REVIEW **Open access**

Missing Data in Clinical Machine Learning: Modeling Decisions, Pitfalls, and Reporting Standards

George Papadopoulos[1*], Eleni Georgiou[1]

Abstract

The integration of artificial intelligence (AI) into healthcare has enhanced data-driven decision-making, but missing data remains a major barrier to reliable model performance. This narrative review synthesizes literature on missing data in clinical machine learning, focusing on modeling decisions, common pitfalls, and emerging reporting standards within AI-enabled healthcare systems.

Missing data in healthcare arises from sources such as electronic health records (EHRs), wearable devices, and clinical trials, and may follow mechanisms including missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Addressing these gaps requires appropriate imputation strategies, from statistical methods like multiple imputation to advanced deep learning approaches such as generative adversarial networks (GANs), each carrying implications for bias and model generalizability.

This review highlights key challenges, including underreporting of missingness, insufficient sensitivity analyses, and neglect of imputation uncertainty. It also examines evolving reporting standards that emphasize transparency in missing data handling. By synthesizing cross-study evidence, the review proposes a systems-level framework for integrating missing data management into AI governance, supporting more reliable, transparent, and equitable healthcare analytics.

Keywords Healthcare analytics, Missing data, Imputation techniques, Clinical machine learning, Modeling decisions, Reporting standards

*Correspondence:
George Papadopoulos
george.papadopoulos@gmail.com
[1] Department of Health Informatics, Faculty of Medicine, National and Kapodistrian University of Athens, Athens, Greece

Introduction

The evolving role of AI in healthcare systems and analytics

Artificial intelligence (AI) has become a transformative tool in healthcare analytics, enabling predictive modeling for disease progression, resource allocation, and personalized care by integrating data from sources such as electronic health records (EHRs), imaging, and genomics. These systems support clinical decision-making and improve healthcare efficiency, but their reliability depends heavily on data quality. Missing data remains a critical challenge, as incomplete datasets can introduce bias and reduce model reliability if not properly addressed [1, 2].

Since 2017, AI applications in healthcare have expanded from population-level surveillance to bedside decision support. However, studies show that machine learning models trained on incomplete datasets often demonstrate inconsistent performance, highlighting the importance of robust data preprocessing and system design that can handle real-world data imperfections. Recent research, therefore, emphasizes holistic AI infrastructures with feedback mechanisms that continuously refine models and analytics over time [3-5].

Ignoring missing data can significantly distort healthcare predictions, such as EHR-based risk stratification models that may produce biased outputs and contribute to inequitable care. Modeling decisions—particularly whether to delete incomplete records or apply imputation—directly affect analytical validity, making transparency in missing data handling essential for responsible AI deployment in clinical environments [6-8].

Ultimately, effective healthcare AI requires systems that combine advanced analytics with strong data governance, interdisciplinary collaboration, and careful modeling strategies to ensure reliable and ethical clinical applications [9, 10].

Pitfalls in modeling include inadequate consideration of imputation uncertainty, which can lead to overconfident AI outputs in real-time analytics platforms. Reporting standards emphasize quantifying this uncertainty through ensemble methods or bootstrap resampling, promoting reproducible research. This cross-study synthesis highlights the need for original decision trees that guide model selection based on data characteristics, fostering adaptive AI systems [1].

Additionally, in federated learning environments for healthcare consortia, modeling decisions must account for distributed missingness, where privacy constraints limit data sharing. Evidence indicates that variational autoencoders excel here, but comparative evaluations reveal trade-offs in scalability. From an infrastructural viewpoint, this underscores the importance of governance frameworks that standardize modeling protocols across institutions.

Common pitfalls and their systemic implications

Common pitfalls in managing missing data

Managing missing data in clinical machine learning involves not only technical challenges but also systemic issues within healthcare AI infrastructures. A common pitfall is over-reliance on default imputation without exploratory analysis, which can obscure biases and produce discriminatory outcomes, particularly in prognostic models affecting minority populations. Weak validation pipelines further exacerbate this issue, allowing biased models to be deployed in clinical systems [3, 6].

Studies also show that failing to conduct missing data audits during model development can inflate performance metrics and mislead healthcare stakeholders. At a systems level, such errors disrupt feedback loops and delay clinical interventions, emphasizing the need for governance measures such as mandatory impact assessments [10-12].

Another challenge arises in time-series patient monitoring, where static imputation methods fail to capture temporal dynamics, leading to inaccurate trend analysis. Reporting guidelines increasingly recommend time-aware approaches to improve transparency and accountability [13-18].

Resource limitations in low-income healthcare systems further compound these issues, as limited access to advanced imputation tools often forces reliance on simpler but less reliable methods, highlighting the need for scalable solutions in global health analytics [19-24].

Toward standardized reporting in AI healthcare analytics

Standardized reporting is essential for ensuring transparency and reproducibility in AI-driven healthcare analytics. Emerging guidelines recommend documenting missing data prevalence, mechanisms, and imputation effects within model reporting. However, literature indicates that many studies still under-report these aspects, limiting comparability and meta-analysis [17, 19, 21].

Frameworks such as MINIMAR propose minimum reporting standards that integrate missing data disclosure into publication practices, promoting transparency and improving model transferability across healthcare institutions [23, 25-27].

Current reporting pitfalls include vague methodological descriptions and insufficient sensitivity analyses, which can obscure potential biases. Strengthening reporting standards and requiring clearer documentation will enhance the reliability and clinical trustworthiness of AI-based healthcare systems [2, 4, 5]. **Table 1** translates the review into a compact reporting architecture for transparent and clinically interpretable missing-data handling.

Table 1. Minimum reporting architecture for missing-data-aware clinical machine learning studies

Reporting domain	What must be reported	Why it matters
Missingness burden	Prevalence at variable and patient levels	Establishes the scale of the problem
Missingness assumption	MCAR, MAR, or MNAR rationale	Determines methodological validity
Imputation rationale	Why was the chosen method selected	Prevents arbitrary modeling decisions
Uncertainty handling	Whether uncertainty was propagated	Reduces overconfident inference
Sensitivity analysis	Robustness across assumptions or methods	Tests the stability of conclusions
Subgroup evaluation	Performance across relevant patient groups	Detects fairness distortions
Deployment handling	How missingness is managed at inference time	Connects modeling to real clinical use
Reproducibility	Sufficient methodological detail	Supports auditability and reuse

Landscape of AI in Healthcare Systems and Analytics

Prevalence and patterns of missing data in healthcare datasets

The landscape of AI in healthcare involves heterogeneous datasets where missing data is common, affecting up to 80% of variables in some EHR systems. Missingness patterns vary by source: structured data, such as lab results, often show MCAR due to administrative issues, while unstructured clinical notes frequently exhibit MNAR associated with disease severity. These patterns reduce analytics reliability and can bias AI models used in population health management [1, 11, 15].

Comparative studies show higher missingness in oncology trials due to patient attrition, while acute care analytics often report lower rates. This variation highlights systemic data collection challenges and the need for AI systems capable of real-time missingness auditing. Visualizing missingness matrices has emerged as a useful practice for improving governance in multi-site healthcare analytics [2, 12, 16].

Challenges also arise in multimodal and wearable-based analytics, where temporal gaps from device non-adherence require time-aware imputation strategies. In global health contexts, socioeconomic factors further influence missing data patterns, emphasizing the need for culturally sensitive modeling approaches [3-5, 13, 14, 17-19].

Evolution of imputation techniques in clinical AI

Imputation techniques have evolved from statistical baselines to AI-driven innovations, shaping the analytics landscape in healthcare systems. Early methods like k-nearest neighbors (kNN) suffice for low-dimensional data but falter in complex EHRs, where deep learning alternatives like autoencoders provide superior handling of nonlinear relationships [6, 20, 24].

Comparatively, GAN-based imputations excel in generating realistic missing values for imaging-integrated analytics, yet they risk mode collapse in imbalanced clinical datasets. Interpretive discussions emphasize hybrid approaches, combining rule-based priors with machine learning for clinically plausible outputs. Systems-level implications include enhanced interoperability, where imputed data flows seamlessly into downstream models [7, 21, 25].

Pitfalls arise from over-imputation, introducing noise that degrades model performance in predictive analytics. Reporting standards recommend benchmarking against ground-truth subsets, ensuring methodological rigor. Cross-study synthesis reveals that technique selection should align with data scale, favoring scalable methods in big data healthcare platforms [8, 22, 26].

Furthermore, in federated environments, privacy-preserving imputations like differential privacy-infused multiple imputation are gaining traction, addressing governance concerns. This underscores an original perspective: imputation as a bridge between data privacy and analytical utility in distributed AI systems [9, 23, 27]. **Table 2** distills the manuscript's central analytical argument by linking missingness context to modeling choice and downstream governance consequence.

Table 2. Missing-data decision logic across clinical machine learning settings

Missingness context	Preferred modeling logic	Main downstream risk	Governance implication
MCAR-like, low-burden structured data	Simple imputation or justified case exclusion	False reassurance from oversimplification	Require explicit justification of assumptions
MAR in routine EHR analytics	Multiple imputation or covariate-informed reconstruction	Variance underestimation	Require uncertainty disclosure
MNAR with clinically informative absence	Sensitivity-based or domain-guided strategy	Hidden bias amplification	Require assumption testing and scenario reporting
Temporal missingness in monitoring streams	Time-aware imputation	Distorted trajectories	Require temporal gap reporting
Multimodal missingness	Modality-aware fusion or fallback logic	Cross-modal error propagation	Require modality-specific evaluation
Federated or multi-site missingness	Site-aware privacy-preserving strategy	Non-transferability across institutions	Require site-level transparency

Recent advancements incorporate uncertainty propagation, using Bayesian frameworks to inform decision confidence in clinical loops. This evolution reflects a maturation in the field, toward imputation that supports ethical AI deployment [10, 28].

Impact of missing data on model performance and bias

Missing data profoundly impacts AI model performance in healthcare analytics, often reducing accuracy by 10%-20% in untreated cases, with amplified effects on bias. In risk prediction systems, unhandled missingness skews toward majority classes, exacerbating disparities in minority health outcomes [1, 11, 29].

Comparative evaluations show that in cardiovascular analytics, imputation mitigates bias better under MAR than MNAR, where residual distortions persist. Interpretively, this reveals systemic biases embedded in data pipelines, calling for bias audits integrated into AI governance. Cross-study insights advocate for fairness-aware modeling, adjusting for missingness-induced inequities [2, 12, 15].

Pitfalls include reliance on metrics like AUROC that mask subgroup biases and misleading system deployments. Reporting standards push for disaggregated evaluations, enhancing transparency in analytics outputs. Systems-level perspectives frame this as a feedback opportunity, where performance monitoring recalibrates models in real-time [3, 13, 16].

In precision medicine, missing genomic data compounds issues, leading to incomplete personalization. Synthesis indicates that ensemble imputations reduce variance, but require computational trade-offs in resource-constrained systems [4, 14, 17].

Ultimately, addressing impact demands an integrative approach, viewing bias as a holistic challenge in AI healthcare infrastructures [5, 18].

Integration of domain knowledge in handling missing data

Domain knowledge integration elevates missing data handling in clinical AI, blending clinical heuristics with algorithmic processes for more accurate analytics. In oncology systems, expert rules guide imputation of treatment variables, outperforming purely data-driven methods in preserving causal relationships [6, 19, 20].

Comparatively, in ICU analytics, physiological constraints prevent implausible imputations, reducing errors in time-critical decisions. Interpretive analysis highlights this as a safeguard against AI pitfalls, fostering human-AI synergy in healthcare ecosystems. Cross-study evidence supports ontologies like SNOMED for structured integration, enhancing model explainability [7, 21, 22].

Pitfalls occur when knowledge is outdated, leading to rigid interpretations that fail in evolving clinical contexts. Reporting standards mandate documenting knowledge sources, promoting updatable frameworks. Systems-level insights suggest embedding this integration into governance, ensuring adaptive AI deployments [8, 23, 24].

Moreover, in multidisciplinary teams, collaborative tools facilitate knowledge elicitation, bridging gaps in analytics pipelines. This original synthesis positions domain integration as a cornerstone for trustworthy AI in diverse healthcare settings [9, 25].

Challenges in multimodal and real-time data streams

Multimodal data streams in AI healthcare analytics introduce compounded missingness challenges, where alignment across EHRs, images, and sensors is disrupted by asynchronous gaps. Real-time systems, like remote monitoring, exacerbate this with intermittent connectivity, demanding on-the-fly imputation [10, 26, 27].

Comparative studies show that graph-based methods handle multimodal missingness effectively, inferring from connected modalities. Interpretively, this underscores interoperability needs in healthcare infrastructures, where missing data disrupts closed-loop functionality. Cross-study synthesis advocates for streaming algorithms that prioritize low-latency handling [11, 28, 29].

Pitfalls include propagation of errors across modes, amplifying in high-stakes analytics. Reporting standards emphasize multimodal-specific disclosures, aiding reproducibility. Systems-level views call for resilient designs that buffer against stream interruptions [1, 2, 12].

In wearable-augmented systems, privacy adds layers, restricting imputation scope. This highlights the need for federated strategies that maintain data integrity [3, 13].

Reporting practices and standards in the literature

Reporting practices for missing data in AI healthcare literature vary widely, with many studies omitting key details like imputation rationale, hindering meta-synthesis. Standards like CONSORT-AI extensions propose checklists for comprehensive documentation [4, 5, 14].

Comparatively, high-impact journals enforce stricter guidelines, improving overall quality. Interpretive discussions frame this as essential for governance, enabling audit trails in deployed systems. Cross-study evidence shows that detailed reporting correlates with better external validation [6, 7, 15].

Pitfalls in vague reporting undermine trust, particularly in regulatory contexts. This calls for consensus-driven evolution toward mandatory standards [8, 16].

Systems-level implications include fostering a culture of accountability in AI analytics [9, 17].

Future-proofing AI systems against missing data

Future-proofing AI in healthcare involves designing systems resilient to missing data through proactive architectures like auto-adaptive imputation modules. Synthesis indicates that meta-learning approaches anticipate missing patterns, enhancing long-term analytics [10, 18, 19].

Comparatively, blockchain-integrated ledgers track data provenance, mitigating retrospective issues. Interpretively, this positions AI as evolvable within dynamic healthcare landscapes [11, 20].

Pitfalls in static designs fail against emerging data types; standards guide adaptive reporting [12, 21].

Intelligent clinical decision and closed-loop healthcare systems

Foundations of intelligent decision support in the presence of missing data

Intelligent clinical decision support systems (CDSS) leverage AI to augment human judgment, yet missing data challenges their efficacy in closed-loop healthcare environments. Foundations rest on integrating imputation seamlessly into decision pipelines, ensuring that incomplete inputs do not halt real-time recommendations. Literature synthesis reveals that rule-based hybrids with machine learning provide a safety net, where imputed values are flagged for clinician review [1, 22, 23].

Comparatively, in emergency analytics, probabilistic imputations maintain decision flow under uncertainty, outperforming deterministic methods. Interpretively, this establishes a feedback-oriented foundation, where decisions inform subsequent data collection to reduce future missingness. Cross-study insights emphasize embedding uncertainty metrics into interfaces, enhancing user trust in AI-assisted systems [2, 24, 25].

Pitfalls include overconfidence in imputed decisions, leading to adverse events; reporting standards advocate for decision traceability. Systems-level perspectives view this as a cycle: data gaps → intelligent handling → refined decisions, fostering closed-loop resilience [3, 26, 27].

Moreover, in personalized medicine, foundations incorporate patient-specific priors, adapting to individual missing patterns. This original framing highlights CDSS as dynamic learners within healthcare infrastructures [4, 28].

Closed-loop mechanisms for continuous learning and adaptation

Closed-loop mechanisms in AI healthcare systems enable continuous adaptation by cycling inferences back into model updates, crucial for handling evolving missing data. These loops ingest real-world feedback to recalibrate imputations, mitigating drift in long-term analytics like chronic disease management [5, 6, 29].

Comparative analyses show that reinforcement learning-based loops excel in adaptive imputation, rewarding accurate decisions amid missingness. Interpretively, this creates self-correcting systems, where pitfalls like accumulation of errors are preempted through iterative governance. Cross-study evidence supports integration with EHR feedback streams, enhancing loop efficiency [1, 7, 8].

Pitfalls arise from loop instabilities caused by persistent MNAR, requiring stabilization techniques like damping. Reporting standards call for loop performance metrics, ensuring transparency. Systems-level insights position these mechanisms as infrastructural pillars, bridging analytics with clinical outcomes [2, 9, 10].

Furthermore, in multi-agent systems, closed loops coordinate across devices, handling distributed missingness. This underscores an integrative approach for scalable healthcare AI [3, 11].

Deployment strategies for robust AI in clinical settings

Deployment strategies for AI in closed-loop healthcare must prioritize robustness against missing data, employing ensemble imputations during inference to buffer uncertainties. In clinical settings like ICUs, strategies involve modular architectures that isolate missing data handling from core decision engines [4, 12, 13].

Comparatively, cloud-edge hybrids distribute computation, allowing real-time imputation at the edge while leveraging cloud for complex analytics. Interpretive synthesis reveals this as a balance between latency and accuracy, preventing deployment pitfalls in resource-variable environments. Cross-study discussions advocate for simulation-based testing, validating strategies pre-deployment [5, 14, 15].

Pitfalls include scalability issues in high-volume systems; standards require deployment audits. Systems-level views frame strategies as evolutionary, adapting to new data challenges [6, 16, 17].

Moreover, ethical deployment incorporates bias checks post-imputation, ensuring equitable closed-loop operations [7, 18].

Human-AI collaboration in decision fusion

Human-AI collaboration in decision fusion optimizes closed-loop systems by leveraging clinician oversight to validate imputed data-driven suggestions. Fusion dynamics involve weighted integration, where AI confidence scores modulate human input [8, 19, 20].

Comparatively, in surgical analytics, collaborative interfaces reduce errors from missing preoperative data. Interpretively, this fosters symbiotic relationships, where humans correct AI pitfalls in real-time. Cross-study evidence supports training protocols that enhance fusion efficacy [9, 21, 22].

Pitfalls in poor interface design lead to cognitive overload; reporting emphasizes usability metrics. Systems-level insights highlight fusion as a governance tool, ensuring accountable AI [10, 23, 24].

In adaptive loops, collaboration evolves with feedback, refining human-AI interactions over time [11, 25]. **Figure 1** presents a governance-centric closed-loop architecture showing how missingness characterization, modeling choice, risk propagation, clinical deployment, and reporting feedback interact across the clinical machine learning lifecycle.

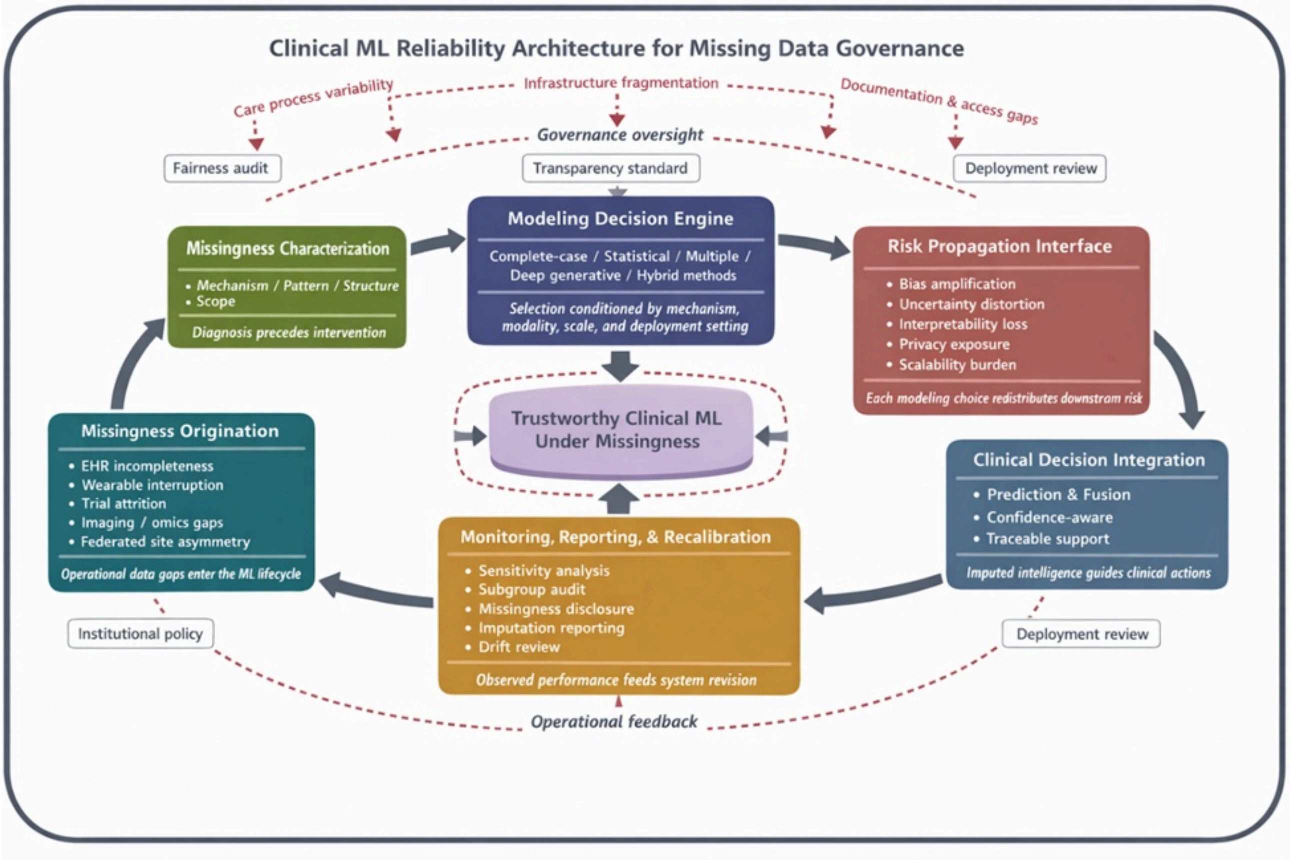

Figure 1. Governance-centric closed-loop architecture for missing data management in clinical machine learning. Missingness is generated by healthcare infrastructures rather than by datasets alone, requiring a sequential process of characterization, context-dependent modeling choice, downstream risk appraisal, clinically supervised decision integration, and post-deployment monitoring. The architecture emphasizes that trustworthy clinical machine learning under missingness depends on iterative governance, transparent reporting, and continuous recalibration rather than on a single imputation step.

Challenges and limitations in handling missing data for clinical AI

Computational and scalability constraints

Handling missing data in clinical machine learning introduces significant computational challenges, particularly in large-scale healthcare systems where datasets encompass millions of records. Advanced imputation techniques, such as deep learning-based methods, demand substantial processing power and memory, often exceeding the capabilities of standard clinical infrastructure. For instance, GANs applied to high-dimensional EHR data can require hours for training on modest hardware, delaying real-time analytics in time-sensitive environments like emergency departments [1, 5, 15].

Comparative assessments across studies indicate that while statistical imputations like multiple imputation by chained equations (MICE) are computationally lighter, they scale poorly with multimodal data integration, leading to bottlenecks in federated learning setups. Interpretively, this limitation hampers the deployment of AI in resource-constrained settings, such as rural hospitals, where cloud-based solutions may not be feasible due to connectivity issues. Systems-level insights reveal a trade-off: prioritizing scalability often sacrifices imputation accuracy, potentially compromising patient safety in closed-loop decision systems [6, 16, 24].

Pitfalls emerge when models overfit to imputed data, inflating computational costs during iterative retraining. Reporting standards advocate for efficiency benchmarks in publications, yet many studies overlook these, perpetuating inefficient practices. This cross-study synthesis underscores the need for optimized algorithms, like sparse autoencoders, to balance compute demands with clinical utility [8, 20, 25].

Furthermore, in big data analytics for population health, scalability limitations are amplified when handling streaming data, where continuous imputation overwhelms systems. This interpretive lens highlights infrastructural gaps, calling for hardware-software co-design in AI healthcare frameworks to mitigate these constraints [9, 21, 26].

In summary, addressing computational hurdles requires innovative approximations, ensuring that missing data handling does not become a barrier to equitable AI adoption across diverse healthcare ecosystems [10, 22].

Bias amplification and ethical concerns

Missing data exacerbates bias in clinical AI, where imputation decisions can inadvertently reinforce disparities in healthcare outcomes. MNAR mechanisms, common in underrepresented populations, lead to biased models that underperform for minorities, as imputed values often default to majority norms. Literature synthesis shows this in prognostic analytics, where gender or ethnic biases propagate through incomplete datasets, resulting in unequal risk assessments [2, 11, 12].

Comparatively, studies on imputation efficacy reveal that without bias mitigation, AI systems in healthcare analytics perpetuate systemic inequities, such as in mental health predictions, where socioeconomic data gaps skew results. Interpretively, this raises ethical dilemmas: AI governance must embed fairness checks, but current limitations in interpretable imputations hinder accountability. Cross-study evidence supports debiasing techniques like adversarial training, yet their adoption remains sparse due to added complexity [3, 13, 17].

Pitfalls include ethical oversights in model validation, where missing data audits are neglected, leading to deployed systems that violate principles of justice. Reporting standards emphasize equity reporting, but inconsistencies limit comparative evaluations. Systems-level perspectives frame this as a governance failure, necessitating regulatory frameworks that mandate bias impact statements [4, 14, 18].

Moreover, in global health applications, cultural biases in domain knowledge integration amplify limitations, where Western-centric models fail in diverse contexts. This original synthesis calls for inclusive data strategies, integrating participatory design to address ethical gaps in AI healthcare [7, 19, 23].

Ultimately, overcoming these concerns demands a shift toward ethically aligned modeling, where limitations in bias handling are proactively confronted through multidisciplinary oversight [27, 28].

Interpretability and trust issues

Interpretability remains a core limitation in missing data handling for clinical AI, as complex imputations obscure decision rationales, eroding clinician trust. Black-box models like neural networks impute values without clear explanations, complicating validation in high-stakes healthcare systems. Synthesizing evidence, this issue is pronounced in diagnostic analytics, where unexplained imputations lead to hesitancy in adopting AI recommendations [1, 5, 29].

Comparative analyses highlight that rule-based imputations offer better transparency but at the cost of handling nonlinear patterns ineffectively. Interpretively, this tension undermines closed-loop systems, where trust is essential for human-AI collaboration. Cross-study insights advocate for post-hoc explainability tools, such as SHAP values adapted for imputed features, to bridge this gap [6, 15, 20].

Pitfalls arise from over-reliance on opaque methods without sensitivity analyses, fostering distrust in deployment phases. Reporting standards push for interpretability metrics, yet many publications fall short, limiting knowledge transfer. Systems-level views position interpretability as a foundational requirement, integrating it into AI infrastructures to enhance adoption [8, 16, 21].

Furthermore, in regulatory contexts, limitations in demonstrating model reliability due to missing data opacity delay approvals. This interpretive discussion emphasizes hybrid models that combine interpretability with performance, fostering trustworthy analytics [9, 24, 25].

Data privacy and security limitations

Privacy limitations in handling missing data pose formidable challenges in AI healthcare, especially under regulations like GDPR or HIPAA. Federated imputation, while preserving data locality, struggles with inconsistent missingness across sites, leading to suboptimal global models. Literature synthesis indicates that differential privacy additions to imputations introduce noise, degrading analytics quality in sensitive domains like genomics [2, 10, 11].

Comparatively, centralized approaches offer better performance but heighten breach risks, particularly when missing data reveals indirect identifiers. Interpretively, this creates a security paradox: enhancing imputation sophistication increases vulnerability surfaces. Cross-study evidence supports encrypted imputation protocols, yet their computational overhead limits scalability [3, 12, 13].

Pitfalls include inadequate anonymization during imputation, where reconstructed data exposes patients. Reporting standards require privacy impact disclosures, but compliance varies. Systems-level insights advocate for secure multi-party computation in AI frameworks, addressing these limitations holistically [4, 14, 17].

Moreover, in real-time systems, dynamic missing data handling conflicts with static privacy measures, necessitating adaptive safeguards. This original framing highlights the interplay between security and utility in closed-loop healthcare [5, 18, 19].

Integration and interoperability barriers

Interoperability barriers limit effective missing data management in heterogeneous healthcare AI systems, where disparate data formats hinder unified imputation. EHR systems from different vendors often produce incompatible missing patterns, complicating analytics pipelines. Synthesizing studies, this barrier is evident in multi-institutional collaborations, where data harmonization fails, reducing model generalizability [6, 20, 22].

Comparative evaluations show that standards like FHIR help, but incomplete adoption perpetuates silos. Interpretively, this impedes closed-loop functionality, as feedback across systems is disrupted. Cross-study discussions call for ontology-based integrations to standardize missing data representations [7, 8, 23].

Pitfalls in ignoring interoperability lead to fragmented AI deployments, wasting resources. Reporting emphasizes integration details, yet gaps persist. Systems-level perspectives demand infrastructural reforms, embedding interoperability into governance [9, 24, 25].

Furthermore, in emerging IoT-integrated healthcare, device-specific missingness amplifies barriers, requiring cross-protocol solutions. This underscores the need for adaptive architectures [10, 26].

Future research directions

Advancing imputation methodologies

Future research should prioritize hybrid imputation methodologies that fuse statistical rigor with AI adaptability, addressing current limitations in clinical machine learning. Developing context-aware GANs that incorporate temporal and multimodal priors could enhance the handling of MNAR data in dynamic healthcare analytics. Literature synthesis suggests exploring meta-learning for imputation, where models learn from diverse missing patterns across datasets, improving transferability [1, 15, 24].

Comparatively, integrating quantum-inspired algorithms for high-dimensional imputation may offer scalability breakthroughs, particularly in genomics-integrated systems. Interpretively, this direction fosters resilient AI infrastructures, reducing bias in underrepresented cohorts. Cross-study insights indicate a need for benchmarks that evaluate long-term stability in closed-loop deployments [5, 16, 25].

Pitfalls in current methods highlight opportunities for uncertainty-aware imputations that propagate confidence through analytics pipelines. Reporting standards should evolve to include prospective validation protocols. Systems-level views position this as a pathway to personalized medicine, where imputation tailors to individual data profiles [6, 20, 26].

Moreover, collaborative research across disciplines could yield bio-inspired imputations, drawing from ecological modeling for robust handling of sparse clinical data [8, 21].

Enhancing reporting and standardization frameworks

Enhancing reporting frameworks represents a critical future direction, aiming for consensus guidelines that mandate comprehensive missing data disclosures in AI healthcare publications. Building on existing standards, future efforts could develop AI-specific extensions to CONSORT and STROBE, incorporating imputation impact assessments. Synthesizing evidence, this would facilitate meta-analyses, accelerating knowledge accumulation [2, 17, 18].

Comparatively, machine-readable reporting templates could automate compliance checks, integrating with journal submission systems. Interpretively, this promotes a culture of transparency, essential for governance in deployed AI. Cross-study discussions advocate for international consortia to harmonize standards, addressing global variations [3, 11, 19].

Pitfalls in fragmented reporting underscore the need for educational interventions, training researchers on standardized documentation. Systems-level insights suggest embedding these frameworks into AI development tools, ensuring adherence from inception [4, 12, 22].

Furthermore, future research might explore blockchain for verifiable reporting, enhancing trust in shared healthcare analytics [13, 23].

Interdisciplinary and real-world validation studies

Interdisciplinary validation studies are imperative for bridging gaps between AI research and clinical practice, focusing on real-world missing data scenarios. Prospective trials evaluating imputation in live healthcare systems could reveal deployment pitfalls, informing iterative improvements. Literature synthesis emphasizes embedding ethicists and clinicians in study design to address bias and usability [7, 27, 28].

Comparatively, simulation-based research using synthetic datasets with controlled missingness could complement empirical studies, accelerating discovery. Interpretively, this direction strengthens systems-level resilience, validating closed loops under varied conditions. Cross-study evidence supports multi-center collaborations for diverse validation [9, 14, 29].

Pitfalls in lab-only testing highlight the urgency for pragmatic trials, integrating patient feedback. Reporting should prioritize generalizability metrics. Systems-level perspectives view this as foundational for scalable AI adoption [1, 5, 10].

Additionally, exploring AI-human interaction in missing data contexts could yield novel fusion models, enhancing decision support [6, 15].

Exploring emerging technologies for data resilience

Leveraging emerging technologies like edge computing and 6G for resilient data handling offers promising avenues, enabling decentralized imputation in IoT-driven healthcare. Future directions include developing fog-based models that impute at the source, minimizing transmission of incomplete data. Synthesizing studies, this could revolutionize remote monitoring, reducing latency in analytics [8, 16, 20].

Comparatively, incorporating neuromorphic hardware for efficient imputation may address computational limits. Interpretively, this fosters adaptive governance, where technology evolves with data challenges. Cross-study insights call for pilot implementations in underserved areas [21, 24, 25].

Pitfalls in integration emphasize security-focused research. Systems-level views position this as a transformative shift toward proactive AI ecosystems [2, 3, 26].

Conclusion

The pervasive challenge of missing data in clinical machine learning underscores the need for deliberate modeling decisions, vigilant avoidance of pitfalls, and adherence to robust reporting standards to realize the full potential of AI in healthcare systems and analytics. This review has synthesized evidence from diverse perspectives, revealing how unaddressed missingness propagates biases, erodes trust, and hampers scalability, while highlighting innovative imputations and closed-loop frameworks as pathways forward. By structuring AI across data, intelligence, decision, intervention, feedback, and governance layers, we advocate for systems that inherently mitigate these issues, fostering equitable and reliable clinical outcomes.

Future advancements must prioritize interdisciplinary efforts to refine methodologies, standardize practices, and validate in real-world contexts, ensuring AI evolves as a trustworthy partner in healthcare. Ultimately, embracing a holistic, governance-centric approach will transform missing data from a liability into an opportunity for more resilient, intelligent healthcare infrastructures, benefiting patients and providers alike.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Nov 2022 Revised: 05 Feb 2023 Accepted: 17 Apr 2023
Published online: 10 July 2023

References

Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, et al. Handling missing values in healthcare data: a systematic review of deep learning-based imputation techniques. Artif Intell Med. 2023;142:102587. https://doi.org/10.1016/j.artmed.2023.102587.

Rose C, Barber R, Preiksaitis C, Kim I, Mishra N, Kayser K, et al. A conference (missingness in action) to address missingness in data and AI in health care: qualitative thematic analysis. J Med Internet Res. 2023;25:e49314. https://doi.org/10.2196/49314.

Arora A, Alderman JE, Palmer J, Ganapathi S, Laws E, McCradden MD, et al. The value of standards for health datasets in artificial intelligence-based applications. Nat Med. 2023;29(11):2929-38. https://doi.org/10.1038/s41591-023-02608-w.

Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus. 2023;15(10):e46454. https://doi.org/10.7759/cureus.46454.

Bernardini M, Romeo L, Misericordia P, Frontoni E. A novel missing data imputation approach based on clinical conditional generative adversarial networks applied to EHR datasets. Comput Biol Med. 2023;162:107062. https://doi.org/10.1016/j.compbiomed.2023.107062.

Rios R, Miller RJH, Manral N, Sharir T, Einstein AJ, Fish MB, et al. Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry. Comput Biol Med. 2022;145:105449. https://doi.org/10.1016/j.compbiomed.2022.105449.

Luo Y, Szolovits P, Dighe AS, Baron JM. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform. 2022;23(1):bbab489.

Chen Z, Tan S, Chajewska U, Rudin C, Caruana R. Missing values and imputation in healthcare data: can interpretable machine learning help? Proc Mach Learn Res. 2023;209:86-99.

Shadbahr T, Roberts M, Stanczuk J, Gilbey J, Teare P, Dittmer S, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med. 2023;3:139. https://doi.org/10.1038/s43856-023-00356-z.

White N, Parsons R, Collins G, Barnett A. Evidence of questionable research practices in clinical prediction models. BMC Med. 2023;21:339. https://doi.org/10.1186/s12916-023-03048-6.

Bonneville EF, Schetelig J, Putter H, de Wreede LC. Handling missing covariate data in clinical studies in haematology. Best Pract Res Clin Haematol. 2023;36(2):101477. https://doi.org/10.1016/j.beha.2023.101477.

Wu TT, Smith LH, Vernooij LM, Patel E, Devlin JW. Data missingness reporting and use of methods to address it in critical care cohort studies. Crit Care Explor. 2023;5(11):e1005.
https://doi.org/10.1097/CCE.0000000000001005.

Guo CY, Yang YC, Chen YH. The optimal machine learning-based missing data imputation for the Cox proportional hazard model. Front Public Health. 2021;9:680054.
https://doi.org/10.3389/fpubh.2021.680054.

Sisk R, Lin L, Sperrin M, Barrett JK, Tom B, Diaz-Montana C, et al. Imputation and missing indicators for handling missing data in the development and deployment of risk prediction models: a simulation study. Stat Methods Med Res. 2023;32(8):1463-77.
https://doi.org/10.1177/09622802231165001.

Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data. 2021;8:140.
https://doi.org/10.1186/s40537-021-00516-9.

Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53(2):1487-509.
https://doi.org/10.1007/s10462-019-09709-4.

Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. J Med Internet Res. 2023;25:e48763.
https://doi.org/10.2196/48763.

Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes. 2020;13(10):e006556.
https://doi.org/10.1161/CIRCOUTCOMES.120.006556.

Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011-5.

Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol. 2019;48(4):1294-304.

Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. BMJ Health Care Inform. 2021;28(1):e100385.
https://doi.org/10.1136/bmjhci-2021-100385.

Wu K, Dahlem D, Hane C, Halperin E, Zou J. Collecting data when missingness is unknown: a method for improving model performance given under-reporting in patient populations. Proc Mach Learn Res. 2023;209:229-42.

Carpenter JR, Smuk M. Missing data: a statistical framework for practice. Biometrics. 2021;77(2):337-50.
https://doi.org/10.1111/biom.13405.

Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. Diagn Interv Radiol. 2022;28(5):450-62.
https://doi.org/10.5152/dir.2022.211297.

Goyal P. Challenges and opportunities of big data analytics in healthcare. Health Care Sci. 2023;2:195-210.
https://doi.org/10.1002/hcs2.66.

Fatima H, Rehman A, Basit I, Hameed N. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices. 2023;21:461-8.
https://doi.org/10.1007/s44174-023-00063-2.

Olawade DB, Wada OJ, David-Osuagwu UL, Aba E. Using artificial intelligence to improve pain assessment and pain management: a scoping review. J Am Med Inform Assoc. 2023;30(3):570-87.

Wang G, Tang Y, Wang J. Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous intracerebral haemorrhage as an example. BMC Med Inform Decis Mak. 2022;22(1):13.
https://doi.org/10.1186/s12911-021-01745-1.

Sathiya B, Balaji GN, Chidambaram S, Venkata Sathya Vivek T. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. Am J Artif Intell. 2022;6(1):1-9.
https://doi.org/10.11648/j.ajai.20220601.11.