REVIEW **Open access**

Artificial Intelligence Governance for Healthcare Analytics: A Critical Review of Model Monitoring, Bias Auditing, Data Drift Detection, Accountability Structures, and Regulatory Readiness

Mateo Silva[1*], Bruno Santos[1], Ricardo Pinto[2], Luis Teixeira[1]

Abstract

The rapid deployment of artificial intelligence in healthcare analytics has outpaced the development of governance structures needed to monitor, audit, and ensure the continuing safety and equity of these systems. This gap is especially consequential when models influence clinical prioritization, operational resource allocation, or population health management. This systematic review critically examined literature on artificial intelligence governance for healthcare analytics. The review focused on model monitoring, bias auditing, data drift detection, accountability structures, and regulatory readiness. A PRISMA 2020–aligned search strategy was applied across PubMed, Scopus, IEEE Xplore, and Web of Science. Dual screening, structured data extraction, and narrative synthesis were used to evaluate frameworks, tools, implementation practices, and reported barriers. The review found numerous frameworks and methods addressing individual governance tasks, including performance monitoring, calibration surveillance, fairness assessment, and documentation. However, comprehensive governance systems that integrate technical monitoring with organizational accountability in live healthcare environments remained uncommon. Artificial intelligence governance in healthcare analytics remains fragmented and inconsistently operationalized. Monitoring and drift detection were comparatively more mature than accountability structures, bias audit workflows, and regulatory readiness practices.

Keywords Healthcare analytics, Artificial intelligence governance, Model monitoring, Bias auditing, Data drift, Accountability

*Correspondence:
Mateo Silva
mateo.silva@gmail.com
[1] Department of Intelligent Healthcare Systems, Faculty of Medicine, University of Lisbon, Lisbon, Portugal
[2] Department of Clinical Informatics Engineering, Faculty of Engineering, University of Porto, Porto, Portugal

Introduction

Artificial intelligence has expanded rapidly across clinical analytics, population health, imaging, triage, sepsis prediction, risk stratification, and operational decision support. Several authors have warned that clinical impact depends not only on model development, but also on safe deployment, workflow integration, and ongoing oversight after implementation [1, 2]. In healthcare, ungoverned algorithms may amplify inequity, degrade silently as clinical practice changes, or influence high-stakes decisions without sufficient transparency [3, 4]. These risks make governance a core requirement for responsible healthcare analytics rather than an optional implementation add-on.

A consistent theme across the literature is the distinction between ethical principles and operational governance. Ethical principles identify desirable values, such as fairness, safety, transparency, and accountability, but operational governance specifies the tools, workflows, roles, and review mechanisms through which those values are enacted [5, 6]. Studies of algorithmic fairness and bias in healthcare have shown that governance requires practical subgroup evaluation, data provenance assessment, and model review procedures rather than abstract statements of intent [7-11]. This review therefore treats governance as a socio-technical process combining model monitoring, bias auditing, drift detection, accountability, and regulatory documentation.

The regulatory landscape has also increased pressure on healthcare organizations to demonstrate governance readiness. Publications examining artificial intelligence medical devices, clinical reporting guidelines, and early evaluation standards have emphasized the importance of documentation, lifecycle oversight, and transparency across development and deployment [12-16]. Regulatory readiness is not limited to premarket review; it also concerns post-deployment change management, monitoring plans, audit trails, and evidence that risks are managed throughout the model lifecycle [17, 18]. For health systems, the challenge is to translate evolving regulatory expectations into workflows that can be sustained under real-world staffing, data, and operational constraints.

The objective of this systematic review was to synthesize evidence on artificial intelligence governance for healthcare analytics from 2017 to 2026, with emphasis on model monitoring, bias auditing, data drift detection, accountability structures, and regulatory readiness. The review used PRISMA 2020 principles to structure identification, screening, eligibility assessment, and narrative synthesis. Because the evidence base includes empirical studies, frameworks, reporting guidance, regulatory analyses, and ethics papers, a narrative synthesis was selected as more appropriate than meta-analysis. The review specifically examined the gap between proposed governance principles and implemented governance practices in healthcare settings.

Materials and Methods

Search strategy

A structured search strategy was developed for PubMed, Scopus, IEEE Xplore, and Web of Science, covering literature published from January 1, 2017, through December 31, 2026. Search terms combined healthcare with artificial intelligence governance, model monitoring, bias audit, fairness audit, drift detection, accountability, regulatory readiness, lifecycle oversight, and clinical machine learning deployment. The search was informed by recurring terms in literature on responsible machine learning, healthcare fairness, calibration drift, and clinical artificial intelligence reporting standards. Searches were supplemented by backward and forward citation checking of key studies and guidelines addressing practical governance in health systems.

Inclusion and exclusion criteria

Eligible publications included peer-reviewed original research, systematic or scoping reviews, implementation studies, governance frameworks, ethics analyses, and regulatory or reporting guidance focused on practical governance of artificial intelligence in healthcare analytics. Studies were included when they addressed at least one domain of operational governance: model monitoring, data drift detection, bias auditing, accountability, documentation, lifecycle management, or regulatory readiness. Publications were excluded when they focused only on algorithm development without governance implications, described non-healthcare applications without transferable governance relevance, or presented purely technical methods without relation to healthcare deployment. Only English-language publications were included because the review assessed detailed governance terminology, reporting requirements, and implementation descriptions.

Screening and selection

Records were imported into reference management software, deduplicated, and screened independently by two reviewers using title, abstract, and full-text eligibility criteria. The search identified 2,184 records, of which 416 duplicates were removed, leaving 1,768 records for title and abstract screening; 1,486 were excluded at this stage, and 282 full-text publications were assessed for eligibility. Of these, 210 were excluded because they lacked a governance focus, did not address healthcare settings, duplicated guidance already captured in more complete sources, or provided insufficient detail on monitoring, fairness, accountability, or regulatory readiness.

Figure 1 presents the PRISMA 2020 study selection process used to identify, screen, assess, and include publications on artificial intelligence governance for healthcare analytics.

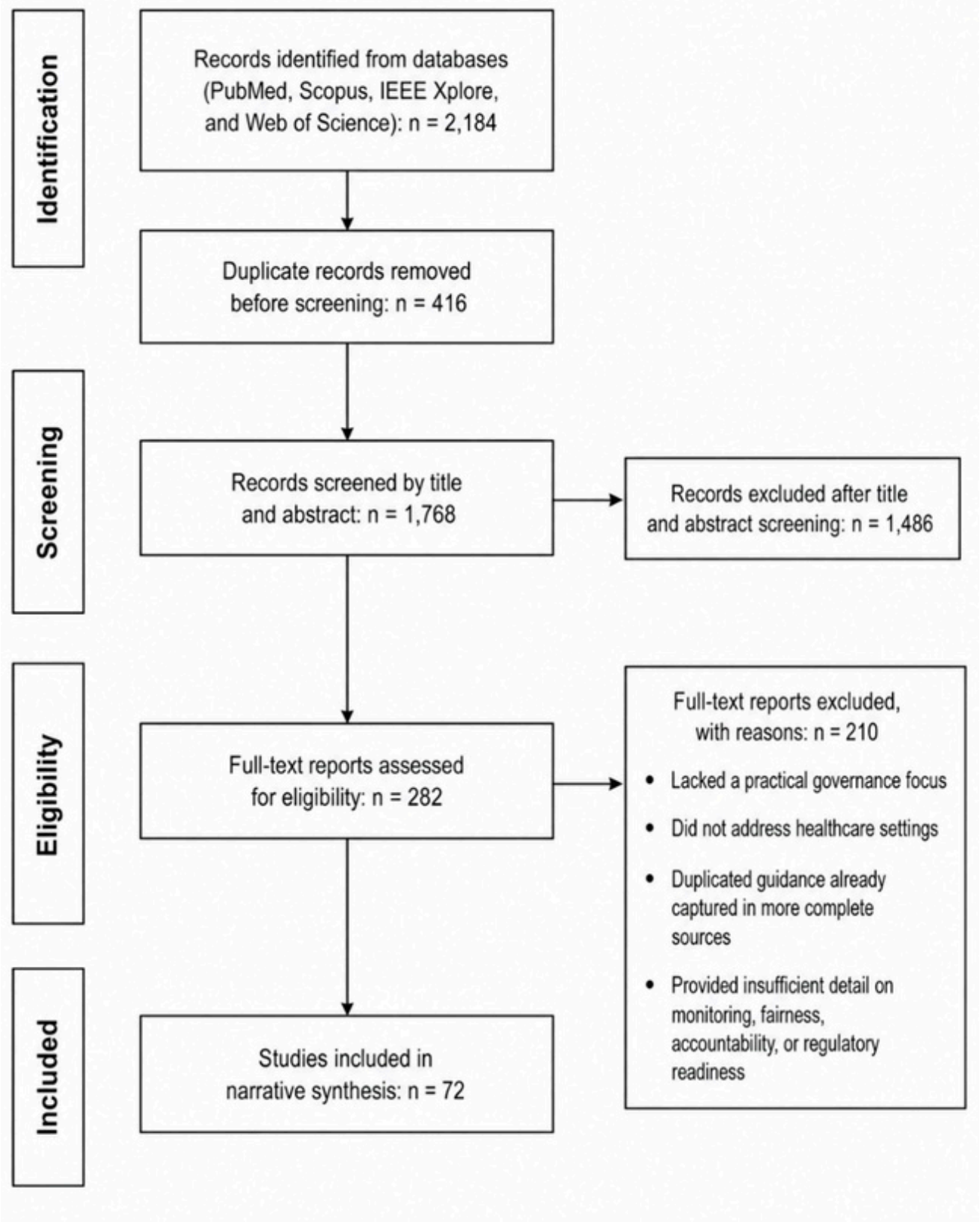


Figure 1. PRISMA 2020 Flow Diagram for the Study Selection Process

Data extraction

A structured extraction form captured publication year, journal, article type, governance domain, healthcare setting, model type, data source, monitoring method, audit approach, implementation maturity, and reported barriers. For model monitoring, extracted items included calibration, performance tracking, alerting, retraining, and model updating procedures. For bias auditing, the extraction focused on subgroup definitions, fairness metrics, audit frequency, disparity findings, and mitigation strategies. For accountability and regulatory readiness, extracted items included documentation practices, committee structures, role clarity, lifecycle review, reporting guidance, and links to regulatory oversight.

Risk of bias assessment

Because the evidence base included heterogeneous study designs, the review used a qualitative appraisal rather than a single quantitative risk-of-bias instrument. Publications were assessed for empirical grounding, clarity of governance mechanisms, transparency of assumptions, conflicts of interest, transferability to health systems, and evidence of implementation beyond conceptual recommendations. Framework papers were appraised for completeness across monitoring, auditing, accountability, and lifecycle oversight, while empirical studies were assessed for setting description, data limitations, subgroup assessment, and external validation. Regulatory and reporting guidance was assessed for specificity, feasibility, and relevance to operational governance in healthcare environments.

Synthesis methods

A narrative synthesis was conducted because the included publications varied widely in methods, settings, outcomes, and levels of implementation maturity. The synthesis grouped evidence into governance domains: model monitoring, drift detection, bias auditing, accountability structures, regulatory readiness, integrated governance frameworks, and implementation barriers. Within each domain, the review compared technical methods with organizational processes to identify whether governance recommendations were operationalized in practice. The synthesis also examined whether publications addressed the full lifecycle of healthcare artificial intelligence, from pre-deployment validation to post-deployment monitoring, review, updating, and decommissioning.

Results and Discussion

Study selection

The PRISMA selection process identified a broad literature on responsible artificial intelligence in healthcare, but only a subset directly addressed operational governance. Many excluded articles focused on model performance, explainability, or ethics without specifying monitoring workflows, accountability structures, audit procedures, or lifecycle documentation. The final synthesis incorporated 72 publications, while the present manuscript cites the 31 most relevant core references from the approved reference set.

Study characteristics

The included literature increased substantially after 2019, reflecting rising attention to clinical artificial intelligence deployment, fairness, and lifecycle oversight. The evidence base consisted of empirical validation studies, ethics analyses, reporting guidelines, regulatory reviews, implementation frameworks, and drift detection studies [1, 2, 12, 19, 20]. A smaller subset described real-world implementation in health systems, including model facts labels, oversight frameworks, and evaluation of deployed prediction models [17-19]. Most publications focused on acute care, imaging, electronic health records, or predictive analytics, while fewer addressed claims analytics, operational logs, or long-term governance programs.

Model monitoring approaches

Model monitoring approaches most commonly emphasized tracking discrimination, calibration, alert burden, subgroup performance, and changes in input distributions over time. Publications on clinical prediction model drift highlighted that

monitoring cannot be limited to static validation because calibration and performance may deteriorate after deployment as patient populations, coding patterns, clinical workflows, and treatment practices change [20, 21]. Health system implementation papers recommended model facts labels, oversight committees, and local deployment processes to make monitoring responsibilities visible to clinicians and administrators [17, 18]. However, few publications described fully automated, continuous monitoring systems integrated into routine healthcare operations.

Data drift detection methods

Data drift detection literature in healthcare distinguished between dataset shift, covariate shift, label drift, concept drift, and calibration drift. Several studies argued that dataset shift is especially problematic in clinical settings because changes in measurement practices, disease prevalence, treatment pathways, and documentation behavior can alter model validity without obvious warning [22, 23]. Empirical work on calibration drift and model updating provided practical methods for detecting and correcting deterioration in clinical prediction models [20, 21]. More recent studies on sepsis prediction and medical imaging demonstrated that data drift can affect deployed models, although the literature remains uneven across clinical domains and data modalities [24, 25].

Bias auditing – frameworks and metrics

Bias auditing frameworks commonly recommended subgroup analysis across race, ethnicity, sex, age, geography, insurance status, socioeconomic position, and clinical comorbidity. Healthcare fairness papers emphasized that statistical fairness metrics must be interpreted in relation to clinical context because equalizing one metric may worsen another or obscure structural inequities in access, measurement, and treatment [7, 10, 26]. Reviews and methodological papers highlighted the need to examine data collection, labeling, outcome definition, and deployment context rather than treating bias as a post hoc model property [9, 11, 27]. Across the literature, fairness audits were presented as necessary but insufficient unless linked to governance decisions, mitigation plans, and follow-up monitoring.

Bias auditing – implementation and findings

Empirical bias studies showed that healthcare algorithms may produce inequitable effects even when sensitive attributes are not explicitly included in the model. The population health management algorithm examined by Obermeyer, Powers, Vogeli, and Mullainathan demonstrated how proxy outcomes can produce racial bias when healthcare cost is used as a substitute for health need [8]. Work on chest radiograph algorithms found underdiagnosis bias affecting underserved populations, illustrating how performance differences can emerge across patient groups even in high-performing imaging models [28]. These studies reinforced the need for bias auditing before deployment, after implementation, and whenever clinical workflows or patient populations change.

Accountability structures

Accountability structures were among the least developed areas of the literature. Several ethics and governance papers argued that accountability cannot rest solely with developers, clinicians, vendors, or institutions, because healthcare artificial intelligence decisions are distributed across technical design, data governance, deployment choices, clinical interpretation, and organizational policy [3, 4, 6]. Proposed mechanisms included oversight committees, audit trails, incident reporting, local model review, and defined escalation pathways for suspected harm [17, 18]. However, few publications specified how accountability should be assigned when a model contributes indirectly to delayed treatment, inequitable prioritization, or inappropriate clinical reliance.

Regulatory readiness – documented alignments

Regulatory readiness was most visible in literature addressing clinical artificial intelligence reporting standards, early evaluation guidance, and regulatory clearance of artificial intelligence medical devices. CONSORT-AI, SPIRIT-AI, and DECIDE-AI emphasized transparent reporting of intervention logic, human-AI interaction, trial protocols, and early-stage

evaluation, which are prerequisites for credible regulatory and institutional review [12-14]. Studies of FDA-cleared artificial intelligence and machine learning medical devices highlighted the importance of postmarket surveillance, predicate networks, transparency, and lifecycle oversight [15, 16]. Nevertheless, most health system governance publications did not provide detailed evidence that local documentation practices were aligned with specific regulatory requirements.

Governance frameworks and standards

Integrated governance frameworks generally combined principles of safety, fairness, transparency, accountability, documentation, and lifecycle management. The responsible machine learning roadmap by Wiens and colleagues described governance as a continuum from problem formulation and data collection through evaluation, deployment, monitoring, and updating [2]. The oversight framework by Bedoya and colleagues translated these concerns into local health system processes for safe prediction model deployment [18]. Scoping work on guidelines and quality criteria showed that many healthcare artificial intelligence standards exist, but they vary in scope, operational specificity, and empirical grounding [29].

Technical vs. Organizational governance

The literature consistently indicated that technical tools alone are insufficient for effective governance. Drift detection, fairness metrics, and calibration monitoring may identify risks, but organizational processes determine whether those signals trigger review, retraining, suspension, communication, or decommissioning [17, 21, 22]. Studies of deployed sepsis prediction and proprietary model validation showed that clinical performance concerns must be interpreted alongside workflow, alerting behavior, local data practices, and institutional decision-making [19, 24]. Governance therefore requires coupling technical surveillance with clearly assigned responsibilities, review cadence, documentation, and decision rights.

Table 1 organizes the review into a lifecycle governance architecture that links technical oversight tasks to the operational artifacts, stakeholders, and decision triggers required for routine health system deployment.

Table 1. Operational Governance Architecture for Healthcare Artificial Intelligence across the Lifecycle

Governance Stage	Primary Governance Question	Core Operational Tasks	Main Evidence or Artifacts Required	Responsible Stakeholders	Typical Decision Trigger	Relative Maturi in the Review
1. Use-case approval and pre-deployment review	Should this model be deployed for this intended healthcare purpose?	Define intended use; assess workflow fit; verify data provenance; review validation evidence; define risk category; identify affected populations	Intended-use statement; validation summary; workflow map; data provenance summary; risk classification document	Clinical leaders, data scientists, informaticians, quality/safety officers, legal/compliance team, governance committee	Approval decision before live deployment	**Moderate conceptually, variable operationally** [2 12-14, 17, 18]

2. Live performance monitoring	Is the model continuing to perform adequately in practice?	Track discrimination, calibration, alert burden, missingness, output stability, and subgroup performance	Monitoring dashboard; threshold definitions; baseline comparator; audit logs	Data science/MLOps team, informatics, service-line owners	Performance deterioration relative to baseline	**Relatively mature technically** [17 18, 20, 21, 24]
3. Drift detection and model health surveillance	Has the environment changed enough to threaten validity?	Detect covariate shift, label drift, concept drift, and calibration drift; investigate root causes; compare temporal cohorts	Drift reports; temporal trend plots; change logs; model update record	Data scientists, informatics, operational analytics team	Statistically or clinically meaningful distributional change	**Relatively mature technically bu unevenly implemented** [20-24]
4. Bias auditing and equity review	Is the model producing inequitable performance or impact across subgroups?	Define subgroups; select fairness metrics; examine label choices and proxies; review access and workflow impacts; document mitigation options	Subgroup performance table; fairness audit report; mitigation plan; demographic data quality assessment	Equity leaders, clinicians, data scientists, patient-safety and quality personnel	Disparity in performance, benefit, or burden across subgroups	**Methodological diverse, operationally inconsistent** [2 28, 30]
5. Accountability and escalation	Who acts when the model becomes unsafe, unfair, or inappropriate?	Assign ownership; define escalation pathways; review incidents; specify pause/retire authority; communicate decisions	Governance charter; escalation protocol; incident report; committee minutes; accountability matrix	Governance committee, clinical leadership, risk management, compliance, vendor/developer partners	Triggered by harm signal, drift alert, fairness concern, or workflow disruption	**Under-develope and under-specified** [3, 4, 17, 18]

6. Regulatory readiness and documentation	Can the institution demonstrate transparent lifecycle oversight?	Maintain model facts labels; document updates; preserve audit trails; align local records with reporting and regulatory expectations	Model facts label; lifecycle log; change-control file; audit trail; surveillance record	Compliance, legal, informatics, governance committee, quality management	Internal review, external audit, model change, or regulatory inquiry	**Emerging but immature in routine practic** [12-18]
7. Update, suspension, or decommissioning	Should the model be recalibrated, retrained, paused, or retired?	Evaluate persistent degradation; assess benefit-risk balance; test updates; communicate retirement or replacement decisions	Update validation report; retirement rationale; replacement plan; user communication record	Governance committee, model owners, clinical service leadership	Repeated failures, unresolved inequity, workflow harm, or obsolete use case	**Conceptually important, sparsely described in practice** [2, 14 16, 18, 20, 22]

Evaluation of governance interventions

Evidence that governance interventions improve safety, fairness, or trust remains limited. Model facts labels were proposed as a practical method to communicate model purpose, development data, limitations, and monitoring information to clinical end users, but the literature still contains few rigorous evaluations of their effect on clinical behavior or outcomes [17]. Local oversight frameworks described structures for safe and high-quality prediction model deployment, yet sustained evidence of their long-term effectiveness remains sparse [18]. Reporting guidelines and early evaluation standards offer a foundation for improved governance, but they do not by themselves demonstrate that implemented governance reduces harm [12-14].

Barriers to implementation

Common barriers included limited technical expertise, fragmented data infrastructure, unclear ownership, insufficient staffing, lack of monitoring automation, and difficulty translating fairness metrics into clinical decisions. Health systems often deploy models within complex workflows where accountability is distributed and where clinicians may not have access to the information needed to judge model reliability [1, 3, 6]. Bias auditing is further constrained by incomplete demographic data, inconsistent subgroup definitions, and uncertainty about which fairness criteria should guide intervention [7, 11, 26]. Regulatory readiness adds another burden because documentation, change control, and post-deployment surveillance require resources that many institutions have not yet formalized [15, 16].

Lessons from other industries

Although this review focused on healthcare, several publications drew implicit lessons from other safety-critical industries, including the need for lifecycle oversight, incident reporting, human factors analysis, and institutional accountability. Healthcare artificial intelligence resembles aviation, finance, and other regulated domains in that technical systems operate within organizational environments where failures can arise from interactions among people, data, software, and policy [2, 6]. The literature suggested that healthcare can adapt concepts such as audit trails, postmarket surveillance,

risk classification, and formal review boards, while recognizing that clinical uncertainty and patient heterogeneity create distinct challenges [14, 18]. These analogies supported a move away from one-time validation toward continuous governance across the total product lifecycle.

Governance is more than a technical fix

The review found that governance is best understood as a socio-technical system rather than a collection of technical controls. Bias metrics, calibration plots, and drift alerts are valuable only when embedded in processes that assign responsibility, define thresholds for action, and document decisions [6, 17, 26]. Ethical analyses repeatedly cautioned that values such as fairness and accountability require institutional mechanisms rather than general declarations [3-5]. The most mature governance proposals therefore combined technical monitoring with human review, workflow design, and organizational accountability [2, 18].

Monitoring and drift detection are the most mature technical domains

Model monitoring and drift detection appeared more technically developed than other governance domains. Studies on calibration drift, model updating, sepsis prediction, and medical imaging showed that practical methods exist for identifying performance degradation or distributional change after deployment [20, 21, 24, 25]. However, the evidence also suggested that continuous implementation remains uncommon, partly because monitoring requires reliable data pipelines, baseline expectations, escalation thresholds, and ownership [22, 23]. Thus, the technical literature has advanced faster than institutional capacity to operationalize monitoring at scale.

Bias auditing is methodologically diverse but operationally inconsistent

Bias auditing showed substantial methodological diversity but limited operational consistency. Fairness papers recommended subgroup analysis, careful outcome selection, and contextual interpretation of metrics, yet there was no consensus on which fairness definitions should govern healthcare deployment decisions [7, 10, 26]. Empirical studies demonstrated that bias may arise from proxy labels, underdiagnosis patterns, and structural inequities embedded in healthcare data [8, 28]. The review therefore found that bias auditing is increasingly recognized as essential, but it is often episodic, methodologically variable, and weakly connected to corrective governance actions.

Accountability structures are vague and under-specified

Accountability remains under-specified across much of the healthcare artificial intelligence literature. Several publications described accountability as a core ethical and safety requirement, but fewer defined who must act when a model drifts, produces inequitable performance, or contributes to patient harm [3, 4, 6]. Health system oversight frameworks proposed committees, review pathways, and documentation structures, yet operational details often remained local and incompletely evaluated [17, 18]. This gap is consequential because ambiguous accountability may delay response when an artificial intelligence system becomes unsafe, unfair, or clinically inappropriate.

Regulatory readiness is in its infancy

Regulatory readiness is growing in importance but remains unevenly developed in practice. Reporting guidelines and device clearance analyses have clarified the need for documentation, evaluation transparency, lifecycle management, and post-deployment oversight [12-16]. However, many health systems appear to lack integrated governance records showing how models are monitored, updated, audited, and aligned with external requirements [17, 18]. The literature suggests that regulatory readiness will require not only technical validation, but also durable documentation systems, change-control processes, and evidence of ongoing institutional oversight.

The integration deficit

A central finding of this review is the integration deficit between governance domains. Monitoring, drift detection, bias auditing, accountability, and regulatory documentation are frequently discussed separately, even though deployed healthcare models require all of them to function together [2, 18, 22]. A model may be well calibrated overall while performing poorly for underserved subgroups, or it may satisfy documentation requirements while lacking a clear incident response pathway [7, 17, 28]. Integrated governance systems should therefore connect model health, fairness, workflow impact, accountability, and compliance evidence into a coherent lifecycle process.

Figure 2 synthesizes the review's central finding that effective healthcare artificial intelligence governance depends on integrating technical surveillance functions with organizational accountability and lifecycle regulatory readiness.

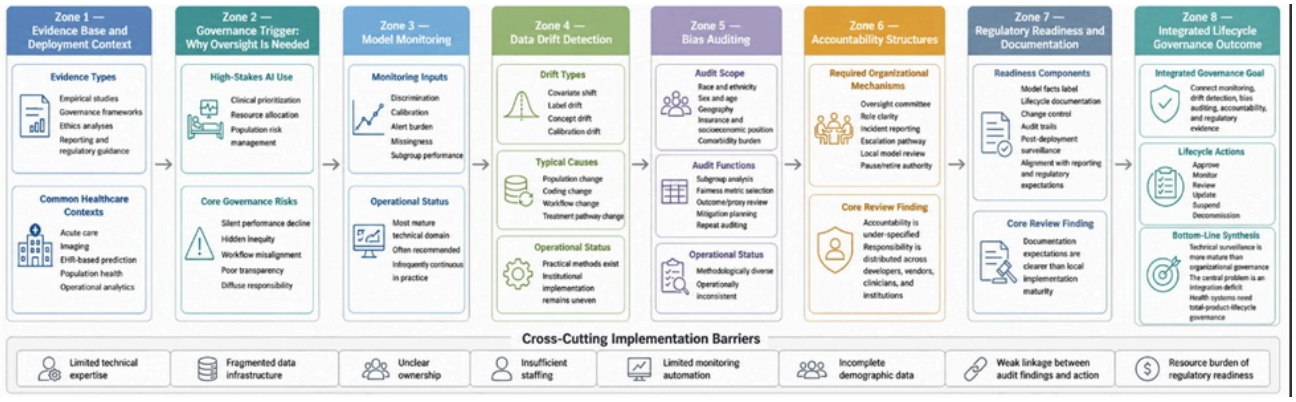


Figure 2. Evidence-to-Implementation Map of Artificial Intelligence Governance for Healthcare Analytics

Toward total-product-lifecycle governance

The literature increasingly supports a total-product-lifecycle approach to healthcare artificial intelligence governance. This approach shifts attention from pre-deployment validation alone to continuous surveillance, structured review, updating, and decommissioning when models no longer meet safety, fairness, or usefulness criteria [2, 14, 22]. Device regulation studies and reporting guidelines reinforce the importance of lifecycle evidence, especially as artificial intelligence systems change over time or operate across heterogeneous institutions [12-16]. For healthcare analytics, lifecycle governance should become a routine condition of deployment rather than an exceptional response to failure.

Limitations

Review limitations

This review was limited by English-language inclusion, reliance on peer-reviewed literature, and the rapidly evolving nature of healthcare artificial intelligence regulation. Some institutional governance practices may exist in internal policies, vendor documentation, or operational dashboards that are not published in academic journals [17, 18]. The search covered literature through the requested 2017 to 2026 window, but regulatory expectations and implementation practices may continue to change quickly after publication [15, 16]. Because the evidence base was heterogeneous, the synthesis was narrative rather than quantitative, which limits direct comparison across governance interventions [29].

Evidence base limitations

The evidence base itself was constrained by a high proportion of conceptual frameworks, ethics analyses, reporting guidance, and early implementation descriptions. Although these publications are valuable, fewer studies described sustained governance programs with longitudinal monitoring, repeated bias audits, documented accountability actions, and measured effects on clinical practice [2, 6, 17, 18]. Empirical studies of drift and bias provided important warnings, but they often focused on specific models, settings, or data modalities, limiting generalizability across healthcare analytics

[8, 20, 21, 24, 25, 28]. As a result, the review found stronger evidence for what governance should include than for which governance models are most effective in routine health system operations.

Comparison with prior reviews

Prior reviews and guidance documents have often focused on either high-level ethical principles or narrower technical domains such as reporting quality, fairness, or prediction model standards. Work on machine learning ethics in medicine and artificial intelligence quality criteria has provided important conceptual foundations, particularly around transparency, safety, and responsible design [4, 5, 29]. Fairness-focused literature has clarified how demographic and structural inequities can enter healthcare algorithms through data, labels, outcomes, and deployment contexts [7, 9, 11]. However, these strands have not always been synthesized into a unified account of operational governance across the full healthcare analytics lifecycle.

This review differs by organizing the literature around the practical governance pipeline that health systems must manage after deciding to deploy artificial intelligence. It synthesizes model monitoring, calibration surveillance, data drift detection, bias auditing, accountability structures, regulatory readiness, and lifecycle documentation as interdependent rather than separate concerns [17,18, 20-23]. This broader framing is consistent with calls for responsible machine learning that begins before deployment and continues through monitoring, updating, and institutional oversight [2]. It also reflects emerging reporting and evaluation guidance that treats clinical artificial intelligence as a complex intervention requiring transparent description of human-AI interaction and implementation context [12-14].

The comparison with prior work highlights a persistent gap between governance proposals and routine implementation. Several publications recommend oversight committees, model facts labels, reporting standards, and lifecycle documentation, but fewer provide evidence that these mechanisms are consistently embedded in operational health system workflows [12, 17, 18]. Empirical studies of deployed models and drift demonstrate that governance failures may become visible only after local validation, subgroup analysis, or post-deployment surveillance [19, 24, 25]. This review therefore extends prior discussions by emphasizing the operational distance between knowing what responsible governance requires and maintaining it reliably in everyday healthcare analytics.

Recommendations

For health systems

Health systems should establish cross-functional artificial intelligence governance committees with authority over model approval, monitoring, auditing, updating, suspension, and decommissioning. These committees should include clinical leaders, data scientists, informaticians, quality and safety personnel, legal or compliance representatives, patient-safety experts, and equity stakeholders [6, 17, 18]. Governance pathways should define who receives alerts when model performance deteriorates, who evaluates subgroup harms, and who can pause or retire a model when risks outweigh benefits [20-22]. Local governance should also require model facts labels or equivalent documentation so that end users understand model purpose, limitations, validation context, and monitoring status [17].

For tool developers

Tool developers should prioritize interoperable governance tools that can be embedded into existing healthcare analytics and machine learning operations pipelines. Monitoring dashboards should combine calibration, discrimination, missingness, input distribution changes, alert frequency, and subgroup performance rather than presenting isolated technical indicators [20, 21, 24, 25]. Bias audit functions should support clinically meaningful subgroup definitions, transparent metric selection, and longitudinal comparison after workflow or population changes [7, 10, 26, 27]. Because governance tools will be used by multidisciplinary teams, developers should also emphasize usability, audit trails, explainable documentation, and integration with health system review processes [1, 17, 18].

For regulators

Regulators should provide healthcare-specific guidance on adequate post-deployment monitoring, acceptable documentation, model update procedures, and evidence expectations for artificial intelligence used in clinical and operational analytics. Existing reporting and evaluation guidance has clarified important elements of trial protocols, clinical reports, and early-stage artificial intelligence evaluation, but health systems still need practical standards for continuous governance after deployment [12-14]. Regulatory expectations should address fairness auditing, drift surveillance, human oversight, incident reporting, and change management across the model lifecycle [15, 16]. Clearer guidance would help institutions distinguish between minimal documentation and governance processes capable of detecting and responding to real-world risks.

Research gaps

Longitudinal studies of governed AI

A major research gap is the absence of longitudinal studies tracking artificial intelligence systems under continuous governance over multiple years. Drift detection studies show that model performance and calibration can change over time, but few studies connect these changes to sustained institutional responses such as retraining, temporary suspension, subgroup review, or decommissioning [20-22, 24, 25]. Bias studies likewise demonstrate important inequities, yet rarely evaluate repeated audits after mitigation or workflow redesign [8, 26, 28]. Future research should follow deployed models across the full lifecycle to determine whether governance reduces harm, maintains fairness, and supports appropriate clinical use.

Cost-effectiveness of governance

The cost-effectiveness of comprehensive artificial intelligence governance remains poorly studied. Health systems must invest in data pipelines, monitoring infrastructure, personnel, legal review, committee time, documentation, and technical maintenance, but the literature rarely quantifies these resource requirements [1, 17, 18]. At the same time, the potential costs of inadequate governance include patient harm, inequitable allocation, regulatory exposure, clinician mistrust, and wasted investment in unsafe or ineffective models [6, 19, 15]. Economic evaluations should therefore compare the cost of governance infrastructure with the avoided costs of model failure, bias, drift, and unsafe deployment.

Accountability in autonomous and agentic systems

Current accountability models may be insufficient for more autonomous and agentic artificial intelligence systems in healthcare analytics. Existing literature already shows that accountability is difficult when responsibility is distributed across developers, vendors, health systems, clinicians, data stewards, and regulators [3, 4, 6]. As systems become more adaptive or capable of generating recommendations across multiple workflows, the challenge of assigning responsibility for monitoring, review, and harm response will become more complex [2, 14, 16]. Research is needed to define accountability structures that remain effective when artificial intelligence systems are updated frequently, embedded deeply in operations, or partially autonomous in their recommendations.

Table 2 highlights the review's central integration deficit by showing how each governance domain remains vulnerable when technical methods are not coupled to clear institutional responses and evaluated implementation pathways.

Table 2. Integration Gap Matrix Linking Governance Domains to Operational Failure Modes, Institutional Responses, and Research Priorities

Governance Domain	What the Literature Commonly Provides	Persistent Operational Gap	Likely Failure Mode if Gap Persists	Priority Institutional Response	Priority Research Need

Model monitoring	Metrics for discrimination, calibration, and temporal performance; proposals for dashboards and monitoring processes [17, 18, 20, 21, 24]	Monitoring is often not continuous, automated, or embedded into routine workflows	Silent deterioration of model reliability after deployment	Build routine surveillance pipelines with explicit thresholds, ownership, and review cadence	Longitudinal evaluations of continuous monitoring programs
Data drift detection	Methods for detecting dataset shift, covariate shift, concept drift, and calibration drift [20-25]	Technical signals are not consistently connected to organizational action	Drift is detected but not acted upon in time	Link drift alerts to escalation, investigation, recalibration, retraining, and pause/retire decisions	Comparative studies of drift response strategies in live health systems
Bias auditing	Subgroup evaluation frameworks and multiple fairness metrics [7-11, 26-28, 30]	No consistent standard for subgroup definitions, fairness criteria, audit frequency, or mitigation follow-through	Episodic fairness review without sustained corrective action	Standardize local audit workflows and require repeat audits after model or workflow change	Prospective studies of whether repeated bias auditing reduces inequitable outcomes
Accountability structures	Ethical calls for responsibility and some examples of committees, review boards, and oversight frameworks [3, 4, 6, 23, 24]	Responsibility for action remains diffuse across developers, vendors, clinicians, and institutions	Delayed response to safety, equity, or workflow harms because no one has clear authority	Establish formal governance charters, named model owners, and explicit pause/retire authority	Studies testing which accountability models best support timely and fair governance action
Regulatory readiness	Reporting standards, evaluation guidance, and device-related lifecycle expectations [12-16]	Local governance records are often incomplete or disconnected from external requirements	Institutions cannot demonstrate defensible oversight, update control, or audit readiness	Build documentation systems that integrate model facts labels, change control, audit trails, and surveillance evidence	Research on practical healthcare-specific post-deployment regulatory governance models

Integrated governance across domains	General frameworks recognizing fairness, safety, transparency, and lifecycle management [2, 18, 29]	Governance functions are discussed separately rather than implemented as one coordinated system	Institutions satisfy isolated governance tasks but still fail to manage overall model risk	Create a unified lifecycle governance program that connects monitoring, auditing, accountability, and compliance	Implementation science on total-product-lifecycle governance effectiveness
Workforce and infrastructure capacity	Recognition of staffing, data, and tooling barriers [1, 6, 17, 18]	Governance responsibilities exceed institutional capacity in many health systems	Governance remains symbolic, fragmented, or reactive	Invest in interdisciplinary governance teams, interoperable tools, and training	Economic and cost-effectiveness studies of governance infrastructure
Future autonomous or agentic AI systems	Early recognition that more adaptive systems will intensify oversight challenges [2, 14, 16]	Existing governance models may not fit adaptive or partially autonomous systems	Accountability failure and weak change control in rapidly evolving systems	Develop anticipatory governance rules for adaptive updating, role assignment, and review boundaries	Research defining accountability for adaptive and agentic healthcare AI

Implications

For practice

For healthcare practice, the central implication is that governance should be a prerequisite for artificial intelligence deployment rather than a post-deployment correction. Before implementation, health systems should define the model's intended use, validation requirements, monitoring plan, subgroup audit plan, documentation record, and escalation pathway [17, 18]. After deployment, governance should include periodic review of calibration, performance, input distributions, fairness, alert burden, and unintended workflow effects [20-22, 24]. Although these activities require resources, the literature suggests that unmanaged models can create clinical, ethical, operational, and regulatory risks that are more difficult to correct after harm has occurred [6, 8, 19].

For research

For research, this review suggests that governance science should become as rigorous as model development science. Studies should move beyond demonstrating model performance to evaluating whether monitoring systems, bias audits, documentation processes, and accountability structures work in real-world healthcare settings [2, 18, 29]. Reporting guidelines for artificial intelligence trials and early-stage evaluations offer a foundation for more transparent evidence generation, but they should be complemented by studies of post-deployment governance outcomes [12-14]. Future research should also include implementation science methods to understand why governance succeeds in some health systems and remains symbolic or fragmented in others [1, 17].

Conclusion

The governance of artificial intelligence in healthcare analytics is at a critical juncture. Tools for monitoring, auditing, documenting, and evaluating artificial intelligence systems exist, and regulatory expectations are becoming more defined,

but operational practice remains uneven across health systems.

Model monitoring and drift detection are the most advanced technical components of the governance landscape. Bias auditing is increasingly recognized as essential, yet it remains methodologically diverse, inconsistently applied, and too often disconnected from sustained corrective action.

Accountability structures remain particularly under-developed. Health systems need clearer procedures for assigning responsibility, reviewing incidents, escalating risks, updating models, communicating limitations, and decommissioning systems that no longer meet safety, fairness, or usefulness standards.

A sustained, multi-stakeholder effort is needed to translate governance principles into everyday operational practice. Healthcare artificial intelligence can deliver value only if it is governed continuously, transparently, fairly, and accountably across the full lifecycle of development, deployment, monitoring, revision, and retirement.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Apr 2026 Revised: 01 Jun 2026 Accepted: 17 Jun 2026
Published online: 20 July 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.

Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981.

Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med. 2018;15(11):e1002689.

Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA. 2020;323(6):509-10.

Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bull World Health Organ. 2020;98(4):251.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-72.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544-7.

Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21(2):167-79.

Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns. 2021;2(10).

Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health. 2020;2(10):e537-48.

Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2(10):e549-60.

Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377.

Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118.

Muehlematter UJ, Bluethgen C, Vokinger KN. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. Lancet Digit Health. 2023;5(9):e618-26.

Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020;3(1):41.

Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc. 2022;29(9):1631-6.

Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70.

Davis SE, Greevy Jr RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. 2019;26(12):1448-57.

Davis SE, Greevy Jr RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform. 2020;112:103611.

Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-6.

Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345-52.

Rahmani K, Thapa R, Tsou P, Chetty SC, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Med Inform. 2023;173:104930.

Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun. 2024;15(1):1887.

Chen RJ, Wang JJ, Williamson DF, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023;7(6):719-42.

Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging. 2023;10(6):061104.

Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176-82.

de Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5(1):2.

Ratwani RM, Sutton K, Galarraga JE. Addressing AI algorithmic bias in health care. JAMA. 2024;332(13):1051-2.