frontiers in Blockchain

TYPE Perspective
PUBLISHED 17 August 2026
DOI 10.3389/fbloc.2026.1809318

Check for updates

OPEN ACCESS

EDITED BY
Szymon Łukaszyk,
Łukaszyk Patent Attorneys, Poland

REVIEWED BY
Ingrid Vasiliu Feltes,
University of Miami, United States
Edouard Lansiaux,
Université de Lille, France
Steven Williamson,
University of North Texas, United States

*CORRESPONDENCE
Pei-Fan Shih,
petershih5@gmail.com

RECEIVED 11 February 2026
REVISED 20 June 2026
ACCEPTED 22 June 2026
PUBLISHED 17 August 2026

CITATION
Shih P-F (2026) The invisible risk trust framework: a decentralized consensus protocol for validating medical AI.
Front. Blockchain 9:1809318.
doi: 10.3389/fbloc.2026.1809318

COPYRIGHT
© 2026 Shih. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The invisible risk trust framework: a decentralized consensus protocol for validating medical AI

Pei-Fan Shih *

Independent Researcher, Taoyuan, Taiwan

Background: In precision medicine, the convergence of ionizing radiation, algorithmic bias, and carbon emissions creates a triad of "invisible risks" that stalls AI adoption. Traditional centralized validation fails to bridge this "trust gap" due to data silos and opacity. A new paradigm is required to audit these risks without compromising privacy.
Methods: This paper proposes the Invisible Risk Trust Framework, a conceptual decentralized architecture serving as a complementary audit layer to regulatory bodies. The framework integrates three core components: (1) Privacy-Preserving Audit utilizing Federated Evaluation and Zero-Knowledge Proofs (ZKPs) to validate models without moving data; (2) Distributed Expert Consensus by board-certified specialists; and (3) Cryptographic Incentives featuring an "Eco-priority Queue" for sustainable compute.
Results: Applied to a theoretical nuclear medicine use case (PET/CT denoising), architectural analysis suggests that this protocol is projected to substantially reduce the redundancy of centralized data warehousing and partially automate audit workflows. This is supposed to offer a scalable, cost-efficient validation pathway, providing an immutable audit trail of model performance, bias, and carbon footprint.
Conclusion: As healthcare transitions to AI-augmented decision-making, trust must be engineered into the digital infrastructure. By rendering invisible risks visible and verifiable, this conceptual framework generates testable hypotheses for a scalable, privacy-preserving, and sustainable validation pathway.

KEYWORDS
algorithmic audit, decentralized science (DeSci), federated evaluation, nuclear medicine AI, sustainable computing, trust framework

Introduction: the convergence of invisible risks

The digitization of healthcare has ushered in a new era of precision medicine; yet, paradoxically, it has simultaneously introduced a triad of "invisible risks" that threaten both patient safety and systemic sustainability. Unlike distinct clinical symptoms, these risks are imperceptible to the naked eye but carry profound, cumulative consequences (as shown in Figure 1).

First, the Physical Risk involves the cumulative burden of ionizing radiation. While the "As Low As Reasonably Achievable" (ALARA) principle guides nuclear medicine, the clinical imperative for high-fidelity imaging often conflicts with dose reduction goals (Miwa et al., 2025; Clement David-Olawade et al., 2025; Balaji et al., 2024; Zhou et al., 2023). AI-driven denoising algorithms promise to resolve this tension, yet their validation remains sporadic and institution-specific (Miwa et al., 2025; Clement David-Olawade et al., 2025; Balaji et al., 2024; Zhou et al., 2023).

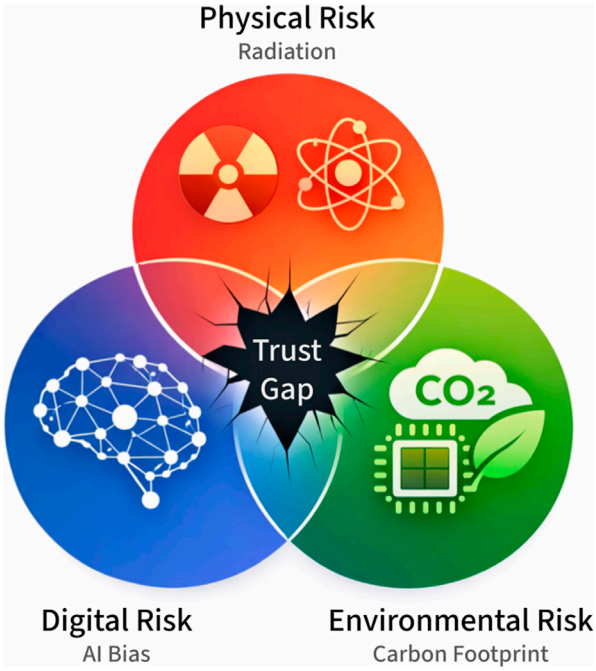

FIGURE 1
The "invisible risk triad" in modern nuclear medicine: The diagram illustrates the convergence of three systemic threats: physical risk (ionizing radiation concerns), digital risk (algorithmic "black box" bias), and environmental risk (carbon footprint of compute). The intersection of these domains represents the "trust gap," a critical implementation barrier that traditional centralized validation methods are structurally ill-equipped to bridge.

Second, the Digital Risk stems from the "black box" nature of deep learning. As noted by recent scholarship, AI diagnostic systems frequently suffer from a "generalizability crisis"—performing exceptionally well in controlled *in silico* environments but failing in heterogeneous clinical settings due to algorithmic bias (Oyeniyi, 2025). This opacity erodes the fiduciary trust between physicians and technology, creating a critical implementation gap (Arhouni et al., 2025; Nouis et al., 2025; Kurt et al., 2025; Goisauf et al., 2025).

Third, the Environmental Risk is the rapidly expanding carbon footprint of high-performance computing. The computational demands for training state-of-the-art medical AI models have escalated exponentially, raising ethical concerns about the energy sustainability of digital healthcare (Prajwal et al., 2025). The recurring energy consumption of inference in clinical settings necessitates an urgent shift toward energy-efficient "Green AI." (Prajwal et al., 2025; Agarwal and Chakraborti, 2025; Strubell et al., 2019).

Finally, Trust within this sociotechnical framework is defined as the expectation of a stakeholder that the AI system will perform its intended function reliably, safely, and ethically, supported by verifiable evidence of its integrity. In the context of decentralized governance, we specifically operationalize 'Verifiable Trust' as a property where the reliance on a system is grounded in cryptographic proofs and transparent consensus rather than the reputation of a centralized intermediary or 'blind faith' in opaque algorithms (Jacovi et al., 2021; Kassem et al., 2025; Latif et al., 2026; Baradwaj et al., 2024).

This paper posits that current centralized validation mechanisms are structurally insufficient to manage these converging risks. This article proposes the Invisible Risk Trust Framework—a unified "Proof-of-Trust" infrastructure. By leveraging decentralized consensus and cryptographic verification, this framework aims to render these invisible risks visible, measurable, and verifiable. This manuscript is a conceptual framework and architectural analysis, intended to generate testable hypotheses and design principles rather than to report an implemented system.

The problem: the limits of centralized validation

The prevailing paradigm of medical AI regulation relies on centralized authorities to approve static software versions. However, this model faces inherent structural limitations when applied to the dynamic, adaptive nature of modern AI.

Data silos and the generalizability gap

Robust validation requires diverse datasets to detect bias. However, stringent privacy regulations (e.g., HIPAA, GDPR) create data silos, preventing the aggregation of multi-institutional data necessary for external validation (Shahsavari et al., 2026; Barbaria et al., 2025). Consequently, centralized repositories introduce single points of failure and privacy bottlenecks (Barbaria et al., 2025).

Scalability and the "equity paradox"

A single centralized authority cannot scale to validate the thousands of AI models and their continuous updates generated by the market. Recent analyses observe that regulatory approvals often lack rigorous multi-site external validation, leading to an "equity paradox" where models underperform in underrepresented populations due to biased training data. The lack of continuous, real-time monitoring means that model drift and performance degradation go undetected until adverse clinical outcomes occur (Oyeniyi, 2025).

Trust deficit

The lack of transparent audit trails creates a "trust deficit." Without verifiable proof of provenance and performance, stakeholders are forced to rely on "blind trust" in the aggregator (Sharma et al., 2025). A potential paradigm shift is required: moving from institutional reputation-based trust to "verifiable trust" based on cryptographic proofs (Sharma et al., 2025).

The proposed architecture: the invisible risk trust framework

The author introduces a conceptual architecture designed to validate medical AI through a Decentralized AI Validation Protocol (DAVP). It is critical to note that this framework is not intended to

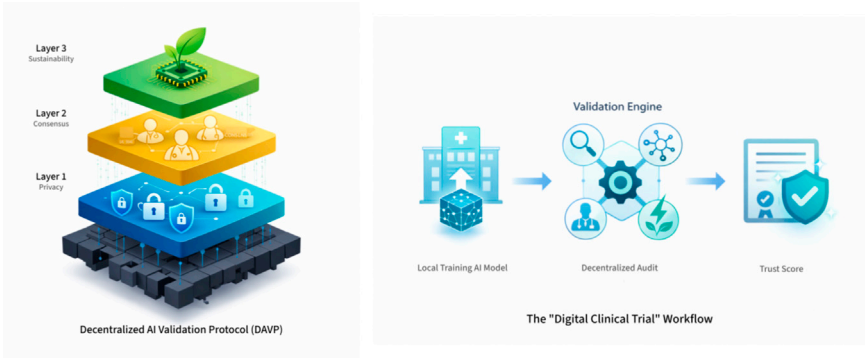

FIGURE 2
Comprehensive overview of the decentralized AI validation protocol (DAVP): (left) Architectural stack: the system is built on three synergistic layers: a privacy layer utilizing zero-knowledge proofs (ZKPs) and federated evaluation, a consensus layer for distributed expert adjudication, and a sustainability layer implementing an "eco-priority queue." (right) Operational workflow: the lifecycle of an AI model from local training to clinical deployment. The unverified model undergoes a three-step decentralized audit process (privacy, consensus, energy audit) to generate a multi-dimensional trust score (TS), ensuring that only models meeting the governance threshold are certified on the blockchain.

replace statutory regulators (like the FDA) but to serve as a complementary technological audit layer—providing continuous, mathematically verifiable evidence to support regulatory decision-making.

The architecture consists of three synergistic layers: (as illustrated in Figure 2).

Layer 1: privacy (federated learning and zero-knowledge proofs)

To dismantle data silos, the framework utilizes Federated Learning (FL) and External Validation, enabling model validation across distributed institutions without moving raw patient data (McMahan et al., 2016; Yang et al., 2024). To mitigate the risk of model poisoning, the DAVP integrates Zero-Knowledge Proofs (ZKPs). ZKPs allow a participating institution to mathematically prove that a model update was computed correctly using authorized data, without revealing the data itself (Commey et al., 2025). This ensures "verifiability-by-default" and protects against model poisoning while maintaining strict compliance with privacy regulations (Sharma et al., 2025).

For a representative 3D U-Net (≈22 million parameters) performing PET/CT denoising on 128 × 128 × 128 volumes, we provide order-of-magnitude estimates extrapolated from recent zkML benchmarks on NVIDIA A100 architecture (e.g., ezkl framework utilizing Halo2 backend) (Akor et al., 2026; Chen et al., 2024; South et al., 2024). Under these projections, proof generation requires approximately 7–12 min per full volumetric evaluation batch (10 studies). Verification is constant-time (<15 ms) with proof sizes of 85–120 KB. These illustrative figures align with current state-of-the-art reports showing that modern zk-SNARK proving times scale predictably with model size. As a pragmatic transitional architecture, the framework supports hybrid ZKP + Trusted Execution Environments (TEEs, e.g., Intel SGX or AWS Nitro) for the initial 6–12 months, progressively migrating to pure ZKP as hardware-accelerated provers mature. This hybrid approach maintains full auditability while keeping per-validation latency under 20 min.

Layer 2: consensus (distributed expert network)

Trust cannot be purely algorithmic; it requires human oversight. This layer implements a Distributed Expert Consensus mechanism where board-certified specialists (e.g., nuclear medicine physicians) act as validator nodes. Through blockchain-based smart contracts, these experts reach consensus on the clinical utility of model outputs (Barbaria et al., 2025). This establishes a robust, immutable infrastructure for clinical trust, recording professional consensus on an immutable ledger.

Within the proposed framework, expert consensus is not based on direct access to raw patient data or original imaging datasets. Instead, specialists evaluate a standardized, privacy-preserving evaluation report generated through federated external validation. This report includes predefined quantitative and qualitative metrics, such as discrimination performance (e.g., AUC), calibration indices, subgroup performance stratified by demographic and technical variables, documented failure modes, and an environmental performance summary (e.g., estimated computational energy consumption per inference or training cycle).

The role of the expert network is to assess whether the model meets predefined clinical and safety thresholds within its intended use context. Based on this structured evaluation dossier, the consensus output consists of three components: (1) a determination of whether minimum performance and safety criteria have been satisfied; (2) recommendations for additional external testing, subgroup analysis, or real-world monitoring where uncertainty remains; and (3) an interpretable multi-dimensional Trust Score reflecting performance robustness, safety considerations, and sustainability indicators. Thresholds are pre-specified by professional societies/regulators (e.g., SNMMI/EANM/FDA guidance) and version-controlled on-chain. Importantly, the consensus process operates exclusively on aggregated and cryptographically verifiable evaluation outputs, thereby preserving patient privacy and institutional data sovereignty while maintaining clinical oversight.

The consensus protocol utilizes a Verifiable Random Function (VRF) for expert sortition to ensure cryptographic randomness. The consensus logic is formally defined as:

$$\text{Decision} = \left\{1, \text{if } \Sigma_{j=1}^{N} v_j \geq \left\lfloor 2N/3 \right\rfloor + 1; 0, \text{otherwise}\right\}$$

The expert consensus employs a Byzantine Fault Tolerant (BFT) supermajority rule. With N experts randomly selected via Verifiable Random Function (VRF) sortition, approval requires at least ⌊2N/3⌋ + 1 affirmative votes ($v_j \in \{0,1\}$). This threshold guarantees integrity even if up to one-third of the validator nodes are compromised.

Layer 3: sustainability (cryptographic incentives)

To address environmental risks, the framework incorporates a Cryptographic Incentive Mechanism. This verifiable model rewards contributors who develop energy-efficient models or utilize low-carbon computational resources. The incentive mechanism is intended for governance and participation alignment rather than financial speculation. The system implements an "**Eco-priority Validation Queue**," which prioritizes the processing of models that demonstrate high diagnostic performance with minimal energy expenditure, thereby aligning clinical excellence with environmental responsibility (Prajwal et al., 2025).

Case study: nuclear medicine, with its quantitative nature and stringent ALARA requirements, represents a particularly well-suited testbed for the DAVP

The Need for Trust: In theranostics and radiopharmaceutical therapy, dosimetry calculations must be precise to spare organs at risk while treating tumors (Montes et al., 2026). A "black box" AI error in this context carries immediate physical risks, making transparent validation non-negotiable (Arhouni et al., 2025; Nouis et al., 2025; Kurt et al., 2025).

Theoretical workflow: (as demonstrated in Figure 2).

Model Submission: A research hospital trains a local AI model for low-count PET denoising (aiming to reduce radiation dose) (Zhou et al., 2023).

Decentralized Validation: The model weights are submitted to the DAVP. The network uses ZK-FL to validate the model against diverse, private datasets held by consortium hospitals to test for generalizability and bias (Commey et al., 2025).

Green Audit: The system measures the inference energy cost during validation.

Trust Score generation: The protocol generates a composite "Trust Score" comprising three metrics.

- Accuracy: Performance across diverse demographics (Bias check) (Oyeniyi, 2025).
- Safety: Adherence to ALARA principles (Radiation check).
- Sustainability: Energy efficiency rating (Carbon check) (Agarwal and Chakraborti, 2025).

To operationalize the abstract concept of "trust," the Trust Score adopts a non-linear, thresholded formulation, formally defined as:

$$\text{TS} = \text{I}(\text{P_equity} \geq \tau\text{_eq}) \cdot \text{I}(\text{S_safety} \geq \tau\text{_safe}) \cdot \text{I}(\text{C_eco} \geq \tau\text{_eco}) \cdot (\text{w}_1\text{P_equity} + \text{w}_2\text{S_safety} + \text{w}_3\text{C_eco})$$

$$TS = \left(\prod_{k \in \{eq,safe,eco\}} \| \ (X_k \geq \tau_k) \right) \left(w_1 P_{equity} + w_2 S_{safety} + w_3 C_{eco} \right)$$

Where P_{equity}, S_{safety}, and C_{eco} are strictly normalized, dimensionless parameters bounded within the interval [0,1], uniformly engineered with a "higher-is-better" orientation.

- P_{equity} (Performance): Measures clinical accuracy and fairness (e.g., normalized subgroup-specific AUC).
- S_{safety} (Safety): Quantifies the clinical risk-benefit ratio (e.g., normalized Structural Similarity Index [SSIM] relative to dose reduction).
- C_{eco} (Sustainability): Represents environmental efficiency. To maintain the positive mathematical orientation, raw environmental costs (e.g., gCO2e per inference or kWh) are inversely normalized against a pre-registered worst-case baseline, such that lower energy consumption yields a C_{eco} score approaching 1.

I (·) is the indicator function (1 if true, 0 otherwise). This forces TS = 0 if any single dimension fails, eliminating substitutability between safety and performance. The weights w_i (e.g., an illustrative baseline of 0.4 for equity, 0.35 for safety, and 0.25 for sustainability) are deliberately modular. This specific hypothetical distribution reflects a clinical logic that marginally prioritizes diagnostic fairness and radiation safety over environmental cost. However, these are not empirical constants; they are governed by a multi-signature smart contract controlled by professional societies (e.g., SNMMI/EANM) and updated annually via public consultation. Thresholds τ are pre-registered clinical minimums set by the same societies.

Clinical deployment: Models exceeding the Trust Score threshold are immutably recorded on the ledger for clinical use.

This hypothetical scenario serves as an illustrative prototype rather than an implemented pilot; empirical testing for further validation is needed.

Discussion

Towards a complementary decentralized audit infrastructure

The Invisible Risk Trust Framework proposes a potential paradigm shift from retrospective regulation to continuous, decentralized auditing. This infrastructure acts as a robust, immutable foundation for clinical trust, ensuring that the software defining modern medicine is as reliable as the biological principles it analyzes.

Complementary regulation

The DAVP functions as a continuous Post-market Monitoring (PMM) infrastructure, not a replacement for statutory bodies. Just as the FDA requires clinical trials for drugs, the DAVP provides the infrastructure for "digital clinical trials" that are continuous, multi-institutional, and tamper-proof (Oyeniyi, 2025; Riaz et al., 2026). This empowers regulators to oversee adaptive systems with mathematically guaranteed integrity (Oyeniyi, 2025; Arhouni et al., 2025; Joshi et al., 2024).

Alignment with NIST AI risk management framework (AI RMF 1.0)

The proposed DAVP architecture is designed to operationalize the core functions defined in the NIST AI RMF 1.0—specifically the Govern, Map, Measure, and Manage functions (Tabassi, 2023). While NIST provides the organizational process standards for AI trustworthiness, DAVP provides the technical infrastructure to automate the 'Measure' function through decentralized quantitative Trust Scores. For instance, the framework's focus on algorithmic bias and radiation safety directly addresses NIST's requirement for systems to be 'Fair' and 'Safe.' Furthermore, by utilizing decentralized expert consensus for version-controlled threshold setting, the protocol enforces the 'Govern' function, ensuring that accountability is not localized within a single institution but distributed across a professional community (Costanza-Chock et al., 2022). Thus, DAVP serves as a technical realization layer that complements NIST's high-level guidance, transitioning from static compliance checklists to continuous, verifiable risk auditing.

Alignment with ISO/IEC 42001 and international standards

Beyond the NIST framework, the DAVP architecture specifically aligns with ISO/IEC 42001:2023 (Artificial Intelligence Management System) (International Organization for Standardization/International Electrotechnical Commission, 2023). While ISO/IEC 42001 provides the high-level management system requirements (Clauses six on Risk Treatment and nine on Performance Evaluation), DAVP serves as a technical enforcement layer that generates the "documented evidence" required for certification. By providing an immutable Trust Score, the framework satisfies ISO requirements for transparency and accountability in AI decision-making. However, DAVP diverges from traditional ISO auditing, which is often periodic and retrospective, by offering real-time, decentralized verification. This ensures that compliance is not a point-in-time snapshot but a continuous property of the AI lifecycle, aligning with ISO/IEC 24028:2020's objectives for establishing trustworthiness in AI systems through technical controls (International Organization for Standardization and International Electrotechnical Commission, 2020).

Generalizability across medical domains

While nuclear medicine serves as the primary testbed due to stringent ALARA constraints, the DAVP architecture is structurally designed to scale across broader medical AI use cases. The "Invisible Risk Triad" maps consistently to other high-stakes specialties. In diagnostic radiology, for instance, the Physical Risk translates from ionizing radiation to contrast agent toxicity (e.g., gadolinium retention) or prolonged MRI scan times, where AI acceleration models must be validated for diagnostic non-inferiority (Chen et al., 2025; Zhang et al., 2025; Deng et al., 2025; Tsui et al., 2024). Similarly, in digital pathology, the analysis of gigapixel whole-slide images (WSIs) demands massive computational resources, exacerbating the Environmental Risk. Here, the DAVP's Eco-priority Queue can audit energy consumption, while ZK-Federated Evaluation assesses algorithmic bias across varying laboratory staining protocols without the need to centralize multi-terabyte WSI datasets (Vafaei Sadr et al., 2024; Lan et al., 2024; Jahanifar et al., 2025; Jin et al., 2026). By adjusting the domain-specific zero-tolerance thresholds (τ_i) and weighting coefficients (w_i) within the smart contract, the protocol can dynamically adapt to the unique risk profiles of oncology, cardiology, or genomics.

Projected scalability and economic feasibility (hypotheses)

On a permissioned Polygon zkEVM L2 (or equivalent), per-validation gas cost is estimated at 1.2–2.5 M gas (≈US$0.03–0.08 at current L2 prices). End-to-end time-to-validation: 15–45 min (ZK proof 7–12 min + expert voting window 20 min + on-chain finality <1 min).

Cold-start is addressed by initial consortium seeding (10–15 major academic hospitals) with subsidized first 500 validations. Hospitals bear only local inference cost (≈0.5–2 kWh per model, covered by existing clinical compute budgets), while receiving immutable Trust Score as regulatory-grade evidence. We hypothesize 60%–75% reduction in multi-site validation administrative overhead compared to centralized CRO models (based on MedRec scaling benchmarks). These figures are architectural projections and will be empirically validated in future pilot implementations.

Projected benefits

Based on architectural modeling, automating data integrity verification via smart contracts is projected to potentially reduce the administrative burden and costs associated with multi-site validation compared to traditional centralized data warehousing (Barbaria et al., 2025). The use of ZKPs and blockchain may alleviate the need for expensive third-party data reconciliation and audits, streamlining the path from development to clinical implementation (Sharma et al., 2025; Riaz et al., 2026).

Ethical safeguards

To address the ethical imperative of preventing "distributed consensus" from devolving into "distributed bias" or cartel-like collusion, the framework employs cryptographic sortition—a stochastic mechanism that randomly selects diverse subsets of qualified experts for each validation task (Gilad et al., 2017). This is reinforced by "Reputation Staking," where validators stake their professional standing on the accuracy of their reviews, facing

TABLE 1 Comparative analysis of the invisible risk trust framework against leading AI governance and healthcare blockchain architectures.

Framework	Privacy mechanism	Expert consensus	Sustainability audit	Regulatory role	AI model validation
NIST AI RMF	Policy-based	Internal	None	Guidance	None
EU AI Act	Compliance-based	Third-party	None	Statutory	High-risk conformity
ISO/IEC 42001	Management system	Internal audit	None	Certifiable	AI-specific
MedRec	Blockchain (permissioned)	None	None	Data management	None
Gem health	Blockchain data exchange	None	None	Interoperability	None
Patientory	Patient-centric PHR	None	None	Data ownership	None
DAVP (proposed)	ZK-federated	Decentralized expert	Eco-priority queue	Complementary audit	Full (performance + bias + carbon)

Table 1 expands the comparison to include additional prominent blockchain-in-healthcare platforms (Gem Health, Patientory) and the leading AI governance standard (ISO/IEC 42001), highlighting DAVP's unique integration of cryptographic AI validation with clinical expert consensus and sustainability auditing.

algorithmic penalties for malicious or scientifically unfounded voting (Zamani et al., 2018). By democratizing the validation process across a global, permissionless network, the protocol actively combats the "equity paradox," ensuring that AI models are vetted by a broad spectrum of clinical perspectives—including those from underrepresented regions—rather than reflecting the biases of a single institution or demographic (Seyyed-Kalantari et al., 2021; Raji et al., 2020).

Security and threat model analysis

The author adopts the standard STRIDE + DREAD model for decentralized systems. Primary attack vectors and mitigations.

- Sybil Attacks: Mitigated by non-transferable Soulbound Tokens (SBTs) bound to verifiable medical licenses (via DID + national medical board oracle). Only board-certified specialists with Proof-of-Professional-Standing can register.
- Collusion in small expert pools (e.g., highly specialized sub-committees or specific validator networks comprising fewer than ~200 nuclear medicine physicians within a given validation pool): Double-blind evaluation + VRF sortition + post-consensus reputation slashing (losing 10%–50% staked reputation points for statistically anomalous voting patterns).
- Adversarial model submission: ZKP-verified evaluation pipeline prevents poisoned inputs.
- Oracle gaming (self-reported energy): Mandatory TEE attestation (Intel SGX quote) at hardware level; false reporting triggers automatic blacklist.

This formal threat model is embedded in the core smart-contract architecture, not merely in limitations (See Table 1).

Limitations and challenges

To provide a rigorous assessment, it is critical to distinguish between the methodological limitations of this manuscript and the systemic challenges inherent in implementing the proposed framework.

Limitations of the study

This manuscript presents a conceptual architecture and theoretical workflow rather than a deployed software system. While technical feasibility is supported by state-of-the-art cryptographic benchmarks, the framework currently lacks empirical validation through a live, multi-institutional clinical pilot. Consequently, the projected reductions in administrative costs and improvements in generalization remain hypotheses derived from architectural modeling, which must be quantitatively tested in future interventional studies.

Implementation challenges of the framework

The realization of this framework faces hurdles. The computational overhead of generating Zero-Knowledge Proofs for large medical images remains a technical challenge, though hardware acceleration is mitigating this (Sharma et al., 2025). In the near term, secure enclaves (e.g., Trusted Execution Environments) combined with tamper-evident audit logs may serve as pragmatic transitional solutions, with ZKPs initially applied to pipeline integrity verification rather than complete model computation, and progressively expanded as verifiable computation technologies mature. Furthermore, regulatory adoption requires a shift in governance models to recognize decentralized validation as legally sufficient evidence (Oyeniyi, 2025; Arhouni et al., 2025; Joshi et al., 2024). This involves overcoming significant institutional inertia and establishing new legal standards for cryptographically verified expert consensus. Additionally, bootstrapping the decentralized network faces a "cold-start" problem; attracting a critical mass of board-certified validators requires carefully calibrated economic or academic incentives to ensure the sustainability of the audit infrastructure.

Conclusion

As healthcare transitions from human-centric to AI-augmented decision-making, trust can no longer be assumed; it must be

engineered into the digital infrastructure itself. The proposed Invisible Risk Trust Framework offers a scalable, transparent, and sustainable path to bridge the current "trust gap" by rendering the invisible risks of radiation, algorithmic bias, and carbon emissions mathematically verifiable. By establishing a decentralized audit layer that complements existing regulatory oversight, this protocol lays the foundation for a new standard of digital health governance, ensuring that the future of medical AI is safe, effective, and environmentally responsible by design.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

P-FS: Conceptualization, Methodology, Formal Analysis, Writing – original draft, Writing – review and editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. During the preparation of this work, the author used Large Language Models (Gemini-3) exclusively for the purpose of language editing, grammatical refinement, and structural organization to improve readability. The author reviewed and edited the content as needed and takes full responsibility for the content of the publication. The core scientific concepts, architectural design, and logical arguments are the original work of the author. Figures 1, 2 were conceptualized by the author and visualized using Chatgpt DALL-E 3, OpenAI, subsequently refined and finalized by the author. The figures contain no patient data and are purely schematic. The author provided specific prompts and architectural descriptions to guide the generation process and verified the accuracy of the resulting visual representations. The original prompt is available upon reasonable request.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agarwal, S., and Chakraborti, M. (2025). The hidden AI race: tracking environmental costs of innovation. *arXiv*. doi: 10.48550/arXiv.2511.22781

Akor, G. C., Ahakonye, L. A. C., Lee, J. M., and Kim, D. S. (2026). Benchmarking CNN components in EZKL: a layer-level analysis for EVM-compatible deployment. In: *2026 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. Tokyo, Japan: IEEE; 645–650. doi: 10.1109/ICAIIC68212.2026.11454315

Arhouni, F. E., Zitouni, I., Ouakkas, S., Abdo, M. A. S., Boukhair, A., Bouhssa, M. L., et al. (2025). "Comprehensive review on the impact of artificial intelligence on diagnosis and personalized treatment in nuclear medicine,". Editors B. Cherradi, A. Jamea, A. Boukhair, and C. Deschepper, 214. doi: 10.1051/shsconf/202521401006

Balaji, V., Song, T. A., Malekzadeh, M., Heidari, P., and Dutta, J. (2024). Artificial intelligence for PET and SPECT image enhancement. *J. Nucl. Med.* 65 (1), 4–12. doi: 10.2967/jnumed.122.265000

Baradwaj, S. S., Gilliland, D., Rincon, J., Hermjakob, H., Yan, Y., Adam, I., et al. (2024). Building an ethical and trustworthy biomedical AI ecosystem for the translational and clinical integration of foundation models. *Bioengineering* 11 (10), 984. doi: 10.3390/bioengineering11100984

Barbaria, S., Jemai, A., Ceylan, H. İ., Muntean, R. I., Dergaa, I., and Boussi Rahmouni, H. (2025). Advancing compliance with HIPAA and GDPR in healthcare: a blockchain-based strategy for secure data exchange in clinical research involving private health information. *Healthcare* 13 (20), 2594. doi: 10.3390/healthcare13202594

Chen, B. J., Waiwitlikhit, S., Stoica, I., and Kang, D. (2024). ZKML: an optimizing system for ML inference in zero-knowledge proofs. In: *Proceedings of the Nineteenth European Conference on Computer Systems*. Athens Greece: ACM; 560–574. doi: 10.1145/3627703.3650088

Chen, Y., Weber, R. A., Neve, O. M., Romeijn, S. R., Hensen, E. F., Wolterink, J. M., et al. (2025). A deep learning model to reduce agent dose for contrast-enhanced MRI of the cerebellopontine angle cistern. *Eur. Radiol.* 36 (5), 3710–3721. doi: 10.1007/s00330-025-12187-8

Clement David-Olawade, A., Olawade, D. B., Vanderbloemen, L., Rotifa, O. B., Fidelis, S. C., Egbon, E., et al. (2025). AI-Driven advances in low-dose imaging and Enhancement-A review. *Diagn. (Basel)* 15 (6), 689. doi: 10.3390/diagnostics15060689

Commey, D., Appiah, B., Klogo, G. S., and Crosby, G. V. (2025). ZKP-FedEval: verifiable and privacy-preserving federated evaluation using zero-knowledge proofs. *arXiv*. doi: 10.48550/arXiv.2507.11649

Costanza-Chock, S., Raji, I. D., and Buolamwini, J. (2022).Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In: *2022 ACM Conference on Fairness Accountability and Transparency. Seoul Republic of Korea*: ACM; 1571–1583. doi: 10.1145/3531146.3533213

Deng, Y., Huang, Y., Wu, H., He, D., Qiu, W., Jing, B., et al. (2025). Establishment of a deep-learning-assisted recurrent nasopharyngeal carcinoma detecting simultaneous tactic (DARNDEST) with high cost-effectiveness based on magnetic resonance images: a multicenter study in an endemic area. *Cancer Imaging* 25 (1), 39. doi: 10.1186/s40644-025-00853-5

Gilad, Y., Hemo, R., Micali, S., Vlachos, G., and Zeldovich, N. (2017). Algorand: scaling byzantine agreements for cryptocurrencies. In: *Proceedings of the 26th Symposium on Operating Systems Principles*. Shanghai China: ACM; 51–68. Available online at: https://dl.acm.org/doi/10.1145/3132747.3132757 (Accessed February 11, 2026).

Goisauf, M., Cano Abadía, M., Akyüz, K., Bobowicz, M., Buyx, A., Colussi, I., et al. (2025). Trust, trustworthiness, and the future of medical AI: outcomes of an interdisciplinary expert workshop. *J. Med. Internet Res.* 27, e71236. doi: 10.2196/71236

International Organization for Standardization/International Electrotechnical Commission (2023). *Information Technology — Artificial Intelligence — Management System*. Geneva, Switzerland: ISO. Available online at: https://www.iso.org/standard/42001 (Accessed April 20, 2026).

International Organization for Standardization, International Electrotechnical Commission (2020). "Information technology — artificial intelligence — overview of trustworthiness in artificial intelligence,". Geneva. Available online at: https://www.iso.org/standard/77608.html (Accessed April 20, 2026).

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM; 624–635. doi: 10.1145/3442188.3445923

Jahanifar, M., Raza, M., Xu, K., Vuong, T. T. L., Jewsbury, R., Shephard, A., et al. (2025). Domain generalization in computational pathology: survey and guidelines. *ACM Comput. Surv.* 57 (11), 1–37. doi: 10.1145/3724391

Jin, H., Liu, S., Cong, C., Feng, Q., Liu, Y., Huang, L., et al. (2026). FedWSIDD: Federated whole slide image classification via dataset distillation. In: J. C. Gee, D. C. Alexander, J. Hong, J. E. Iglesias, C. H. Sudre, A. Venkataraman, et al. editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Cham: Springer Nature Switzerland; 178–188. doi: 10.1007/978-3-032-05185-1_18

Joshi, G., Jain, A., Araveeti, S. R., Adhikari, S., Garg, H., and Bhandari, M. (2024). FDA-approved artificial intelligence and machine learning (AI/ML)-Enabled medical devices: an updated landscape. *Electronics* 13 (3), 498. doi: 10.3390/electronics13030498

Kassem, H., Banks, O., Cansiz, S., Edwards, B., Foley, P., Hagestedt, I., et al. (2025). A technical policy blueprint for trustworthy decentralized AI. *arXiv*. doi: 10.48550/ARXIV.2512.11878

Kurt, K. K., Timurtaş, M., Pınar, S., Ozaydin, F., and Türkeli, S. (2025). Smart contracts, blockchain, and health policies: past, present, and future. *Information* 16 (10), 853. doi: 10.3390/info16100853

Lan, Y. C., Strauch, M., Pilva, P., Schmitz, N. E. J., Sadr, A. V., Niggemeier, L., et al. (2024). Ecologically sustainable benchmarking of AI models for histopathology. *Npj Digit. Med.* 7 (1), 378. doi: 10.1038/s41746-024-01397-x

Latif, R. M. A., Ullah, F., Raza, U., Mostarda, L., and Cacciagrano, D. (2026). Blockchain-enabled ethical governance framework for transparent, accountable, and trustworthy AI systems: from transparency to ethical AI governance. In: A. Y. Yasin, F. Ullah, and R. Fatima, editors. *Advances in Computational Intelligence and Robotics*. IGI Global Scientific Publishing; 93–122. doi: 10.4018/979-8-3373-6038-6.ch004

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y (2016). Communication-efficient learning of deep networks from decentralized data. Available online at: https://arxiv.org/abs/1602.05629 (Accessed February 11, 2026).

Miwa, K., Yamao, T., Hashimoto, F., Miyaji, N., Kamitaka, Y., Masubuchi, M., et al. (2025). Innovations in clinical PET image reconstruction: advances in bayesian penalized likelihood algorithm and deep learning. *Ann. Nucl. Med.* 39 (9), 875–898. doi: 10.1007/s12149-025-02088-7

Montes, A. L., Yousefirizi, F., Chen, Y., Salimi, Y., Seifert, R., Afshar-Oromieh, A., et al. (2026). Artificial intelligence for simplified patient-centered dosimetry in radiopharmaceutical therapies. *Pet. Clin.* 21 (1), 73–88. doi: 10.1016/j.cpet.2025.09.010

Nouis, S. C., Uren, V., and Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Med. Ethics* 26 (1), 89. doi: 10.1186/s12910-025-01243-z

Oyeniyi, J. G. (2025). From lab to clinic: addressing bias and generalizability in AI diagnostic systems. *World J. Adv. Res. Rev.* 28 (3), 2134–2179. doi: 10.30574/wjarr.2025.28.3.4249

Prajwal, R., Pawan, S. J., Nazarian, S., Heller, N., Weight, C. J., Duddalwar, V., et al. (2025). A study on energy consumption in AI-Driven medical image segmentation. *J. Imaging* 11 (6), 174. doi: 10.3390/jimaging11060174

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM; 33–44. Available online at: https://dl.acm.org/doi/10.1145/3351095.3372873 (Accessed February 11, 2026).

Riaz, U., Talha, M., Ali, U., Hassan, M., and Mahato, R. K. (2026). Blockchain integration for rare disease clinical trials. *Ann. Med. Surg. (Lond).* 88 (1), 1134–1135. doi: 10.1097/ms9.0000000000004542

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., and Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* 27 (12), 2176–2182. doi: 10.1038/s41591-021-01595-0

Shahsavari, Y., Baseri, Y., Hafid, A., Dambri, O. A., and Makrakis, D. (2026). Integration of federated learning and blockchain in healthcare: a tutorial on medical data, architectures, privacy, security, and regulatory compliance. *J. Med. Internet Res.* 28, e80178. doi: 10.2196/80178

Sharma, S., Petrovic, G., and Kaushik, S. (2025). zkFL-Health: blockchain-enabled zero-knowledge federated learning for medical AI privacy. *arXiv*. doi: 10.48550/arXiv.2512.21048

South, T., Camuto, A., Jain, S., Nguyen, S., Mahari, R., Paquin, C., et al. (2024). Verifiable evaluations of machine learning models using zkSNARKs. *arXiv*. doi: 10.48550/ARXIV.2402.02675

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics; 3645–3650. Available online at: https://aclanthology.org/P19-1355 (Accessed February 11, 2026).

Tabassi, E. (2023). "Artificial intelligence risk management framework (AI RMF 1.0),". Gaithersburg, MD: National Institute of Standards and Technology U.S. doi: 10.6028/NIST.AI.100-1

Tsui, B., Calabrese, E., Zaharchuk, G., and Rauschecker, A. M. (2024). Reducing gadolinium contrast with artificial intelligence. *Magn. Reson. Imaging* 60 (3), 848–859. doi: 10.1002/jmri.29095

Vafaei Sadr, A., Bülow, R., Von Stillfried, S., Schmitz, N. E. J., Pilva, P., Hölscher, D. L., et al. (2024). Operational greenhouse-gas emissions of deep learning in digital pathology: a modelling study. *Lancet Digital Health* 6 (1), e58–e69. doi: 10.1016/S2589-7500(23)00219-4

Yang, T., Yu, X., McKeown, M. J., and Wang, Z. J. (2024). When federated learning meets medical image analysis: a systematic review with challenges and solutions. *APSIPA Trans. Signal Inf. Process.* 13 (1), 1–55. doi: 10.1561/116.20240048

Zamani, M., Movahedi, M., and Raykova, M. (2018). RapidChain: scaling blockchain via full sharding. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. Toronto Canada: ACM; 931–948. doi: 10.1145/3243734.3243853

Zhang, Y., Sheng, R., Qian, X., Wang, H., Wu, F., Dai, H., et al. (2025). Deep learning empowered gadolinium-free contrast-enhanced abbreviated MRI for diagnosing hepatocellular carcinoma. *JHEP Rep.* 7 (5), 101392. doi: 10.1016/j.jhepr.2025.101392

Zhou, B., Xie, H., Liu, Q., Chen, X., Guo, X., Feng, Z., et al. (2023). FedFTN: personalized federated learning with deep feature transformation network for multi-institutional low-count PET denoising. *Med. Image Anal.* 90, 102993. doi: 10.1016/j.media.2023.102993