REVIEW **Open access**

Explainable Artificial Intelligence in Clinical Systems: Interpretability, Transparency, and Deployment Constraints

Sanjay Kulkarni[1], Meenal Joshi[1*], Rohan Patil[2]

Abstract

The integration of artificial intelligence (AI) into healthcare systems has revolutionized clinical analytics, enabling enhanced diagnostic accuracy, predictive modeling, and personalized treatment pathways. However, the opacity of many AI models poses significant challenges to their clinical adoption, necessitating advancements in explainable AI (XAI) to ensure interpretability and transparency. This narrative review synthesizes the literature on XAI within clinical systems, focusing on interpretability mechanisms, transparency frameworks, and deployment constraints in healthcare analytics. Drawing from high-impact studies, we examine how XAI addresses the “black box” nature of machine learning models in high-stakes medical decisions, particularly in contexts where performance has traditionally been prioritized over explainability. Key themes include the shift toward inherently interpretable models for critical applications, such as diagnostic imaging and predictive analytics, where post-hoc explanations often fall short. We explore the ethical imperatives for responsible AI deployment, including strategies for mitigating harm through transparent systems that align with clinical workflows. The review integrates perspectives on XAI in clinical diagnostics, emphasizing challenges in balancing model complexity with user trust. Transparency is framed not merely as a technical feature but as a systemic requirement, incorporating structured reporting practices for AI interventions and standardized modeling approaches. Deployment constraints are analyzed through the lens of real-world integration, including regulatory considerations, data privacy concerns, and human–AI interaction dynamics in healthcare infrastructures. We synthesize evidence from diverse applications, such as lung cancer diagnosis via explainable models and radiographic assessments, underscoring the need for multidisciplinary approaches to XAI. Furthermore, the review highlights biases in AI systems, particularly sex and gender disparities, and advocates for inclusive analytics to foster equitable healthcare. Clinical applications beyond the black box are discussed, with calls for standardized reporting to enhance reproducibility and trust. We position XAI as essential for closed-loop systems that incorporate feedback mechanisms, ensuring ongoing model recalibration in dynamic clinical environments. The synthesis reveals persistent gaps in current XAI deployments, such as overreliance on surrogate explanations that may mislead clinicians. Ultimately, this review proposes a systems-level framework for XAI in healthcare, integrating data ingestion, inference, decision support, and governance loops to overcome transparency barriers. This comprehensive overview informs the development of future AI-enabled healthcare infrastructures, emphasizing interpretability as a cornerstone for safe and effective clinical analytics.

Keywords Healthcare analytics, Explainable AI, Clinical interpretability, AI transparency, Deployment constraints, Clinical decision systems

*Correspondence:
Meenal Joshi
meenal.joshi@gmail.com
[1] Department of AI in Healthcare Systems, School of Medicine, Savitribai Phule Pune University, Pune, India
[2] Department of Clinical Data Engineering, School of Engineering, Indian Institute of Technology Bombay, Mumbai, India

Introduction

The evolution of AI in healthcare: from promise to practicality

Artificial intelligence has emerged as a transformative force in healthcare, reshaping how clinical data is analyzed and utilized for patient care. Over the past decade, AI technologies have advanced from experimental tools to integral components of healthcare systems, driven by exponential growth in computational power and data availability [1-5]. In clinical analytics, AI enables the processing of vast datasets—from electronic health records (EHRs) to imaging modalities—facilitating predictive insights that were previously unattainable through traditional statistical methods [4]. For instance, machine learning models have demonstrated superior performance in tasks such as disease detection and risk stratification, often surpassing human experts in controlled settings. Yet, this evolution is tempered by persistent concerns over the interpretability of these models, particularly in clinical environments where decisions carry life-altering consequences [1, 2].

The push for explainable AI in healthcare stems from the inherent opacity of deep learning architectures, which dominate modern clinical applications [1, 6-11]. Black-box models, while powerful, obscure the reasoning behind their outputs, eroding clinician trust and complicating regulatory approval [2, 3]. This has led to a paradigm shift toward XAI, where interpretability is prioritized to align AI with clinical imperatives [6]. Literature underscores the need for transparency in AI-driven systems, not only to comply with ethical standards but also to enhance collaborative human-AI decision-making [9, 12, 13].

Defining explainability in clinical contexts

Explainability in AI refers to the ability to articulate how a model arrives at its predictions in terms understandable to end-users, such as clinicians and patients [1, 6]. In healthcare, this encompasses both local explanations (for individual predictions) and global insights (into overall model behavior) [6, 7]. Transparency extends beyond explanations to include the openness of data pipelines, model training processes, and deployment protocols [14-17]. Deployment constraints further complicate this, involving factors like computational resources, integration with legacy systems, and adherence to privacy regulations such as HIPAA or GDPR [3, 9].

Recent perspectives emphasize multidisciplinary approaches to XAI, integrating insights from computer science, medicine, and ethics [6, 9]. For example, challenges in clinical diagnostic models highlight the tension between model accuracy and interpretability, advocating for hybrid systems that combine black-box efficiency with explainable interfaces [6, 8]. This section synthesizes these definitions, framing XAI as a bridge between technological innovation and clinical utility.

Historical and regulatory milestones

The trajectory of AI in healthcare can be traced to early applications in medical imaging, where deep learning surveys revealed potential for automated analysis [11]. By 2018, clinically applicable models for retinal disease diagnosis showcased the feasibility of AI in real-world settings [10]. Regulatory bodies have responded with frameworks like the FDA's AI/ML-based Software as a Medical Device (SaMD) guidelines, which mandate transparency for approval [3, 5].

Consensus statements and reporting guidelines have proliferated, including SPIRIT-AI and CONSORT-AI extensions for clinical trials involving AI [17-19]. These standards ensure that AI interventions are documented with sufficient detail on interpretability, addressing gaps in traditional reporting [16, 20]. The MI-CLAIM checklist further promotes minimum information standards for AI modeling, emphasizing transparency in clinical contexts [20].

Ethical imperatives and bias mitigation

Ethical challenges in AI deployment are paramount, with literature calling for responsible practices to prevent harm [3, 9]. Biases in training data can perpetuate disparities, as seen in gender and sex differences in biomedicine AI [14]. XAI

serves as a tool for bias detection, enabling clinicians to interrogate model decisions and mitigate inequities [13, 14].

Warnings against overreliance on post-hoc explanations underscore the risk of false confidence in AI outputs [2, 13]. Instead, inherently interpretable models are advocated for high-stakes scenarios, reducing the ethical burden on users [1].

Scope and synthesis logic of this review

This review adopts a systems-level perspective on XAI in clinical healthcare, synthesizing literature across data analytics, model interpretability, and deployment ecosystems. Unlike prior reviews that focus narrowly on technical methods, we integrate cross-study analyses to propose an original framework for AI-enabled clinical systems. The synthesis logic organizes evidence into landscapes of current applications, followed by architectures for intelligent decision-making. We emphasize interpretive structuring that links interpretability to transparency and constraints, positioning XAI as foundational for sustainable healthcare innovation.

Landscape of AI in healthcare systems and analytics

Data-driven foundations and analytics pipelines

The landscape of AI in healthcare is underpinned by robust data analytics pipelines that transform raw clinical data into actionable intelligence [4, 5]. Electronic health records, genomic sequences, and multimodal imaging form the bedrock, processed through machine learning to yield predictive analytics [11, 21-24]. Surveys on deep learning in medical image analysis illustrate how convolutional neural networks extract features from radiographs and ultrasounds, enabling automated diagnostics [11]. However, the integration of XAI is crucial to demystify these pipelines, ensuring that data transformations are traceable and biases are identifiable [6, 14].

In analytics, transparency begins at data ingestion, where standards like MI-CLAIM advocate for detailed reporting of data sources and preprocessing [20]. This mitigates issues like shortcut learning in radiographic models, where AI exploits non-clinical artifacts rather than true signals [12]. Literature synthesizes these challenges, highlighting the need for explainable preprocessing to foster trust in downstream analytics [7, 15].

Interpretability mechanisms in clinical models

Interpretability mechanisms vary from inherent model simplicity to post-hoc techniques [1, 6]. Inherently interpretable models, such as decision trees or rule-based systems, are favored for clinical transparency, avoiding the pitfalls of black-box deep networks [1]. For complex tasks like lung cancer diagnosis, explainable models using radial endobronchial ultrasound demonstrate how feature attributions enhance clinician understanding [7, 8].

Post-hoc methods, including saliency maps and counterfactuals, are critiqued for their potential unreliability in healthcare [2, 13]. Studies warn that these explanations may not faithfully represent model reasoning, leading to misguided clinical decisions [2]. Instead, a multidisciplinary perspective advocates for context-specific interpretability, tailored to clinical workflows [6].

Transparency frameworks and reporting standards

Transparency frameworks are essential for bridging AI development and clinical deployment [15, 16]. Guidelines like SPIRIT-AI and CONSORT-AI extend traditional trial reporting to include AI-specific details on interpretability and validation [17-19]. These frameworks ensure that AI interventions are transparently documented, facilitating peer review and regulatory scrutiny [21].

In diagnostic accuracy studies, STARD-AI proposes standards for AI assessments, emphasizing transparent reporting of model limitations [21]. Comparative analyses reveal that AI often matches or exceeds clinician performance in imaging

tasks, but only when transparency allows for fair evaluation [23, 24]. This landscape synthesizes these standards as foundational for trustworthy AI analytics.

Deployment constraints in healthcare infrastructures

Deployment constraints encompass technical, regulatory, and human factors [3, 5]. Computational demands of AI models strain hospital infrastructures, necessitating efficient, explainable alternatives [4]. Privacy constraints under data protection laws require transparent handling of sensitive health information [9].

Human-AI interaction poses another constraint, where the lack of interpretability hinders adoption [22]. Systematic reviews show that machine learning decision support systems improve diagnostic performance when explanations align with clinician expertise [22]. Moreover, ethical challenges in direct-to-consumer AI applications underscore the need for transparent governance [25-28].

Bias and equity in AI analytics

Bias remains a pervasive issue, with AI systems often amplifying disparities in healthcare [14]. Sex and gender biases in biomedicine AI highlight the need for explainable analytics to detect and correct imbalances [14]. Transparent models enable auditing for equity, ensuring that analytics serve diverse populations [13].

Literature synthesizes these concerns, advocating for inclusive data practices and explainable frameworks to promote fair AI deployment [3, 9]. This section integrates evidence from chest radiograph diagnostics and breast cancer screening, where transparent AI has revealed biases in model training [25, 26].

Emerging applications and systemic integration

Emerging applications include AI-powered digital medicine, where transparency enables personalized analytics [29]. From retinal diagnostics to skin cancer classification, XAI facilitates systemic integration, linking analytics to clinical outcomes [10, 27]. The landscape reveals a maturing field, where interpretability and transparency are increasingly viewed as prerequisites for scalable healthcare systems [5, 28].

Intelligent clinical decision and closed-loop healthcare systems

Architectures for AI-enabled decision support

Intelligent clinical decision systems leverage AI to augment human judgment, forming architectures that integrate analytics into real-time workflows [4, 22]. These systems encompass predictive modeling for risk assessment, diagnostic aiding, and treatment recommendation, all underpinned by XAI to ensure reliable outputs [6, 8]. In closed-loop configurations, AI not only predicts but also monitors outcomes, creating feedback mechanisms that refine decisions over time [3, 10].

Core architectures involve modular components: data aggregation, model inference, explanation generation, and decision fusion [15, 17]. For instance, in radiographic diagnostics, explainable models provide attributions that clinicians can validate against domain knowledge [7, 12]. This synthesis highlights hybrid architectures where interpretable layers interface with complex neural networks, balancing efficacy and transparency [1, 11].

Human-AI collaboration dynamics

Human-AI collaboration is central to intelligent systems, where transparency fosters trust and effective integration [13, 22]. Studies demonstrate that explainable decision support enhances clinician performance, particularly in ambiguous cases [22, 23]. However, constraints like cognitive overload from poor explanations can impede this synergy [2, 6].

In closed-loop systems, collaboration extends to ongoing monitoring, where AI alerts trigger human intervention [3, 5]. Ethical frameworks emphasize shared decision-making, with XAI enabling clinicians to override or refine AI suggestions

[9, 13].

Feedback and recalibration mechanisms

Closed-loop healthcare systems incorporate feedback loops to adapt to evolving clinical data [10, 20]. These mechanisms involve continuous model recalibration based on outcome data, ensuring sustained performance [3, 17]. Transparency in feedback processes allows for auditing drifts in model behavior, critical in dynamic environments like intensive care [4].

Literature synthesizes these as essential for deployment, with guidelines mandating reporting of recalibration protocols [18, 19]. This prevents degradation over time, maintaining interpretability amid data shifts [12, 15].

Governance and constraint management

Governance in intelligent systems addresses deployment constraints through structured oversight [3, 21]. This includes regulatory compliance, ethical auditing, and resource allocation, all facilitated by transparent architectures [9, 16]. Constraints like interoperability with EHR systems are mitigated via standardized interfaces, ensuring seamless integration [5, 28].

Synthesis reveals that effective governance relies on XAI to expose vulnerabilities, such as biases or uncertainties, enabling proactive management [14, 20].

Conceptual formalization of clinical intelligence cycles

To synthesize these architectures, we introduce a conceptual formula for the clinical intelligence pipeline:

$$\begin{aligned} & I\left(t\right) \\ & = f\left(D\left(t\right), M, E, H\right) \\ & \rightarrow D\left(t+1\right) \end{aligned} \quad (1)$$

Where I(t) represents the intelligent intervention at time t, derived from function f applied to current data D(t), model M, explanations E, and human input H. This feeds back to update D(t+1), formalizing the closed-loop nature without empirical metrics.

A second formula captures human-AI decision fusion:

$$\begin{aligned} & Dec \\ & = \alpha \\ & \cdot AIpred \\ & + (1-\alpha) \\ & \cdot Clinjudg \\ & + \beta \\ & \cdot Explval \end{aligned} \quad (2)$$

Here, Dec is the fused decision, weighted by α (AI confidence), clinician judgment Clinjudg , and explanation validity Explval modulated by β, emphasizing interpretive dynamics. **Figure 1** illustrates the systems-level architecture of explainable AI integration across closed-loop clinical decision ecosystems.

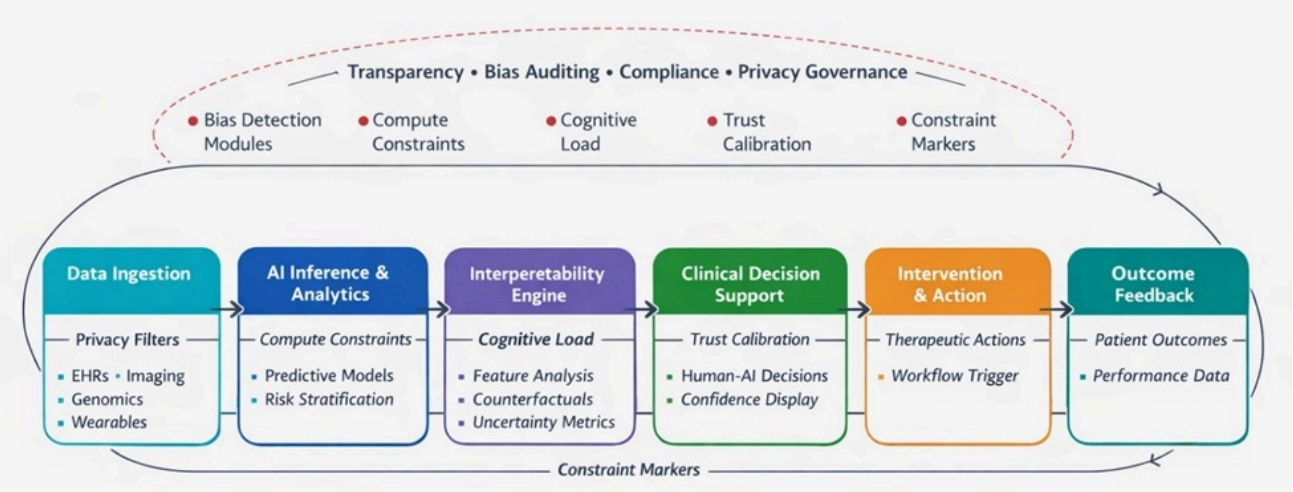

Figure 1. Systems-level architecture of explainable AI integration in closed-loop clinical decision ecosystems.

The figure depicts an end-to-end clinical intelligence cycle in which multimodal healthcare data are processed through explainable AI inference engines and interpretability modules to support human-AI collaborative decision-making. Clinical interventions generate outcome data that feed back into system recalibration, forming a continuous learning loop. A governance superstructure overlays the architecture, embedding transparency auditing, bias surveillance, privacy enforcement, and regulatory compliance across all operational layers. Deployment constraints—including infrastructural, computational, and human-factor limitations—are annotated at key system interfaces to reflect real-world implementation conditions.

Results and Discussion

Integrating interpretability into clinical practice

The discourse on explainable AI in clinical systems reveals a critical juncture where technological advancements must align with practical healthcare delivery [1, 3]. Synthesizing the literature, interpretability emerges not as an add-on but as an integral component of AI architectures, facilitating seamless integration into clinical workflows [6, 15]. For instance, in diagnostic models for lung cancer and retinal diseases, explainable mechanisms have demonstrated potential to enhance clinician confidence by providing verifiable rationales [7, 8, 10]. This integration extends to analytics platforms, where transparency enables iterative improvements in predictive accuracy without sacrificing user trust [4, 11].

However, the discussion underscores disparities between AI performance in controlled studies and real-world efficacy [23, 24]. Systematic reviews indicate that while AI often rivals clinicians in specific tasks, such as image-based disease detection, the lack of standardized interpretability metrics hinders broad adoption [24, 25]. Multidisciplinary perspectives advocate for a holistic view, where interpretability intersects with ethical governance to address systemic biases [6, 9, 14]. This synthesis posits that effective discussion around XAI must prioritize user-centric designs, ensuring explanations are tailored to diverse clinical stakeholders—from physicians to policymakers [13, 22].

Transparency as a systemic enabler

Transparency in AI healthcare systems transcends technical explanations, encompassing end-to-end visibility in data handling and model deployment [15, 17]. Reporting guidelines like CONSORT-AI and SPIRIT-AI exemplify this, mandating detailed disclosures that foster reproducibility and accountability [18, 19]. In the context of closed-loop systems, transparency enables robust feedback loops, where model outputs are continuously scrutinized against clinical outcomes [20, 21].

Literature highlights how transparency mitigates deployment risks, such as algorithmic shortcuts that compromise diagnostic reliability [12, 13]. By framing transparency as a systemic enabler, this discussion integrates evidence from high-impact applications, advocating for infrastructures that embed auditability at every stage [5, 28]. Ultimately, these elements converge to support resilient healthcare analytics, where AI augments rather than supplants human expertise [4, 22].

Challenges and limitations

Technical challenges in achieving interpretability

A primary challenge in XAI for clinical systems lies in the trade-off between model complexity and interpretability [1, 6]. Deep learning models, prevalent in medical image analysis, offer high performance but inherent opacity, complicating the generation of faithful explanations [11, 12]. Post-hoc interpretability methods, while popular, are critiqued for their potential to provide misleading insights, as they may not accurately reflect internal model logic [2, 13]. For example, in radiographic COVID-19 detection, AI models have been shown to rely on non-clinical cues, underscoring the limitations of surrogate explanations [12].

Furthermore, computational constraints in resource-limited healthcare settings hinder the deployment of explainable models, which often require additional processing for attribution maps or counterfactuals [4, 7]. Literature synthesizes these technical hurdles, emphasizing the need for scalable interpretability without degrading inference speed [8, 11].

Regulatory and ethical limitations

Regulatory frameworks pose significant limitations, with varying standards across jurisdictions impeding global AI adoption [3, 5]. Guidelines such as STARD-AI and MI-CLAIM address reporting, yet inconsistencies in interpretability requirements lead to deployment delays [20, 21]. Ethically, the opacity of AI exacerbates issues like bias amplification, particularly in underrepresented populations, as evidenced by gender disparities in biomedicine [14].

Challenges extend to data privacy, where transparent systems must balance explainability with compliance to regulations like GDPR, often restricting access to sensitive training data [9, 16]. This synthesis reveals ethical dilemmas in high-stakes decisions, where insufficient transparency can result in harm, advocating for stricter governance [3, 13].

Integration and human factors constraints

The integration of explainable artificial intelligence (XAI) into established healthcare infrastructures introduces multi-layered technical, organizational, and cognitive constraints that extend beyond algorithmic design. Legacy electronic health record (EHR) systems often operate on heterogeneous architectures with limited interoperability, creating friction points for embedding real-time explainability modules within clinical workflows [4, 28]. Many EHR platforms were not architected for modular AI integration, resulting in fragmented data pipelines, latency in explanation rendering, and limited contextual alignment between model outputs and clinician-facing interfaces. Consequently, XAI systems must negotiate both infrastructural rigidity and regulatory compliance frameworks that govern data provenance, auditability, and clinical traceability.

Human factors further constrain effective deployment. Clinicians frequently evaluate AI-generated explanations through the lens of domain expertise, experiential heuristics, and institutional norms. When explanations diverge from established medical reasoning patterns, they may be perceived as unreliable—even when statistically valid—leading to skepticism and underutilization [2, 22]. Conversely, overly persuasive or visually authoritative explanations risk inducing automation bias, whereby clinicians defer excessively to algorithmic outputs without adequate scrutiny [22, 23]. This dual risk of overreliance and distrust underscores the delicate epistemic balance required for effective human-AI collaboration.

Systematic reviews reveal heterogeneous impacts of XAI on diagnostic accuracy and workflow efficiency, with performance gains often mediated by explanation clarity, task complexity, and clinician training level [22, 23]. In high-

acuity contexts such as emergency departments or intensive care units, these dynamics are amplified. Models must adapt to rapidly evolving clinical states while maintaining stable interpretability outputs that support time-sensitive decisions [10, 15]. The tension between adaptability and transparency becomes particularly salient in adaptive or continuously learning systems, where evolving model parameters may alter explanation structures over time.

Educational gaps compound these challenges. Many healthcare professionals receive limited formal training in AI literacy, reducing their capacity to critically interpret probabilistic outputs, feature attributions, or uncertainty intervals [5, 6]. Without structured education programs, explanation interfaces risk functioning as superficial compliance artifacts rather than meaningful decision-support tools. Addressing integration constraints, therefore, requires a systems-level approach encompassing infrastructural modernization, workflow redesign, and longitudinal professional training.

Evaluation and standardization gaps

Despite rapid advances in model interpretability techniques, the field lacks unified evaluation frameworks for measuring explainability in clinical contexts. Unlike predictive accuracy—supported by well-established benchmarks such as AUROC, sensitivity, and calibration—interpretability remains assessed through heterogeneous, often subjective metrics [1, 25]. Studies variably operationalize explainability through proxy measures such as clinician satisfaction, alignment with expert reasoning, or task completion time, complicating cross-study comparison and evidence synthesis [16, 24].

The absence of standardized interpretability benchmarks impedes regulatory clarity and slows translation from experimental prototypes to real-world clinical systems. In multicenter deployments, additional complexity arises from data heterogeneity, including demographic variation, institutional practice patterns, and imaging protocol differences [11, 26]. These factors affect not only model generalizability but also the stability and consistency of explanation outputs across sites. An explanation deemed intuitive in one institutional context may be ambiguous or misleading in another.

Moreover, post-hoc explanation techniques often lack formal guarantees regarding faithfulness to underlying model mechanisms. As a result, explanatory visualizations may reflect approximation artifacts rather than genuine causal drivers, undermining epistemic transparency [1, 25]. The literature increasingly emphasizes the need for rigorous validation frameworks that integrate technical fidelity assessments with human-centered usability evaluations [16, 24].

Synthesizing these limitations, the field requires robust real-world evaluation pipelines that test XAI systems under operational conditions rather than controlled experimental settings [3, 27]. Such pipelines should incorporate longitudinal monitoring to assess explanation drift, bias propagation, and clinician adaptation over time. Without standardized and context-sensitive evaluation protocols, interpretability risks remaining conceptually appealing but operationally inconsistent. **Table 1** synthesizes operational constraints and transparency enablers across the explainable clinical AI deployment lifecycle.

Table 1. Operational constraints and transparency enablers in clinical XAI deployment

Domain layer	Key constraints	Transparency mechanisms	Clinical impact	Governance implications
Data ingestion	Privacy regulations; data heterogeneity	Data provenance tracking; audit trails	Improved data trustworthiness	GDPR / HIPAA compliance enforcement
Model inference	Black-box opacity; compute load	Interpretable architectures; model reporting	Safer predictive analytics	Regulatory approval facilitation
Interpretability interfaces	Cognitive overload; explanation fidelity	Saliency mapping; counterfactuals; uncertainty metrics	Enhanced clinician understanding	Explanation validation standards

Decision support	Automation bias; trust variability	Confidence scoring; human override systems	Balanced human-AI collaboration	Liability allocation frameworks
Intervention execution	Workflow disruption	Transparent recommendation logic	Adoption in care pathways	Clinical accountability mapping
Feedback and recalibration	Outcome drift; data lag	Continuous monitoring dashboards	Sustained performance	Post-deployment surveillance
Governance oversight	Ethical risk; bias propagation	Bias auditing; fairness analytics	Equitable care delivery	Policy and compliance alignment

Future research directions

Advancing inherently interpretable models

Future research should prioritize inherently interpretable architectures that embed transparency within model structure rather than relying exclusively on post-hoc explanation layers [1, 6]. Models such as attention-based frameworks with clinically meaningful feature hierarchies, generalized additive models, or rule-augmented neural networks may offer improved alignment between algorithmic reasoning and medical cognition. Hybrid paradigms combining symbolic AI with deep learning present a promising pathway, particularly for multimodal fusion tasks integrating imaging, genomics, and longitudinal EHR data [11, 15].

Domain-specific interpretability research is especially warranted. For example, genomic analytics require explanations that reflect biological pathways, while wearable-based monitoring demands temporally coherent narrative outputs [4, 14]. Tailoring interpretability mechanisms to clinical subspecialties will enhance contextual relevance and adoption.

Enhancing transparency through standardization

The development of standardized XAI evaluation metrics constitutes a critical research priority. Building upon emerging guidelines, the field must establish benchmark datasets and reproducible protocols that quantify interpretability fidelity, stability, and usability [16, 20]. Regulatory agencies increasingly call for demonstrable transparency in AI-driven medical devices, necessitating harmonized global standards for explainability reporting [3, 21].

Automated auditing tools represent another frontier. Real-time bias detection systems capable of monitoring demographic performance disparities and explanation consistency could enhance equitable deployment [13, 14]. Embedding such auditing layers within clinical AI pipelines would operationalize transparency rather than relegating it to retrospective analysis.

Exploring human-AI symbiosis

Human-AI interaction research remains central to advancing explainable systems. Empirical investigations should examine how explanation modality—textual, visual, counterfactual, or probabilistic—affects clinician trust calibration and diagnostic performance [22, 23]. Adaptive explanation interfaces that personalize the detail level according to clinician expertise or task complexity may improve usability and mitigate cognitive overload [6, 9].

Longitudinal studies of closed-loop clinical systems are particularly important. Evaluating how sustained exposure to XAI affects clinician learning, reliance patterns, and patient outcomes will provide insight into long-term system dynamics [10, 17]. Such research should incorporate behavioral analytics to quantify trust evolution and decision confidence over time.

Addressing deployment in resource-constrained settings

Resource-limited environments introduce additional infrastructural and computational constraints. Lightweight XAI frameworks optimized for edge computing and low-bandwidth settings are essential to ensure equitable global

deployment [5, 28]. Simplified interpretable models or compressed architectures may provide transparency without excessive hardware requirements.

The integration of federated learning with explainability mechanisms represents a promising strategy for preserving patient privacy while enabling collaborative model improvement across institutions [9, 16]. Embedding interpretable components within federated aggregation protocols could enhance transparency in decentralized training environments.

Interdisciplinary and ethical innovations

Interdisciplinary collaboration between AI researchers, clinicians, ethicists, and policy experts will shape the next generation of governance-aware XAI systems. Explainable AI may extend beyond clinical decision support to inform policy modeling, resource allocation, and digital health governance structures [3, 13]. Embedding ethical reasoning frameworks within AI architectures can operationalize principles such as fairness, accountability, and transparency.

Emerging domains—including digital therapeutics, personalized medicine, and adaptive behavioral interventions—offer fertile ground for XAI innovation [27, 29]. In these contexts, interpretability is not merely supportive but foundational to patient engagement and regulatory approval.

Conclusion

The evolving landscape of explainable artificial intelligence in clinical systems reflects a broader transition from performance-centric machine learning to trust-centered healthcare analytics. This review synthesizes evidence demonstrating that interpretability is not an auxiliary feature but a structural requirement for responsible AI deployment. From infrastructural integration challenges and human-factor constraints to evaluation gaps and regulatory uncertainties, XAI implementation demands coordinated technical, organizational, and ethical strategies.

Key insights underscore the necessity of inherently interpretable architectures, standardized transparency metrics, and human-centric interface design. While trade-offs between model complexity and explainability persist, emerging hybrid and governance-aware frameworks suggest viable pathways forward. Future progress will depend on rigorous real-world validation, interdisciplinary collaboration, and sustained attention to equity and bias mitigation.

Ultimately, embedding explainable AI within healthcare infrastructures is essential for achieving trustworthy, accountable, and clinically meaningful analytics. By aligning algorithmic reasoning with medical cognition and regulatory oversight, XAI can enable a new generation of transparent decision-support systems that enhance patient safety, professional autonomy, and system-wide resilience.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Feb 2024 Revised: 14 Mar 2024 Accepted: 08 Apr 2024
Published online: 20 July 2024

References

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.
https://doi.org/10.1038/s42256-019-0048-x.

Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750.
https://doi.org/10.1016/S2589-7500(21)00208-9.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.
https://doi.org/10.1038/s41591-019-0548-6.

Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58.
https://doi.org/10.1056/NEJMra1814259.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
https://doi.org/10.1038/s41591-018-0300-7.

Antoniadi AM, Du Y, Gu Y, Klüwer T, Wilk M, Galvin M, et al. Current challenges and future opportunities for XAI in machine learning-based clinical diagnostic models. Artif Intell Med. 2021;114:102044.
https://doi.org/10.1016/j.artmed.2021.102044.

Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-time radial endobronchial ultrasound miniprobe imaging for lung cancer diagnosis. J Biomed Inform. 2020;103:103406.
https://doi.org/10.1016/j.jbi.2020.103406.

Noh J, Park J, Park H, Lee H. Explainable AI models using real-time radial endobronchial ultrasound miniprobe imaging for lung cancer diagnosis. npj Digit Med. 2021;4:150.
https://doi.org/10.1038/s41746-021-00526-0.

Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLOS Med. 2018;15(11):e1002689.
https://doi.org/10.1371/journal.pmed.1002689.

De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342-50.
https://doi.org/10.1038/s41591-018-0107-6.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
https://doi.org/10.1016/j.media.2017.07.005.

DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3(7):610-9.
https://doi.org/10.1038/s42256-021-00338-7.

Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-6.
https://doi.org/10.1126/science.abg1834.

Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Rizzoli S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digit Med. 2020;3:81.
https://doi.org/10.1038/s41746-020-0288-5.

Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:l886.
https://doi.org/10.1136/bmj.l886.

Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577-9.
https://doi.org/10.1016/S0140-6736(19)30037-6.

Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63.
https://doi.org/10.1038/s41591-020-1037-7.

Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2(10):e549-e560.
https://doi.org/10.1016/S2589-7500(20)30219-3.

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74.
https://doi.org/10.1038/s41591-020-1034-x.

Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4.
https://doi.org/10.1038/s41591-020-1041-y.

Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. Nat Med. 2020;26(6):807-8.
https://doi.org/10.1038/s41591-020-0941-1.

Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, et al. Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review. JAMA Netw Open. 2021;4(3):e211276.
https://doi.org/10.1001/jamanetworkopen.2021.1276.

Liu X, Faes L, Kale AU, Challa S, Wagner SK, Fu DJ, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297.
https://doi.org/10.1016/S2589-7500(19)30123-2.

Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.
https://doi.org/10.1136/bmj.m689.

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
https://doi.org/10.1038/s41586-019-1799-6.

Rajpurkar P, Irvin J, Ball RL, Zhu K, Mehta H, Yang B, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Med. 2018;15(11):e1002686.
https://doi.org/10.1371/journal.pmed.1002686.

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.
https://doi.org/10.1186/s12916-019-1426-2.

Babic B, Gerke S, Evgeniou T, Cohen IG. Direct-to-consumer medical machine learning and artificial intelligence applications. Nat Mach Intell. 2021;3(4):283-7.
https://doi.org/10.1038/s42256-021-00331-0.

Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. npj Digit Med. 2018;1:5.
https://doi.org/10.1038/s41746-017-0002-6.