From Ethical Principles to Auditable Systems: A Regulatory Operational Assurance Framework for Trustworthy Healthcare AI

Amna Akram[1], Kinza Riaz[2], Prakash K. Singh[3], Safwan Abdul Rahim[4], Aseel Smerat[5], Salbia Abbas[6]

[1] School of Life and Health Sciences, Birmingham City University, Birmingham, UK.
Email: Amna.akram3@mail.bcu.ac.uk

[2] Bachelors in Dentistry, University of Health Sciences, Pakistan

[3] Virginia Commonwealth University, USA. Email: SINGH1776@HOTMAIL.COM

[4] SRMO Fatmiyah Hospital. Email: safwanabdulrahim1@gmail.com

[5] Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602105, India; Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan. Email: smerat.2020@gmail.com

[6] Lecturer (Associate), Department of Psychology, GC Women University Sialkot.
Email: salbia.abbas@gcwus.edu.pk | ORCID: https://orcid.org/0000-0003-4524-3447

ABSTRACT

Artificial intelligence (AI) has proven significant capabilities of enhancing the accuracy of diagnoses, clinical decision making, and efficiency of care. Nevertheless, the successful application of AI in safety critical clinical contexts has not been well implemented in practice resulting in persistent concerns of fairness, accountability, transparency, and regulatory compliance. Current ethical frameworks define desirable principles of trustful AI, whereas technical approaches offer metrics and interpretability means, but neither offers a single, operational channel of regulatory assurance. This paper introduces the Trustworthy AI Assurance Framework (TAIAF), a regulatory operational model that redefines trustworthiness as a system assurance property, similar to safety or reliability. TAIAF incorporates fairness, accountability and transparency as verifiable and auditable properties throughout the AI lifecycle to align system design, implementation, and monitoring to U.S. regulatory expectations. The framework consists of five assurance layers integrated: Governance & Compliance, Data and Model Development, Decision and Control, Deployment and Runtime Monitoring and Audit and Accountability producing inspection read artifacts that can be certified, audited, and monitored in the after market. The empirical evidence of TAIAF is based on publicly available healthcare datasets on the U.S. such as NIH ChestX-ray14, MIMIC-III, and OpenFDA. Instead of benchmarking predictive performance, the case studies demonstrate how the three components of demographic risk stratification, data governance signals, provenance records, and post market transparency artifacts are operationalized as the components of trustworthiness. Findings demonstrate that fairness may be viewed as equitable exposure to diagnostic risk, accountability as reconstructions of traceability, and transparency as auditability other than explainability. This work bridges the gap between ethical intent, technical implementation, and regulatory oversight by actually mapping TAIAF to FDA guidance, NIST AI Risk Management Framework, and ISO 14971, as well as U.S. quality system regulations. TAIAF offers a viable, lifecycle grounded roadmap to the development, assessment, and governance of secure, disproportionate, and answerable healthcare AI systems within current regulatory frameworks.

Keywords: Trustworthy artificial intelligence, healthcare AI regulation, fairness and bias in medical AI, AI governance and accountability, regulatory assurance framework, lifecycle risk management

How to cite this article: Akram A, Riaz K, Singh PK, Rahim SA, Smerat A, Abbas S. From Ethical Principles to Auditable Systems: A Regulatory Operational Assurance Framework for Trustworthy Healthcare AI. Int J Drug Deliv Technol. 2026;16(23s): 24-36. DOI: 10.25258/ijddt.16.23s.3

Source of support: Nil.

Conflict of interest: None

1. Introduction

Artificial intelligence (AI) has become a revolutionary phenomenon in modern healthcare with unequivocal opportunities to increase diagnostic accuracy, streamline clinical decision-making, and increase the efficiency of the underlying systems. These

innovations solve acute problems such as the shortage of workforce, the rising cost of healthcare, and the increasing cost of chronic diseases (Fahim et al., 2025; Chustecki, 2024; Alghareeb and Aljehani, 2025). Recent achievements in machine learning (ML), deep learning (DL), and multimodal AI systems have shown performance on the level of experts in radiological interpretation, disease detection, and prognostic modeling, which also promises a new era of AI-enhanced clinical practice (Islam et al., 2025; Chinta et al., 2025; Zhang et al., 2026).

Regardless of these technological accomplishments, the implementation of AI technologies into clinical practice is significantly constrained (Hassan et al., 2024; Abdelwanis et al., 2026; El Arab et al., 2025; Qi et al., 2025). Ongoing issues related to the fairness of algorithms, transparency, accountability, interpretability, and governing systems still hinder accountable and sustainable implementation (Khan et al., 2025).

The ability of AI to decrease the time interval of diagnosis, support early disease diagnosis, and personalized treatment plans has been widely reported (Fahim et al., 2025; Borkar and De, 2025). The systematic reviews highlight its potential of transformation in medical imaging, predictive analytics, and clinical decision support systems (Bader, 2025; Olawade et al., 2024). Nevertheless, high level technical performance alone does not guarantee responsible usage in a clinical setting. Interpretable AI outputs, rigorously validated, regulated, and equitably applied to different patient groups are crucial to the realization of clinical value (Chong et al., 2025; Goktas and Grzybowski, 2025).

One of the issues that have been present in the literature has to do with the black box nature of most deep learning systems, whose non transparent computing mechanisms are not readily interpretable by any clinician or regulator (Xu and Shuttleworth, 2024; Sadeghi et al., 2024). This is a fundamental defect of algorithmic opacitance that contributes to clinical trust, difficulties regulatory assessment, and obscures possible error routes that can jeopardize patient safety (Tun et al., 2025; Saoudi et al., 2025). Therefore, building a sense of trust in AI in healthcare requires the shift toward the idea of putting forth the metrics of accuracy and combining them with the overall technical, ethical, and regulatory protection.

The issues of algorithmic bias and fairness are essential in the application of AI in healthcare. Unrepresentative training samples, flawed annotation procedures, or objectives guiding optimization may contribute to biases (Gao et al., 2025; Chinta et al., 2025). Empirical studies show that models being trained on demographically imbalanced samples will often perform poorly on underrepresented populations, thus worsening the currently experienced health inequity (Cross et al., 2024; Joseph, 2025). Modern research highlights the need to have a demographically balanced dataset, fairness-conscious algorithmity, and interdisciplinary mitigation models (Ueda et al., 2024; Sasseville et al., 2025). The lack of active equity tools puts AI systems at the risk of reinforcing diagnostic and prognostic inequities. Bioethical studies also contextualize fairness as a moral premise based on the guiding principles of justice, autonomy, and beneficence that support the field of medicine (Park et al., 2025; Gorelik et al., 2025).

Both fairness and clinical trust cannot exist without transparency and interpretability. The use of explainable AI recommendations is increasingly demanded by clinicians and regulatory bodies, especially the high stakes diagnostic decisions (Abgrall et al., 2024; Abbas et al., 2025). The opacitance of algorithms prevents strict validation, kills the process of informed consent, and hides the mechanisms of accountability. Explainable AI (XAI) research has shown that interpretability is another key to the safe clinical integration of AI, which provides opportunities to detect errors, make inferences, and align the derived results with accepted clinical reasoning models (Nasarian et al., 2024). Still, there are still notable gaps between existing interpretability practices and regulatory demands to auditability and compliance, which highlights the need to make transparency a measurable and testable system characteristic instead of a desired trait.

The healthcare AI regulatory environment is still in its early developmental phases (Schmidt et al., 2024). Although the U.S. Food and Drug Administration (FDA) has provided guidance to AI/ML based medical devices and is developing Software as a Medical Device (SaMD) frameworks (Warraich et al., 2025), current regulatory frameworks are disjointed and underdeveloped (Rahimzadeh, 2023). The National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) explicitly outlines the principles expected such as fairness, transparency, accountability, privacy, and robustness; but its voluntary adoption framework does not include any mechanisms to enforce these principles on individual practitioners (Tabassi, 2023). Likewise, the World Health Organization (WHO) has outlined ethical and governance requirements of healthcare AI and has

highlighted frameworks that can protect the benefit of the population but not undermine human rights or health equity (World Health Organization, 2024; 2021). Although such developments are made in regulation, there are still large gaps between policy frameworks and engineering implementation. Current standards mainly present visionary goals of credible AI without defining operationalization approaches, measurable metrics, and check procedures of regulated systems (Alelyani, 2025). As an example, although the FUTURE-AI principle outlines fairness, traceability, robustness, and explainability as the core pillars, it lacks enough tangible mechanisms aimed at compliance control and post market tracking (Lekadir et al., 2025).

The literature has three gaps that are interrelated:

1. Effectiveness vs. Trustworthiness: Research focuses on the performance measures and overlooks fairness, mitigation of bias, and interpretability that is needed to deploy the research safely (Amadi and Ojo, 2025).

2. Principles vs. Compliance Pathways: Ethical frameworks provide a description of desirable qualities, but do not provide auditable standards in accordance with regulatory requirements (Khan et al., 2025).

3. Lifecycle Integration: Fairness, transparency, and accountability have been considered as discrete challenges, and not as lifecycle properties (Nastoska et al., 2025).

These clinical gaps have a physical clinical impact. Lack of operational mechanisms, including model cards, demographic audits, explainability layers, and compliance trace logs, may cause regulatory rejection, mistrust among clinicians and un-safe patient outcomes. This paper will overcome these limitations by developing the Trustworthy AI Assurance Framework (TAIAF) an integrated lifecycle model, which entails ensuring fairness, accountability, and transparency as verifiable properties consistent with regulatory concepts like the FDA guidance and NIST AI RMF. TAIAF incorporates practical assurance strategies, such as equity auditing, unified documentation, and interpretability criteria; and satisfies regulatory demands of certification and surveillance; and incorporates reliability throughout the AI lifecycle. The model is empirically tested with public U.S. datasets, NIH ChestX-ray14 to analyze fairness, MIMIC-III to govern risks, and OpenFDA adverse event data to demonstrate pharmacovigilance transparency, which proves the functionality of the framework. In general, the research fills the gap between theoretical principles of high level trustworthiness and practical, regulated clinical AI systems and provides a systematic roadmap to safe, fair, and responsible healthcare innovation.

2. Literature Review

Artificial intelligence (AI) has grown rapidly in the medical field, triggering interdisciplinary research in computer science, medicine, ethics, and regulation. Although these initiatives have developed theoretical insights into reliable AI, they are still disjointed among ethical principles, technical metrics and regulatory directions. This section critically appraises existing literature in three spaces, namely, (i) credible AI schemes in healthcare, (ii) ways of fairness, accountability, and transparency, and (iii) regulatory and standards based policies, with gaps that persist in encouraging the desire to have a coherent, regulatory operational assurance scheme.

2.1 Trustworthy AI in Healthcare

The aspect of trustworthy AI has emerged as a core part of healthcare because clinical decision making is a matter of safety. The initial theoretical approaches conceptualized trustworthiness in terms of bioethical concepts, including beneficence, non maleficence, autonomy, justice, and explicability (Afroogh et al., 2024; Hildt, 2025). These foundations have developed into multidimensional constructs that include fairness, transparency, accountability, robustness, privacy, and human oversight (Ahadian et al., 2026). The studies also state that clinical trust is not simply a question of accuracy but also of reliability, equity, and transparency in different contexts (Astobiza et al., 2025; Tun et al., 2025; Khan et al., 2025). Nonetheless, conceptual definitions tend to have no tangible processes of engineering or auditing credibility within real systems.

Ethics frameworks are still one of the most impactful models of trustful AI (Papagiannidis et al., 2025). The international institutions and policy bodies have come up with principles of inclusiveness, transparency, human agency, and accountability (WHO, 2025; OECD, 2024; UNESCO, 2021). The principles of ethical and governance of AI in health are described by the WHO, and the use of AI in health is promoted by value based and human centered AI design that is proclaimed by UNESCO and national strategies. Academic models also support participatory growth, knowledgeable permission, and participation (Tzimourta, 2025; van Leersum and Maathuis, 2025).According to scholars, fairness and transparency are highlighted, but there are few measurable indicators and enforceable controls, which restricts the use of indicators in a regulated clinical setting (Chan et al., 2025; Solanki et al., 2023; Ahadian et al., 2026).

Governance studies are concerned with the institutional and organizational processes of managing AI in healthcare (Kim et al., 2025). The models suggested focus on shared responsibility among clinicians, developers, regulators, and patients (Rozenblit et al., 2025). Some of the strategies will be human in the loop oversight and organizational structures like AI oversight committees and internal audits (Freeman, 2025). Nevertheless, governance models are frequently not integrated with technical assurance procedures, but instead, they are procedural instead of being embedded in AI system engineering.

2.2 Fairness, Accountability, and Transparency Methods

One of the dimensions of trustworthy AI that have been researched widely is fairness. The quantitative fairness measures include demographic parity, equalized odds, and calibration to test the disparities by sex, age, ethnicity, and socioeconomic status (Rabonato and Berton, 2025). Audits of bias and subgroup analysis of performance are now widely used, especially in diagnostic imaging (Xu et al., 2024). However, they are analytical and not operational instruments of determining inequities but do not specify how fairness is to be established, recorded, and continued over time.

One of the dimensions of trustworthy AI that have been researched widely is fairness. The quantitative fairness measures include demographic parity, equalized odds, and calibration to test the disparities by sex, age, ethnicity, and socioeconomic status (Rabonato and Berton, 2025). Audits of bias and subgroup analysis of performance are now widely used, especially in diagnostic imaging (Xu et al., 2024). However, they are analytical and not operational instruments of determining inequities but do not specify how fairness is to be established, recorded, and continued over time.

Trust in healthcare AI requires transparency and explainability. The explainable AI (XAI) algorithms include SHAP, LIME, saliency maps, Grad-CAM, and counterfactual explanations, which aim to make predictions by a model understandable (Sadeghi et al., 2024). XAI helps in clinical validation, trust calibration, and shared decision making in healthcare (Houssein et al., 2025). Nonetheless, explanations might be not in line with regulatory requirements of auditability and documentation, and most approaches tend to be volatile or situational (Seth and Sankarapu, 2025). Therefore, explainability increases knowledge but not formal responsibility and readiness to comply. The majority of strategies are post training and not integrated with governance, deployment, or post market monitoring (Batool et al., 2025), which makes it impossible to ensure continuity in changing clinical environments.

2.3 Regulatory and Standards Based Approaches

The United States regulates AI-powered medical devices under domain like Software as a Medical Device (SaMD) and Good Machine Learning Practice (GMLP), and is focused on transparency, data quality, and risk management (Sandalow et al., 2025). FDA guidance accepts adaptive system difficulties such as performance drift but is deliberately loose to promote innovation (Sandalow et al., 2025). The flexibility also makes developers in charge of the operationalization of fairness, accountability, and transparency without standardized mechanisms or templates.

National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) provides a methodologically structured way of defining and addressing AI risks at system lifecycle stages, which plans activities as Govern, Map, Measure and Manage (Tabassi, 2023). It offers a common set of terms of reliance but it is clearly non prescriptivem (Dotan et al., 2024).These standards include ISO 14971 and the U.S. Quality System Regulation (21 CFR Part 820) that outline the standards of hazard identification, risk management, and traceability (ISO, 2019). Medical device safety is based on these standards, although they were not created to support adaptive, data-driven systems. They need to be reinterpreted when applied to AI in terms of data governance, model revision and bias control (Jenko et al., 2025). As they state the purpose of risk control, they fail to specify implementation of such in AI-based architectures.

In all regulatory and standards based frameworks, the recurring gap is the definition of expectations by standards, but not design mechanisms. They provide compliance goals without providing lifecycle built assurance procedures to fairness, accountability, and transparency. This has resulted in compliance being reactive and paper-intensive, but not proactively designed. Such ambiguity is a barrier to developers and regulators to introduce operationalizable trustworthy AI in healthcare.

The literature demonstrates a disorganized disposition of ethical, technical, and regulatory landscapes. Ethics systems tend to lay stress on values, but are not operationally specific. The technical means provide measures and instruments but lack systemic guarantees. Regulatory standards present requirements but not architectures that can be implemented.

Therefore, although the literature espouses principles of trustworthy AI, none of them offer a lifecycle sensitive, unified, regulatory operational framework,

whose different attributes of fairness, accountability, and transparency are implemented as verifiable system properties. In order to close the gap, the Trustworthy AI Assurance Framework (TAIAF) presented in this paper is proposed which combines ethical will, technical execution, and compliance with the law.

Table 1. Comparison of existing trustworthy AI approaches and TAIAF

Approach	Primary Focus	Verifiability	U.S. Regulatory Alignment	Lifecycle Coverage
Ethical AI Frameworks	Normative principles and values	No	No	Limited
Fairness / XAI Methods	Metrics and interpretability tools	Partial	No	Partial
Regulatory Guidance (FDA, NIST)	Oversight expectations	Partial	Yes	Partial
TAIAF (This Work)	Regulatory-operational assurance	By design	Explicit	Architectural

3. Research Methodology

3.1 Research Design

The proposed study applies a design science research approach in order to create and test a regulatory operational assurance framework of trustful healthcare artificial intelligence (AI). In contrast to performance oriented machine learning research, this work is based on the engineering of trustworthiness as verifiable system property consistent with regulatory oversight. Design science strategies are especially suited to the situation, when one aims to develop and justify well structured artifacts filling the gaps in theory and practice that have been identified.

The methodology consists of four phases: conceptual reframing of trustworthiness, architectural design of Trustworthy AI Assurance Framework (TAIAF), systematic regulatory mapping, and empirical validation by using publicly available U.S. healthcare data. This strategy is a reaction to reported disunity in ethical principles, technical measures, and regulatory advice in healthcare AI governance (Khan et al., 2025; Chan et al., 2025; Batool et al., 2025). The methodology is a translation of abstract trustworthiness principles into auditable lifecycle processes by implementing assurance mechanisms into system architecture.

3.2 Conceptual Modeling of Trustworthiness

3.2.1 Trustworthiness as a System Assurance Property

Healthcare AI is often viewed through the lens of clinician confidence or ethical intent (Afroogh et al., 2024; Astobiza et al., 2025). But these interpretations cannot be concluded to be adequate to safety critical systems that are to be reviewed by the regulatory authority. A conceptualisation of trustworthiness in this work is a system assurance property, similar to safety or reliability, the ability of a system to act in a predictable, transparent and responsible manner given operational and audit conditions.

This explanation is in line with the regulatory paradigms of medical technologies, under which approval and post-market surveillance are predetermined by the reported evidence of risk mitigation and traceability but not normative ethical concurrence (Rahimzadeh, 2023; Warraich et al., 2025). Correspondingly, fairness, accountability and transparency are not considered as aspirational, but rather verifiable lifecycle properties that are implemented in system architecture.

3.2.2 Fairness as Equitable Risk Exposure

Healthcare AI fairness has been extensively discussed in the context of statistical indicators, including demographic parity or equalized odds (Rabonato and Berton, 2025; Ueda et al., 2024). Although these metrics are helpful diagnostic tools, there is no intrinsic relationship between bias mitigation and regulatory risk management. Empirical studies prove that the occurrence of algorithmic bias can be based on the imbalance of the set of data, annotation, or optimizing goals, and the demographic groups can have differentiated performance (Chinta et al., 2025; Cross et al., 2024; Hasanzadeh et al., 2025).

This paper redefines fairness as equitable exposure to diagnostic risk in the prism of safety-engineering. In this sense, the difference in false negative or misclassification rates is considered to be inequitable exposure of certain groups to safety risks. The re-framing aligns the fairness assessment with the ISO 14971 risk management principles, in which diagnostic

errors are conceptualized as the hazards, which need to be identified, estimated, and controlled (ISO, 2019; Jenko et al., 2025). Fairness is consequently not a purely statistical outcome but a measurable safety property.

3.2.3 Accountability as Traceability

The concept of accountability in AI systems is frequently addressed through the prism of liability attribution after the negative results occur (Novelli et al., 2024; Habli et al., 2020). However, in multifaceted healthcare facilities, a distributed responsibility among the clinicians, developers, and institutions is necessary, which requires a system level traceability mechanism (Terranova et al., 2024).

In TAIAF, accountability is described as the ability to have decision provenance and reconstruction. A responsible system should have recorded versioning of model and datasets, retain input output connections and explicit definitions of human AI responsibility borders. Retrospective review and regulatory investigation is assisted by audit trails and documentation that is version controlled (Ojewale et al., 2026). Responsibility is therefore constituted into design and not activated once failure occurs.

3.2.4 Transparency as Auditability

Transparency is often confused with explainable AI (XAI), which is concerned with model prediction interpretability (Sadeghi et al., 2024; Abbas et al., 2025). Although explainability can improve clinician knowledge and enhance clinician trust calibration (Nasarian et al., 2024; Hildt, 2025), it does not necessarily meet regulatory standards of traceable documentation and control.

In this paper, explainability is differentiated with auditability, which facilitates regulatory verification by use of structured artifacts, including logs, validation reports, and lifecycle documentation. Transparency, in turn, includes interpretability along with long-lasting evidence creation that could be inspected and certified.

3.2.5 Lifecycle Integration

Reliability cannot be built at one point in development. Studies have shown that fairness, transparency, and governance systems are often implemented after the fact instead of being incorporated in system lifecycles (Singhal et al., 2024; Nastoska et al., 2025). TAIAF thus instantiates assurance measures throughout data governance, model development, deployment, runtime monitoring and post incident review, to provide continuity in safety and compliance.

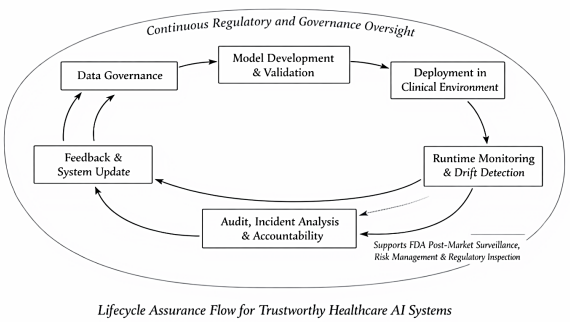

Figure 1. Lifecycle assurance perspective for trustworthy healthcare AI.

3.3 TAIAF Framework Design

Trustworthy AI Assurance Framework (TAIAF) constitutes the operationalization of the conceptual bases above, with the help of a layered lifecycle architecture. Instead of proposing new predictive algorithms, TAIAF outlines assurance mechanisms that produce inspection-ready artifacts which match U.S. regulatory expectations.

This framework follows five interrelated layers that are Governance and Compliance, Data and Model Development, Decision and Control, Deployment and Runtime Monitoring, and Audit and Accountability. All layers generate formatted documentation and traceable products that aid certification, tracking and investigating an incident.

The design complies with the NIST AI Risk Management Framework lifecycle functions (Tabassi, 2023; Dotan et al., 2024), FDA Good Machine Learning Practice, and ISO 14971 risk management processes (ISO, 2019) as well as the U.S. Quality System Regulation requirements (Rahimzadeh, 2023). Using this correspondence, TAIAF is a translation device between regulatory language and implementable system artifacts.

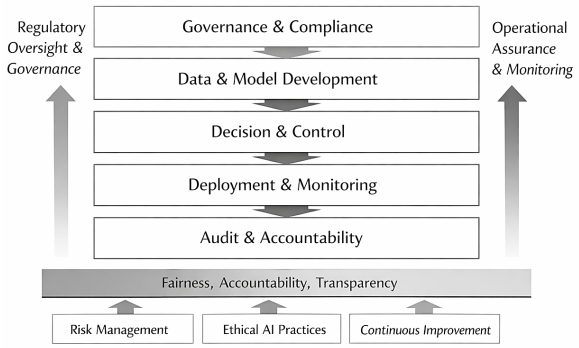

Figure 2. TAIAF Layered Architecture

Table 2. TAIAF layers, objectives, and assurance artifacts

TAIAF Layer	Primary Objective	Representative Assurance Artifacts
Governance & Compliance	Regulatory alignment and responsibility allocation	Regulatory applicability assessment, risk management plan (ISO 14971–aligned), governance charter
Data & Model Development	Fairness, bias identification, and pre-deployment risk control	Dataset datasheets, data provenance records, fairness and bias audit reports, model cards
Decision & Control	Clarity of decision authority and human–AI responsibility	Decision logic specifications, explainability documentation, human override and escalation protocols
Deployment & Runtime Monitoring	Continuous safety and risk monitoring in real-world use	Performance and fairness monitoring reports, drift detection summaries, incident alerts
Audit & Accountability	Traceability, auditability, and post-incident accountability	End-to-end audit logs, provenance records, incident investigation reports, CAPA documentation

3.4 Regulatory Mapping Method

There was a systematic regulatory mapping analysis of the alignment of TAIAF assurance artifacts to the existing U.S. healthcare regulatory frameworks. The mapping process entailed establishing regulatory expectations in terms of risk management, transparency, traceability, and lifecycle documentation and mapping them into architectural assurance mechanisms.

To illustrate, the Govern-Map-Measure-Manage functions of the NIST AI Risk Management Framework are operationalized in TAIAF in the form of governance charters, demographic stratification audits, performance monitoring logs, and formalized procedures of escalation. FDA Good Machine Learning Practice recommendations on transparency and lifecycle documentation (Warraich et al., 2025) are applied in the form of model cards, dataset datasheets and version-controlled validation reports. The ISO 14971 hazard identification and risk estimation criteria (ISO, 2019) are expanded to the aspects of fairness risk exposure analysis between demographic groups.

The mapping of the high-level regulatory principles into the system architecture through this mapping turns the reproducible assurance processes into the high-level regulatory principles.

3.5 Empirical Evaluation Design

Three exemplary case studies were undertaken to assess operational feasibility based on publicly available datasets of U.S. healthcare including NIH ChestX-ray14, MIMIC-III and OpenFDA. Such datasets are diagnostic imaging, clinical risk prediction, and post-market pharmacovigilance, respectively.

Predictive benchmarking was not the objective but showing that fairness, accountability, and transparency can be transformed into verifiable assurance artifacts. Demographic stratification, data availability relevant to governance, provenance documentation, and structured records of transparency analyses were undertaken, in line with regulatory demands of lifecycle management.

4. Results

In this section, the empirical demonstration of the Trustworthy AI Assurance Framework (TAIAF) is provided in three healthcare AI scenarios, including diagnostic imaging (NIH ChestX-ray14), clinical risk governance (MIMIC-III), and post-market transparency (OpenFDA). It is not a goal of predictive benchmarking but rather validation of the capability of TAIAF to produce structured and inspection-ready assurance artifacts that are consistent with U.S. regulatory expectations.

4.1 Diagnostic Fairness Risk Stratification: NIH ChestX-ray14

4.1.1 Dataset Context

The NIH ChestX-ray14 dataset includes more than 100,000 frontal chest radiographs with named thoracic diseases and the related demographic information. Being one of the most popular algorithms in the field of radiology AI evolution, it has an adequate environment to conduct fairness assessment at the dataset governance level.

In TAIAF, the fairness analysis commences with demographic stratification as an approximation of the

potential risk of exposure to diagnostic risks, which is in line with the literature on the association between imbalance in the dataset and downstream performance differences (Chinta et al., 2025; Cross et al., 2024).

4.1.2 Demographic Distribution Findings

Sex stratification showed variation of representation of the different demographic groups. As illustrated in Table 3, male samples dominate the dataset than female samples with slight variation in the mean age in the groups.

Table 3. Demographic Stratification and TAIAF Fairness Risk Classification (NIH ChestX-ray14)

Demographic Group	Number of Samples	Mean Age (years)	Dataset Proportion	TAIAF Fairness Risk Classification
Female	48,780	46.5	0.44	Low
Male	63,340	47.2	0.56	Medium

Note: Risk classifications are illustrative and reflect relative dataset representation; they do not imply measured clinical performance disparities.

The imbalance proportion of Table 3 is not in itself an indication of algorithmic bias. Nevertheless, demographic imbalance under TAIAF is taken as an organized risk exposure signal that needs to be documented and possibly stratified on validation. The method correlates fairness evaluation with ISO 14971 type hazard identification (ISO, 2019) and regulation anticipations of bias transparency through the FDA Good Machine Learning Practice guidance (Warraich et al., 2025). Figure 3 represents the demographic distribution in a graphical way showing the proportional representation of the sex groups. Together, Table 3 and Figure 3 demonstrate how demographic metadata can be transformed into a formal fairness assurance artifact suitable for regulatory review.

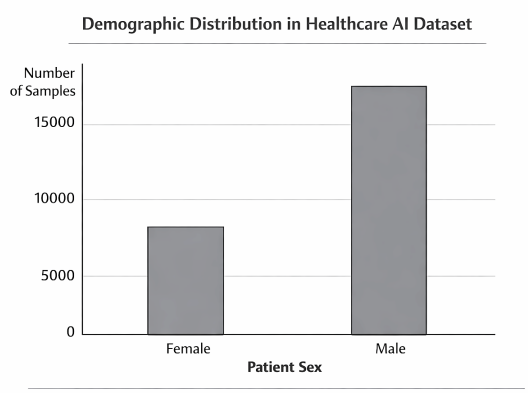

Figure 3. Demographic distribution by sex in the NIH ChestX-ray14 dataset as an indicator of potential diagnostic risk exposure.

The figure visualizes proportional differences between demographic groups. Under TAIAF, such imbalances are treated as fairness risk exposure indicators requiring structured documentation and monitoring rather than immediate claims of bias.

4.2 Governance Risk Identification: MIMIC-III

4.2.1 Dataset Context

MIMIC-III database is the de-identified electronic health record data of intensive care unit admissions in the United States. Since the fairness and accountability gauge is based on demographic completeness, TAIAF uses data availability as an indicator of governance.

4.2.2 Demographic Data Availability Findings

The selected MIMIC-III subsets exploratory review disclosed inconsistencies in the demographic structuring. Some variables needed to complete the fairness auditing were here and there or in pieces as they are summarized in Table 4.

Table 4. Governance Relevant Data Availability Assessment (MIMIC-III Subset)

Governance Indicator	Observed Status	TAIAF Risk Classification	Assurance Implication
Sex variable completeness	Partial	Medium	Requires cautious stratified validation
Age documentation consistency	Present	Low	Suitable for subgroup monitoring
Structured ethnicity variable availability	Fragmented	High	Fairness audit constraint; documenta

			tion escalation required
Provenance documentation clarity	Moderate	Medium	Additional traceability documentation recommended

As seen in Table 4, the demographic attributes that are incomplete limit the ability of fairness auditing. In TAIAF, these loopholes are not seen as technical minor concerns but governance and transparency threats that could restrict the provision of compliance preparedness. This interpretation can be aligned to the literature of governance that underlines institutional responsibility and formal control in healthcare AI systems (Kim et al., 2025; Batool et al., 2025).These results reveal that the reliability of the algorithmic results is not only based on the upstream data governance infrastructure but also on the outputs presented by the algorithm.

4.3 Post Market Transparency and Accountability: OpenFDA

4.3.1 Surveillance Data Context

OpenFDA offers organized access to adverse event reports that were reported to the U.S. Food and Drug Administration. The records are standardized with metadata in support of post-market surveillance.

4.3.2 Transparency Artifact Evaluation

Survey of sampled adverse event reports revealed structured fields that allowed the traceability and inspection. These artifacts are available and regulatorily relevant as summarized in Table 5.

Table 5. Transparency Artifact Evaluation (OpenFDA Adverse Event Sample)

Transparency Artifact Type	Availability	TAIAF Assurance Function	Regulatory Alignment
Event timestamp documentation	Present	Incident reconstruction	FDA post-market surveillance
Device identification records	Present	Traceability	Quality System Regulation
Patient demographic indicators	Partial	Fairness monitoring support	NIST AI RMF risk identification
Outcome classification fields	Present	Risk trend analysis	FDA adverse event monitoring
Machine-readable structured format	Present	Auditability	Transparency principles

The structured and machine-readable format of OpenFDA records, as revealed in Table 5, is auditable, but not explainable. Through TAIAF, these artifacts aid retrospective investigation, trend analysis, and regulatory inspection, as required by lifecycle documentation standards under U.S. regulatory frameworks (Rahimzadeh, 2023; Warraich et al., 2025).Figure 4 shows the lifecycle role of these transparency artifacts in TAIAF, and gives a picture of cross-domain application of assurance mechanisms.

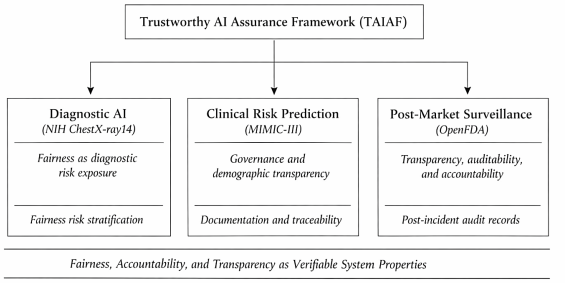

Figure 4. Application of TAIAF assurance mechanisms across diagnostic, clinical risk prediction, and post-market healthcare AI domains.

4.4 Cross Domain Synthesis

Table 6 provides a summary of the cross-domain operationalization of trustworthiness comparing assurance mechanisms in the three empirical contexts.

Table 6. Cross Domain Operationalization of Trustworthiness under TAIAF

Domain Context	Fairness Mechanism	Accountability Mechanism	Transparency Mechanism
Diagnostic Imaging (NIH)	Demographic risk exposure stratification	Dataset and validation documentation	Structured fairness audit artifact
Clinical Risk (MIMIC)	Demographic completeness	Provenance and governance	Data availability reporting

	assessment	documentation	
Pharmacovigilance (OpenFDA)	Population-level event analysis	Incident traceability	Structured post-market surveillance logs

As it is shown in Table 7, TAIAF is a domain-neutral assurance architecture. In every context, fairness is achieved through risk exposure analysis, accountability through traceable provenance and transparency through lifecycle auditability.The empirical results affirm that TAIAF makes it possible to transform abstract ethical principles into tangible and measurable assurance artifacts. Instead of the predictive performance, the results justify the ability of the framework to produce inspection ready documentation in line with the FDA instructions, the NIST AI Risk Management Framework, the principles of risk management in ISO 14971, and the Quality System Regulation guidelines.

TAIAF helps decrease ambiguity in compliance and enhance compliance regulations readiness to safety critical healthcare AI systems by implementing fairness, accountability, and transparency throughout the AI lifecycle.

7. Discussion

The main question discussed in this research was: *How can fairness, accountability, and transparency be modeled as verifiable system properties of safety-critical healthcare AI within U.S. regulatory frameworks?* The results establish that trustworthiness can be developed as a lifecycle-cutting assurance property instead of being viewed as an after hoc ethical ideal or performance proxy. Fairness, accountability, and transparency are implemented through the Trustworthy AI Assurance Framework (TAIAF), which is an auditable artifact, such as outputs of demographic risk stratification, provenance documentation, and structured transparency logs, built in the system design, deployment, and monitoring.

The empirical findings on NIH ChestX-ray14, MIMIC-III, and OpenFDA data support one key theoretical argument, namely, trustworthiness in diagnostic AI should be approached similarly to safety and reliability, meaning that it should be empirically demonstrated in a manner that is ready to be inspected as opposed to being supported by an assertion of interpretability or ethical alignment. This framing is a direct response to criticism that healthcare AI ethics frameworks tend to be normative and too operational (Chan et al., 2025; Gunasekara et al., 2025). Although the principles of ethics expressed by the international organisations like the World Health Organization (WHO, 2021; 2024), the Organisation for Economic Co-operation and Development (OECD, 2024), and UNESCO (2021) define some valuable values, they lack any engineering mechanisms that can facilitate a regulatory inspection or post-marketing accountability. TAIAF supplements these normative initiatives by converting the spirit of ethics into the knowledge of governance and guarantee procedures.

The NIH ChestX-ray14 case study illustrates that fairness may be restructured as equitable exposure in the diagnostic risk (in the dataset preparation). The previous studies have demonstrated that the presence of dataset imbalance and representational bias can result in unequal performance between the demographic groups (Chinta et al., 2025; Cross et al., 2024; Hasanzadeh et al., 2025). Nevertheless, the discourse of fairness has often been weighed against such statistical measures as demographic parity or equalized odds (Rabonato and Berton, 2025; Ueda et al., 2024). TAIAF further develops this work with fairness analysis and medical device risk management logic, and in line with the ISO 14971 principles (ISO, 2019; Jenko et al., 2025). Demographic imbalance is a possible hazard indicator in this understanding that needs to be reported and mitigation measures planned. Fairness is in this way a safety-relevant property that is entrenched in the regulatory risk management processes.

The MIMIC-III demonstration also indicates that trustworthiness risks can be frequent before the model training, especially in data governance infrastructure. Missing demographic variables limited the fairness auditing and accountability analysis, which demonstrates how transparency problems can be caused by documentation gaps instead of the algorithm design weaknesses. This conclusion is consistent with the governance literature that responsible AI brings about the need of institutional structures and documentation practices, rather than just technical metrics (Kim et al., 2025; Batool et al., 2025). TAIAF clearly categorizes demographic incompleteness as a governance risk, which supports the argument that accountability is determined by traceability and documentation even more than by model behavior.

The OpenFDA case study highlights the difference between explainability and auditability. Although explainable AI (XAI) methods can enhance the clinical interpretability (Sadeghi et al., 2024; Abbas et al., 2025), they not necessarily provide the sustainable evidence needed to oversee the operations of the

regulator. OpenFDA structured adverse event reports represent a classic example of artifacts of post-market transparency that can be reviewed and reformed after the incident. This lifecycle perspective of the transparency is in line with regulatory requirements in terms of FDA supervision and documentation-based compliance schemes (Rahimzadeh, 2023; Warraich et al., 2025). TAIAF develops the discussion beyond the user-exposure of interpretability to the regulator facing assurance, by quietly making auditability the priority, as opposed to the explanation.

Together, these results form the part of ongoing discussions on the issue of the disjuncture between the principles of ethical AI and enforcing compliance channels. Researchers have observed that the current systems often present high level objectives without defining implementation systems that can be measured (Khan et al., 2025; Alelyani, 2025). In the same way, the NIST AI Risk Management Framework offers a logical vocabulary of the AI governance but is consciously non-prescriptive (Tabassi, 2023; Dotan et al., 2024). TAIAF bridges this gap, as it serves as a translation layer between regulatory language risk management, transparency, traceability and tangible architectural artifacts incorporated into system design.

Notably, no model training or benchmarking was done in this study since it was not part of the study design. It did not aim at measuring predictive superiority but to confirm the assurance possibility. This model is indicative of increased awareness of how high algorithmic accuracy does not promote fair or reliable deployment (Goisauf et al., 2025; Goktas and Grzybowski, 2025). Focusing on the governance evidence instead of the performance metrics, TAIAF transforms the evaluation criteria to regulatory readiness and lifecycle accountability.

The key strength of TAIAF is its clear adherence to the U.S. regulatory paradigms, such as guidance on Good Machine Learning Practice developed by FDA, NIST AI Risk Management Framework, ISO 14971 risk management, and Quality System Regulation (Rahimzadeh, 2023; Warraich et al., 2025; ISO, 2019). In contrast to regulatory models where the emphasis is on ex ante conformity checks, TAIAF can be used in the context of the U.S. system of regulation based on documentation and monitoring. This congruency makes it more practical to federal agencies, healthcare facilities, and AI developers that do not break the current regulatory frameworks.

However, there are a number of limitations that should be discussed. First, the empirical analyses are demonstrative, not exhaustive, they show that they can be done, but not extensive in terms of statistical analysis. Second, TAIAF neglects to give universal fairness standards because the tolerable levels of risks should be set contextually by regulatory authorities or healthcare organizations. Third, the framework presupposes the presence of well-organized documentation and monitoring infrastructure, which is not equally distributed in healthcare settings. Further research on TAIAF in deployed clinical AI systems and automation of assurance artifacts such as continuous fairness checking and dynamically generated documentation should thus form part of future work.

Practically, the implication is multiple. To developers, regulatory preparedness can be enhanced by incorporating assurance mechanisms early during the AI lifecycle and decrease the risk of remediation after market introduction. To regulators, TAIAF presents systematic guidelines according to which trustworthiness assertions are assessed based on already needed evidence as stipulated by current policies. In the case of healthcare organizations, the framework offers tangible governance standards in vendor selection, deployment assessment and post deployment observation. The results at the policy level justify the transition to non-principles based AI governance and through practical assurance practices that do not require completely new regulatory requirements.

Finally, reliable healthcare AI does not arise based on predictive accuracy or moral ambition. It emerges out of verifiable, lifecycle based assurance measures integrated into system architecture and in line with regulatory controls. TAIAF can provide a methodical approach to building safe, equitable, and regulator-ready healthcare AI systems in the United States by operationalizing fairness in terms of equitable risk exposure, accountability in terms of traceability, and transparency in terms of auditability.

Conclusion

This paper suggested a regulation-operational framework, the Trustworthy AI Assurance Framework (TAIAF), which conceived trustworthiness in healthcare AI as a verifiable system assurance property and not an ethical or performance-based notion. TAIAF demonstrated empirically how fairness, accountability, and transparency, when implemented as lifecycle-spanning, auditable mechanisms, can be operationalized, as equitable risk exposure, traceability, and auditability, respectively. On the whole, TAIAF is a viable roadmap that facilitates the designing, testing, and regulation of safe, equitable, and regulation ready

healthcare AI systems under the current oversight frameworks.

References

Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare (Basel), 13*(17), 2154. https://doi.org/10.3390/healthcare13172154

Abdelwanis, M., Simsekler, M. C. E., Gabor, A. F., Sleptchenko, A., & Omar, M. (2026). Artificial intelligence adoption challenges from healthcare providers' perspectives: A comprehensive review and future directions. *Safety Science, 193*, 107028. https://doi.org/10.1016/j.ssci.2025.107028

Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K., & Monnet, X. (2024). Should AI models be explainable to clinicians? *Critical Care, 28*, 301. https://doi.org/10.1186/s13054-024-05005-y

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*, 1568. https://doi.org/10.1057/s41599-024-03555-3

Ahadian, P., Xu, W., Liu, D., & Guan, Q. (2026). Ethics of trustworthy AI in healthcare: Challenges, principles, and practical pathways. *Neurocomputing, 661*, 131942. https://doi.org/10.1016/j.neucom.2025.131942

Alelyani, T. (2025). A validated framework for responsible AI in healthcare autonomous systems. *Scientific Reports, 15*, 44432. https://doi.org/10.1038/s41598-025-25266-z

Alghareeb, E., & Aljehani, N. (2025). AI in health care service quality: Systematic review. *JMIR AI, 4*, e69209. https://doi.org/10.2196/69209

Astobiza, A. M., Alonso, M., & Ortega Lozano, R. (2025). Trust and AI in healthcare: A systematic review. *Monash Bioethics Review*. https://doi.org/10.1007/s40592-025-00272-z

Bader, T. (2025). Artificial intelligence in healthcare diagnostics: A literature review. *Open Journal of Applied Sciences, 15*(12), 4110–4133. https://doi.org/10.4236/ojapps.2025.1512266

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics, 5*, 3265–3279. https://doi.org/10.1007/s43681-024-00447-9

Borkar, S. R., & De, A. (2025). Artificial intelligence in internal medicine: A study on reducing diagnostic errors and enhancing efficiency. *European Journal of Cardiovascular Medicine, 15*(9), 105–111. https://doi.org/10.61336/ejcm/25-09-19

Chan, A., Rahimi-Ardabilli, H., Rogers, W. A., & Coiera, E. (2025). The real-world impact of artificial intelligence ethics frameworks across a decade in healthcare: A scoping review. *Journal of the American Medical Informatics Association, 32*(11), 1767–1777. https://doi.org/10.1093/jamia/ocaf167

Chinta, S. V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M. A., Liu, J., & Zhang, W. (2025). AI-driven healthcare: A review on ensuring fairness and mitigating bias. *PLOS Digital Health, 4*(5), e0000864. https://doi.org/10.1371/journal.pdig.0000864

Chong, P. L., Vaigeshwari, V., Reyasudin, B. K. M., Noor Hidayah, B. R. A., Tatchanaamoorti, P., Yeow, J. A., et al. (2025). Integrating artificial intelligence in healthcare: Applications, challenges, and future directions. *Future Science OA, 11*(1). https://doi.org/10.1080/20565623.2025.2527505

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health, 3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving AI risk management: A maturity model based on the NIST AI Risk Management Framework. *arXiv*. https://doi.org/10.48550/arXiv.2401.15229

Fahim, Y. A., Hasani, I. W., Kabba, S., & Ragab, W. M. (2025). Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. *European Journal of Medical Research, 30*, 848. https://doi.org/10.1186/s40001-025-03196-w

Goisauf, M., Cano Abadía, M., Akyüz, K., et al. (2025). Trust, trustworthiness, and the future of medical AI. *Journal of Medical Internet Research, 27*, e71236. https://doi.org/10.2196/71236

Goktas, P., & Grzybowski, A. (2025). Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine, 14*(5), 1605. https://doi.org/10.3390/jcm14051605

Gunasekara, L., El-Haber, N., Nagpal, S., et al. (2025). A systematic review of responsible artificial intelligence principles and practice. *Applied System Innovation, 8*(4), 97. https://doi.org/10.3390/asi8040097

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization, 98*(4), 251–256. https://doi.org/10.2471/BLT.19.237487

Hasanzadeh, F., Josephson, C. B., Waters, G., et al. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *npj Digital Medicine, 8*, 154. https://doi.org/10.1038/s41746-025-01503-7

Hassan, M., Kushniruk, A., & Borycki, E. (2024). Barriers to and facilitators of artificial intelligence adoption in health care. *JMIR Human Factors, 11*, e48633. https://doi.org/10.2196/48633

Hildt, E. (2025). What is the role of explainability in medical artificial intelligence? *Bioengineering, 12*(4), 375. https://doi.org/10.3390/bioengineering12040375

ISO. (2019). *Medical devices—Application of risk management to medical devices (ISO 14971:2019)*. https://www.iso.org/standard/72704.html

Jenko, S., Papadopoulou, E., Kumar, V., et al. (2025). Artificial intelligence in healthcare: How to develop and implement safe, ethical and trustworthy AI systems. *AI, 6*(6), 116. https://doi.org/10.3390/ai6060116

Khan, M. M., Shah, N., Shaikh, N., et al. (2025). Towards secure and trusted AI in healthcare. *International Journal of Medical Informatics, 195*, 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780

Kim, J. Y., Hasan, A., Kueper, J., et al. (2025). Establishing organizational AI governance in healthcare. *npj Digital Medicine, 8*, 522. https://doi.org/10.1038/s41746-025-01909-3

Lekadir, K., Frangi, A. F., Porras, A. R., et al. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388*, e081554. https://doi.org/10.1136/bmj-2024-081554

Nasarian, E., Alizadehsani, R., Acharya, U. R., & Tsui, K.-L. (2024). Designing interpretable ML system to enhance trust in healthcare. *Information Fusion, 108*, 102412. https://doi.org/10.1016/j.inffus.2024.102412

Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. (2025). Evaluating trustworthiness in AI. *Electronics, 14*(13), 2717. https://doi.org/10.3390/electronics14132717

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & Society, 39*, 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

OECD. (2024). *OECD AI principles*. https://www.oecd.org/en/topics/sub-issues/ai-principles.html

Rahimzadeh, V. (2023). US regulation of medical artificial intelligence and machine learning (AI/ML) research and development. In *Research handbook on health, AI and the law*. Edward Elgar Publishing. https://www.ncbi.nlm.nih.gov/books/NBK613220/

Rabonato, R. T., & Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics, 5*, 1943–1954. https://doi.org/10.1007/s43681-024-00577-5

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., et al. (2024). A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering, 118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

Sandalow, M., Adams, K., & Loud, G. (2025). FDA oversight: Understanding the regulation of health AI tools. *Bipartisan Policy Center*.

Singhal, A., Neveditsin, N., Tanveer, H., & Mago, V. (2024). Toward fairness, accountability, transparency, and ethics in AI. *JMIR Medical Informatics, 12*, e50048. https://doi.org/10.2196/50048

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Warraich, H. J., Tazbaz, T., & Califf, R. M. (2025). FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA, 333*(3), 241–247. https://doi.org/10.1001/jama.2024.21451

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

World Health Organization. (2024). WHO releases AI ethics and governance guidance for large multi-modal models.

Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem. *Intelligent Medicine, 4*(1), 52–57. https://doi.org/10.1016/j.imed.2023.08.001

Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., & Zhou, S. K. (2024). Addressing fairness issues in deep learning-based medical image analysis. *npj Digital Medicine, 7*, 286. https://doi.org/10.1038/s41746-024-01276-5