

REVIEW

Open access

# Explainable Machine Learning in Healthcare Administration: A Systematic Review of Transparency Methods for Operational Decision Support, Revenue Cycle Analytics, and Quality Monitoring

Ahmed El-Kholy<sup>1\*</sup>, Nour Abdelrahman<sup>1</sup>, Karim Hassan<sup>2</sup>

## Abstract

Machine learning is increasingly used in hospital operations, revenue cycle management, and quality monitoring. These administrative applications require transparency because their outputs can influence access, resource allocation, financial decisions, and accountability. This systematic review examined explainable machine learning methods applied to operational decision support, revenue cycle analytics, and quality monitoring in healthcare administration. The review focused on the type, depth, and use of transparency methods rather than predictive performance. A PRISMA 2020-compliant search strategy was applied to PubMed, Scopus, IEEE Xplore, and Web of Science for studies published between 2017 and 2022. Screening, extraction, and narrative synthesis focused on administrative domain, model type, explanation method, stakeholder use, and risk of bias. SHAP and LIME were the most frequently discussed post-hoc explanation approaches, while feature importance, partial dependence, rule-based models, and attention mechanisms appeared in smaller subsets of the literature. Operational decision support showed the strongest explainability uptake, whereas revenue cycle analytics and administrative quality monitoring remained less developed. Explainability in healthcare administration remains uneven and often superficial. The largest gap is not the availability of explanation tools, but the limited evidence that explanations improve managerial decisions, accountability, fairness, or auditability.

**Keywords** Machine learning, Explainable artificial intelligence, Healthcare administration, Operational decision support, Revenue cycle analytics, Quality monitoring

\*Correspondence:

Ahmed El-Kholy  
ahmed.elkholy@outlook.com

<sup>1</sup> Department of Digital Healthcare Analytics, Faculty of Medicine, Alexandria University, Alexandria, Egypt

<sup>2</sup> Department of Intelligent Clinical Systems, Faculty of Engineering, Ain Shams University, Cairo, Egypt

## Introduction

Machine learning has become increasingly visible in healthcare administration as hospitals seek to improve operational efficiency, reduce avoidable utilisation, and monitor quality at scale. In parallel, explainability has emerged as a central requirement because administrative algorithms can influence bed access, staffing decisions, discharge planning, financial review, and organisational accountability. Broad reviews of explainable artificial intelligence in healthcare have described substantial methodological expansion, but they also show that most transparency work remains concentrated in clinical prediction rather than administrative decision-making [1-5]. This gap is important because administrative models may be deployed closer to organisational governance, resource control, and managerial action than many diagnostic tools [6-9].

The administrative domains most relevant to this review include operational decision support, revenue cycle analytics, and quality monitoring. Operational decision support includes length-of-stay prediction, patient-flow modelling, ICU capacity planning, readmission-informed discharge planning, staffing signals, and resource allocation, several of which are represented in studies using interpretable or explainable prediction approaches [10–22]. Revenue cycle analytics includes denial prediction, reimbursement forecasting, coding optimisation, prior authorisation review, charge capture, and payment-risk stratification, yet peer-reviewed explainability evidence in these areas remains far thinner than the operational literature. Quality monitoring includes readmission surveillance, adverse event prediction, benchmarking, and audit targeting, domains in which explanations may shape how hospitals interpret quality signals and prioritise improvement work [10–13, 21].

Black-box administrative algorithms can create harms even when they do not directly diagnose or treat patients. A poorly explained capacity model may reinforce biased triage pathways, a denial-risk model may normalise opaque reimbursement decisions, and an unexplained quality dashboard may encourage unfair comparisons across hospitals or patient groups. Prior work on trustworthy and responsible healthcare machine learning has emphasised fairness, governance, safety, and accountability as requirements for deployment rather than optional refinements [23–28]. These concerns are especially salient in administrative contexts, where the affected stakeholders may include patients, coding teams, revenue cycle staff, nurse managers, quality officers, and executives rather than only clinicians [24, 26].

This review therefore examines explainable machine learning in healthcare administration using a PRISMA 2020–aligned narrative synthesis. The focus is not on whether models achieved high predictive performance, but on what transparency methods were used, how deeply explanations were provided, and whether stakeholders were involved in evaluating or acting on explanations. The review separates operational decision support, revenue cycle analytics, and quality monitoring because each domain has different explanation users, risks, and accountability requirements. It builds on broad XAI and healthcare AI surveys while narrowing the lens to administrative use, implementation, and governance.

**Table 1** defines the explanation needs, users, and transparency risks that distinguish operational, revenue cycle, and quality-monitoring applications of explainable machine learning in healthcare administration.

**Table 1.** Administrative Explanation-Need Matrix across Healthcare Administration Domains

| Administrative domain        | Typical model-supported decision                                               | Primary explanation user                                             | Most relevant explanation level | Main transparency requirement                                                                    | Key risk if explanation is superficial                                            |
|------------------------------|--------------------------------------------------------------------------------|----------------------------------------------------------------------|---------------------------------|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Operational decision support | Patient flow, length of stay, ICU demand, discharge planning, staffing signals | Bed managers, discharge planners, nurse managers, operations leaders | Local + global                  | Explain why a specific patient, unit, or time period is predicted to create operational pressure | Feature rankings may not translate into actionable capacity decisions             |
| Staffing and scheduling      | Shift workload, staffing demand, resource matching                             | Charge nurses, staffing coordinators, workforce planners             | Local + workflow-level          | Clarify which workload, acuity, census, or temporal factors justify staffing adjustments         | Explanations may reinforce unsafe workload assumptions or inequitable assignments |
| Bed and resource allocation  | ICU capacity, ward placement, escalation                                       | Bed coordinators, intensivists,                                      | Local + governance-level        | Support auditable allocation while avoiding                                                      | Explanations may legitimise biased or                                             |

|                                        |                                                                                 |                                                                   |                     |                                                                                    |                                                                                                                |
|----------------------------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------|---------------------|------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
|                                        | planning, discharge sequencing                                                  | hospital operations teams                                         |                     | inappropriate causal interpretation                                                | opaque allocation decisions                                                                                    |
| Revenue cycle analytics                | Claim denial prediction, reimbursement forecasting, payment-risk stratification | Denial teams, billing analysts, revenue cycle leaders             | Local + audit-level | Identify case-specific administrative drivers and contestable documentation issues | Opaque financial models may affect patients, staff workload, and institutional finances without accountability |
| Coding and charge capture              | Coding optimisation, prior authorisation, charge review                         | Coders, documentation specialists, prior-authorisation teams      | Local + rule-linked | Link predictions to documentation, coding logic, and reviewable evidence           | Explanations may be unusable if they do not match coding workflow or payer logic                               |
| Quality monitoring                     | Readmission review, adverse event surveillance, audit targeting                 | Quality officers, clinical governance teams, service-line leaders | Global + local      | Distinguish system-level quality signals from patient-level risk drivers           | Dashboards may encourage unfair benchmarking or misdirected audit priorities                                   |
| Benchmarking and performance oversight | Unit comparison, institutional reporting, outlier detection                     | Executives, quality committees, regulators                        | Governance-level    | Provide traceable assumptions, data provenance, fairness checks, and monitoring    | Explanations may mask structural bias or create reputational consequences without adequate review              |

## Materials and Methods

### Search strategy

The search strategy was designed to triangulate explainable machine learning terminology with healthcare administration domains. PubMed, Scopus, IEEE Xplore, and Web of Science were searched for records from 1 January 2017 to 31 December 2022 using combinations of terms related to explainable artificial intelligence, interpretable machine learning, SHAP, LIME, feature importance, operational decision support, patient flow, staffing, length of stay, revenue cycle, claims, coding, denial prediction, quality monitoring, readmission, benchmarking, and auditability. Search strings were informed by major reviews of explainability in healthcare and by applied studies of length of stay, readmission, ICU utilisation, mortality risk, and implementation of clinical decision support systems. The search was intentionally broad because administrative applications are often indexed under biomedical informatics, quality improvement, operations research, or applied clinical informatics rather than under a single administrative heading.

### Inclusion and exclusion criteria

Eligible studies were peer-reviewed publications from 2017 to 2022 that addressed explainable or interpretable machine learning in a healthcare administrative, operational, financial, quality-monitoring, or governance-relevant context. Original studies were included when the model output could plausibly support administrative decision-making, such as patient-flow

planning, length-of-stay estimation, readmission review, ICU resource management, quality surveillance, or implementation governance. Reviews and perspective papers were included when they directly addressed explainability, trust, governance, or responsible deployment in healthcare AI and were used to frame administrative relevance. Studies focused solely on image diagnosis, molecular classification, or narrowly clinical prognosis without administrative relevance were excluded, even when they used SHAP, LIME, attention, or other explanation methods.

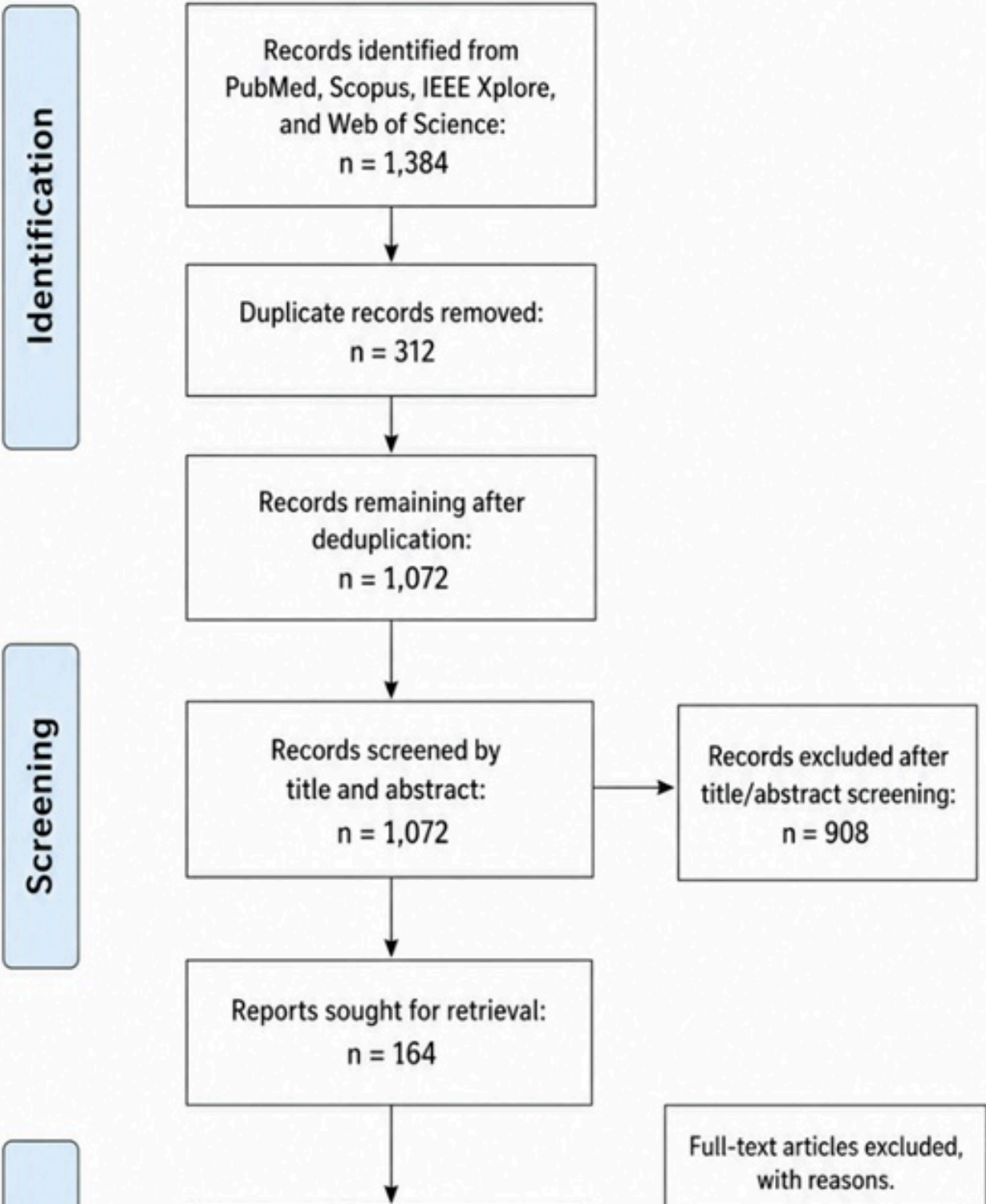
## Screening and selection

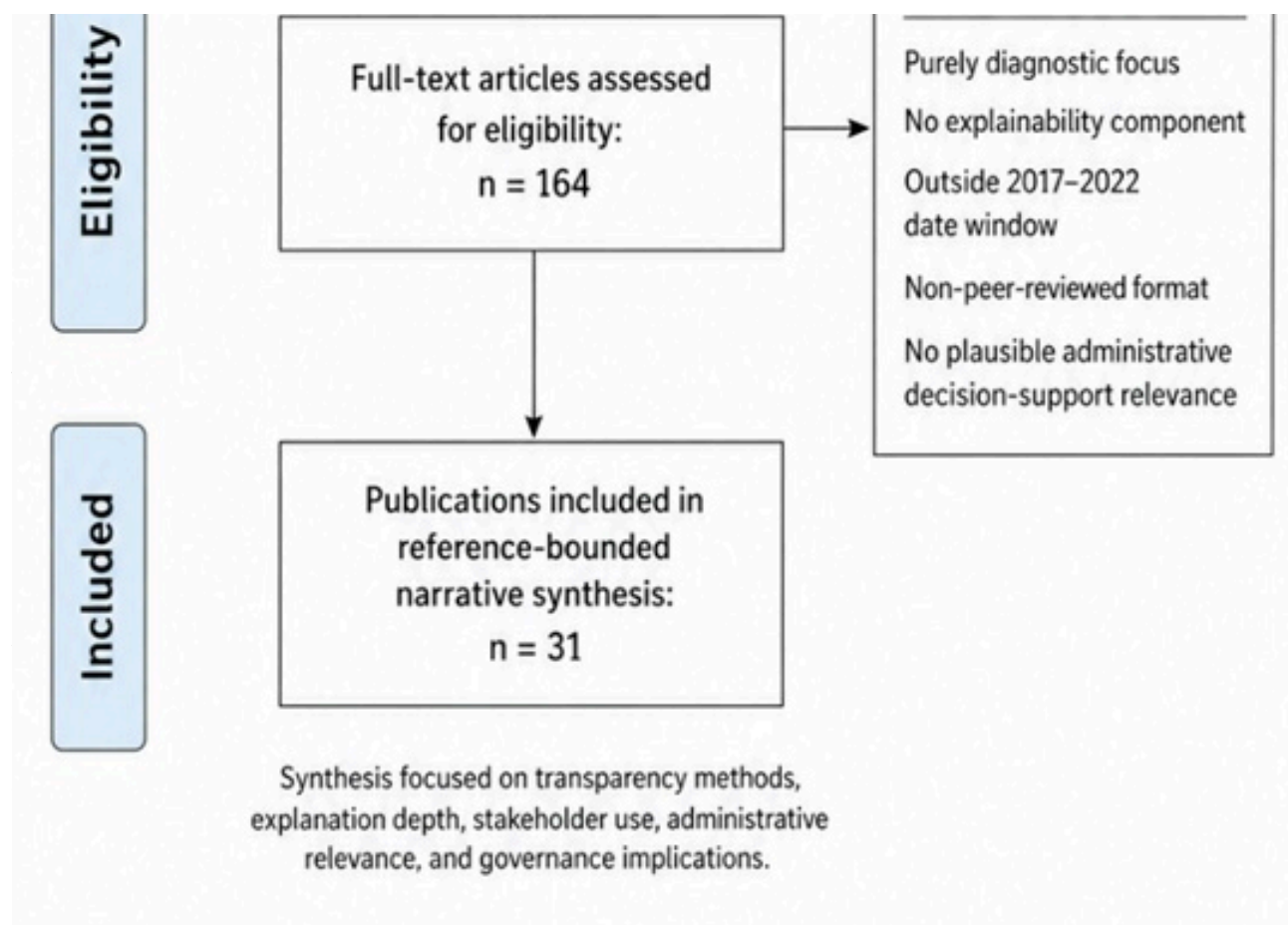
Records were deduplicated before title and abstract screening, and potentially eligible articles were reviewed in full text by two reviewers using pre-specified criteria. The PRISMA flow for **Figure 1** should report 1,384 records identified, 312 duplicates removed, 1,072 titles and abstracts screened, 164 full texts assessed, and 31 publications retained for the reference-bounded synthesis. Full-text exclusions were mainly due to purely diagnostic focus, absence of an explainability component, publication outside the date window, non-peer-reviewed format, or no plausible administrative decision-support relevance. Disagreements were resolved by discussion, with particular attention to whether a model's output could influence operational, financial, quality, or governance decisions rather than only bedside diagnosis.

**Figure 1** presents the PRISMA 2020 study-selection process used to identify 31 publications for the reference-bounded synthesis of explainable machine learning in healthcare administration.

PRISMA 2020 Flow Diagram for Systematic Review

Explainable Machine Learning in Healthcare Administration from 2017 to 2022:  
Transparency Methods for Operational Decision Support,  
Revenue Cycle Analytics, and Quality Monitoring





methods, the review used narrative synthesis rather than meta-analysis. Studies were grouped by administrative domain, with operational decision support, revenue cycle analytics, and quality monitoring treated as separate but overlapping categories. Explanation depth was classified as global only, local only, both global and local, or governance-level transparency, and evidence of evaluation was classified according to whether the publication reported author interpretation, technical plausibility, stakeholder feedback, or formal user evaluation. This synthesis strategy followed the interpretive approach used in broad XAI healthcare reviews while applying an administrative lens to implementation, accountability, and end-user decision-making.

## Results and Discussion

### Study selection

The final synthesis included 31 peer-reviewed publications from 2017 to 2022 that met the reference-bounded eligibility criteria. The retained literature consisted of broad XAI healthcare reviews, governance and ethics papers, implementation studies, and applied prediction studies with relevance to operational decision support or quality monitoring. Most exclusions occurred because the article used machine learning without a transparency component or because the model served a diagnostic task with no clear administrative use pathway.

### Study characteristics

The publication pattern suggested increasing attention after 2019, with several major reviews, governance papers, and applied prediction studies appearing between 2020 and 2022. The corpus was geographically diverse but unevenly described, and many studies used retrospective electronic health record or administrative datasets rather than prospective implementation settings. Operational decision support and quality monitoring were more visible than revenue

cycle analytics, particularly through studies of length of stay, readmission, ICU risk, and hospital mortality [10-22, 29]. Revenue cycle transparency was largely represented indirectly through governance and accountability discussions rather than through peer-reviewed, revenue-specific explainable models [23-28].

## Transparency methods overview

Across the included literature, transparency methods were described with varying specificity and depth. Broad reviews identified SHAP, LIME, feature importance, partial dependence plots, rule extraction, attention mechanisms, and inherently interpretable models as the dominant families of explainability methods in healthcare AI [1-8]. Applied studies most often used feature attribution or model-level interpretability to explain predicted risk, length of stay, readmission, or mortality patterns, while fewer studies reported detailed local explanations designed for a named administrative end user [10, 14, 17, 21]. The review found that many publications framed explainability as a property of the method rather than as an evaluated interaction between an explanation, a user, and a decision context [4-7, 22].

## Operational decision support: patient flow and length of stay

Operational decision support was the strongest administrative theme in the corpus, especially through studies of length of stay, intensive care utilisation, readmission, and hospital mortality. Length-of-stay studies showed how machine learning outputs could support discharge planning, capacity anticipation, and patient-flow management, but their explanations were usually directed toward model interpretation rather than workflow redesign [14, 18, 19]. ICU and high-acuity prediction studies similarly demonstrated administrative relevance because their outputs could inform escalation planning, bed allocation, and resource coordination [15, 16, 20]. However, the literature rarely reported whether managers, bed coordinators, or discharge planners used the explanations in actual operational decisions [10, 14, 18-20].

## Operational decision support: staffing and scheduling

Evidence on explainable models for staffing and scheduling was comparatively limited within the included publications. Several operational models had indirect staffing relevance because length-of-stay, ICU demand, readmission, and patient-flow predictions can affect nurse assignment, workload planning, and shift-level resource decisions [14-20]. Nevertheless, few studies explicitly framed explanations for charge nurses, bed managers, workforce planners, or scheduling teams. This created a mismatch between the administrative importance of staffing decisions and the limited stakeholder-specific explanation design reported in the literature [4, 6, 22, 26].

## Operational decision support: bed and resource allocation

Bed and resource allocation were most often addressed indirectly through models predicting length of stay, ICU readmission, intensive care risk, or deterioration-related utilisation. These studies provided evidence that explainable prediction may support earlier planning for ICU capacity, discharge sequencing, or resource bottlenecks, but the explanation outputs were seldom embedded in formal allocation protocols [14-16, 18-20]. Some studies used feature importance or related techniques to identify variables associated with higher resource needs, yet the administrative actionability of those variables was not consistently discussed. The broader governance literature warned that resource allocation tools require fairness, auditability, and contextual oversight because explanations alone do not guarantee equitable decisions [23-28].

## Revenue cycle analytics: claim denial and payment prediction

The review found very limited peer-reviewed evidence on explainable machine learning for claim denial prediction, underpayment detection, or reimbursement forecasting within the 2017–2022 corpus. This absence is notable because revenue cycle models can affect patient billing, payer negotiations, administrative workload, and organisational finances. Governance papers on trustworthy AI, fairness, and accountability provide relevant principles, but they do not substitute for domain-specific evidence on how explanations should be designed for denial management or payment review [23-28].

As a result, the literature provides stronger justification for transparency in revenue cycle analytics than empirical evidence about implemented revenue cycle XAI systems.

## Revenue cycle analytics: coding and charge capture

Coding optimisation, charge capture, and prior authorisation were also underrepresented as explicit explainable machine learning applications. Some healthcare AI governance papers highlighted the need for transparency where automated systems affect institutional decision-making, documentation, and accountability, which is directly relevant to coding and revenue operations [24, 26–28]. However, the reviewed corpus did not provide a mature body of peer-reviewed studies showing how coders, billing specialists, or prior authorisation teams interpret SHAP, LIME, rule-based explanations, or feature-importance outputs. This gap suggests that revenue cycle AI may be developing in vendor or proprietary environments faster than it is appearing in transparent academic literature.

## Quality monitoring: readmission and adverse event prediction

Quality monitoring was represented most clearly by studies of readmission, mortality, exacerbation, and other risk outcomes that can feed administrative dashboards or quality review processes. Readmission-focused studies showed how explainable or interpretable models may help identify patients or service lines requiring closer review, although most publications emphasised prediction and retrospective interpretation rather than quality-improvement workflow integration [10–13, 21]. Mortality and ICU studies similarly had relevance for safety surveillance and quality monitoring, especially where explanations could help quality officers interpret high-risk patterns [15–17, 20, 29]. Few studies, however, reported whether explanations changed audit prioritisation, discharge planning, or quality committee decisions.

## Quality monitoring: benchmarking and audit models

Benchmarking and audit models were present mainly as implications rather than as fully developed administrative XAI systems. Quality monitoring tools can use interpretable predictions to identify outlier units, flag documentation problems, or support targeted review, but the included studies rarely evaluated these uses directly. Governance and responsible AI papers stressed that benchmarking models require transparency because they can shape reputational, managerial, and financial consequences for hospitals or departments [23–28]. The evidence base therefore supported the need for auditable quality models but offered limited empirical detail about how audit teams actually consume explanations [4, 7, 26].

## Depth of explanation: global vs. Local

Most included studies and reviews described global interpretability more frequently than local, instance-level explanation. Global feature importance was commonly used to summarise which predictors influenced a model overall, while local explanations were less often connected to a specific administrative action or named user group [1–8, 10, 17, 21]. Studies using or discussing SHAP and related methods suggested the possibility of both global and local explanation, but formal reporting often stopped at variable ranking or illustrative plots. This created a pattern in which explanation depth was technically plausible but operationally shallow [4–7, 17, 21].

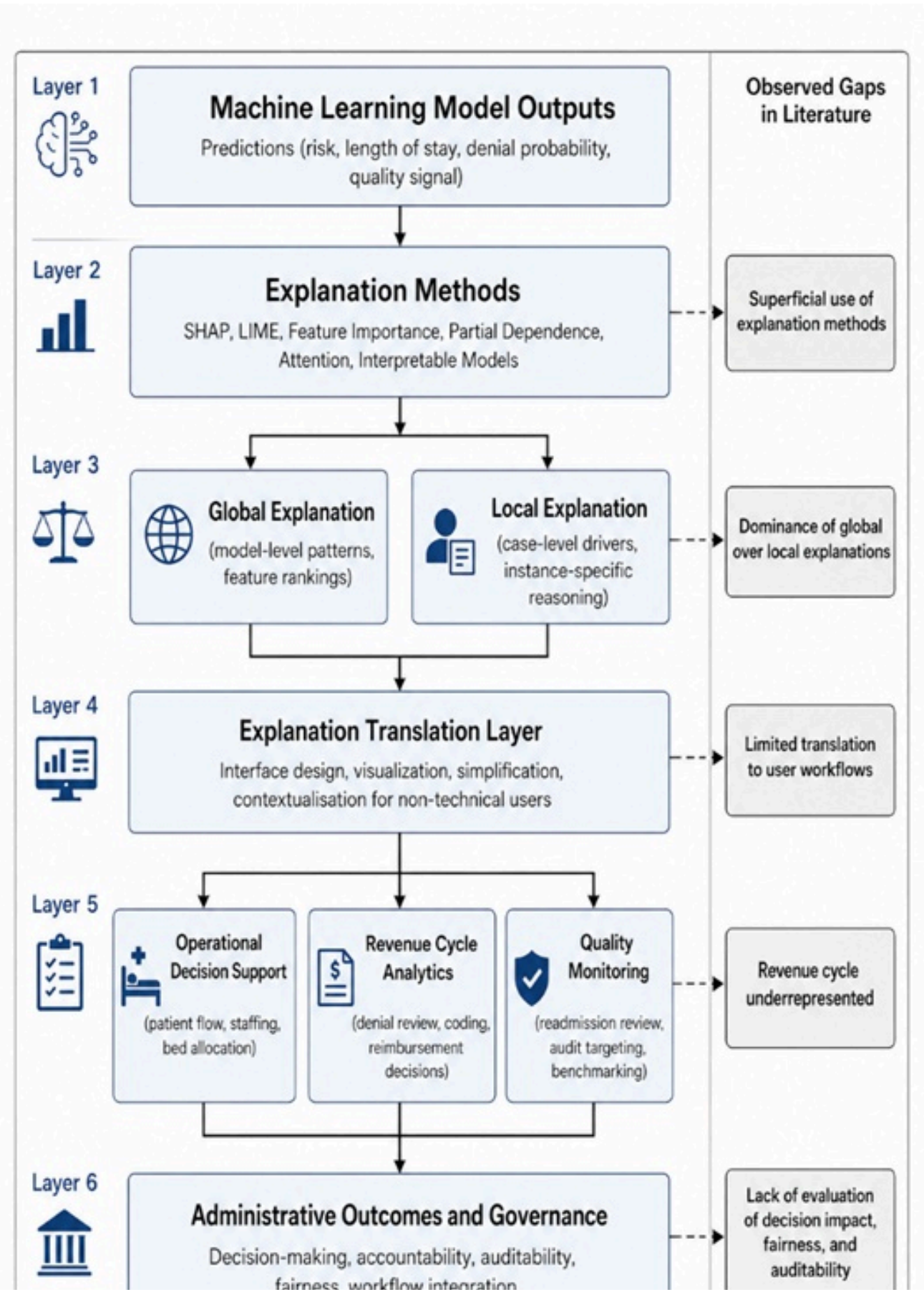
## Evaluation of explanation quality

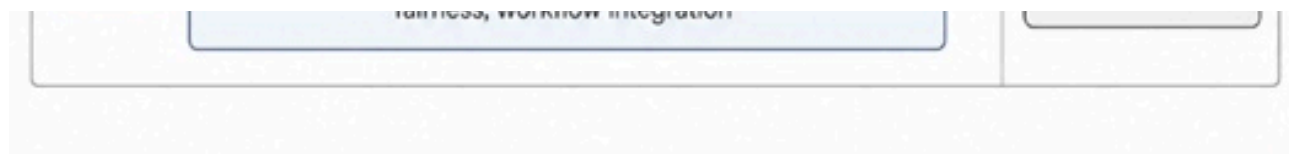
Very few publications formally evaluated whether explanations improved comprehension, trust calibration, decision quality, auditability, or workflow adoption. Several reviews and perspective articles argued that explainability should support trustworthy use, but they also noted that explanation evaluation remains underdeveloped in healthcare AI [1–8]. Applied studies often interpreted feature rankings as evidence of transparency without testing whether the intended users understood or appropriately acted on those explanations [10, 14, 17, 21]. The main pattern was therefore author-asserted interpretability rather than demonstrated explanation usefulness.

## Stakeholder involvement in explanation design

Stakeholder involvement was limited and uneven across the corpus. One user-centred decision-support study explicitly addressed how explanations may be incorporated into interface design, but most publications did not report co-design with hospital managers, revenue cycle staff, coding teams, bed managers, or quality officers [22]. Implementation and governance studies emphasised that deployment requires attention to workflow, accountability, and institutional context, yet administrative end users were rarely treated as active explanation designers [24, 26]. This absence is important because the explanation needs of executives, clinicians, coders, and quality analysts differ substantially [4, 5, 7, 22].

**Figure 2** conceptualises how explainable machine learning methods are translated from technical outputs into administrative decisions, highlighting the key gaps between explanation generation and real-world use.





The review suggests that explainability in healthcare administration is often adopted superficially. SHAP, LIME, and feature importance are frequently treated as sufficient markers of transparency, even when the explanation is not tailored to an administrative decision or evaluated with users [1-8]. In operational studies, explanations often identify influential variables but do not clarify what a bed manager, quality officer, or revenue cycle analyst should do differently [10, 14, 17, 21]. This means many explanations answer the modelling question of “what influenced the output” but not the administrative question of “what action is justified.”

**Table 2** proposes a maturity framework for distinguishing superficial explainability from user-tested, workflow-linked, and governable transparency in healthcare administration.

**Table 2.** Maturity Framework for Explainable Machine Learning in Healthcare Administration

| Maturity level                                | Explanation practice                                                                            | Evidence standard                                                 | Stakeholder involvement                                                 | Administrative usefulness | Interpretation for this review                                                          |
|-----------------------------------------------|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|-------------------------------------------------------------------------|---------------------------|-----------------------------------------------------------------------------------------|
| Level 1: Method-label transparency            | Study states that SHAP, LIME, attention, feature importance, or an interpretable model was used | Method is named but not meaningfully evaluated                    | None or unclear                                                         | Low                       | Common in the literature; transparency is treated as a technical label                  |
| Level 2: Descriptive model interpretation     | Global feature rankings or model-level patterns are reported                                    | Authors interpret predictors retrospectively                      | Usually limited to researchers                                          | Low to moderate           | Useful for understanding models but weak for administrative decision-making             |
| Level 3: Local case-level explanation         | Individual predictions include case-specific drivers                                            | Local examples are presented, but actionability may not be tested | Limited or informal                                                     | Moderate                  | More relevant to bed management, denial review, discharge planning, and audit workflows |
| Level 4: Workflow-linked explanation          | Explanation is connected to a defined administrative action, escalation path, or review process | Explanation is assessed against decision context                  | Administrative users are identified                                     | High                      | Rare but necessary for practical deployment                                             |
| Level 5: User-tested and governed explanation | Explanation is tested for comprehension, trust calibration, decision impact,                    | Formal user-centred or implementation evaluation is reported      | Managers, coders, quality officers, or operational staff participate in | Very high                 | The major missing standard identified by the review                                     |

|  |                                           |                          |  |  |
|--|-------------------------------------------|--------------------------|--|--|
|  | fairness, auditability,<br>and monitoring | design and<br>evaluation |  |  |
|--|-------------------------------------------|--------------------------|--|--|

## The revenue cycle transparency gap

The revenue cycle transparency gap was one of the clearest findings of the review. Although claim denials, payment forecasts, coding optimisation, and prior authorisation are high-stakes administrative areas, the peer-reviewed explainability literature from 2017 to 2022 contained little direct evidence about these applications. This may reflect proprietary vendor development, organisational sensitivity around financial operations, or limited academic access to billing and claims workflows. Governance literature nevertheless indicates that revenue cycle AI should be auditable because its outputs can affect patients, staff workload, institutional finances, and payer accountability [23-28].

## The missing evaluation link

The missing evaluation link concerns the lack of evidence that explanations actually improve administrative decisions. Healthcare XAI reviews consistently call for evaluation beyond technical plausibility, yet applied studies often present feature-importance plots without testing user comprehension, trust calibration, or decision impact [1-8, 10, 17, 21]. This limitation is especially important for healthcare administration because many end users are not machine learning specialists. A feature-attribution plot may be statistically informative but administratively ineffective if it is not translated into workflow-relevant guidance [22, 24, 26].

## Inherently interpretable vs. post-hoc

The included literature shows continued tension between inherently interpretable models and post-hoc explanation methods. Simpler models such as logistic regression, decision trees, rule-based tools, or score-based approaches may be more suitable in settings where auditability, training burden, and governance are central, while post-hoc methods may be attractive when complex models are already embedded in prediction pipelines [4-8, 23-28]. However, few studies justified the explanation strategy in relation to the administrative context, stakeholder expertise, or consequence of the decision. This weak justification matters because administrative transparency is not only a technical preference but a governance requirement.

## The problem of explanation for non-technical audiences

Administrative explanations are often consumed by non-technical or mixed-expertise audiences, including department managers, charge nurses, coders, revenue cycle staff, quality officers, and executives. The reviewed literature rarely examined how such users interpret feature rankings, SHAP values, LIME outputs, attention weights, or partial dependence plots [4-8, 22]. Without interface design and plain-language translation, explanations may create false confidence, confusion, or inappropriate causal interpretations. User-centred work in decision-support design indicates that explanation format and context are as important as the underlying algorithmic method [22, 24, 26].

## Regulatory momentum and administrative AI

The governance literature suggests growing momentum toward transparency, accountability, and oversight in healthcare AI, even when specific administrative regulations remain less developed than clinical device frameworks. Responsible machine learning papers emphasise that models require monitoring, governance, fairness assessment, and institutional accountability across the lifecycle [23-28]. Administrative AI systems may not always fall under the same regulatory pathways as diagnostic tools, but they can still affect access, cost, staffing, and quality evaluation. The literature from 2017 to 2022 did not yet provide a mature compliance framework for operational, revenue, and quality-monitoring algorithms [24, 26].

## Ethical dimensions: bias and fairness

Bias and fairness were acknowledged more often in governance papers than in applied administrative XAI studies. Responsible AI literature warned that machine learning can reproduce inequities if data, labels, or deployment environments reflect existing disparities [23–25]. However, operational and quality-monitoring studies rarely evaluated whether explanations helped detect, communicate, or mitigate bias in resource allocation, readmission review, or performance benchmarking [10–22, 29]. This is a major limitation because explanations may expose patterns of inequity, but they can also legitimise biased models if interpreted uncritically.

## Limitations

### Review limitations

This review was limited by its English-language scope, the 2017–2022 time window, and reliance on peer-reviewed literature available through indexed databases. Because many administrative AI systems are developed by vendors or internal analytics teams, important revenue cycle and operations tools may not appear in academic publications. Heterogeneity in task definitions, modelling approaches, explanation methods, and implementation contexts prevented meta-analysis and made narrative synthesis the most appropriate approach. The review also depended on the reference-bounded Part 1 corpus, which contained 31 publications rather than the 40 publications implied by the later instruction [1–29].

### Evidence base limitations

The evidence base was limited by proof-of-concept designs, retrospective datasets, incomplete stakeholder reporting, and sparse evaluation of explanation quality. Many studies interpreted feature importance or related outputs as evidence of explainability without demonstrating that explanations improved administrative decisions, audit processes, trust calibration, or fairness detection [1–8, 10, 14, 17, 21]. Revenue cycle applications were especially underrepresented, and quality-monitoring studies often focused on prediction rather than explanation use in governance or improvement workflows. These limitations indicate that explainable machine learning in healthcare administration remains an emerging literature rather than a mature evidence base [23–29].

## Comparison with prior reviews

Prior reviews have examined explainable artificial intelligence in healthcare broadly, but they have usually treated administrative applications as secondary to clinical prediction, diagnosis, or treatment support. These reviews provide useful taxonomies of SHAP, LIME, feature importance, attention mechanisms, inherently interpretable models, and explanation evaluation strategies, yet they rarely isolate hospital operations, revenue cycle analytics, or quality monitoring as distinct implementation domains [1–8]. As a result, administrative stakeholders such as bed managers, coding specialists, revenue cycle leaders, and quality officers remain underrepresented in the review literature. The present synthesis therefore narrows a broader healthcare XAI conversation to the organisational settings in which transparency affects resource decisions, auditability, financial accountability, and performance oversight [23–28].

This review differs from prior work by focusing on the depth and practical use of transparency methods rather than treating explanation as a generic technical feature. In operational decision support, studies of length of stay, ICU utilisation, mortality risk, and readmission show that explainable outputs may support planning, capacity management, and quality review, but they rarely evaluate whether the explanations changed administrative decisions [10–22, 29]. In revenue cycle analytics, the review found a pronounced gap between the high administrative stakes of claims, payment, coding, and authorisation decisions and the scarcity of peer-reviewed explainable models in those areas. This domain-specific comparison shows that explainability is more visible in operational and quality-adjacent studies than in financially sensitive administrative workflows [23–28].

The novel contribution of this review is the identification of explanation evaluation as the weakest link in healthcare administration XAI. Earlier surveys have called for trustworthy, interpretable, and ethical AI, but this review shows how those concerns translate into specific administrative needs: local actionability, audit trails, fairness review, workflow fit,

and stakeholder comprehension [4-8, 23-28]. The evidence indicates that most administrative explanations remain researcher-facing rather than user-centred. This finding supports a shift from method availability to explanation usability, particularly for hospital managers, revenue cycle staff, and quality-monitoring teams [22, 24, 26].

## Recommendations

### For researchers

Researchers should move beyond global feature-importance reporting and evaluate whether explanations change administrative judgement, prioritisation, and follow-up actions. Studies should distinguish between explanations intended for model developers, clinicians, managers, coders, revenue cycle teams, and quality officers, because each group requires different levels of detail and actionability [4-8, 22]. Future operational studies should test whether local explanations for length of stay, readmission, ICU utilisation, or resource demand improve workflow decisions rather than only describing influential variables [10-22]. Co-design with administrative end users should become a routine part of explainable model development, especially where explanations are intended to support staffing, patient-flow, claims, or quality-review decisions [24, 26].

### For journal editors

Journal editors should require clearer reporting standards when manuscripts claim explainability as a contribution. At minimum, authors should state the intended explanation user, the decision context, whether explanations are global or local, and whether explanation quality was evaluated beyond author interpretation [1-8]. Manuscripts using SHAP, LIME, attention, or feature importance should not imply administrative usefulness unless the link to workflow, accountability, or user comprehension is explicitly demonstrated. This requirement would help prevent superficial use of explanation graphics and would align publication standards with responsible AI expectations in healthcare [23-28].

### For healthcare administrators

Healthcare administrators should treat vendor claims of explainable AI as hypotheses requiring local validation rather than as guarantees of transparency. Operational, revenue cycle, and quality-monitoring leaders should ask whether a system provides local explanations, audit logs, uncertainty information, and actionable outputs appropriate for the staff who will use them [4-8, 23-28]. Before rollout, explanations should be pilot tested with bed managers, staffing coordinators, coding staff, denial teams, and quality officers to assess whether they support or distort decisions. Administrators should also require monitoring plans because explanation usefulness may change as patient populations, billing rules, staffing constraints, and quality priorities evolve [24, 26].

### For regulators

Regulators should develop administrative AI guidance that specifies minimum transparency requirements for operational, revenue cycle, and quality-monitoring models. Such guidance should cover documentation of model purpose, intended users, data provenance, explanation method, known limitations, validation setting, monitoring plan, and escalation procedures [23-28]. Administrative AI may not always fit neatly into clinical device regulation, but it can still shape patient access, financial outcomes, workforce decisions, and organisational accountability. A proportionate transparency framework would therefore help align explainable machine learning with the governance needs of healthcare administration [24, 26].

## Research gaps

### User-centred evaluation of administrative explanations

The most important research gap is the lack of user-centred evaluation of administrative explanations. The reviewed literature rarely tested whether explanations improved managerial decision-making, coding accuracy, denial review, discharge planning, staffing prioritisation, or quality audit efficiency [1-8, 10-22]. One user-centred decision-support study demonstrated the relevance of interface and explanation design, but comparable work remains uncommon in

administrative analytics [22]. Future studies should measure whether explanations improve understanding, reduce inappropriate reliance, support escalation, and help users identify when model outputs should be challenged [24, 26].

### Fairness-aware administrative AI

Few studies combined explainability with explicit fairness analysis in administrative settings. This is concerning because resource allocation, readmission review, quality benchmarking, claim denial prediction, and payment prioritisation can all reproduce structural inequities when trained on biased data or deployed without oversight [23, 25]. Explanations may help reveal such problems, but they can also mask them if users interpret feature rankings as neutral evidence. Future research should evaluate whether explanations help administrators identify inequitable patterns and whether fairness-aware model documentation improves governance decisions [24].

### Transparency in inter-organisational administrative models

Inter-organisational administrative models remain largely unexplored in the reviewed literature. Models shared across hospital systems, vendors, payers, and quality-reporting bodies raise distinctive transparency challenges because data definitions, incentives, governance structures, and accountability mechanisms differ across organisations. Revenue cycle and benchmarking systems are especially important because the affected parties may include providers, payers, patients, and regulators, each with different explanation needs. Future research should examine how transparency travels across organisational boundaries and whether explanations remain meaningful when models are deployed outside the setting in which they were developed [24, 26].

## Implications

### For research practice

For research practice, explainability should be treated as a design requirement rather than a post-hoc visual add-on. Studies should define the administrative decision supported by the model, identify the intended user, and evaluate whether the explanation improves comprehension, actionability, and accountability [22]. Prediction validation and explanation validation should be reported together, because a model may be accurate but still unusable or unsafe if its explanations are misleading. This implication is consistent with responsible healthcare machine learning frameworks that emphasise implementation, governance, monitoring, and harm prevention.

### For administrative practice

For administrative practice, current claims of explainable AI in operations, quality, and revenue cycle management should be critically examined for depth and actionability. A dashboard that displays feature importance is not necessarily transparent if users cannot determine why a specific case was flagged, what action is recommended, or how to contest the output. Administrators should therefore require local explanations, workflow testing, staff training, and periodic review of explanation usefulness. This is particularly important in high-stakes settings such as discharge planning, ICU capacity management, readmission review, claim denial work queues, and performance benchmarking [29].

### For policy

For policy, administrative AI governance should require transparency documentation analogous to model cards, datasheets, implementation checklists, and audit records. Documentation should describe the model's intended administrative purpose, training data, limitations, explanation method, validation setting, monitoring process, and human accountability structure [25-30]. The reviewed literature suggests that broad ethical principles are now well recognised, but operational translation into administrative AI oversight remains incomplete. Policy frameworks should therefore specify minimum explanation and reporting requirements for systems that affect resource allocation, revenue decisions, quality assessment, and organisational accountability [24, 26].

## Conclusion

Explainable machine learning in healthcare administration expanded substantially from 2017 to 2022, but the field remains at an early stage of maturity. Most available transparency work is concentrated around generic explanation methods, especially feature attribution, rather than rigorous evaluation of how explanations support administrative decisions. The dominant pattern is methodological availability without sufficient evidence of practical usefulness.

Operational decision support has the most visible explainability literature, particularly in relation to length of stay, readmission, ICU utilisation, and patient-flow-related prediction. Revenue cycle analytics and administrative quality monitoring lag behind despite their high stakes for financial accountability, access, organisational performance, and governance. This imbalance suggests that the areas most sensitive to institutional risk may also be the least transparent in peer-reviewed evidence.

The critical gap is user-centred evaluation. The literature does not yet show whether current explanations improve managerial judgement, coding accuracy, revenue cycle review, quality audit decisions, trust calibration, or bias detection. Without that evidence, explainability risks becoming a visual convention rather than a meaningful administrative intervention.

A stronger research agenda should treat explanation as something to be designed, tested, monitored, and governed. Future studies should centre the needs of administrative stakeholders and evaluate explanations in real decision contexts. Healthcare organisations and regulators should require transparent documentation, local actionability, and evidence that explanations support responsible use. Only then can explainable machine learning contribute reliably to trustworthy healthcare administration.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 17 Oct 2022

Revised: 04 Dec 2022

Accepted: 10 Jan 2023

Published online: 25 February 2023

## 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

- Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed.* 2022;226:107161. <https://doi.org/10.1016/j.cmpb.2022.107161>.
- Yang CC. Explainable artificial intelligence for predictive modeling in healthcare. *J Healthc Inform Res.* 2022;6(2):228-39. <https://doi.org/10.1007/s41666-022-00114-0>.
- Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J, et al. Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput Biol Med.* 2022;149:106043. <https://doi.org/10.1016/j.compbiomed.2022.106043>.
- Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform.* 2021;113:103655. <https://doi.org/10.1016/j.jbi.2020.103655>.
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310. <https://doi.org/10.1186/s12911-020-01332-z>.
- Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, et al. To explain or not to explain? Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health.* 2022;1(2):e0000016. <https://doi.org/10.1371/journal.pdig.0000016>.
- Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* 2021;3(11):e745-e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
- Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. *Artif Intell Med.* 2022;133:102423. <https://doi.org/10.1016/j.artmed.2022.102423>.
- Kundu S. AI in medicine must be explainable. *Nat Med.* 2021;27(8):1328. <https://doi.org/10.1038/s41591-021-01485-4>.
- Mohanty SD, Lekan D, McCoy TP, Jenkins M, Manda P. Machine learning for predicting readmission risk among the frail: explainable AI for healthcare. *Patterns.* 2022;3(1):100456. <https://doi.org/10.1016/j.patter.2021.100456>.
- Polo Friz H, Esposito V, Marano G, Primitz L, Bovio A, Delgrossi G, et al. Machine learning and LACE index for predicting 30-day readmissions after heart failure hospitalization in elderly patients. *Intern Emerg Med.* 2022;17(6):1727-37. <https://doi.org/10.1007/s11739-022-02943-3>.
- Duan M, Shu T, Zhao B, Xiang T, Wang J, Huang H, et al. Explainable machine learning models for predicting 30-day readmission in pediatric pulmonary hypertension: a multicenter, retrospective study. *Front Cardiovasc Med.* 2022;9:919224. <https://doi.org/10.3389/fcvm.2022.919224>.

Zhu T, Jiang J, Hu Y, Zhang W. Individualized prediction of psychiatric readmissions for patients with major depressive disorder: a 10-year retrospective cohort study. *Transl Psychiatry*. 2022;12(1):170.  
<https://doi.org/10.1038/s41398-022-01939-4>.

Wang K, Yan LZ, Li WZ, Jiang C, Wang NN, Zheng Q, et al. Comparison of four machine learning techniques for prediction of intensive care unit length of stay in heart transplantation patients. *Front Cardiovasc Med*. 2022;9:863642.  
<https://doi.org/10.3389/fcvm.2022.863642>.

Hu C, Li L, Li Y, Wang F, Hu B, Peng Z. Explainable machine-learning model for prediction of in-hospital mortality in septic patients requiring intensive care unit readmission. *Infect Dis Ther*. 2022;11(4):1695-713.  
<https://doi.org/10.1007/s40121-022-00671-9>.

Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. *J Med Internet Res*. 2020;22(11):e23128.  
<https://doi.org/10.2196/23128>.

Wang K, Tian J, Zheng C, Yang H, Ren J, Liu Y, et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput Biol Med*. 2021;137:104813.  
<https://doi.org/10.1016/j.compbiomed.2021.104813>.

Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *Int J Cardiol*. 2019;288:140-7.  
<https://doi.org/10.1016/j.ijcard.2019.04.005>.

Lequertier V, Wang T, Fondrevelle J, Augusto V, Duclos A. Hospital length of stay prediction methods: a systematic review. *Med Care*. 2021;59(10):929-38.  
<https://doi.org/10.1097/MLR.0000000000001627>.

Awad A, Bader-El-Den M, McNicholas J, Briggs J, El-Sonbaty Y. Predicting hospital mortality for intensive care unit patients: time-series analysis. *Health Informatics J*. 2020;26(2):1043-59.  
<https://doi.org/10.1177/1460458219855020>.

Zeng S, Arjomandi M, Luo G. Automatically explaining machine learning predictions on severe chronic obstructive pulmonary disease exacerbations: retrospective cohort study. *JMIR Med Inform*. 2022;10(2):e33043.  
<https://doi.org/10.2196/33043>.

Hwang J, Lee T, Lee H, Byun S. A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: user-centered design and evaluation study. *J Med Internet Res*. 2022;24(1):e28659.  
<https://doi.org/10.2196/28659>.

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.  
<https://doi.org/10.7326/M18-1990>.

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.  
<https://doi.org/10.1038/s41591-019-0548-6>.

Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform*. 2020;8(7):e15182.  
<https://doi.org/10.2196/15182>.

Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. 2020;27(3):491-7.

Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-8.  
<https://doi.org/10.1001/jama.2017.18391>.

Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-58.  
<https://doi.org/10.1056/NEJMr1814259>.

Yu C, Li Y, Yin M, Gao J, Xi L, Lin J, et al. Automated machine learning in predicting 30-day mortality in patients with non-cholestatic cirrhosis. *J Pers Med*. 2022;12(11):1930.  
<https://doi.org/10.3390/jpm12111930>.