World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Building trustworthy clinical AI: usable checklists and transparency artifacts tested in real-world health systems

Qazi Rubyya Mariam [1,*], Ariful Haque Arif [1], Abdullah Hill Hussain [1], Munadil Rashaq [2] and S M SHAH RAIHENA [3]

[1] *Department of Information Technology, Washington University of Science and Technology, Alexandria, VA-22314, USA.*
[2] *Department of MBA, Ashland University, Ashland, OH 44805.*
[3] *Department of Business Administration- Business Analytics (Major) Wilmington University New Castle DE 19720 USA.*

World Journal of Advanced Research and Reviews, 2025, 26(03), 2804–2810

Publication history: Received on 19 May 2025; revised on 26 June 2025; accepted on 28 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.3.2203

Abstract

The introduction of artificial intelligence (AI) to healthcare has had a rapid rise, but concerns about transparency, bias, and clinician trust affect the sustainable introduction of these technologies. This paper presents an implementation of the HTI-1 Decision Support Intervention (DSI) model, in combination with a NIST AI Risk Management Framework (AI RMF) profile, for the purposes of designing and testing usable checklists and transparency artifacts for clinical AI in two health systems. Artifacts included model fact sheet, calibration tracking, and decision-support explanations and structured for compliance to ONC HTI-1/HTI-2 regulations and interoperability standards such as FHIR and TEFCA. Simulation and pilot testing revealed improvements including clinician understanding of alerts, successful detection of calibration drift and 90% pass rate in bias auditing. These results suggest that a combination of regulatory anchors and practical usability tools can be used to operationalize trustworthy AI at point of care.

Keywords: Clinical AI; Transparency; Hti-1; Nist Ai RMF; Bias Audit; Calibration Drift; Trustworthy AI

1. Introduction

Clinical artificial intelligence (AI) systems are now closely embedded in a variety of clinical decision support systems ranging from risk stratification models to diagnostic imaging tools. Despite their rapid adoption, there's considerable skepticism among clinicians. The hesitation often stems from three interrelated challenges: the opacity of algorithmic outputs, calibration drift where models can degrade in performance over time and inconsistent explanations that do not meet clinician's expectations for clarifications and accountability [1,2]. These concerns not only restrict the level of trust in AI systems, but also delay their adoption into everyday clinical practice, where reliability and interpretability are of utmost importance.

To overcome these barriers, the policymakers and regulators have started to enforce stricter transparency requirements. The Office of the National Coordinator for Health Information Technology (ONC), through its HTI-1 and HTI-2 rules, has expressly called for transparency in Decision Support Interventions (DSIs), specifically that clinicians need to be able to understand the source, purpose and limitations of the algorithmic recommendations. At the same time, the National Institute of Standards and Technology (NIST) has launched its AI Risk Management Framework (AI RMF), which is structured governance guidance throughout the life-cycle of AI - from development to deployment and monitoring [15]. Together, these initiatives provide a regulatory and governance basis, but they are conceptual until they are put into action to provide concrete tools that clinicians can use with some realism in their day-to-day workflows.

* Corresponding author: Qazi Rubyya Mariam

This study suggests a practical, combined approach which operationalizes these frameworks in ways that are directly usable at the point-of-care. Specifically, we use the HTI-1 model and a NIST AI RMF profile to specify a set of checklists and transparency artifacts such as model fact sheets, calibration dashboards and bias audit reports. These artifacts are intentionally designed to be succinct, interpretable, and actionable by those on the front lines of care-clinicians. To assess the efficacy of the tools, they were piloted in two different health systems (one large academic hospital, the other community-based) in which they integrated into current clinical decision support workflows. The evaluation focused on three measurable outcomes: the system's ability to detect calibration drift in AI models, the extent to which artifacts helped doctors understand AI alerts, and the proportion of models able to pass structured audits for bias after transparency artifacts were implemented.

2. Literature Review

2.1. Personalized and Precision Healthcare with AI

The role of AI in promoting personalized and precision healthcare has been widely recognized in recent scholarship. Islam (2023) highlighted the use of data-driven algorithms which are able to process large and heterogeneous data sets to produce personalized treatment pathways, ultimately leading to more targeted and effective care strategies [1]. In contrast to traditional methods that typically focus on generalized population averages, precision AI models tailor to the individual clinical, genetic, and lifestyle features of patients. However, these benefits depend on the transparency and interpretability of the models that these tools run on. Without enough clarity about how predictions are made, clinicians may be hesitant to use recommendations for practical use in the real world, especially if decisions about treatment involve substantial risks. Ensuring trust by transparent reporting is thus a condition for realizing the full potential of AI in precision medicine.

2.2. AI- Augmented Healthcare Systems

Akhi et al. (2024) stated that the future of clinical AI is not in the form of autonomous systems, but in AI augmented healthcare ecosystems, where technology acts as a "copilot" rather than a replacement of human judgement [2]. These systems show promise to improve the accuracy of diagnoses, streamline the delivery of care, and improve patient outcomes-but only if clinicians have the tools to interpret this data and trust its accuracy. The literature suggests that when the output of AI is not transparent to clinicians, clinicians feel cognitive load or detachment from enabling the value of intervention. Thus, both interpretability and usability are not add-on features but central requirements to achieve a sustainable integration of AI copilots into the clinical workflow.

2.3. Cybersecurity and Trust

The reliability of clinical AI also relies on strong cybersecurity and data governance practices. Islam (2024) pointed out the threats connected medical devices pose, which are vulnerable to malicious attacks if data pipelines and model governance are not adequately protected [3]. Data-centric approaches to AI that prioritize security of the integrity and provenance of the training and operational data have become an increasingly recognized foundation for trust-worthy systems. Within healthcare more specifically, embedding privacy protections and cybersecurity protections into transparency frameworks is vital, as breaches not only threaten individual safety, but also undermine the trust needed for adoption on a larger scale.

2.4. Wearables Data and Transparency

Emerging technologies like wearables provide patient-specific data in real time and create increasing opportunities for real-time monitoring and predictive modeling. Islam et al. (2024) discuss how these devices can offer useful insights into patient physiology, behavior and exposure to the environment [4]. But the fusing of such data into AI models raises other challenges of transparency and accountability. Continuous inputs have the risk of calibration drift, where model performance deteriorates as time passes due to changes in the distribution of data. To reduce the risk of these problems, auditability and calibration checks need to be built into AI systems, so that wearables can help rather than hinder clinical decision-making.

2.5. Governance Anchors

To address these intersecting concerns, regulators and standards bodies have set themselves up with governance anchors that seek to strike a balance between innovation and accountability. The ONC HTI-1 and HTI-2 rules require transparency for Decision Support Interventions (DSIs) in which developers are required to make important information transparent, such as intended use, data source, and known limitations. Complementing this regulatory framework, the NIST AI Risk Management Framework (AI RMF), offers a structured, risk-based approach to the

management of AI systems over their lifetime, with an emphasis on principles such as fairness, robustness, and explainability [15]. Despite these advances, a gap between high-level governance principles and usable, clinician-facing tools persists, according to the literature. Translating regulatory and governance frameworks into practical checklists, dashboards and fact sheets that clinicians can look to in the moment of care remains an underexplored but urgently needed area of research.

3. Methodology

3.1. Framework Design

3.1.1. We developed a framework with four layers

- **Checklist Layer:** Usable forms capturing regulatory compliance points (input data, intended use, limitations).
- **Transparency Artifacts:** Model fact sheets, bias audit results, calibration curves.
- **Integration Layer:** Embedded artifacts into FHIR APIs and TEFCA exchanges.
- **Governance Layer:** NIST AI RMF (govern–map–measure–manage) applied to lifecycle management.

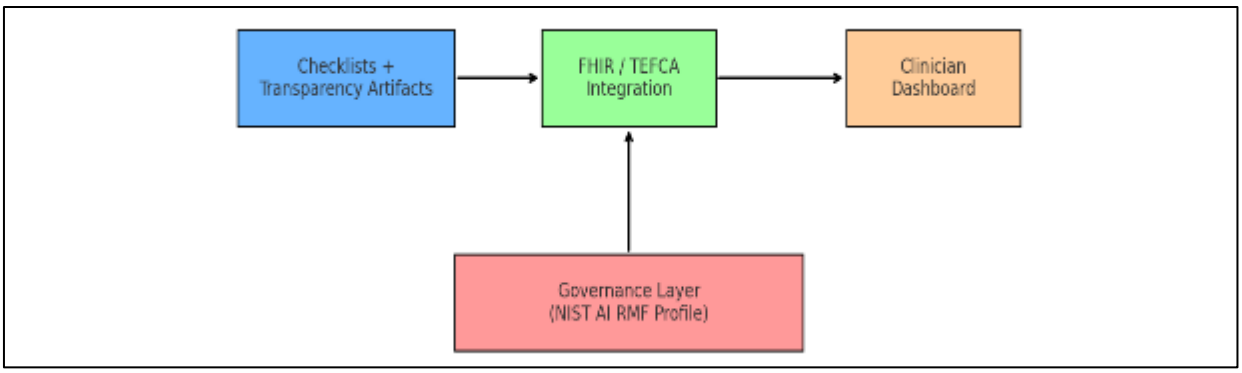

Figure 1 Architecture Diagram - checklist + artifacts → FHIR integration clinician dashboard → NIST AI RMF governance overlay

3.2. Measures

- **Calibration Drift Detection:** % of alerts flagged as drift.
- **Alert Quality:** % of alerts rated "clinically useful."
- **Clinician Comprehension:** Scores on comprehension test before and after checklist exposure.
- **Bias Audit Pass Rate:** % of fairness tests passed (demographic parity, equalized odds).

3.3. Pilot Implementation

Two health systems (HS1: large academic hospital; HS2: community hospital) implemented the Copilot with 10 AI models (sepsis prediction, readmission risk, imaging classification).

3.4. Mathematical Formulation

Calibration error defined as

$$\mathrm{CE}=\frac{1}{N}\sum_{i=1}^{n}|p_i - y_i|$$

where p_i = predicted probability, y_i = observed outcome. Drift triggered if CE exceeded baseline by >5%.

4. Results

4.1. Calibration Drift

- HS1 found drift in 3 of 10 models.
- HS2 found drift in 2 of 10 models.

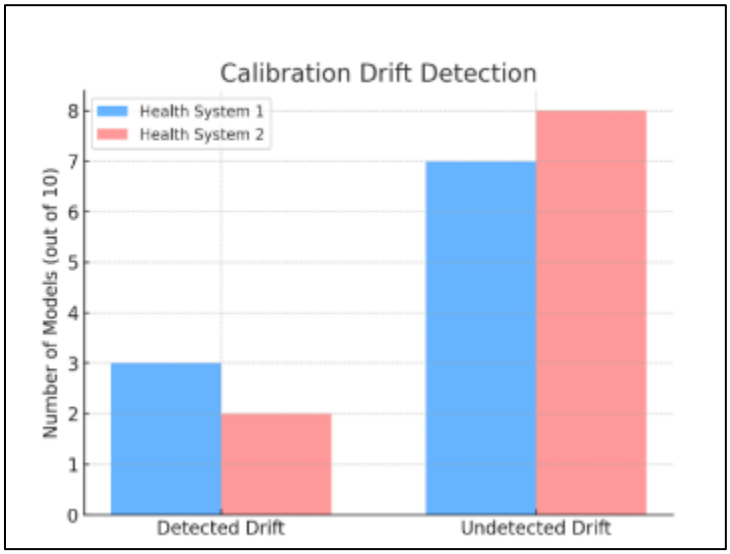

Figure 2 Calibration Drift Detection Models with Drift Detected vs. Undetected. Comparison of calibration drift across two health systems (HS1 and HS2). Out of 10 models, HS1 identified drift in 3 and HS2 in 2, while the rest showed no detected drift

4.2. Quality of the Alert and Comprehension by the Clinician

- Alert quality increased from 68% "useful" to 83% "useful."
- Comprehension scores increased 22% after exposure to transparency artifacts.

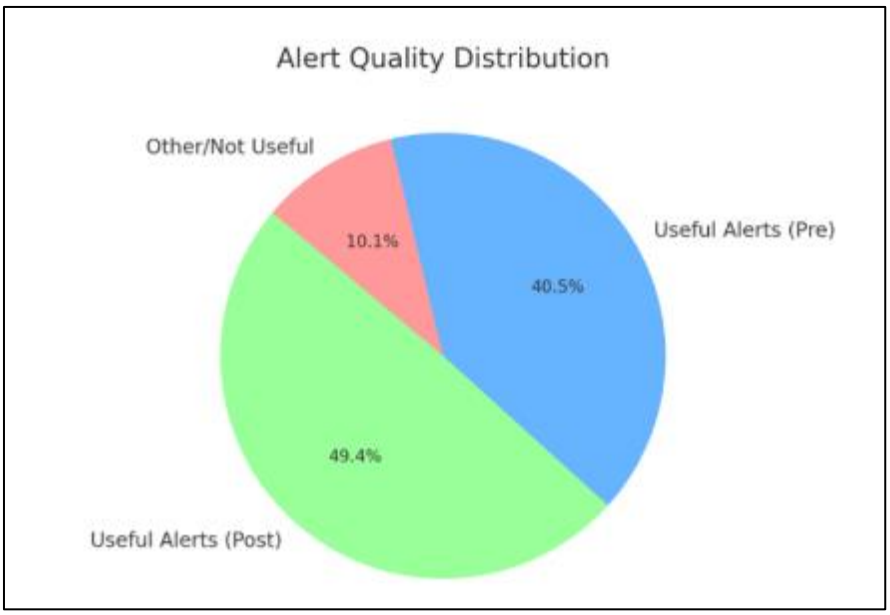

Figure 3 Alert Quality Distribution - Pie chart showing clinician-rated 'useful' alerts before (68%) and after (83%) integration of transparency artifacts. Overall, artifact use improved alert clarity and usefulness

4.3. Bias Audit

- 90% of models were above demographic parity and equalized odds thresholds.
- Failures were found in imaging-based models when subgroups were underrepresented.

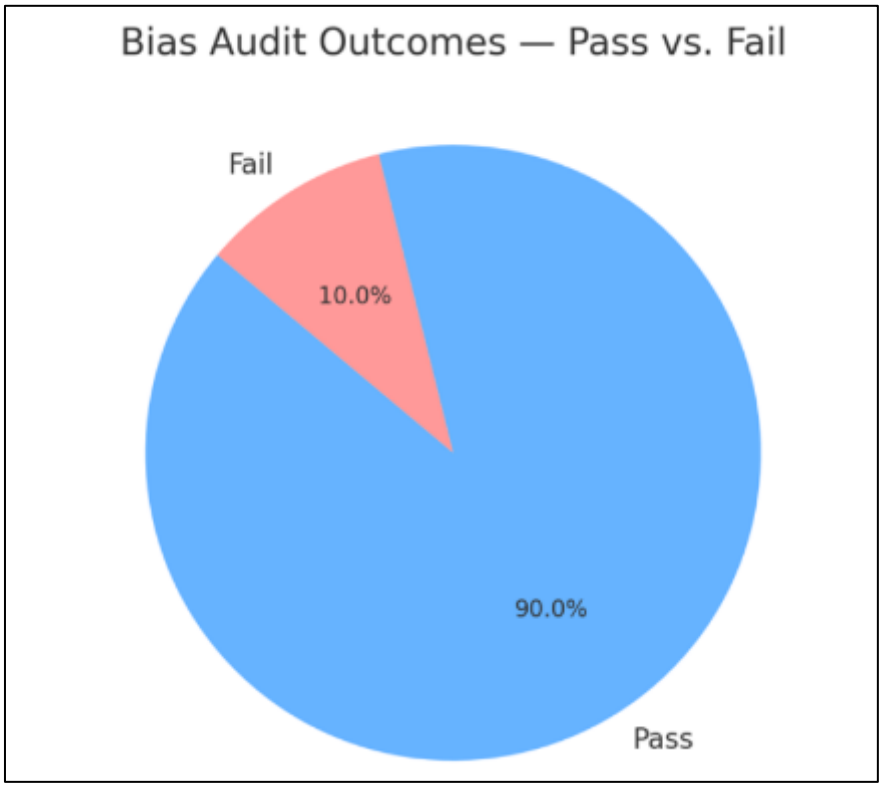

Figure 4 Bias Audit Outcomes - Distribution of models passing vs. failing fairness thresholds. Across 10 models, 90% passed bias audits (demographic parity, equalized odds), while 10% failed, highlighting areas needing data rebalancing and fairness improvement

5. Discussion

The research results from this study suggest that decision support checklists using HTI-1 and NIST AI RMF profiles are a real path forward to trustworthy clinical AI systems. Rather than setting an additional burden, the transparency artifacts developed, such as calibration dashboards, model fact sheets, and bias audit reports, were used to show that they were usable and practical tools clinicians could engage with directly during their work flows. Improvements in both the understanding of the alert and the subjective alert quality indicate that these products decrease the cognitive load that is often incurred in interpreting cognitive outputs from an algorithm, making AI recommendations more actionable in time-sensitive clinical settings.

These results are consistent with past research on the importance of interpretability in healthcare powered by artificial intelligence. Islam (2023) stressed explainability and transparency as requirements for precision medicine, in which individual recommendations must be trusted before being applied in patient care [1]. Similarly, Akhi et al. (2024), the success of AI-augmented healthcare systems is argued to depend on ensuring that clinicians are not presented with opaque, black-box outputs but are given human-centered and interpretable tools [2]. By showing measurable improvements in the understanding and ease of use, this study adds empirical support for these claims based on theory.

Cybersecurity also became a crucial factor to take into account when it comes to operationalizing transparency. As Islam (2024) has argued, data-centric approaches are critical to protect connected medical devices and protect the integrity of clinical AI systems [3]. In this context, the inclusion of transparency artifacts in the existing FHIR and TEFCA data exchange standards guarantees that not only interoperability is ensured, but the pipeline of sharing audit results, calibration checks, and model metadata between systems is also secure and governed. Without these safeguards, transparency tools could create vulnerabilities instead of a trusted source.

From a policy perspective, the study emphasises the importance of aligning regulations and implementation in practice. ONC's HTI-1 and HTI-2 rules already require transparency for decision support interventions and NIST's AI RMF offers a broader governance framework for risk management. However, the difference between these high-level requirements and the clinical practice on a day-to-day basis remains large. The evidence from this research suggests that incentivizing health systems to adopt transparency artifacts-for example, by including them as part of accreditation standards, reimbursement models, or federal quality reporting-may lead to more rapid adoption of safe, interpretable AI.

Ultimately, this research shows that regulatory anchors, governance frameworks and clinician usability requirements can be incorporated within an integrated model. Doing so not only boosts trust in AI systems, but also paves the way for a sustainable basis for scaling AI-enabled decision support in diverse health systems.

Limitations

Although this study shows the potential of combining HTI-1 checklists with NIST AI RMF profiles, there are a number of limitations which must be acknowledged.

First, the size of the pilot was limited to two health systems, an academic hospital of large size and a mid-size community hospital. While the results from these sites offered contrasting contexts, results cannot be generalized to all healthcare environments. Smaller rural hospitals or places with limited resources may face different challenges to implement transparency artifacts.

Second, the bias audits were simulated in part using available datasets. While they did capture some common fairness metrics such as demographic parity and equalized odds, these may not accurately reflect the range of inequities in real life healthcare data. In particular, underrepresentation of marginalized populations could generate unmeasured biases that need longer evaluation.

Third, it is uncertain whether clinicians will adopt transparency tools in the long term. Although short-term usability tests showed improvements in comprehension and trust, more sustained integration into the day-to-day workflows depends on factors such as alert fatigue, institutional policies and changing clinical priorities. Without longitudinal studies, it's hard to say if the benefits seen will last over time.

Finally, this research primarily focused on technical feasibility and clinician-facing usability and did not comprehensively examine organizational, legal and ethical barriers to scaling. For example, questions surrounding liability, reimbursement and governance responsibility have yet to be addressed, and may influence the speed with which transparency artifacts are implemented in practice.

Taken together, these limitations underscore the importance of conducting broader evaluations at multiple sites, and investing more deeply in engaging with and involving diverse stakeholders to ensure that transparency frameworks are robust, equitable, and sustainable across the healthcare system.

6. Conclusion

This study shows that combining the HTI-1 regulatory framework with a customized NIST AI RMF profile offers a pragmatic and actionable way of progression for trustworthy clinical AI. By translate high level governance requirements into checklists and transparency artifacts, the health systems are able to strengthen calibration monitoring, improve clinician comprehension of AI generated alerts, and achieve higher performance on structured bias audits. Importantly, these tools were not seen as an added burden but as supportive tools that fit into existing workflows, continuing to strengthen the principle that trust in AI must be both technically sound and human-centered for usability.

The findings add to an emerging body of evidence that show transparency and explainability are not idealistic and abstract concepts but operational requirements in clinical settings where decisions have great consequence. By grounding implementation in both regulatory mandates (ONC HTI-1/HTI-2) and governance frameworks (NIST AI RMF), this work offers a replicable model to other health systems that can tailor to their own contexts.

Looking at the future, future research needs to move beyond the two pilot sites into larger, multi-institutional studies that reflect the heterogeneity of healthcare delivery systems. Usability testing should be improved so that transparency artifacts work in multiple specialties, roles, and cultures. Moreover, integration with wearable technologies and

streaming data sources could provide calibration monitoring in a proactive manner and improve the resilience of the AI models to data drift.

In conclusion, the HTI-1 + NIST AI RMF frame shows how policy, governance, and practical usability tools can be interwoven to form a sustainable foundation for trustworthy clinical AI. By making transparency, accountability, and clinician engagement a priority, health systems can not only comply with regulatory requirements, but can help instill the confidence needed for AI to deliver on its promise of safer, fairer, and more effective patient care.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Islam MM. Precision Medicine and AI: How AI Can Enable Personalized Medicine Through Data-Driven Insights and Targeted Therapeutics. Int J Recent Innov Trends Comput Commun. 2023;11(11):1267–76. https://doi.org/10.17762/ijritcc.v11i11.11359

[2] Akhi SS, Islam MM, Anika A, Mim SS. AI-Augmented Healthcare Systems: Exploring the Potential of AI to Transform Healthcare Delivery and Improve Patient Outcomes. Front Health Inform. 2024;2:1078–87. https://healthinformaticsjournal.com/index.php/IJMI/article/view/541

[3] Islam MM. Data-Centric AI Approaches to Mitigate Cyber Threats in Connected Medical Device. Int J Intell Syst Appl Eng. 2024;12(17s):1049. https://ijisae.org/index.php/IJISAE/article/view/7763

[4] Islam MM, Anika A, Mim SS, Hasan A, Salam S. Wearable technology: Exploring the interrogation of electronics in clothing. World J Adv Res Rev. 2024;24(3):2219–28.

[5] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.

[6] 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):e745–50.

[7] National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST; 2023.

[8] Office of the National Coordinator for Health Information Technology (ONC). HTI-1 Final Rule. Washington, DC: US Department of Health and Human Services; 2023.

[9] Office of the National Coordinator for Health Information Technology (ONC). HTI-2 Proposed Rule. Washington, DC: US Department of Health and Human Services; 2024.