Digital Twin-Driven Risk Assessment and Security Optimization of Large Language Model Agents in Personalized Healthcare

Mikkel Bell
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
bellmikkel@buffalo.edu

Gerald L. Becker
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
geraldmail@ucf.edu

Clifford Perry
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
cliffordmail@unr.edu

Abstract

The integration of large language model agents into personalized healthcare promises transformative improvements in clinical decision support, patient engagement, and treatment customization. However, these autonomous software entities introduce unprecedented risks stemming from adversarial manipulation, data biases, and systemic failures that can compromise patient safety and privacy. This paper presents a novel framework that leverages digital twin technology to systematically assess and optimize the security profile of LLM agents operating in healthcare environments. The digital twin constructs a high-fidelity virtual replica of the agent, its interaction context, and the underlying care ecosystem, enabling continuous simulation of threat scenarios without endangering real patients. Through a dual-loop architecture, risk assessment outputs inform proactive security optimization, while the optimized configurations are fed back into the digital twin for validation. We examine the structural trade-offs between fidelity, computational cost, and real-time responsiveness, and discuss the architectural requirements for integrating digital twin pipelines with clinical data infrastructures. The governance implications of such simulation-driven risk management are analyzed, including questions of regulatory accountability, cross-institutional data sharing, and algorithmic fairness. By situating the discussion at the system level, we avoid narrow technical fixes and instead advocate for a holistic socio-technical approach that treats LLM agent security as an emergent property of continuous, evidence-based co-adaptation between the digital and physical realms. The paper concludes with a forward-looking perspective on how federated digital twin ecosystems could underpin a new class of resilient healthcare AI infrastructures, balancing innovation with rigorous safety guarantees.

Keywords

digital twin; large language model agents; risk assessment; security optimization; personalized healthcare; adversarial robustness; system-level governance.

1. Introduction

The rapid advancement of large language models has catalyzed a paradigm shift in the design of intelligent software agents capable of performing complex, context-aware tasks in open-ended environments. In the domain of personalized healthcare, such agents are increasingly being positioned as mediators between massive clinical knowledge bases, real-time patient data streams, and the nuanced decision-making needs of clinicians and patients [1, 2]. An LLM agent can, for instance, ingest a patient's entire electronic health record, reason over the latest medical literature, and generate a tailored treatment recommendation that accounts for genetic markers, lifestyle factors, and drug interaction risks. The seductive power of this vision, however, masks a deep and expanding risk surface. Unlike traditional deterministic clinical decision support systems, LLM agents are fundamentally probabilistic and learn their behaviors from vast corpora that are impossible to fully audit. Their chain-of-thought reasoning processes remain partially opaque, and their interaction with external tools and data sources creates multiple injection points for adversarial interference [3, 4]. A manipulated prompt, a subtly poisoned knowledge graph, or a crafted set of patient observations could steer an agent toward harmful conclusions without triggering conventional anomaly detection mechanisms. As healthcare moves toward increasingly autonomous agent-based workflows, the question of how to rigorously assess and continuously optimize the security of these agents becomes not merely a technical concern but an urgent matter of patient safety and public trust.

Traditional software security approaches, predicated on clearly specified requirements and testable invariants, are largely inadequate for LLM agents whose behavior emerges from statistical training. Penetration testing and red-teaming exercises offer valuable point-in-time snapshots but cannot capture the dynamic evolution of threats or the cascading effects of agent decisions across longitudinal care pathways [5]. Furthermore, proposing security fixes after an agent has been deployed in a live clinical setting is ethically fraught and operationally disruptive. To address this gap, we introduce the concept of a digital twin-driven risk assessment and security optimization framework. A digital twin is a living virtual model that mirrors a physical system with sufficient fidelity to support experimentation, prediction, and continuous improvement. Originally developed in manufacturing and aerospace, digital twin technology has more recently been adopted in healthcare for equipment monitoring, surgical planning, and hospital workflow optimization [6, 7]. Extending this paradigm to LLM agents involves constructing a high-resolution replica of the agent's cognitive architecture, its operational environment, and the patient profiles it is likely to encounter. Within this sandbox, we can systematically inject adversarial perturbations, simulate rare but catastrophic failure modes, and measure the resulting clinical impact in a risk-free manner. The outputs of this simulation loop inform a security optimization engine that adjusts model parameters, prompt templates, guardrail configurations, and orchestration policies to harden the agent against identified weaknesses.

This paper is organized as follows. Section 2 delineates the architectural foundations of LLM agents in personalized healthcare, highlighting the modular components and integration points that define their vulnerability surface. Section 3 develops the digital twin paradigm for risk assessment, detailing the twin's constitutive models, data requirements, and simulation orchestration. Section 4 surveys the security threat landscape, categorizing adversarial attacks and systemic failure modes specific to healthcare agent deployments. Section 5 presents the dual-loop integration of risk assessment and security optimization, discussing feedback mechanisms, trade-off surfaces, and validation strategies. Section 6 expands the analysis to system-level governance, exploring regulatory alignment, fairness, sustainability, and the

policy infrastructure needed to realize such frameworks at scale. A concluding section synthesizes the contributions and charts directions for future research.

2. Architectural Foundations of LLM Agents in Personalized Healthcare

An LLM agent in healthcare is not a monolithic model but a composite system comprising a core language model, a prompt orchestration layer, a memory module, and a suite of tool-use interfaces that connect to external databases, clinical calculators, and communication channels. The core model is typically a transformer-based neural network pre-trained on web-scale corpora and then fine-tuned on domain-specific medical texts, clinical guidelines, and de-identified patient records [8]. The prompt orchestration layer manages the assembly of context windows from diverse sources: the patient's current query, retrieved guideline snippets, recent lab results, and the agent's internal chain-of-thought history. Memory modules enable the agent to maintain longitudinal awareness of a patient's evolving condition, often implemented via vector databases or structured summaries that persist across sessions. Tool-use interfaces allow the agent to execute API calls to drug interaction checkers, dosage calculators, and scheduling systems, grounding abstract reasoning in verifiable computational outputs.

This modular architecture, while enabling flexibility and extensibility, creates a fragmented security perimeter. Attacks can target the language model's inherent brittleness through adversarial prompts that exploit its tendency for sycophantic or hallucinatory completions [4]. Alternatively, they can poison the retrieval corpus by injecting fraudulent medical articles into the knowledge base, thereby biasing the agent's evidence synthesis. The tool-use interfaces represent another critical attack surface: a compromised external service could return manipulated values, leading the agent to recommend a toxic dosage. Moreover, the memory module, which stores sensitive patient histories and reasoning traces, becomes a high-value target for data exfiltration or unauthorized manipulation. Any security framework must therefore account for the interplay between these components rather than treating them in isolation.

The operational context of personalized healthcare adds further complexity. Agents are expected to adapt to individual patient variability in genetics, comorbidities, and social determinants of health. This personalization imperative often requires the agent to access fine-grained, identifiable data, increasing the privacy stakes and regulatory scrutiny under frameworks such as HIPAA and GDPR. The temporal dimension is equally critical: an agent that performs safely for a single consultation may still introduce cumulative risks over weeks or months of continuous care, as small biases compound or as the agent's internal state drifts. Capturing these longitudinal dynamics is essential for any realistic risk assessment, and this is precisely where the digital twin paradigm offers unique leverage.

3. Digital Twin Paradigm for Risk Assessment

A digital twin for an LLM agent in healthcare is a multiscale virtual construct that replicates the agent's behavioral dynamics, its interaction environment, and the simulated patient population. The twin comprises three interconnected sub-models. The agent behavioral model captures the input-output mapping of the LLM, including its prompt sensitivity, reasoning pathways, and tool-use patterns. This is not a simple replay of the original model weights but a statistical surrogate that can be efficiently queried and perturbed. Techniques such as knowledge distillation, behavioral cloning, and Monte Carlo tree search over token likelihoods allow the twin to approximate the agent's decision distribution while remaining

computationally manageable for large-scale simulation [9, 10]. The environment model simulates the clinical IT ecosystem: the electronic health record system, the messaging platforms, the knowledge retrieval APIs, and any device interfaces. Crucially, this environment can be instrumented with configurable failure modes, such as latency spikes, data corruption, or malicious third-party sensor feeds. The patient model generates a diverse synthetic cohort spanning demographic strata, disease trajectories, and psychosocial profiles, ensuring that risk assessment covers edge cases and underrepresented groups [11]. By modulating cohort composition, analysts can measure differential security impacts across subpopulations, a fundamental fairness requirement.

The digital twin operates in a continuous simulation loop. Scenarios are constructed by combining patient trajectories with threat vectors drawn from a library of known and hypothesized attacks. For each scenario, the twin records the agent's actions, the clinical outcomes as judged by embedded outcome evaluators, and any violations of predefined safety constraints. The outcome evaluators themselves can be LLM-based judges, structured clinical rule engines, or hybrid systems that compare the agent's recommendations against a panel of expert-curated gold standards. The aggregation of thousands of such simulations produces a risk profile that quantifies not only the frequency of harmful events but also their severity, reversibility, and propagation patterns across the care continuum.

A major structural trade-off in digital twin design is between fidelity and tractability. A perfect replica of the agent, running the full model at production scale, would yield the most accurate risk estimates but would be prohibitively expensive to simulate across the combinatorial space of patients and attacks. Conversely, overly simplified surrogates may miss subtle vulnerabilities that arise from the agent's higher-order reasoning. A pragmatic resolution involves tiered simulation: low-fidelity twins are used for rapid screening of attack vectors, while high-fidelity twins are invoked for detailed investigation of flagged risks. This tiered architecture requires careful orchestration and governance to ensure that the screening filters do not introduce systematic blind spots.

Data provenance presents an additional challenge. Building a realistic patient model demands access to high-fidelity clinical data, which is sensitive and often siloed across institutions. Federated learning and synthetic data generation algorithms, such as generative adversarial networks trained on distributed datasets, can mitigate privacy concerns by ensuring that no raw patient data leaves its originating institution [12]. The digital twin, in such a federated setup, becomes a constellation of institution-specific replicas whose aggregated insights inform a shared security model. The governance of such federated digital twin networks raises profound questions about data sovereignty, liability, and the harmonization of risk assessment standards, which we revisit in Section 6.

4. Security Threat Landscape and Adversarial Vulnerabilities

The threat landscape for LLM agents in personalized healthcare is characterized by a fusion of classical AI security concerns and domain-specific clinical hazards. We can taxonomize threats into three broad categories: direct prompt injection and adversarial input attacks, data and knowledge poisoning, and systemic multi-agent failures. Direct prompt injection attacks manipulate the agent's context window by embedding malicious instructions within user messages or retrieved documents [3]. In a healthcare setting, a patient might inadvertently or deliberately include a hidden instruction in a symptom description that causes the agent to downplay a serious condition or recommend an unnecessary procedure. More sophisticated variants exploit the agent's tool-use patterns, crafting prompts that trigger an API call with

crafted parameters, leading to data leakage or unauthorized actions. Adversarial input attacks, on the other hand, introduce imperceptible perturbations to clinical texts or lab values that shift the model's internal representations, steering its reasoning toward attacker-chosen conclusions. The robustness of medical LLMs against such perturbations has been shown to vary significantly across demographic groups, raising alarming equity concerns [13].

Data and knowledge poisoning attacks target the corpora and databases on which the agent relies. If an adversary can inject fraudulent case reports or meta-analyses into a widely used medical literature repository, the agent's retrieval-augmented generation pipeline may systematically incorporate these sources, biasing treatment recommendations. The long-tailed nature of medical knowledge exacerbates this risk: rare diseases or emerging treatments are documented in relatively few sources, making it easier for a small number of poisoned articles to dominate the retrieved evidence set. Moreover, the feedback loops inherent in continuous learning systems can amplify initial biases. If the agent's recommendations influence clinical practice and those outcomes are subsequently fed back into training data, the poison becomes self-reinforcing, a dynamic that digital twins are uniquely positioned to simulate before it manifests in the real world.

Systemic multi-agent failures arise when multiple LLM agents, or combinations of human and AI decision-makers, interact in ways that produce emergent pathology. For instance, a prescribing agent might recommend a drug that interacts harmfully with a supplement recommended by a separate wellness agent, with neither agent having full visibility into the other's reasoning. The lack of a coordinating super-agent or a shared safety ledger creates gaps that attackers can exploit by manipulating one agent to indirectly compromise another. These systemic risks cannot be assessed by examining single agents in isolation; they demand an environment simulation that captures the web of interdependent decisions, a requirement for which the digital twin paradigm is inherently well-suited.

In addition to adversarial threats, non-malicious failure modes such as model drift, data staleness, and prompt template degradation over time pose significant security challenges. A model that was rigorously tested at deployment may, after months of interaction, develop idiosyncratic behaviors due to distribution shifts in patient queries or updates to underlying clinical knowledge. Continuous risk assessment through a living digital twin enables the detection of such drift before it translates into patient harm, effectively turning security from a pre-deployment checkpoint into an ongoing operational practice.

5. Security Optimization through Dual-Loop Digital Twin Integration

The separation of risk assessment and security optimization into distinct yet tightly coupled loops is a defining architectural choice of our framework. The inner loop runs at high frequency within the digital twin environment, iteratively applying threat scenarios, measuring risks, and adjusting security parameters. The outer loop operates at a lower cadence, synchronizing the optimized configuration with the physical agent deployment and feeding real-world telemetry back into the twin to correct for simulation-reality gaps. This dual-loop structure draws inspiration from model predictive control and continuous integration paradigms in safety-critical software engineering [14].

Within the inner loop, a library of candidate security interventions is maintained. These interventions span multiple layers of the agent stack: prompt-level guardrails that detect and neutralize injection attempts, retrieval filters that assess the credibility and provenance of fetched documents, output validators that cross-check medication dosages against established

formularies, and orchestration policies that limit the agent's autonomy for high-risk decisions. When the risk assessment engine detects an elevated probability of harm for a particular patient-threat combination, the optimization engine selects and parameterizes a subset of these interventions to minimize a composite cost function that balances safety, clinical efficacy, and user experience. Because exhaustive combinatorial search is infeasible, we rely on surrogate-based optimization, where the digital twin's behavioral model serves as the surrogate, enabling rapid evaluation of intervention candidates. This process may be further accelerated by Bayesian optimization or evolutionary strategies that exploit the structure of the intervention space.

A crucial aspect of the optimization loop is the preservation of clinical utility. Hardening an agent to the point where it refuses all ambiguous requests or escalates every decision to a human clinician would negate the efficiency gains that motivated its deployment. The digital twin thus facilitates a multi-objective optimization that surfaces the Pareto frontier between security and clinical effectiveness, allowing system designers and regulators to make explicit, evidence-based trade-offs. The digital twin can also simulate the patient acceptance and trust implications of different security postures. For example, an agent that frequently interrupts its workflow to request authentication or confirm consent might frustrate patients, reducing adherence and potentially leading to worse outcomes [1]. These socio-behavioral variables are incorporated into the digital twin's patient model, ensuring that security optimization is not pursued in a vacuum.

The outer loop addresses the critical challenge of sim-to-real transfer. No digital twin is a perfect replica; discrepancies between the simulated patient responses, the actual behavior of live LLM inference servers, and the ever-changing real-world threat landscape will inevitably emerge. Continuous monitoring of the deployed agent generates telemetry streams including decision logs, clinician override frequencies, and patient outcome data. These streams are anonymized, aggregated, and compared against the digital twin's predictions. Significant gaps trigger a model reconciliation process in which the twin's behavioral and environment models are updated, and previously optimized security configurations are re-evaluated. This closes the loop, creating a cybernetic organism in which security is perpetually co-evolved with the system's operational reality. The outer loop also serves a vital governance function by providing an audit trail that demonstrates due diligence to regulators, showing that the agent's safety case is supported by continuous, empirically validated simulation evidence rather than a static certification artifact.

6. System-Level Governance and Policy Implications

The introduction of digital twin-driven risk assessment for healthcare LLM agents extends beyond technical architecture into the realms of law, policy, and institutional design. One immediate governance challenge is the determination of liability when an agent whose security was optimized via a digital twin nevertheless causes patient harm. Is the liability borne by the healthcare provider, the digital twin platform operator, the developer of the LLM, or some combination thereof? The digital twin's simulation logs and the outer loop's audit trail could become crucial evidence in legal proceedings, but this raises the stakes on the trustworthiness and non-repudiability of the twin's data infrastructure. Standards for the provenance, integrity, and retention of simulation data will need to be developed in concert with medical device regulators and data protection authorities.

Fairness constitutes a second governance imperative. Digital twins that are trained predominantly on data from well-resourced academic medical centers may systematically

underestimate risks for rural, low-income, or minority populations [11]. If the security optimization loop is then tuned to minimize aggregate harm, it could embed protective measures that are disproportionately responsive to the majority profile while leaving vulnerable groups exposed. Mitigating this requires regulatory mandates for stratified risk reporting, where digital twin outputs are disaggregated by demographic and socioeconomic segments. It also implies a need for independent auditing bodies with the technical capacity to inspect digital twin models, challenge their assumptions, and verify their claims, akin to the role of notified bodies in medical device regulation.

The federated operation of digital twins across multiple healthcare institutions introduces cross-border data governance complexities. While federated learning and differential privacy techniques can reduce the flow of raw patient data, the model gradients, risk profiles, and security configurations that are exchanged still encode sensitive information about institutional practices and patient populations. Trade secret concerns may inhibit the sharing of vulnerability data, leaving each institution to independently rediscover threats that others have already encountered. A trusted third-party coordination mechanism, possibly instantiated as a non-profit consortium or a government-chartered health AI safety institute, could facilitate the pooling of threat intelligence while enforcing strict data minimization and anonymization protocols. The establishment of shared digital twin benchmarks and challenge datasets, analogous to those used in cybersecurity, would further accelerate collective learning.

Sustainability is an often-overlooked dimension of large-scale digital twin infrastructure. Running high-fidelity simulations of LLM agents at the scale of a national health system consumes significant computational resources, with associated energy and carbon costs. The tiered simulation architecture described in Section 3 offers a partial mitigation, but system designers must also weigh the environmental footprint of continuous simulation against its safety benefits. Policy instruments such as mandatory environmental impact disclosures for AI systems could incentivize the development of more efficient surrogate models and hardware-accelerated simulation platforms. In the long term, advances in neuromorphic computing and sparse model architectures may shift this calculus favorably.

Finally, the broader societal implications of normalizing continuous digital surveillance of AI agents merit careful reflection. The same telemetry streams that enable the outer loop's sim-to-real correction could be repurposed for workforce performance monitoring, insurance underwriting, or other forms of algorithmic management that erode clinician and patient autonomy. Governance frameworks must therefore enshrine principles of purpose limitation, data minimization, and meaningful human oversight. The goal is not to create a panopticon of AI surveillance but to cultivate a culture of responsible autonomy where digital twins serve as trusted co-pilots in the ongoing journey toward safer, more equitable AI-augmented healthcare.

7. Conclusion

This paper has argued that the security of LLM agents in personalized healthcare cannot be assured through static, pre-deployment evaluation alone but requires a dynamic, simulation-centric approach rooted in digital twin technology. By constructing high-fidelity virtual replicas that encompass the agent, its environment, and diverse patient populations, we can systematically explore the risk landscape, identify emergent vulnerabilities, and optimize protective measures within a safe sandbox. The dual-loop architecture we propose bridges the gap between simulation and reality, allowing security configurations to co-evolve with real-world feedback. We have emphasized that the structural choices embedded in such systems,

around fidelity versus tractability, centralized versus federated operation, and single-objective versus multi-objective optimization, are not merely engineering decisions but carry profound implications for fairness, accountability, and sustainability. The governance challenges are formidable and demand a coordinated response from technologists, clinicians, regulators, and patient communities. As the healthcare industry accelerates its adoption of LLM agents for tasks ranging from triage to chronic disease management, the frameworks and principles outlined here offer a pathway toward a future in which innovation and safety advance hand in hand, guided by continuous evidence and grounded in a deep respect for the dignity and diversity of those whose health is at stake.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233–1239.

3. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79–90).

4. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043.

5. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

6. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In Transdisciplinary Perspectives on Complex Systems (pp. 85–113). Springer.

7. Corral-Acero, J., Margara, F., Marciniak, M., Rodero, C., Loncaric, F., Feng, Y., ... & Lamata, P. (2020). The 'Digital Twin' to enable the vision of precision cardiology. European Heart Journal, 41(48), 4556–4564.

8. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

9. Takagi, S., & Kato, T. (2021). Surrogate modeling for high-fidelity simulation: A review. Structural and Multidisciplinary Optimization, 64(5), 2689–2717.

10. Yao, S., Yu, D., Zhao, J., Shafto, I., Griffiths, T. L., Huang, T., & Zhu, S. C. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

11. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

12. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1–7.

13. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

14. Schirmer, G., & Denil, M. (2019). Sim-to-real transfer for robotics: A survey. arXiv preprint arXiv:1909.11013.

15. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

16. Carlini, N., & Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 1–7). IEEE.

17. Chowdhury, G. G. (2021). A review of digital twins in healthcare: Towards an integrated approach. Health Informatics Journal, 27(3), 14604582211043159.

18. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

19. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.

20. Mökander, J., & Floridi, L. (2022). From algorithmic accountability to digital due process: The case for a Digital Twin regulatory sandbox. Philosophy & Technology, 35(3), 1–20.

21. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

22. Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980–22012.

23. Wu, C., Wu, W., & Pan, Y. (2022). Security and privacy of digital twin: A survey. IEEE Internet of Things Journal, 9(19), 18397–18412.

24. Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Santamaría, J. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion, 96, 156–191.

25. Schneider, J., & Breitinger, F. (2023). AI security in healthcare: Understanding the threat landscape. ACM Computing Surveys, 55(12), 1–32.