American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Research Article

Clinical Artificial Intelligence as a Sociotechnical System: Structural Failure Modes and Governance Requirements

Julian Borges, MD, MS

Department of Computer Science, Boston University Metropolitan College, USA

***Corresponding author:** Julian Borges, MD, Department of Computer Science, Boston University Metropolitan College, USA.

To Cite This article: *Borges, Julian Y. V.*, Clinical Artificial Intelligence as a Sociotechnical System: Structural Failure Modes and Governance Requirements. Am J Biomed Sci & Res. 2026 30(1) AJBSR.MS.ID.003888,*

DOI: 10.34297/AJBSR.2026.30.003888

Received: February 06, 2026; **Published:** February 18, 2026

Abstract

Background: Clinical artificial intelligence systems increasingly shape healthcare delivery, yet many fail to produce sustained real-world benefit despite acceptable retrospective performance. Existing evaluation paradigms emphasize algorithmic accuracy and discrimination while systematically under accounting for the sociotechnical, operational, and economic structures within which clinical AI systems are deployed.

Objective: To examine clinical artificial intelligence deployment failure through a sociotechnical systems lens and to articulate governance as a system level design requirement integrating workflows, documentation practices, reimbursement incentives, and institutional accountability.

Methods: We conducted a structured qualitative synthesis spanning sociotechnical systems theory, health informatics, health services research, and artificial intelligence governance literature. Recurrent deployment failure modes were identified using an inductive thematic approach and organized into a taxonomy. These failure modes were mapped to governance gaps across the clinical AI lifecycle and used to derive explicit system level requirements, evaluation pathways, and testable propositions.

Results: Five recurrent clinical AI deployment failure modes were identified: workflow incompatibility, documentation mediated data distortion, reimbursement driven behavioural adaptation, audit and monitoring blind spots, and diffusion of accountability. These failures arise from structural properties of healthcare delivery and interact with statistical degradation mechanisms such as covariate shift and calibration drift, producing compounding risk after deployment. A governance-oriented deployment framework specifying actionable checkpoints across pre deployment, deployment, and post deployment phases is proposed.

Conclusions: Clinical artificial intelligence safety and effectiveness are emergent properties of coupled sociotechnical systems rather than attributes of isolated models. Governance frameworks that neglect operational workflows, economic incentives, and accountability structures are structurally insufficient. This work provides a transparent and testable foundation for governance oriented clinical AI deployment and motivates future formalization of governed adaptive decision systems.

Keywords: Clinical Artificial Intelligence, Socio Technical Systems, AI Governance, Adaptive Decision Systems, Healthcare Workflows, Reimbursement Incentives, Auditability, Lifecycle Evaluation, Covariate Shift

Introduction

Artificial intelligence systems are increasingly embedded in healthcare delivery, influencing diagnostic, prognostic, and operational decisions across clinical settings [1-13]. Despite promising retrospective performance, many deployed systems exhibit degraded effectiveness, inequitable outcomes, workflow disruption, or safety concerns in real world use [4-6,13].

Evidence from biomedical informatics and health services research suggests that these failures rarely arise from algorithmic deficiencies alone. Rather, they emerge from interactions between AI systems and the socio technical environments in which care is delivered, including clinical workflows, documentation practices, reimbursement incentives, and institutional accountability structures [1-3]. Healthcare delivery operates as a distributed work system in which decision making is jointly produced by clinicians, information technologies, and organizational processes rather than isolated agents or tools.

This analysis focuses on clinical AI systems that influence care delivery, including predictive risk stratification tools, diagnostic classification systems, and decision support applications that affect ordering, triage, or treatment decisions. The proposed framework applies primarily to systems embedded within routine clinical workflows and does not directly address consumer facing wellness technologies or purely administrative automation.

Electronic health records illustrate the socio technical complexity underlying clinical AI deployment. Health information technologies have long been shown to generate unintended consequences, including workflow fragmentation, cognitive burden, and new sources of error, even when introduced with quality improvement objectives [1,3]. These effects are not anomalies but predictable properties of tightly coupled socio technical systems [2].

Healthcare data used for AI development are further shaped by documentation and reimbursement incentives. Diagnostic codes, problem lists, and structured data elements often reflect billing requirements and institutional practices rather than clinical intent, introducing systematic label noise and distributional distortion [9-11]. AI systems trained on such data may therefore learn patterns associated with documentation behaviour, coding incentives, or institutional norms rather than underlying disease processes [6,9].

Clinical AI system behaviour emerges from interactions among clinical workflows, documentation and reimbursement incentives, and institutional accountability structures rather than from algorithmic properties alone. These tightly coupled system drivers shape data generation, tool use, and responsibility, creating conditions under which deployment failures can arise despite acceptable retrospective model performance. These dynamics position governance not as an external compliance function but as an internal system property that shapes data generation, decision authority, and post deployment behaviour (Figure 1).

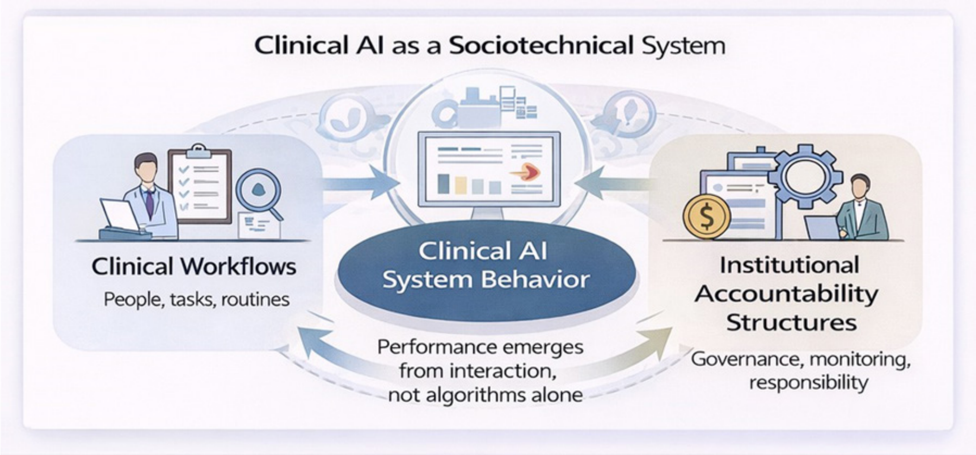

Clinical AI as a Sociotechnical System
Clinical Workflows
People, tasks, routines
Clinical AI System Behavior
Institutional Accountability Structures
Governance, monitoring, responsibility
Performance emerges from interaction, not algorithms alone

Figure 1: Clinical artificial intelligence as a sociotechnical system.

contributes to label shift, reimbursement driven behavioural adaptation induces covariate shift, and selective adoption creates feedback loops that affect calibration, subgroup performance, and downstream clinical behaviour over time [5-8]. Static evaluation and point in time validation obscure these dynamics, limiting detection of emerging risk after deployment.

This paper synthesizes evidence across socio technical systems theory, health informatics, health services research, and AI governance to identify recurrent clinical AI deployment failure modes. Building on this synthesis, we propose a governance-oriented deployment framework with explicit evaluation pathways designed to surface, monitor, and mitigate system level risks across the clinical AI lifecycle [1-15].

Methods

Study Design

This study employed a qualitative systems analysis combined with narrative synthesis to examine structural causes of clinical artificial intelligence deployment failure [1-5]. The analytic objective was not to assess individual algorithms but to identify recurrent system level patterns that undermine clinical AI performance after deployment. This approach is appropriate for identifying structural mechanisms that persist across technologies, institutions, and clinical domains

Literature Identification and Selection

We conducted a targeted review of peer reviewed literature spanning socio technical systems theory, health informatics, health services research, and artificial intelligence governance. Sources were identified through structured searches of PubMed, Web of Science, and Google Scholar, supplemented by backward citation tracking from highly influential publications. Inclusion criteria prioritized seminal or highly cited works addressing clinical workflows, documentation accuracy, reimbursement incentives, deployment failure, monitoring practices, and institutional accountability. Approximately 120 abstracts were screened, with 38 full text articles reviewed in depth. The final synthesis emphasizes sources demonstrating conceptual convergence, empirical grounding, and relevance to real world clinical deployment [1-15].

Analytical Procedure and Taxonomy Construction

An inductive thematic analysis was performed. Deployment failure modes were defined as recurrent patterns in which AI system behaviour diverged from intended clinical or operational objectives despite acceptable retrospective technical validation. Identified themes were iteratively clustered, refined, and stress tested against alternative categorizations to assess robustness.

Thematic saturation was reached when no additional failure categories emerged across independent sources. Five higher order failure mode categories demonstrated consistent explanatory power across clinical contexts, institutional settings, and AI application types.

Conceptual Framework

The analysis was structured using the Systems Engineering Initiative for Patient Safety (SEIPS) work system model, which conceptualizes healthcare delivery as an interaction among people, tasks, technologies, organizations, and environment [2]. Reimbursement incentives and documentation practices were explicitly incorporated as system drivers shaping data generation, clinical behaviour, and downstream AI performance [9-11].

Results

Failure Mode Taxonomy

Five higher-order failure modes recur across clinical AI systems and institutional contexts: workflow incompatibility, documentation-mediated data distortion, reimbursement-driven behavioural adaptation, audit and monitoring blind spots, and diffusion of accountability. Although these failure modes manifest differently across applications, they reflect shared structural mechanisms that undermine sustained effectiveness after deployment (Figure 2).

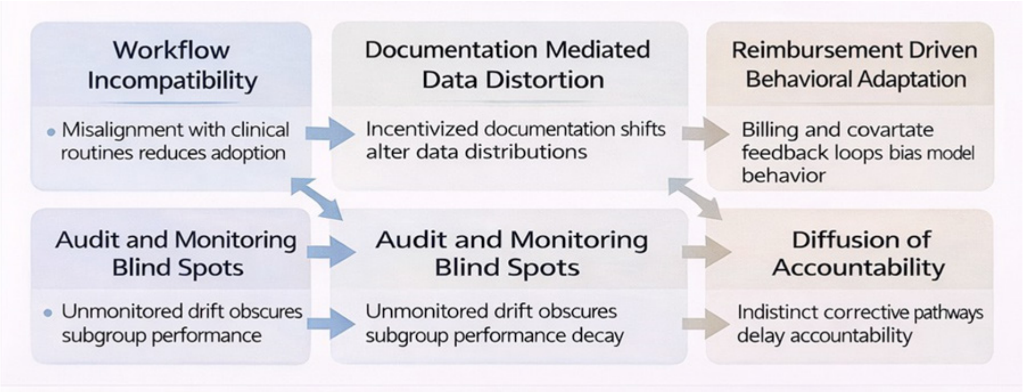

Workflow Incompatibility
Misalignment with clinical routines reduces adoption
Documentation Mediated Data Distortion
Incentivized documentation shifts alter data distributions
Reimbursement Driven Behavioral Adaptation
Billing and covartate feedback loops bias model behavior
Audit and Monitoring Blind Spots
Unmonitored drift obscures subgroup performance
Audit and Monitoring Blind Spots
Unmonitored drift obscures subgroup performance decay
Diffusion of Accountability
Indistinct corrective pathways delay accountability

Figure 2: Taxonomy of recurrent clinical AI deployment failure modes.

Although these failure modes manifest differently across AI application classes, they reflect shared structural mechanisms that recur across institutions, clinical domains, and deployment contexts (Table 1).

Table 1: Taxonomy of Recurrent Clinical AI Deployment Failure Modes.

Failure Mode	Structural Mechanism	Operational Manifestation	Implications for AI Performance and Safety
Workflow Incompatibility	Misalignment between AI system assumptions and established clinical workflows, task sequencing, and cognitive load	Disruption of clinical routines, increased documentation burden, selective or inconsistent tool use	Reduced adoption, biased outcome observation, automation avoidance or misuse, degradation of real-world effectiveness
Documentation-Mediated Data Distortion	Clinical documentation shaped by administrative, billing, or reporting requirements rather than clinical intent	Diagnostic codes and structured data reflect reimbursement incentives instead of disease state	Label noise, label shift, reduced generalizability, spurious correlations learned by models
Reimbursement-Driven Behavioral Adaptation	Financial incentives alter clinician behavior and documentation patterns after deployment	Changes in ordering, coding, or clinical thresholds in response to Al-influenced workflows	Covariate shift, feedback loops, calibration drift, performance decay over time

Audit and Monitoring Blind Spots	Absence of systematic post-deployment surveillance for performance, safety, and equity	Lack of subgroup-level monitoring, delayed detection of degradation or harm	Undetected subgroup performance decay, delayed corrective action, increased patient risk
Diffusion of Accountability	Unclear assignment of responsibility for AI oversight, performance review, and	Ambiguity regarding who monitors, updates, pauses,	Delayed escalation, governance gaps, normalization of unsafe performance

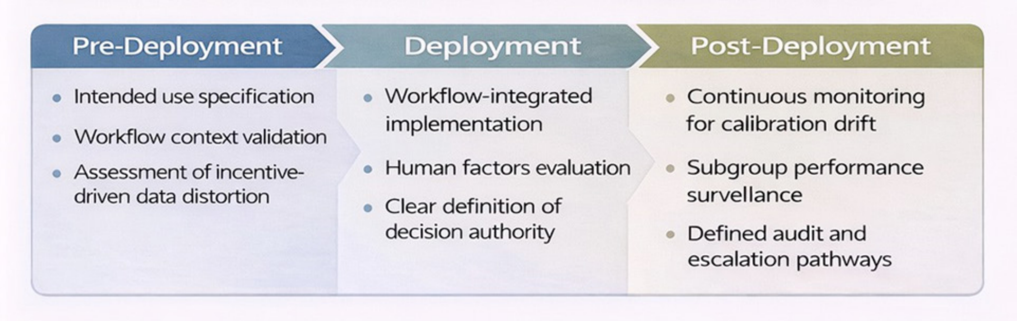
Interaction With Statistical Performance Dynamics

Identified socio technical failure modes interact directly with established statistical degradation mechanisms. Documentation mediated distortion contributes to label shift, reimbursement driven behavioural adaptation induces covariate shift, and workflow incompatibility promotes selective adoption that biases observed outcomes and undermines calibration over time [5-8]. Audit and monitoring blind spots impede timely detection of subgroup specific performance decay, while diffusion of accountability delays corrective intervention.

These interactions produce compounding risk after deployment, wherein socio technical pressures and statistical degradation reinforce one another, leading to divergence between retrospective validation and real-world performance.

Governance Oriented Deployment Framework

A lifecycle framework mapping recurrent sociotechnical failure mode to actionable governance checkpoints across pre-deployment, deployment, and post-deployment phases. The framework emphasizes intended use specification, workflow-integrated implementation, and continuous monitoring with defined accountability and escalation pathways to support sustained safety and effectiveness of clinical AI systems (Figure 3).


Pre-Deployment

- Intended use specification
- Workflow context validation
- Assessment of incentive-driven data distortion

Deployment

- Workflow-integrated implementation
- Human factors evaluation
- Clear definition of decision authority

Post-Deployment

- Continuous monitoring for calibration drift
- Subgroup performance survellance
- Defined audit and escalation pathways

Figure 3: Governance-oriented clinical AI deployment lifecycle.

The proposed governance-oriented deployment framework maps identified failure modes to actionable lifecycle checkpoints spanning pre deployment, deployment, and post deployment phases.

Pre Deployment:

a) Explicit specification of intended clinical use, workflow context, and decision authority [4,5]

b) Assessment of data generating processes for incentive driven distortion, including documentation and reimbursement dependencies [9-11]

Deployment:

a) Workflow integrated implementation incorporating human factors evaluation and usability testing [2-4,10]

b) Role clarity and transparency to mitigate over reliance, automation bias, and inappropriate task substitution [7,13]

Post Deployment:

a) Continuous monitoring for calibration drift, subgroup performance degradation, and adoption bias [5-8]

b) Defined audit trails, responsibility assignment, and escalation pathways for corrective action [7,8,15]

Evaluation and Implementation Pathway

Testable Propositions:

a) Clinical AI systems deployed without workflow integration will demonstrate lower sustained adoption independent of retrospective accuracy.

b) Systems trained on reimbursement sensitive labels will exhibit greater calibration drift over time relative to systems trained on clinically grounded labels.

c) Continuous post deployment monitoring will detect subgroup specific performance degradation earlier than aggregate performance metrics.

Measurable Endpoints:

a) Adoption, override, and deferral rates

b) Documentation burden and workflow interruption metrics

c) Subgroup specific calibration and error rates

d) Incident reports, audit findings, and corrective actions Minimal Implementation Model

e) Multidisciplinary governance committee with defined authority

f) Routine audit cadence with predefined performance and safety thresholds

g) Incident escalation, review, and remediation protocols

Discussion

This study demonstrates that clinical artificial intelligence deployment failures arise from systematic socio technical misalignment interacting with well described statistical degradation mechanisms, rather than from isolated algorithmic deficiencies [1-8]. Retrospective model performance alone is therefore insufficient to ensure sustained safety or effectiveness after deployment. Instead, clinical AI behaviour emerges from the coupling of learning systems with workflows, documentation practices, economic incentives, and institutional accountability structures.

By integrating governance, operational context, and quantitative monitoring within a unified lifecycle framework, this work reframes clinical AI deployment as a systems design problem rather than a validation exercise. Governance is treated not as an external compliance layer but as a necessary system property that shapes data generation, clinical behaviour, and post deployment adaptation. This perspective explains why technically sound models may degrade, produce inequitable outcomes, or generate unintended consequences when introduced into routine care.

The proposed framework aligns with current regulatory expectations for risk management, transparency, and post market surveillance, while simultaneously highlighting persistent gaps in operational ownership, monitoring responsibility, and escalation pathways [R1-R6]. In particular, existing regulatory guidance often presumes static model behaviours and well-defined accountability, assumptions that do not hold for adaptive systems embedded in distributed clinical environments. As a result, responsibility for detecting and mitigating deployment related risk is frequently diffuse, delayed, or informally assigned.

Importantly, the failure modes identified in this analysis are not unique to healthcare. Healthcare serves as an early and visible testbed for socio technical failure because of its complexity, regulatory constraints, and asymmetric harm profiles. However, the structural mechanisms described here generalize to other high stakes domains in which learning systems influence human decision making under uncertainty. This suggests that current artificial intelligence evaluation and governance paradigms remain incomplete for adaptive decision systems operating in real world institutional contexts. These findings motivate further development of formal framework that treat governance, safety, and authority as intrinsic components of adaptive decision systems rather than post hoc controls. Future work should focus on formalizing these properties, defining decision safety objectives beyond accuracy, and modelling human oversight as a dynamic system with explicit constraints. Such advances are necessary to move from reactive governance toward principled design of safe, accountable, and adaptive clinical AI systems.

Regulatory Alignment Statement

The governance-oriented deployment framework proposed in this study aligns with existing regulatory expectations for clinical artificial intelligence systems, including requirements related to risk management, transparency, post market surveillance, and accountability. In particular, the framework is consistent with current guidance emphasizing intended use specification, lifecycle-based oversight, human oversight, and ongoing performance monitoring for software as a medical device and clinical decision support tools.

At the same time, this analysis highlights structural gaps in prevailing regulatory approaches. Existing guidance largely presumes static model behaviours, clearly bounded accountability, and stable data generating processes. These assumptions are frequently violated in real world clinical environments, where adaptive learning systems interact dynamically with workflows, documentation practices, and reimbursement incentives. As a result, regulatory compliance alone may be insufficient to ensure sustained safety and effectiveness after deployment. By explicitly mapping socio technical failure modes to governance checkpoints and evaluation pathways, this framework operationalizes regulatory principles in a manner that supports earlier detection of deployment related risk, clearer assignment of responsibility, and more effective corrective action. The framework is intended to complement, rather than replace, existing regulatory processes by providing a testable systems level approach to post deployment oversight for adaptive clinical AI.

Limitations and Future Directions

This study is conceptual in nature and does not present primary empirical validation of the proposed framework. However, the objective of this work is not to evaluate individual models or deployment instances, but to identify recurrent structural failure mechanisms that persist across clinical AI applications and institutional contexts. As such, the framework is intended to generate testable hypotheses and guide prospective evaluation rather than to substitute for empirical assessment.

Future research should prospectively evaluate governance interventions derived from this framework across heterogeneous clinical AI system classes, organizational settings, and payment environments. In particular, comparative studies assessing workflow integrated deployment, continuous monitoring strategies, and explicit accountability structures are needed to determine their impact on sustained adoption, safety outcomes, and subgroup performance over time. Formalization of governance and decision safety as computational constructs represents an additional priority for advancing theory and informing regulatory design.

Conclusion

Clinical artificial intelligence systems operate within complex socio technical environments shaped by workflows, economic incentives, and accountability structures. Deployment failure is therefore a system level phenomenon rather than a property of individual models. Governance frameworks that treat operational context, monitoring, and responsibility as secondary considerations are structurally insufficient for ensuring sustained safety and effectiveness.

This work provides a transparent, testable, and operationally grounded foundation for governance oriented clinical AI deployment. By framing governance as a system property and linking socio technical dynamics to statistical degradation mechanisms, it offers a principled basis for evaluating, monitoring, and improving clinical AI systems across the deployment lifecycle.

Acknowledgments and Disclosures

The author acknowledges the contributions of the socio technical systems, health informatics, and clinical artificial intelligence research communities whose foundational work on workflow integration, auditability, accountability, and deployment safety informed this analysis. In particular, prior scholarship on electronic health record safety, unintended consequences of health information technology, and post deployment monitoring of clinical decision support systems provided essential conceptual grounding for this study.

The author also acknowledges the broader artificial intelligence governance literature, including work on algorithmic accountability, audit frameworks, and lifecycle oversight, which shaped the analytical lens and synthesis approach adopted in this manuscript.

Funding

This research received no external funding.

Competing Interests

The author declares no competing financial or non-financial interests related to this work.

Ethical Approval

This study did not involve prospective intervention, experimentation, or interaction with human participants. The analysis is based on synthesis of published literature and conceptual systems analysis of clinical artificial intelligence deployment. In accordance with applicable regulatory definitions and institutional policies, this work did not constitute human subjects research and did not require institutional review board approval or informed consent.

Data Availability

No new datasets were generated or analysed for this study. All evidence supporting the analysis is derived from previously published literature cited in the manuscript. As a conceptual and qualitative systems analysis, the work does not rely on proprietary data or patient level information.

Declaration of AI Use

The author affirms that generative artificial intelligence tools were used solely to assist with image generation, language editing and formatting during manuscript preparation. All conceptual framing, methodological design, analysis, interpretation, and conclusions were developed by the author. The author takes full responsibility for the integrity, accuracy, and originality of the work.

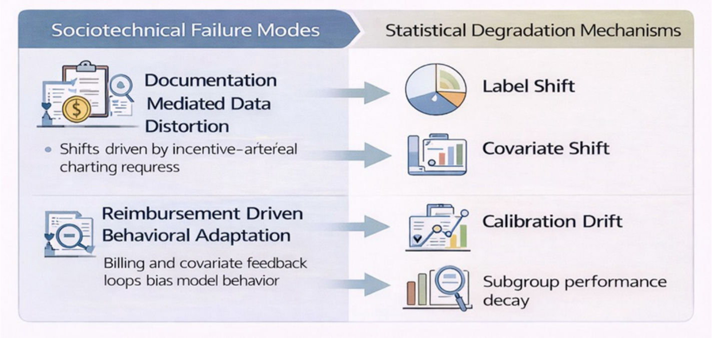
References

1. Sittig DF, Singh H (2010) A socio technical approach to preventing, mitigating, and recovering from electronic health record related safety hazards. J Am Med Inform Assoc 17(6): 597-606.
2. Carayon P, Schoofs Hundt A, Karsh B, AP Gurses, CJ Alvarado, et al. (2006) Work system design for patient safety: the SEIPS model. Qual Saf Health Care 15(suppl 1): i50-i58.
3. Ash JS, Berg M, Coiera E (2004) Some unintended consequences of information technology in health care: the nature of patient care information system related errors. J Am Med Inform Assoc 11(2): 104-112.
4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17(1): 195.
5. Sendak MP, Gao M, Nichols M, Anthony Lin, Suresh Balu, et al. (2019) Machine learning in health care: a critical appraisal of challenges and opportunities. eGEMs 7(1): 1.
6. Chen IY, Szolovits P, Ghassemi M (2019) Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 21(2): E167-E179.
7. London AJ (2019) Artificial intelligence and black box medical decisions: accuracy versus explainability. Hastings Cent Rep 49(1):15-21.
8. Raji ID, Smart A, White RN, Margaret Mitchell, Timnit Gebru, et al. (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. Proceedings of the Conference on Fairness, Accountability, and Transparency.
9. O Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, et al. (2005) Measuring diagnoses: ICD code accuracy. Health Services Res 40(5 Pt 2): 1620-1639.
10. Erickson SM, Rockwern B, Koltov M, McLean RM (2017) Putting patients first by reducing administrative tasks in health care. Ann Internal Med 166(9): 659-660.
11. Cutler DM (2010) How health care reform must bend the cost curve. Health Affairs 29(6):1 131-1135.
12. Topol EJ (2019) High performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1): 44-56.
13. Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA approved medical devices and algorithms: an online database. NPJ Digit Med 3: 118.
14. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB (2016) SMART on FHIR: a standard based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 23(5): 899-908.
15. McGraw D, Mandl KD (2021) Privacy protections to encourage use of health relevant digital data in a learning health system. NPJ Digit Med 4(1): 2.

APPENDIX

Appendix A. Regulatory and Standards References

Sociotechnical factors such as documentation-mediated data distortion and reimbursement driven behavioural adaptation contribute to statistical phenomena including label shift, covariate shift, calibration drift, and subgroup performance decay. These interactions illustrate how operational and institutional dynamics can drive post-deployment performance degradation independent of initial model accuracy. (Appendix Figure A1).


Sociotechnical Failure Modes
Statistical Degradation Mechanisms
Documentation Mediated Data Distortion
Shifts driven by incentive-artereal charting requress
Reimbursement Driven Behavioral Adaptation
Billing and covariate feedback loops bias model behavior
Label Shift
Covariate Shift
Calibration Drift
Subgroup performance decay

Appendix Figure A1: Conceptual linkage between sociotechnical failure modes and statistical performance degradation mechanisms.

R1. U.S. Food and Drug Administration (FDA)

Good Machine Learning Practice for Medical Device Development: Guiding Principles. FDA, Health Canada, and Medicines and Healthcare products Regulatory Agency (MHRA); 2021.

Available at: *https://www.fda.gov/medical-devices/software-medical-device-samd/good- machine-learning-practice-medical-device-development-guiding-principles*

R2. U.S. Food and Drug Administration (FDA)

Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA.

Available at: *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial- intelligence-and-machine-learning-software-medical-device*

R3. U.S. Food and Drug Administration (FDA)

Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence and Machine Learning Enabled Device Software Functions. Draft Guidance; 2023.

Available at: *https://www.fda.gov/regulatory-information/search-fda-guidance- documents/marketing-submission-recommendations-predetermined-change-control-plan- artificial*

R4. World Health Organization (WHO)

Ethics and Governance of Artificial Intelligence for Health. WHO; 2021.

Available at: *https://www.who.int/publications/i/item/9789240029200*

R5. Organisation for Economic Co-operation and Development (OECD)

Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). OECD; 2019.

Available at: *https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449*

R6. U.S. Food and Drug Administration (FDA), Health Canada, Medicines and Healthcare products Regulatory Agency (MHRA)

Transparency Principles for Machine Learning Enabled Medical Devices. 2021.

Available at: *https://www.fda.gov/medical-devices/software-medical-device-samd/transparency- principles-machine-learning-enabled-medical-devices*