	—	High
Aldhafeeri (2025)[62]	Radiology	AI governance frameworks	Systematic review	legal governance	—	High
Ekundayo & Nyavor (2024)[63]	Cardiovascular medicine	Predictive analytics for diagnosis and risk	Narrative review	Early diagnosis	—	Moderate
Neri et al. (2023)[64]	Radiology	Explainable AI (XAI)	Expert consensus	Explainability, safety	—	High
Chen & Asch (2022)[65]	General medicine	Predictive ML models	perspective	Safety	—	Moderate
Wachter et al. (2020)[66]	Data protection	Automated decision-making	Legal and doctrinal analysis	Explainability	—	High
Gu et al. (2025)[67]	Ophthalmology	Deep learning for retinal disease diagnosis	Quantitative diagnostic accuracy study	Safety, generalizability	Multiple datasets	High

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
Garcia Santa Cruz et al. (2021)[68]	Radiology (COVID-19 imaging)	AI models using public X-ray datasets	Systematic review	dataset governance	—	High
Amann et al. (2020)[69]	Healthcare (multidisciplinary)	Explainable AI	interdisciplinary analysis	Explainability	—	High
Ahmed et al. (2023)[70]	Healthcare systems	AI implementation across specialties	Systematic review	governance	—	High
Abbott et al.[23]	Emergency Medicine, USA	AI systems with bias in emergency care	Primer / Review	Bias identification, addressing bias in AI emergency applications	—	Moderate
Di Palma et al.[26]	Healthcare / Cybersecurity	AI-induced cybersecurity risks; blockchain solutions	Narrative review	Cybersecurity vulnerabilities, clinical risk management	—	Moderate
Bailo et al.[71]	Healthcare governance	Fairness, transparency, human oversight in AI	Narrative review	Real-world AI governance; fairness, transparency, oversight coexistence	—	Moderate
Gaffney et al.[31]	Health financing / Public health, USA	COVID-19 health financing; algorithmic resource allocation	Policy analysis	Resource allocation bias; proxy-variable health inequity	—	High
Glocker et al.[32]	Radiology (Chest X-ray)	Deep learning foundation models for chest radiography	Empirical bias analysis	Risk of bias in chest radiography AI; demographic performance disparities	Multiple datasets	High
Hanna et al.[72]	Pathology / AI/ML	Ethical and bias considerations in pathology AI	Review / Perspective	Algorithmic bias, ethical AI deployment in pathology	—	Moderate
Ramadan et al.[34]	Nursing practice	AI adoption in nursing	Qualitative study	Implementation barriers and facilitators; clinician trust and training	N/A	Moderate
Ghassemi et al.[37]	Healthcare (general)	Explainable AI (XAI) methods	Critical review	False hope of current XAI methods; misleading explanations	—	High
Blackman and Veerapen[73]	Clinical AI / Healthcare law	Clinical AI explainability requirements	Legal and ethical analysis	Practical, ethical, legal necessity of AI	—	High

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
				explainability		
Cestonaro et al.[40]	Medical liability / Healthcare law	AI in diagnostic algorithms	Systematic review	Medical liability when AI contributes	—	High
Park[43]	Patient perspectives / Informed consent	Medical AI and patient consent	Web-based experiment	Informed consent for AI; patient awareness and preferences	N/A	Moderate
Kuzio[46]	AI regulation (comparative)	EU vs US AI legal frameworks	Comparative legal analysis	Regulatory divergence; EU AI Act vs US frameworks	—	High
Carvalho et al.[47]	Healthcare regulation / AI-ML	Predetermined change control plans for AI	Policy guidance	Safe, effective AI-ML regulation; post-market adaptive oversight	—	High
Aboy and Vayena[49]	Digital medicine / EU regulation	EU AI Act implications for medical devices	Regulatory analysis	EU AI Act impact on regulated digital medical products	—	High
Malgieri[51]	Data protection / EU member states	Automated decision-making under GDPR	Legal doctrinal analysis	Right to explanation; GDPR Article 22 safeguards across EU nations	—	High
Holzinger et al.[52]	AI oversight (general)	Human oversight of AI systems	Commentary / Perspective	Feasibility of human oversight in complex AI; explainability limits	—	Moderate
Kim et al.[54]	Health information technology	HIT problems and patient outcomes	Systematic review	Implementation failures; HIT errors affecting care delivery	—	Moderate
McGaughey et al.[57]	Acute hospital wards / Early warning systems	AI early warning and rapid response systems	Cochrane systematic review	Implementation of early warning systems; patient safety outcomes	—	High
Beauchemin et al.[74]	Clinical practice guidelines / Nursing	Implementation of clinical practice guidelines	Review / Implementation science	Barriers and facilitators to guideline implementation in healthcare	—	Moderate

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
Ahmad et al.[24]	AI systems (general)	Bias in AI systems	Formal and socio-technical analysis	Integrating formal and socio-technical approaches to AI bias	—	Moderate
Algarni and Thayananthan[27]	Digital health / Cybersecurity	AI-based assistive technology cybersecurity	Review	Cybersecurity analysis of AI assistive systems in digital health	—	Moderate
Bignami et al.[50]	EU AI regulation	EU AI Act in global uncertainty context	Policy commentary	Balancing innovation and regulatory control in EU AI Act	—	Moderate
Chee et al.[75]	Intensive care / Emergency / COVID-19	AI applications for COVID-19 in ICU and emergency	Systematic review	COVID-19 AI applications; safety and deployment readiness	—	Moderate
Cherezov et al.[33]	Chest X-ray / Radiology	Fairness assessment in chest X-ray AI	Quantitative empirical study	Impact of technical and population factors on AI fairness	Multiple datasets	High
Goktas & Grzybowski[25]	Healthcare (general) / Clinical ethics	Ethical challenges in trustworthy AI	Narrative review	Ethical clinical challenges; pathways to trustworthy healthcare AI	—	Moderate
Hadian et al.[55]	Non-communicable disease management / Iran	AI innovations in NCD management	Expert perspectives / Qualitative	AI-driven NCD management challenges based on Iranian health system	N/A	Low–Moderate
Hurley et al.[45]	Patient consent / Bioethics	Patient consent and AI notice/explanation rights	Bioethical analysis	Right to notice and explanation of AI in healthcare; informed consent gaps	—	Low–Moderate
Kocak et al.[29]	Medical imaging AI / Evaluation metrics	Evaluation metrics in medical imaging AI	Methodological review	Fundamentals, pitfalls, misapplications of AI evaluation metrics	—	High
Mehta et al.[48]	FDA-reviewed medical devices	Transparency in AI/ML model characteristics	Empirical regulatory review	Evaluation of transparency in FDA-reviewed AI/ML medical	N/A	High

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
				devices		
Mennella et al.[53]	Healthcare AI (general)	Ethical and regulatory challenges in healthcare AI	Narrative review	Ethical and regulatory AI challenges; governance frameworks	—	Moderate
Nazir et al.[39]	Biomedical imaging / Deep learning	Explainable AI for biomedical imaging	Survey / Review	XAI techniques for biomedical imaging with deep neural network	—	Moderate
Nji[41]	Automated healthcare / Liability	Liability in AI-powered healthcare systems	Framework design / Legal analysis	Multi-stakeholder responsibility framework for AI healthcare liability	—	Low–Moderate
Obuchowicz et al.[30]	Radiology / AI	AI-empowered radiology	Critical review	Current status of AI in radiology; critical evaluation of deployment	—	Moderate
Rosenbacke et al.[36]	Clinical decision support / Trust	Explainable AI and clinician trust	Systematic review	How XAI increases or decreases clinician trust in healthcare AI	—	High
Sanchez et al.[35]	Healthcare AI operationalization	Trustworthy AI design framework	Design framework analysis	Requirements, tradeoffs, challenges for clinical AI adoption	—	Moderate
Smith[42]	Clinical decision-making / Medical law, UK	AI and clinical negligence/manslaughter	Legal analysis	Gross negligence and corporate manslaughter with AI clinical decisions	—	Moderate
Stogiannos et al.[56]	Radiology / AI implementation	Errors in AI implementation in radiology	Narrative review	Recognising and addressing AI implementation errors in radiology	—	Moderate
Subasri et al.[76]	Clinical AI deployment / Data shifts	Detecting harmful data shifts in clinical AI	Empirical study	Data shift detection for responsible clinical AI model	N/A	High

Author, Year	Setting / Specialty	AI Application	Design	Safety / Governance Domain	N	Quality
				deployment		
Vallenas & Parent[44]	Medical ethics / AI	Informed consent in value-laden medical AI	Ethical analysis	Informed consent challenges when AI embeds values in medical decisions	—	Moderate
Zhang et al.[38]	Radiology AI / Saliency methods	Trustworthiness of saliency methods in radiology	Empirical evaluation	Revisiting reliability of XAI saliency methods for radiology AI	N/A	High
Mann et al.[77]	Clinical AI deployment / Data shifts	Detecting harmful data shifts in clinical AI	Empirical study	Data shift detection for responsible clinical AI model deployment	N/A	Moderate

Abbreviations: SR, systematic review; AI, artificial Intelligence; ML, machine learning. Quality assessed using GRADE criteria (see Table 2). **Table 2** presents characteristics of included studies. The 'N' column indicates sample size for empirical studies, systematic reviews and policy analyses, the (—) stands for not applicable and N/A not reported by authors.

Appendix 2: The Safety Risks and Recommended Safeguards

Table 3: Common Safety Risks and Recommended Safeguards

Risk Domain	Specific Risk	Clinical Example	Recommended Safeguard	Evidence Grade
Patient Safety	Diagnostic error from overreliance on AI output	False-negative AI mammography accepted without independent radiologist review	explicit override protocols	Moderate
Algorithmic Bias	Training data underrepresentation of minority groups	Risk-scoring algorithm underestimated illness severity in Black patients	Diverse training datasets	High
Explainability	Black-box models; clinicians unable to justify AI-driven decisions to colleagues or patients	Unexplained AI sepsis-risk flag rejected by clinical team lacking interpretable rationale	Explainability requirements calibrated to decision stakes; XAI standards for high-risk tools	Moderate
Legal Liability	Unclear responsibility when AI-assisted diagnosis causes harm to patient	AI misses fracture	AI disclosure in informed consent	Low–Moderate
Regulatory	Absence of post-market	Cleared AI model silently	Mandatory real-world	Moderate

Risk Domain	Specific Risk	Clinical Example	Recommended Safeguard	Evidence Grade
Oversight	surveillance for continuously learning models	degrades after distribution shift in patient population	performance monitoring	
Data Privacy	Non-compliant data collection	Patient imaging data shared with developer without HIPAA/GDPR compliance	Stringent data governance	Moderate
Implementation	Workflow disruption	AI-generated alerts ignored after repeated false positives in ICU setting	Clinician co-design	Moderate

Abbreviations: AI, artificial Intelligence; XAI, explainable artificial Intelligence; HIPAA, Health Insurance Portability and Accountability Act; GDPR, General Data Protection Regulation; ICU, intensive care unit (Table 3). Evidence grades based on GRADE framework (see Table 4).

Appendix 3: Included Studies Quality Ratings

Table 4: Evidence Quality Ratings by Domain (GRADE Framework)

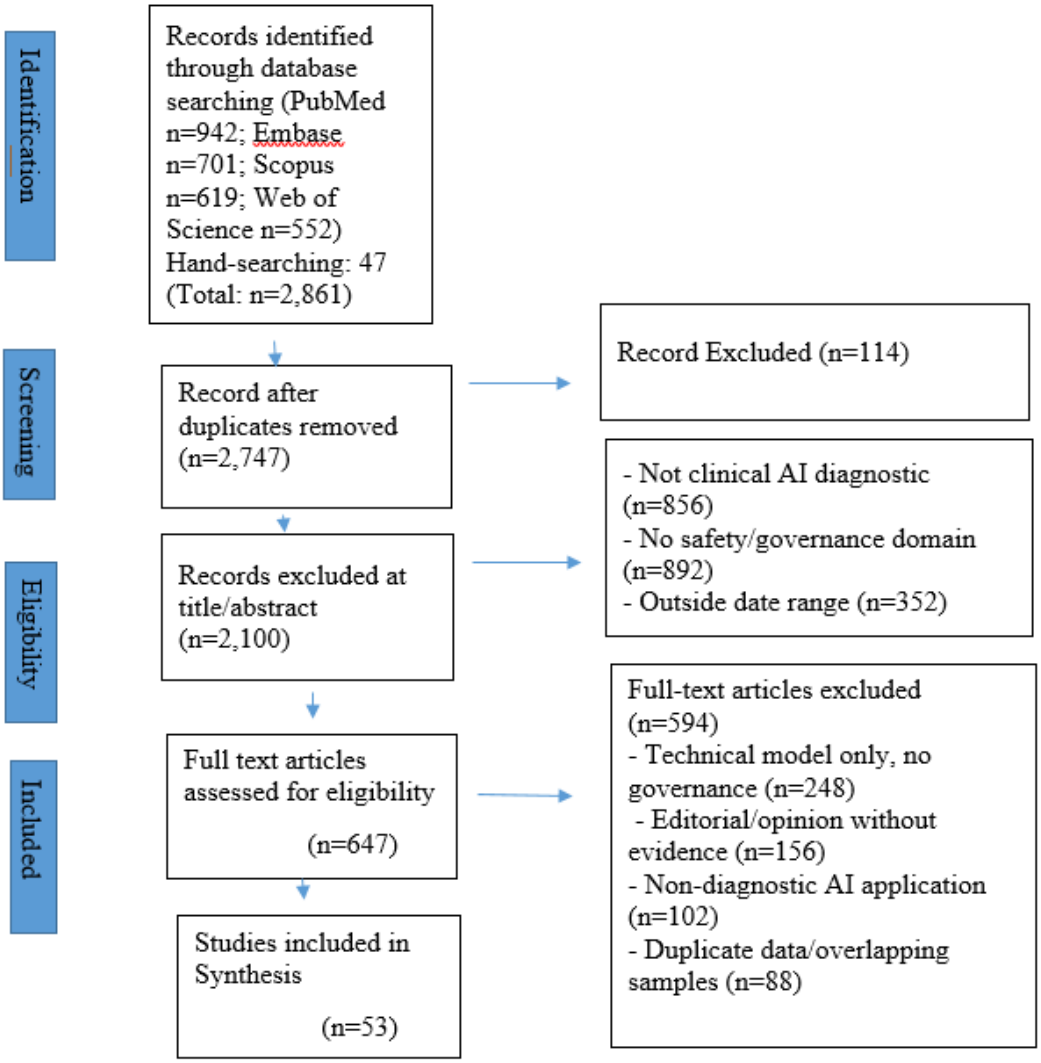
Evidence Domain	No. Studies	Study Designs	GRADE Quality	Rationale for Grade
Algorithmic bias	12	Empirical cohort, SR	Moderate	Consistent, replicated findings across settings
Regulatory frameworks	18	Policy analyses, SR	Moderate	Frameworks well-described and consistent
Legal liability	11	Legal analyses	Low	limited empirical data on adjudicated AI harm cases
Explainability and transparency	16	Empirical, SR	Moderate	Consistent call for XAI requirements
Patient safety	14	SR of AI models	Moderate	Strong signals from model audits
Implementation barriers	18	Qualitative, SR	Moderate	setting-specific variability limits generalizability
Post-market surveillance	8	Policy analyses	Low	Sparse empirical monitoring data

GRADE quality levels: **High** = consistent results from well-designed studies with low risk of bias; **Moderate** = some concern regarding study design, consistency, or precision; **Low** = significant concerns regarding study design, consistency, or applicability to review question.

Abbreviations: SR, systematic review; XAI, explainable artificial Intelligence; GRADE, Grading of Recommendations Assessment, Development and Evaluation.

Appendix 4: The PRISMA Flowchart

Figure 2: The PRISMA Flowchart



Note- This PRISMA flowchart was adapted from Page et al. The chart demonstrates the systematic study selection process from preliminary database searches throughout the final inclusion, with exclusion reasons documented at every stage to ensure transparency

References

References

[1] Duggal M, Chauhan A, Gupta V, et al. Real-world evaluation of AI-driven diabetic retinopathy screening in public health settings: validation and implementation study. *JMIR Medical Informatics*. 2025;13(2):1-9. doi:10.2196/67529

[2] Guo L, Zhou C, Xu J, Huang C, Yu Y, Lu G. Deep learning for chest X-ray diagnosis: competition between radiologists with or without Artificial Intelligence Assistance. *Journal of Imaging Informatics in Medicine*. 2024;37(3):922-934. doi:10.1007/s10278-024-00990-6

[3] Nadour N, Duguet T, Zahedi S, Figoni H, Liard R. Diagnostic accuracy of artificial Intelligence compared to family physicians and dermatologists for skin conditions: a systematic review and meta-analysis. *BMC Primary Care*. 2025;26(1):1-9. doi:10.1186/s12875-025-03073-9

[4] Ogut E. Artificial Intelligence in clinical medicine: challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery. *Clinics and Practice*. 2025;15(9):169-182. doi:10.3390/clinpract15090169

[5] Binder D, Bourgeois M. Direct and indirect effects of group discussion on consensus. *Social Influence*. 2026;1(4):249-264. doi:10.1080/15534510600867098

[6] Ardic N, Dinc R. Artificial Intelligence in healthcare: current regulatory landscape and future directions. *British Journal of Hospital Medicine*. 2025;86(8):1-21. doi:10.12968/hmed.2024.0972

[7] Cajueiro D, Celestino V. A comprehensive review of artificial intelligence regulation: weighing ethical principles and innovation. *Journal of Economy and Technology*. 2026;4(2):77-91. doi:10.1016/j.ject.2025.07.001

[8] Liu F, Niu Y, Zhang Q, et al. A foundational architecture for AI agents in healthcare. *Cell Reports Medicine*. 2025;6(10):10-23. doi:10.1016/j.xcrm.2025.102374

[9] Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. *Nature Medicine*. 2020;26(9):1327-1328. doi:10.1038/s41591-020-1020-3

[10] Jia L-L, Zhao J-X, Pan N-N, et al. Artificial intelligence model on chest imaging to diagnose COVID-19 and other pneumonias: a systematic review and meta-analysis. *European Journal of Radiology Open*. 2022;9(4):1-9. doi:10.1016/j.ejro.2022.100438

[11] Suleman MU, Mursaleen M, Khalil U, et al. Assessing the generalizability of Artificial Intelligence in radiology: a systematic review of performance across different clinical settings. *Annals of Medicine & Surgery*. 2025;87(12):8803-8811. doi:10.1097/ms9.0000000000004166

[12] Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021;13(4):1-9. doi:10.1136/bmj.n160

[13] Page M, Shamseer L, Tricco A. Registration of systematic reviews in Prospero: 30,000 records and counting. *Systematic Reviews*. 2018;7(1):1-9. doi:10.1186/s13643-018-0699-4

[14] Gusenbauer M, Gauster S. How to search for literature in systematic reviews and meta-analyses: a comprehensive step-by-step guide. *Technological Forecasting and Social Change*. 2025;212(4):12-38. doi:10.1016/j.techfore.2024.123833

[15] Valizadeh A, Moassefi M, Nakhostin-Ansari A, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Medical Research Methodology*. 2022;22(1):1-9. doi:10.1186/s12874-022-01631-8

[16] Almarie B, Gonzalez-Gonzalez LF, dos Santos Barbosa LA, Lutz A, Grosse U, Fregni F. Machine learning-enabled medical devices authorized by the US Food and Drug Administration in 2024: regulatory characteristics, predicate lineage, and transparency reporting. *Biomedicines*. 2025;13(12):1-9. doi:10.3390/biomedicines13123005

[17] Mitani AA, Freer PE, Nelson KP. Summary measures of agreement and association between many raters' ordinal classifications. *Annals of Epidemiology*. 2020;27(10):1-9. doi:10.1016/j.annepidem.2017.09.001

[18] Choi G, Kang H. Heterogeneity in meta-analyses: an unavoidable challenge worth exploring. *Korean Journal of Anesthesiology*. 2025;78(4):301-314. doi:10.4097/kja.25001

[19] Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. *Research Synthesis Methods*. 2019;10(4):539-545. doi:10.1002/jrsm.1369

[20] He Z, Yang L, Li X, Du J. Discrepancies in reported results between trial registries and journal articles for AI clinical research. *eClinicalMedicine*. 2025;80(2):1-9. doi:10.1016/j.eclinm.2024.103066

[21] Alqattan H, Morrison Z, Cleland JA. A narrative synthesis of qualitative studies conducted to assess patient safety culture in hospital settings. *Sultan Qaboos University Medical Journal*. 2019;19(2):1-9. doi:10.18295/squmj.2019.19.02.002

[22] Iseko A. Diversity as ethical infrastructure: reimagining AI governance for justice and accountability. *International Journal of Science, Technology and Society*. 2025;13(5):190-204. doi:10.11648/j.ijsts.20251305.13

[23] Abbott EE, Rehman T, Rosania A, et al. Understanding and addressing bias in artificial intelligence systems: a primer for the emergency medicine physician. *JACEP Open*. 2025;7(1):10-32. doi:10.1016/j.acepjo.2025.100311

[24] Ahmad A, Vallès Y, Idaghdour Y. Bias in AI systems: integrating formal and socio-technical approaches. *Frontiers in Big Data*. 2025;8(6):1-9. doi:10.3389/fdata.2025.1686452

[25] Goktas P, Grzybowski A. Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine*. 2025;14(5):1-12. doi:10.3390/jcm14051605

[26] Di Palma G, Scendoni R, Ferorelli D, De Benedictis A, Tambone V, De Micco F. AI-induced cybersecurity risks in healthcare: a narrative review of blockchain-based solutions within a clinical risk management framework. *Risk Management and Healthcare Policy*. 2025;Volume 18(2):3479-3497. doi:10.2147/rmhp.s544523

[27] Algarni A, Thayananthan V. Cybersecurity for analyzing Artificial Intelligence (AI)-based assistive technology and systems in digital health. *Systems*. 2025;13(6):439-452. doi:10.3390/systems13060439

[28] Bernstein MH, Atalay MK, Dibble EH, et al. Can incorrect Artificial Intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology*. 2023;33(11):8263-8269. doi:10.1007/s00330-023-09747-1

[29] Kocak B, Klontzas ME, Stanzione A, et al. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*. 2025;3(2):1-22. doi:10.1016/j.ejrai.2025.100030

[30] Obuchowicz R, Lasek J, Wodziński M, Piórkowski A, Strzelecki M, Nurzynska K. Artificial Intelligence-empowered radiology—current status and critical review. *Diagnostics*. 2025;15(3):282-291. doi:10.3390/diagnostics15030282

[31] Gaffney A, Himmelstein DU, Woolhandler S. Covid-19 and US health financing: perils and possibilities. *International Journal of Health Services*. 2020;50(4):396-407. doi:10.1177/0020731420931431

[32] Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*. 2023;5(6):1-9. doi:10.1148/ryai.230060

[33] Cherezov D, Fu P, Madabhushi A. Quantitative assessment of impact of technical and population-based factors on fairness of AI models for chest X-ray scans. *Computers in Biology and Medicine*. 2025;198(4):1-22. doi:10.1016/j.compbiomed.2025.111147

[34] Ramadan OM, Alruwaili MM, Alruwaili AN, Elsehrawy MG, Alanazi S. Facilitators and barriers to AI adoption in nursing practice: a qualitative study of Registered Nurses' perspectives. *BMC Nursing*. 2024;23(1):1-9. doi:10.1186/s12912-024-02571-y

[35] Sanchez P, Del Ser J, van Gils M, Hernesniemi J. A design framework for operationalizing trustworthy artificial intelligence in healthcare: requirements, trade-offs and challenges for its clinical adoption. *BMC Nursing*. 2025;13(6):1-9. doi:10.2139/ssrn.5249603

[36] Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *JMIR AI*. 2024;3(2):1-12. doi:10.2196/53207

[37] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable Artificial Intelligence in health care. *The Lancet Digital Health*. 2021;3(11):1-21. doi:10.1016/s2589-7500(21)00208-9

[38] Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiology: Artificial Intelligence*. 2024;6(1):1-9. doi:10.1148/ryai.220221

[39] Nazir S, Dickson DM, Akram MU. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*. 2023;156(2):10-36. doi:10.1016/j.compbiomed.2023.106668

[40] Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when Artificial Intelligence is applied to diagnostic algorithms: a systematic review. *Frontiers in Medicine*. 2023;10(9):1-21. doi:10.3389/fmed.2023.1305756

[41] Nji N. Navigating liability in automated healthcare: designing multi-stakeholder framework for responsibility in AI-powered health care systems. *International Journal of Research and Innovation in Social Science*. 2025;IX(VIII):5726-5734. doi:10.47772/ijriss.2025.908000466

[42] Smith H. Artificial Intelligence for clinical decision-making: gross negligence manslaughter and corporate manslaughter. *The New Bioethics*. 2024;30(3):228-242. doi:10.1080/20502877.2024.2416862

[43] Park H. Patient perspectives on informed consent for medical AI: a web-based experiment. *DIGITAL HEALTH*. 2024;10(2):1-9. doi:10.1177/20552076241247938

[44] Vallenas N, Parent B. Informed consent in the age of value-laden medical AI. *Journal of Ethics in Entrepreneurship and Technology*. 2025;13(2):1-12. doi:10.1108/jeet-06-2025-0044

[45] Hurley ME, Lang BH, Kostick-Quenet KM, Smith JN, Blumenthal J. Patient consent and the right to notice and explanation of AI systems used in health care. *The American Journal of Bioethics*. 2024;25(3):102-114. doi:10.1080/15265161.2024.2399828

[46] Kuzior A. Navigating AI regulation: a comparative analysis of EU and US legal frameworks. *Materials Research Proceedings*. 2024;45(2):258-266. doi:10.21741/9781644903315-30

[47] Carvalho E, Mascarenhas M, Pinheiro F, et al. Predetermined change control plans: guiding principles for advancing safe, effective, and high-quality AI-ML technologies. *JMIR AI*. 2025;4(2):1-9. doi:10.2196/76854

[48] Mehta V, Komanduri A, Bhadouriya RS, et al. Evaluating transparency in AI/ML model characteristics for FDA-reviewed medical devices. *npj Digital Medicine*. 2025;8(1):1-9. doi:10.1038/s41746-025-02052-9

[49] Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. *npj Digital Medicine*. 2024;7(1):1-9. doi:10.1038/s41746-024-01232-3

[50] Bignami E, Russo M, Semeraro F, Bellini V. Balancing innovation and control: the European Union AI Act in an ERA of global uncertainty. *JMIR AI*. 2025;4(2):10-19. doi:10.2196/75527

[51] Malgieri G. Automated decision-making in the EU member states: the right to explanation and other “suitable safeguards” in the national legislations. *Computer Law & Security Review*. 2019;35(5):1-9. doi:10.1016/j.clsr.2019.05.002

[52] Holzinger A, Zatloukal K, Müller H. Is human oversight to AI systems still possible? *New Biotechnology*. 2025;85(4):59-62. doi:10.1016/j.nbt.2024.12.003

[53] Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. *Heliyon*. 2024;10(4):1-9. doi:10.1016/j.heliyon.2024.e26297

[54] Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. *Journal of the American Medical Informatics Association*. 2019;24(2):246-250. doi:10.1093/jamia/ocw154

[55] Hadian M, jabbari KM, Mohammadi M, Jafari A. AI-driven innovations in NCD management: challenges and solutions based on expert perspectives in Iran’s health system. *BMC Nursing*. 2025;13(2):1-12. doi:10.21203/rs.3.rs-7225629/v1

[56] Stogiannos N, Cuocolo R, Akinci D’Antonoli T, et al. Recognising errors in AI implementation in radiology: a narrative review. *European Journal of Radiology*. 2025;191(2):11-23. doi:10.1016/j.ejrad.2025.112311

[57] McGaughey J, Fergusson DA, Van Bogaert P, Rose L. Early warning systems and rapid response systems for the prevention of patient deterioration on Acute Adult Hospital wards. *Cochrane Database of Systematic Reviews*. 2021;20(1):1-9. doi:10.1002/14651858.cd005529.pub3

[58] Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*. 2021;109(5):820-838. doi:10.1109/jproc.2021.3054390

[59] Singh V, Cheng S, Kwan AC, Ebinger J. United States Food and Drug Administration regulation of clinical software in the era of Artificial Intelligence and Machine Learning. *Mayo Clinic Proceedings: Digital Health*. 2025;3(3):10-23. doi:10.1016/j.mcpdig.2025.100231

[60] 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 Digital Medicine*. 2020;3(1):1-9. doi:10.1038/s41746-020-0221-y

[61] Pietrobon R, Machiavelli A, Paulsen Rodrigues L, et al. The bias assessment method: evaluating racial inequities in clinical risk prediction models (motivated by Obermeyer et al.'s analysis of cost-driven labeling in healthcare algorithms). *SSRN Electronic Journal*. 2025;13(2):10-21. doi:10.2139/ssrn.5861522

[62] Aldhafeeri F. Governing Artificial Intelligence in radiology: a systematic review of ethical, legal, and regulatory frameworks. *Diagnostics*. 2025;15(18):1-22. doi:10.3390/diagnostics15182300

[63] Ekundayo F, Nyavor H. Ai-driven predictive analytics in cardiovascular diseases: integrating big data and machine learning for early diagnosis and risk prediction. *International Journal of Research Publication and Reviews*. 2024;5(12):1240-1256. doi:10.55248/gengpi.5.1224.3437

[64] Neri E, Aghakhanyan G, Zerunian M, et al. Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology. *La radiologia medica*. 2023;128(6):755-764. doi:10.1007/s11547-023-01634-5

[65] Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *New England Journal of Medicine*. 2022;376(26):2507-2509. doi:10.1056/nejmp1702071

[66] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*. 2020;7(2):76-99. doi:10.1093/idpl/ipx005

[67] Gu X, Zhou Y, Zhao J, et al. Diagnostic performance and generalizability of Deep Learning for multiple retinal diseases using bimodal imaging of fundus photography and optical coherence tomography. *Frontiers in Cell and Developmental Biology*. 2025;13(4):1-9. doi:10.3389/fcell.2025.1665173

[68] Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public COVID-19 X-ray datasets and their impact on model bias – a systematic review of a significant problem. *Medical Image Analysis*. 2021;74(2):10-22. doi:10.1016/j.media.2021.102225

[69] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for Artificial Intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*. 2020;20(1):1-9. doi:10.1186/s12911-020-01332-6

[70] Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of Artificial Intelligence in healthcare. *Cureus*. 2023;13(2):9. doi:10.7759/cureus.46454

[71] Bailo P, Nittari G, Pesel G, Basello E, Spasari T, Ricci G. Governing healthcare AI in the real world: how fairness, transparency, and human oversight can coexist: a narrative review. *Sci*. 2025;8(2):36-45. doi:10.3390/sci8020036

[72] Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*. 2025;38(3):1-12. doi:10.1016/j.modpat.2024.100686

[73] Blackman J, Veerapen R. On the practical, ethical, and legal necessity of clinical Artificial Intelligence explainability: an examination of key arguments. *BMC Medical Informatics and Decision Making*. 2025;25(1):1-9. doi:10.1186/s12911-025-02891-2

[74] Beauchemin M, Cohn E, Shelton RC. Implementation of clinical practice guidelines in the health care setting. *Advances in Nursing Science*. 2019;42(4):307-324. doi:10.1097/ans.0000000000000263

[75] Chee ML, Ong ME, Siddiqui FJ, et al. Artificial intelligence applications for covid-19 in intensive care and emergency settings: a systematic review. *SSRN Electronic Journal*. 2021;13(2):1-23. doi:10.1101/2021.02.15.21251727

[76] Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Network Open*. 2025;8(6):1-12. doi:10.1001/jamanetworkopen.2025.13685

[77] Mann S, Berdahl CT, Baker L, Girosi F. Artificial intelligence applications used in the clinical response to COVID-19: a scoping review. *PLOS Digital Health*. 2022;1(10):1-9. doi:10.1371/journal.pdig.0000132