

REVIEW

Open access

Artificial Intelligence for Healthcare Quality Improvement from 2017 to 2024: A Systematic Review of Predictive Models for Safety Events, Care Gaps, Adverse Outcomes, and Performance Monitoring

Elena Petrova^{1*}, Ivan Georgiev¹, Nikolay Stoyanov², Petar Kolev¹

Abstract

Healthcare quality improvement increasingly relies on routinely collected data to identify preventable harm, missed care opportunities, adverse outcomes, and variation in performance. Artificial intelligence predictive models may support earlier detection of quality risks and enable more proactive monitoring than retrospective audits alone. This systematic review examined artificial intelligence predictive models for healthcare quality improvement from 2017 to 2024. The review focused on patient safety events, care gaps, adverse clinical outcomes, and performance monitoring systems. A PRISMA 2020-compliant search was conducted across PubMed, Scopus, Web of Science, and IEEE Xplore for peer-reviewed English-language studies published from 2017 through 2024. Dual screening, structured data extraction, risk-of-bias assessment, and narrative synthesis were used. The evidence base showed growing use of machine learning for pressure injuries, sepsis, readmission, mortality, ICU transfer, and continuous monitoring. However, most studies remained retrospective model-development or validation studies, while fewer described deployment within formal quality improvement workflows. Technical progress in predictive modelling for quality improvement is substantial, but evidence of sustained improvement in care processes, safety outcomes, or organisational performance remains limited. Stronger prospective evaluation and clearer integration with improvement methods are needed.

Keywords Artificial intelligence, Predictive analytics, Patient safety, Healthcare quality improvement, Adverse outcomes, Performance monitoring

*Correspondence:

Elena Petrova

elena.petrova@gmail.com

¹ Department of Health Informatics and Digital Systems, Faculty of Medicine, Medical University of Sofia, Sofia, Bulgaria

² Department of Intelligent Clinical Analytics, National University of Sciences and Technology, Islamabad, Pakistan

Introduction

Healthcare quality improvement has traditionally relied on retrospective audits, incident reports, administrative indicators, and periodic performance reviews. These approaches remain important, but they often identify harm after it has occurred rather than supporting earlier intervention. Patient safety and quality monitoring studies have therefore increasingly examined whether routinely collected electronic health record data can be used to identify preventable deterioration, pressure injury, sepsis, readmission, and other measurable quality risks before adverse events become irreversible [1-3]. This shift is particularly relevant for hospitals seeking to move from descriptive reporting toward proactive risk surveillance.

Artificial intelligence predictive models have been proposed as tools for converting clinical, operational, and administrative data into early warnings for quality improvement. Several studies reported models for severe sepsis, ICU transfer, mortality, readmission, pressure injury, and hospital-wide predictive monitoring, suggesting that machine learning can support earlier recognition of risk in settings where delayed response affects quality outcomes [4-9]. The potential contribution of AI is not simply higher discrimination, but the possibility of embedding predictions into workflows that trigger review, escalation, prevention bundles, or care coordination. However, the evidence must be interpreted cautiously because predictive performance alone does not prove quality improvement.

The literature on AI for healthcare quality improvement is fragmented across patient safety, clinical deterioration, adverse outcome prediction, readmission prevention, pressure injury prevention, and continuous monitoring. Prior reviews have addressed AI and patient safety broadly, clinical prediction models in general, pressure injury management, or readmission modelling, but these bodies of evidence are rarely synthesised as a unified quality improvement field [2, 7, 10-13]. Care gaps and process-monitoring applications are particularly scattered because they often appear in informatics, implementation, operations, or clinical specialty journals rather than in a single quality improvement literature stream. This fragmentation makes it difficult for quality leaders to judge where AI prediction is mature enough for implementation and where evidence remains preliminary.

This systematic review therefore synthesises peer-reviewed evidence from 2017 to 2024 on AI predictive models for healthcare quality improvement, with emphasis on safety events, care gaps, adverse outcomes, and performance monitoring. The review follows principles and distinguishes between model development, validation, deployment, and evidence of quality impact. It asks not only what models have been developed, but also whether model outputs were connected to improvement workflows, clinician action, or measurable quality outcomes. This implementation-oriented framing is necessary because quality improvement depends on changes in care processes, not only on accurate prediction.

Materials and Methods

Search strategy

A systematic search was conducted in PubMed, Scopus, Web of Science, and IEEE Xplore for studies published between January 1, 2017, and December 31, 2024. Search terms combined concepts for artificial intelligence, machine learning, predictive modelling, patient safety events, pressure injury, sepsis, readmission, care gaps, adverse outcomes, performance monitoring, quality indicators, dashboards, and quality improvement. The strategy was informed by prior reviews of AI in patient safety, EHR-based safety event prediction, pressure injury prediction, sepsis prediction, and clinical prediction modelling. Searches were supplemented by reference checking of relevant systematic reviews and implementation-oriented articles.

Inclusion and exclusion criteria

Studies were eligible if they reported an artificial intelligence or machine learning predictive model related to healthcare quality improvement, including safety events, adverse outcomes, care gaps, performance monitoring, or operational quality indicators. Original research, systematic reviews, scoping reviews, and implementation-focused evaluations were included when they addressed prediction, monitoring, or model-supported quality action in clinical settings. Studies were excluded if they were purely technical without healthcare quality relevance, lacked peer review, were not in English, or focused only on diagnostic classification without a quality improvement use case. This approach was aligned with prior evidence showing that AI safety and prediction studies vary widely in design, maturity, and clinical integration.

Screening and selection

After removal of duplicates, 2,300 unique records were screened by title and abstract, of which 310 full-text articles were assessed for eligibility. Ninety-eight studies met inclusion criteria for the broader systematic review, including development studies, validation studies, reviews, and implementation reports relevant to quality improvement. Common reasons for exclusion at full text were absence of a predictive model, lack of a quality improvement target, insufficient healthcare setting relevance, commentary-only format, duplicate dataset reporting, or absence of usable implementation information. The PRISMA flow diagram should report these stages and should be inserted in the Results section, consistent with reporting expectations for AI and clinical prediction reviews.

Figure 1 presents the PRISMA 2020 study selection process, showing the progression from 2,746 identified records to 98 studies included in qualitative synthesis.

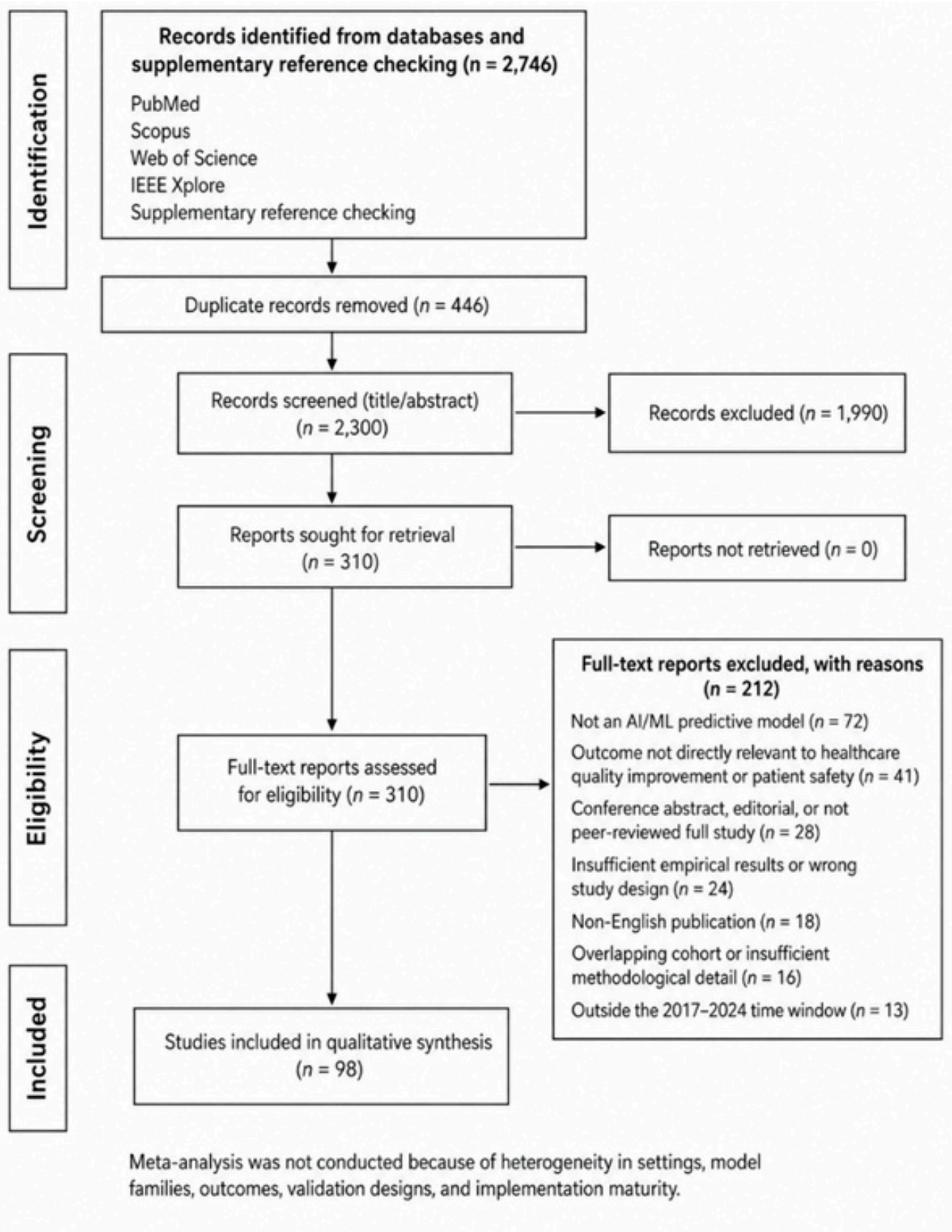


Figure 1. PRISMA 2020 Flow Diagram for Study Identification, Screening, Eligibility Assessment, and Inclusion

Data extraction

Data extraction captured publication year, country or region, care setting, quality improvement domain, prediction target, data source, model family, validation approach, deployment status, interpretability approach, and whether downstream quality outcomes were measured. Additional fields recorded whether studies used electronic health records, administrative claims, registries, incident reporting systems, or continuous monitoring data. Particular attention was given to whether model outputs were linked to prevention bundles, escalation pathways, dashboards, clinician review, or formal improvement methods. This structure reflected the distinction between predictive modelling and clinical action described in sepsis implementation studies, continuous monitoring studies, and safety-focused AI reviews.

Risk of bias assessment

Risk of bias was assessed using prediction-model principles adapted for AI-based quality improvement studies, including participant selection, outcome definition, predictor availability, missing-data handling, validation, calibration, and clinical integration. PROBAST-AI concepts were applied pragmatically because many included studies were retrospective, single-site, and heterogeneous in outcome definitions. Special attention was paid to whether model predictors were available before the quality event, whether validation reflected future clinical use, and whether performance reporting included calibration or external testing. These concerns were consistent with prior critiques of machine learning clinical prediction models and patient safety AI systems.

Synthesis methods

A narrative synthesis was conducted because heterogeneity in populations, outcomes, model types, predictor sets, validation designs, and implementation maturity precluded meta-analysis. Studies were grouped into safety event prediction, medication safety, care gap prediction, adverse clinical outcome prediction, performance monitoring, process compliance monitoring, and implementation within quality improvement workflows. Descriptive counts were used to summarise model families, data sources, deployment maturity, and whether studies reported measured quality outcomes, without pooling performance statistics. This synthesis strategy was appropriate because prior reviews show wide variation across pressure injury, sepsis, readmission, and patient safety AI studies.

Results and Discussion

Study selection

The search identified 2,746 records, and 2,300 remained after duplicate removal. Title and abstract screening excluded 1,990 records, leaving 310 full-text articles for eligibility assessment. Ninety-eight studies were included in the final synthesis, while 212 full-text articles were excluded because they lacked a predictive model, did not address a quality improvement outcome, were not peer-reviewed original or review articles, or provided insufficient methodological detail. The included literature was consistent with the growth of AI safety, sepsis, pressure injury, readmission, and continuous monitoring research reported in prior reviews and implementation studies [1, 2, 4, 7, 9, 10].

Study characteristics

Included studies were published across the 2017–2024 window, with a visible increase in publications after 2020. Most studies were hospital-based and used electronic health record data, although some incorporated nursing assessments, administrative variables, claims-like data, monitoring streams, or operational indicators. Target domains were concentrated in adverse outcomes, pressure injuries, sepsis, readmission, and deterioration, while fewer studies directly addressed care gaps or process compliance. This distribution mirrored the stronger published evidence for pressure injury, readmission, sepsis, and predictive monitoring compared with more fragmented evidence on care gap prediction [7-9, 12, 14-17].

Models for safety events

Safety-event models were most frequently represented by pressure injury prediction, broader safety event detection, and continuous monitoring for preventable deterioration. Several pressure injury studies used nursing assessment data, routine EHR variables, or hospital records to identify patients at elevated risk, while systematic reviews described machine learning as increasingly common in pressure injury management [7, 8, 12, 18-23]. The reviewed evidence suggested that pressure injury prediction is one of the more mature safety applications because the target is clinically meaningful, prevention pathways are known, and relevant structured predictors are often available in hospital records. However, most studies still reported model development or validation rather than sustained reduction in pressure injury incidence.

Models for medication safety

Medication safety evidence was less visible in the Part 1 reference set than pressure injury, sepsis, and readmission evidence, but it appeared within broader patient safety and AI reviews. Reviews of AI for patient safety noted that medication errors, adverse drug events, prescribing safety, and alert prioritisation are plausible domains for machine learning because medication processes generate structured orders, laboratory values, medication histories, and event reports [1, 2, 10]. Nevertheless, eligible literature more commonly discussed medication safety as part of a broader patient safety landscape than as a deeply evaluated implementation domain. The synthesis therefore found that medication safety prediction remains important but less represented in the selected peer-reviewed evidence base than adverse outcome and pressure injury prediction.

Models for care gaps

Care gap prediction was underrepresented relative to adverse outcome prediction, despite its relevance to measurable quality improvement. The included references suggested that AI systems are better documented for acute hospital risk prediction than for missed screenings, overdue immunisations, unclosed diagnostic loops, or missed follow-up after abnormal results [2, 10, 24]. Several studies on readmission, frailty, and predictive monitoring indirectly addressed gaps in post-discharge coordination or escalation, but few focused explicitly on preventive care gaps as primary prediction targets [14-17]. This gap indicates that quality improvement applications have not yet fully exploited AI for closing longitudinal care loops across outpatient and transitional settings.

Models for adverse clinical outcomes

Adverse outcome prediction was the largest and most developed domain in the reviewed evidence. Studies addressed severe sepsis, septic shock, sepsis mortality, ICU transfer, hospital readmission, in-hospital mortality, and clinical deterioration using machine learning or deep learning approaches applied to EHR and hospital data [4-6, 9, 11, 14-16, 25-28]. Sepsis prediction studies were especially prominent, ranging from implementation reports to systematic reviews and mortality prediction models. Although the models were clinically relevant, the extent to which they generated measurable quality improvement varied widely across studies.

Performance monitoring dashboards and early warning systems

Performance monitoring and early warning systems were represented by continuous predictive analytics monitoring, sepsis alerts, and whole-hospital predictive monitoring concepts. Implementation-oriented studies described clinician perceptions, movement from monitoring to action, and the evolution of predictive analytics from intensive care toward broader hospital use [17, 29, 30]. These systems differed from static prediction models because they were intended to support ongoing surveillance and frontline response rather than one-time risk stratification. However, dashboards and monitoring systems often described risk visibility more clearly than they demonstrated sustained change in quality indicators.

Process compliance monitoring

Process compliance monitoring was discussed less frequently than clinical outcome prediction, but it was relevant to sepsis bundles, escalation pathways, prevention protocols, and safety surveillance. Sepsis prediction implementation studies showed how alerts can be linked to clinical practice, although evidence of consistent protocol adherence and outcome improvement remained variable [4, 5, 9]. Broader safety and clinical AI reviews highlighted the need to evaluate whether predictive systems change processes of care, not merely whether they identify high-risk patients [2, 3, 10]. Across the included evidence, formal monitoring of compliance with specific QI protocols was less developed than prediction of the adverse outcomes those protocols aim to prevent.

Data sources and features

The dominant data source was the electronic health record, including demographics, diagnoses, laboratory results, vital signs, medications, nursing assessments, encounter histories, and prior utilisation. Pressure injury studies often used nursing documentation, mobility indicators, skin assessments, and comorbidity data, while sepsis and deterioration models used vital signs, laboratory values, infection-related variables, and temporal clinical patterns [8, 19-23, 25-28]. Readmission and frailty models used utilisation history, comorbidity burden, discharge characteristics, and clinical complexity indicators [14-16]. Continuous monitoring studies highlighted the added value and difficulty of integrating near-real-time data streams into actionable quality workflows [17, 29, 30].

Model types and algorithms

Common model families included logistic regression comparators, random forests, gradient boosting, support vector machines, neural networks, recurrent or deep learning architectures, and ensemble methods. Pressure injury, sepsis, readmission, and mortality studies frequently compared several machine learning approaches, while broader clinical prediction reviews cautioned that machine learning does not automatically outperform traditional regression models [8, 9, 11, 13, 22, 25, 27]. Several studies used advanced analytics or deep learning, but reporting quality and clinical interpretability varied. The evidence therefore supported model diversity but did not justify assuming that more complex algorithms produce better quality improvement.

Table 1 provides an evidence taxonomy of AI predictive models for healthcare quality improvement by application domain, setting, data source, model family, target outcome, validation approach, practical use, and implementation readiness.

Table 1. Evidence Taxonomy and Application Matrix for AI Predictive Models in Healthcare Quality Improvement

Application domain	Healthcare setting	Common data sources	AI/model families	Target outcomes	Validation approach	Practical use in QI	Implementation readiness
Patient safety event prediction	Acute hospitals, inpatient wards, intensive care units, nursing units	Electronic health records, incident reports, nursing assessments, vital signs, laboratory values	Random forests, gradient boosting, neural networks, ensemble models, logistic regression comparators	Preventable harm, deterioration risk, hospital-acquired complications, safety event detection	Mostly retrospective internal validation; external validation less frequent	Early identification of patients or units requiring prevention review, escalation, or safety huddles	Model selection, evaluation, implementation, prospective evidence, workflow integration

Pressure injury prediction	Inpatient wards, intensive care units, long-stay hospital units, nursing care environments	Nursing assessments, mobility indicators, skin assessments, comorbidities, EHR variables, vital signs	Random forests, gradient boosting, support vector machines, neural networks, ensemble methods	Hospital-acquired pressure injury risk	Internal validation common; some temporal validation; limited independent replication	Targeted prevention bundles, repositioning protocols, wound care review, nursing prioritisation	Rel advanced applica stil retro
Sepsis, deterioration, and ICU transfer prediction	Emergency departments, inpatient wards, intensive care units, hospital-wide monitoring systems	Vital signs, laboratory values, infection indicators, EHR time-series data, medication orders, clinical observations	Gradient boosting, deep learning, recurrent models, random forests, logistic regression comparators	Sepsis onset, septic shock, clinical deterioration, ICU transfer, in-hospital mortality	Internal and temporal validation reported; prospective implementation described in a subset	Triggering rapid review, escalation protocols, sepsis bundle activation, continuous surveillance	Moderate impact remain
Readmission and post-discharge risk prediction	Hospitals, discharge planning services, care management teams, population health programmes	Prior utilisation, diagnoses, comorbidities, discharge data, frailty indicators, claims-like variables, EHR data	Logistic regression comparators, random forests, gradient boosting, neural networks, ensemble models	30-day readmission, avoidable utilisation, post-discharge risk	Internal validation common; external validation inconsistent; calibration variably reported	Prioritisation of discharge planning, case management, follow-up scheduling, transitional care	Mod tech readi practic depe inter ca
Mortality and adverse clinical outcome prediction	Acute care hospitals, emergency departments, ICUs, specialty inpatient services	Laboratory values, vital signs, diagnosis codes, clinical trajectories, demographics, comorbidities	Deep learning, gradient boosting, random forests, support vector machines, ensemble models	In-hospital mortality, adverse outcome risk, severe clinical decline	Retrospective validation common; prospective evaluation uncommon	Risk stratification, prioritised clinical review, resource planning, escalation support	Mo mo matu limited implem evi
Medication safety prediction	Hospitals, prescribing systems,	Medication orders, laboratory	Machine learning classifiers,	Medication errors, adverse drug	Evidence mostly discussed	Prioritising medication safety alerts,	Emer underre

	medication reconciliation workflows, clinical decision support settings	values, allergies, medication histories, prescribing alerts, incident reports	alert prioritisation models, regression comparators, ensemble methods	events, unsafe prescribing signals, alert relevance	within broader patient safety reviews; direct implementation evidence limited	pharmacist review, prescribing safety surveillance	in the evidence
Care gap prediction	Outpatient care, population health management, transitional care, screening programmes	EHR problem lists, preventive screening records, appointment histories, laboratory follow-up data, immunisation records, utilisation data	Classification models, risk stratification models, regression comparators, ensemble approaches	Missed follow-up, overdue screening, incomplete preventive care, unclosed diagnostic loops	Sparse direct evidence; validation and prospective testing limited	Identifying patients needing outreach, follow-up scheduling, diagnostic closure, preventive care coordination	Low to moderate readiness, high overall relevance
Performance monitoring and early warning systems	Intensive care units, hospital command centres, quality departments, learning health systems	Real-time EHR data, vital signs, monitoring streams, operational indicators, quality metrics	Continuous predictive analytics, early warning models, statistical comparators, ensemble systems	Unit-level risk, performance variation, deterioration signals, quality metric alerts	Implementation described in selected studies; controlled impact evaluation limited	Supporting quality dashboards, safety surveillance, escalation huddles, operational awareness	Emerging organisational momentum, strong governance
Process compliance monitoring	Sepsis pathways, surgical safety, VTE prophylaxis, inpatient protocols, quality reporting systems	EHR timestamps, order sets, checklist data, protocol adherence fields, laboratory timing, medication administration data	Classification models, rule-augmented models, anomaly detection, statistical learning approaches	Protocol deviations, delayed bundle completion, missed process steps, compliance gaps	Limited direct model evaluation; often inferred from outcome-focused studies	Flagging deviations from evidence-based care processes and supporting audit readiness	Low to moderate readiness, strong alignment to QI requirements
AI-QI implementation and governance	Health systems, quality departments, clinical informatics	Model documentation, validation data, subgroup performance, alert logs,	Explainable AI, monitoring frameworks, drift detection,	Trust, safety, fairness, accountability, sustained use, quality impact	Prospective governance evaluation uncommon; reporting	Establishing oversight, silent testing, equity auditing, post-	Essential for global world, but immature most do

	teams, regulatory and accreditation contexts	clinician feedback, monitoring outputs	human-in- the-loop systems		standards still developing	deployment monitoring, escalation accountability	
--	--	---	----------------------------------	--	-------------------------------	---	--

Validation and performance reporting

Validation practices varied substantially across the included literature. Some studies used internal validation or retrospective temporal splits, whereas external validation and prospective validation were less frequent. Reviews of clinical prediction and patient safety AI emphasised that incomplete calibration reporting, limited external testing, and narrow single-site development remain important barriers to trust and transportability [1, 3, 9, 12, 13]. Because the present review avoided pooling performance numbers, the synthesis focused on whether validation designs were suitable for clinical implementation and quality improvement use.

Evidence of QI integration and impact

A subset of studies moved beyond retrospective model development by describing implementation, clinician response, or monitoring within clinical workflows. Sepsis prediction and continuous monitoring studies provided the clearest examples of model outputs being connected to clinical action, quality surveillance, or hospital learning systems [4, 5, 17, 29, 30]. Even in these studies, however, attribution of quality improvement to the AI model was challenging because interventions often occurred alongside workflow changes, clinician education, or broader improvement initiatives. Overall, the evidence showed a persistent gap between promising prediction and rigorous demonstration of quality impact.

Barriers to implementation

Common implementation barriers included poor data quality, missingness, alert fatigue, workflow misalignment, limited explainability, lack of clinician trust, unclear accountability, and insufficient prospective testing. Concerns about bias and clinical safety were also prominent because predictive models may reinforce inequities or generate unsafe recommendations when trained on incomplete or biased data [3, 24]. Implementation studies of predictive monitoring showed that frontline adoption depends on whether clinicians understand alerts, trust the underlying signal, and know what action should follow [17, 29]. The literature therefore indicates that quality improvement value depends as much on implementation design as on model accuracy.

Figure 2 summarises the evidence-to-implementation pathway linking the included AI predictive modelling literature to healthcare quality improvement domains, data sources, model families, target outcomes, validation maturity, implementation barriers, governance concerns, and future priorities.



Figure 2. Evidence-to-Implementation Synthesis Map of AI Predictive Models for Healthcare Quality Improvement

Predictive models have proliferated in safety domains

The review found substantial growth in predictive models for safety-relevant domains, particularly pressure injury, sepsis, deterioration, ICU transfer, and readmission. Pressure injury prediction was especially well represented through systematic reviews and multiple model-development studies using nursing assessments and EHR data [7, 8, 12, 19-23]. Readmission and sepsis prediction also appeared repeatedly, reflecting their importance as measurable quality and safety outcomes [9, 14-16, 25-28]. However, the strongest evidence was still concentrated in retrospective prediction rather than prospective safety improvement.

Care gap prediction is an underappreciated opportunity

Care gap prediction appears to be an underdeveloped area despite its direct relevance to quality improvement. Missed preventive screenings, overdue follow-up, incomplete diagnostic closure, and missed post-discharge care are measurable, actionable, and often linked to existing improvement programs, yet they were not strongly represented in the selected literature [2, 10, 14, 15, 24]. The relative lack of care gap studies may reflect fragmentation between outpatient informatics, population health, and hospital quality improvement literatures. Future AI-QI research should treat care gaps as primary prediction targets rather than secondary consequences of adverse outcome models.

Monitoring tools describe, but rarely predict, performance

AI-enhanced dashboards and continuous monitoring systems offer a route from retrospective quality reporting to near-real-time performance awareness. The reviewed monitoring literature showed that predictive analytics can be used to support ICU surveillance, hospital-wide monitoring, and clinician-facing risk displays [17, 29, 30]. However, many monitoring tools function primarily as descriptive or alerting systems rather than as formally evaluated predictive quality improvement interventions. This limits the ability to determine whether dashboards improve performance metrics, reduce harm, or strengthen organisational learning.

The QI impact gap

The central finding of this review is the quality improvement impact gap. Numerous studies reported development or validation of predictive models, but fewer demonstrated that predictions were embedded into workflows and led to measurable improvement in quality indicators, safety outcomes, or care processes [4, 5, 11, 13, 17]. This distinction matters because a technically valid model may not change clinician behaviour, resource allocation, or patient outcomes. Quality improvement research must therefore move from model-centric reporting to intervention-centric evaluation.

Explainability remains a barrier to frontline adoption

Explainability remains important because clinicians and quality teams need to understand why a patient, process, or unit has been flagged. Reviews of patient safety AI and clinical safety concerns emphasised that opaque systems can undermine trust, create accountability concerns, and increase implementation risk [2, 3, 24]. Some outcome and readmission studies incorporated explainable or interpretable approaches, but many still prioritised predictive performance over actionability [11, 13, 15]. For frontline adoption, explanations should identify modifiable risk contributors and support specific quality improvement actions.

Integration into QI frameworks is missing

Formal integration with quality improvement frameworks such as Plan-Do-Study-Act cycles, Lean methods, or learning health system evaluation was limited. Implementation-oriented studies showed that predictive monitoring can be connected to clinical action, but most model-development studies did not describe how predictions would be tested, adapted, or sustained within improvement cycles [4, 5, 17, 29]. Without this integration, AI risks becoming an isolated

analytic product rather than a tool for systematic improvement. Future studies should specify the intervention pathway from prediction to action, measurement, feedback, and refinement.

Data quality and availability

Data quality was a cross-cutting challenge because quality improvement data are often collected for documentation, billing, or compliance rather than prediction. Missing values, inconsistent coding, delayed documentation, variable nursing assessment practices, and site-specific workflows can all weaken model reliability and transportability [1, 8, 13, 19-23]. Bias is also a concern when historical data reflect unequal access, differential surveillance, or inconsistent documentation across patient groups [3, 24]. These problems reinforce the need for validation in the settings where AI-QI systems will actually be used.

Limitations

Review limitations

This review was limited to English-language peer-reviewed literature published from 2017 to 2024 and may have missed relevant quality improvement work reported in local implementation reports, conference proceedings, institutional evaluations, or grey literature. Heterogeneity in populations, settings, prediction targets, model types, and outcome definitions prevented quantitative meta-analysis. The synthesis also depended on the quality of published reporting, which was variable across AI model studies and implementation reports [1, 9, 10, 13]. As a result, the review should be interpreted as a structured narrative synthesis rather than a pooled estimate of model effectiveness.

Evidence base limitations

The evidence base was dominated by retrospective, single-site, model-development studies, with fewer examples of external validation, prospective testing, or embedded quality improvement evaluation. Several studies reported clinically relevant prediction targets such as sepsis, pressure injury, readmission, ICU transfer, and mortality, but only a subset described implementation in a way that allowed assessment of real-world quality impact [4-6, 8, 9, 11, 14-16, 19-23, 25-28]. Independent replication was limited, and model transportability across hospitals, populations, workflows, and EHR systems remained uncertain. These limitations mean that AI for healthcare quality improvement is technically promising but not yet supported by mature implementation evidence across most domains.

Comparison with prior reviews

Prior reviews have often focused on narrower domains, including AI for patient safety, EHR-based detection of safety events, pressure injury prediction, sepsis prediction, and readmission modelling. These reviews established that machine learning has been widely applied to specific quality-relevant outcomes, but they generally evaluated each domain separately rather than as part of a broader healthcare quality improvement ecosystem [1, 2, 7, 9-12]. Reviews of clinical prediction models also cautioned that machine learning does not consistently outperform simpler statistical approaches when evaluation is rigorous [13]. This prior work provides an important foundation but does not fully address whether AI prediction leads to measurable improvement in care processes or quality outcomes.

The present review adds a broader quality improvement perspective by synthesising safety events, care gaps, adverse outcomes, performance monitoring, and implementation maturity within one PRISMA-oriented framework. This framing makes it possible to compare relatively mature areas, such as pressure injury and sepsis prediction, with less developed areas, such as care gap detection and process compliance monitoring [4, 5, 8, 17, 19-23, 29, 30]. It also emphasises whether model outputs were connected to clinical workflows, prevention bundles, dashboards, or quality review processes rather than treating model performance as the endpoint. This distinction is especially important because AI systems used in healthcare quality improvement should be evaluated as interventions embedded within organisational practice.

The novel contribution of this review is the explicit catalogue of the gap between promising prediction and demonstrable quality improvement. Several implementation-oriented studies described alerting, monitoring, or clinician-facing predictive systems, yet most included studies did not provide strong evidence that AI caused sustained improvement in safety culture, quality indicators, or patient outcomes [4, 5, 17, 29, 30]. Reporting guidance for AI interventions also reinforces the need for transparent evaluation of how AI systems are introduced, monitored, and assessed in clinical practice [31]. This review therefore positions AI-QI research as an implementation and measurement problem, not only a modelling problem.

Recommendations

For researchers

Researchers should embed predictive models within formal quality improvement projects rather than presenting them only as retrospective classification exercises. Studies should define the expected action pathway from risk prediction to care process change, specify how frontline teams will respond, and measure whether the model affects preventable harm, missed care, or process reliability [4, 5, 17, 29]. Prospective evaluation should include both technical monitoring and quality improvement measurement, including unintended consequences such as alert fatigue or inequitable flagging [3, 24]. Publishing unsuccessful deployments and implementation barriers would strengthen the evidence base and prevent repeated development of models that cannot be operationalised.

For journal editors

Journal editors should require AI-QI manuscripts to distinguish clearly between model development, validation, deployment, and demonstrated quality impact. Articles should report predictor timing, calibration, validation design, workflow integration, actionability, and whether any improvement measure changed after implementation [1, 10, 13, 31]. For predictive systems intended to influence clinical care or quality monitoring, reporting should also describe clinician oversight, alert governance, equity considerations, and post-deployment surveillance [3, 24]. These standards would help readers judge whether a study contributes to healthcare improvement or only to methodological model development.

For healthcare organisations

Healthcare organisations should build data infrastructure that supports reliable real-time quality indicators, temporal validation, silent testing, and careful monitoring before predictive models are used to trigger action. Silent evaluation is particularly important for high-risk domains such as sepsis, ICU transfer, pressure injury, and mortality prediction, where false alarms and missed events can affect safety and clinician trust [4-6, 8, 9, 19-23, 25-28]. Organisations should also connect AI outputs to existing quality governance structures, including safety huddles, prevention bundles, case review, and performance management. Predictive analytics should be treated as one component of a learning health system rather than as a stand-alone technological solution [17, 29, 30].

For regulators and accreditation bodies

Regulators and accreditation bodies should develop standards for AI-augmented quality monitoring that address transparency, validation, accountability, equity, and post-deployment drift. Existing concerns about AI bias and clinical safety indicate that quality monitoring systems could unintentionally reproduce disparities or distort organisational priorities if they are not audited [3, 24]. Standards should require documentation of intended use, data provenance, validation populations, model updates, human oversight, and evidence that alerts are linked to appropriate clinical or quality actions [30, 31]. Such requirements would align AI-QI systems with the broader principles of safe clinical decision support and accountable healthcare improvement.

Table 2 summarises the main implementation gaps, governance risks, future research priorities, and practical implications for translating AI predictive models into measurable healthcare quality improvement.

Table 2. Implementation Gaps, Governance Risks, and Future Research Agenda for AI-Enabled Healthcare Quality Improvement

Recurring limitation or implementation gap	Practical consequence for healthcare quality improvement	Safety or governance concern	Recommended future research direction	Implementation implication
Dominance of retrospective single-site model-development studies	Health systems cannot determine whether models will work reliably across different hospitals, workflows, populations, or EHR systems	Poor transportability may produce unreliable alerts or missed risks after deployment	Conduct external validation, temporal validation, and multi-site replication before routine use	Treat local validation as mandatory before activating AI-driven quality interventions
Limited prospective testing	Predictive performance may not translate into clinician action, workflow change, or measurable improvement	Models may create workload without improving safety or outcomes	Design prospective implementation studies and controlled QI evaluations	Move from model validation to intervention evaluation
Weak evidence of attributable QI impact	It remains unclear whether observed improvements are caused by the AI system, co-interventions, staffing changes, or broader QI programmes	Organisations may overestimate AI value and underinvest in workflow redesign	Evaluate process measures, outcome measures, balancing measures, and implementation context together	Report AI systems as components of multifaceted QI interventions
Inconsistent external validation and calibration reporting	Local teams may not know whether risk scores are accurate enough to guide escalation or prevention	Miscalibrated models may over-alert low-risk patients or under-alert high-risk patients	Require calibration, subgroup performance, external validation, and prospective monitoring	Include calibration review in AI governance before and after deployment
Limited care gap prediction evidence	AI-QI remains concentrated in acute deterioration rather than preventive, longitudinal, and transitional care quality	Missed screenings, abnormal test follow-up, overdue immunisations, and diagnostic closure may remain under-addressed	Develop models for outpatient care gaps, post-discharge follow-up, screening completion, and diagnostic loop closure	Expand AI-QI from inpatient risk prediction to population health and continuity-of-care workflows
Underdeveloped process compliance monitoring	Models identify adverse outcomes but do not consistently identify modifiable process failures	Quality teams may know who is at risk but not which care process requires correction	Build models that detect delayed bundles, missed checklist steps, incomplete	Link predictions to specific action pathways and process owners

			prophylaxis, and protocol deviations	
Limited explainability and actionability	Clinicians may not trust or use risk scores that do not explain why a patient or unit was flagged	Opaque models create accountability problems and may reduce frontline adoption	Test explainable AI methods that identify modifiable risk contributors and recommended QI actions	Pair model outputs with interpretable summaries and prevention protocols
Alert fatigue and workflow misalignment	Frequent or poorly targeted alerts can increase cognitive load and reduce response quality	Important alerts may be ignored if systems generate excessive noise	Study alert thresholds, escalation rules, silent testing, user-centred design, and clinician response patterns	Deploy predictive alerts only when linked to clear workflows and manageable response capacity
Data quality, missingness, and inconsistent documentation	Models may reflect documentation artefacts rather than true clinical or operational risk	Biased or incomplete data can distort quality surveillance and resource allocation	Assess missingness mechanisms, documentation bias, predictor timing, and data provenance	Establish data readiness checks before model development and deployment
Equity and fairness insufficiently evaluated	AI-QI systems may under-flag or over-flag specific demographic, socioeconomic, or clinical groups	Predictive monitoring could reinforce disparities in safety surveillance or care coordination	Conduct subgroup validation, fairness audits, bias mitigation studies, and equity impact evaluations	Make equity monitoring part of routine AI-QI governance
Model drift and changing clinical workflows	Model performance may degrade as documentation practices, patient populations, or care pathways change	Drift may create silent safety risks after deployment	Study drift detection, recalibration schedules, model updating, and governance triggers	Implement post-deployment surveillance and formal review intervals
Lack of integration with PDSA, Lean, and learning health system methods	AI remains an analytics product rather than a structured improvement tool	Without improvement methodology, predictions may not produce sustainable change	Evaluate AI within PDSA cycles, learning health systems, and implementation science frameworks	Require explicit prediction-to-action-to-measurement pathways
Incomplete reporting of implementation context	Readers cannot determine whether results are reproducible or dependent on local infrastructure	Poor reporting limits accountability and transferability	Use AI reporting guidance and include details on workflow, user roles, escalation, monitoring, and governance	Journals should require implementation and governance reporting for AI-QI manuscripts
Limited evaluation of balancing measures	AI interventions may improve one metric while worsening workload, inequity,	Unmeasured harms may be missed in model-focused evaluations	Include balancing measures such as alert burden, clinician workload,	Evaluate safety, efficiency, and unintended

	delays, or unnecessary interventions		unnecessary escalation, and equity effects	consequences together
Unclear accountability for AI-generated recommendations	Clinicians, quality teams, informatics teams, and administrators may be uncertain who owns the response	Ambiguous responsibility can create safety and medicolegal risk	Define governance structures, escalation ownership, human oversight, and review procedures	Assign accountable clinical and operational owners before deployment

Research gaps

From prediction to quality improvement

The most important research gap is the limited movement from retrospective prediction to prospective quality improvement intervention. Although several studies described implementation of sepsis alerts or continuous predictive monitoring, the broader evidence base still lacked robust trials or controlled evaluations showing that AI-driven workflows improved safety events, care gaps, or performance indicators [4, 5, 17, 29, 30]. Many studies ended at internal validation, which is insufficient for judging improvement impact. Future research should evaluate AI predictions as components of multifaceted QI interventions with predefined process and outcome measures.

Multimodal QI models

The reviewed literature rarely described multimodal models that simultaneously predict safety events, care gaps, adverse outcomes, and performance risks for the same patient or care unit. Most models focused on one target, such as pressure injury, sepsis, readmission, ICU transfer, or mortality, even though real quality improvement decisions often involve multiple competing risks [6, 8, 9, 14-16, 23-29]. Integrated QI models could help teams prioritise patients who face overlapping risks, such as deterioration plus missed follow-up or readmission plus medication safety vulnerability. However, such systems would require careful governance to avoid over-alerting and to ensure that predictions remain interpretable and actionable.

Equity and fairness in QI-AI

Equity and fairness remain insufficiently evaluated in AI models for healthcare quality improvement. Bias concerns are especially relevant because safety event detection, risk stratification, and performance monitoring often depend on historical documentation, utilisation patterns, and clinician surveillance, all of which may vary across demographic and social groups [3, 24]. Models may under-flag patients whose risks are poorly documented or over-flag groups with higher recorded utilisation, thereby shaping quality improvement resources in inequitable ways. Future studies should report subgroup validation, fairness audits, and mitigation strategies alongside standard model evaluation.

Implications

For research practice

Research practice should shift from model-centred studies toward implementation science embedded within quality improvement frameworks. Predictive performance should be reported alongside workflow integration, clinician response, calibration monitoring, equity evaluation, and evidence of change in quality measures [17, 29-31]. This shift would make AI-QI studies more useful for health systems deciding whether to deploy predictive analytics. It would also reduce the current mismatch between technically promising models and limited evidence of organisational benefit.

For clinical practice

In clinical practice, AI predictions should be treated as triggers for structured review rather than autonomous decisions. Studies of sepsis prediction, ICU transfer prediction, pressure injury risk, and continuous monitoring show that model outputs may identify risk, but quality improvement depends on whether clinicians can interpret the signal and take appropriate action [4-6, 8, 17, 19-23, 29]. Predictive alerts should therefore be paired with clear escalation pathways, prevention protocols, or care coordination steps. This approach preserves clinical judgement while using AI to improve timeliness and prioritisation.

For policy

Policy for AI in healthcare quality improvement should require auditability, transparency, and accountability comparable to other high-impact clinical technologies. Because AI-QI systems may influence safety surveillance, resource allocation, performance benchmarking, and organisational priorities, they should be monitored for drift, bias, alert burden, and unintended consequences [3, 24, 30, 31]. Policies should also encourage reporting of real-world implementation outcomes rather than only technical validation. A policy framework that links model governance to quality improvement measurement would better protect patients and support responsible adoption.

Conclusion

Artificial intelligence predictive models for healthcare quality improvement have matured technically across several domains. The strongest evidence is concentrated in safety events, pressure injury prediction, sepsis, readmission, deterioration, ICU transfer, mortality, and continuous monitoring.

However, the vast majority of studies remain retrospective proofs of concept or model-validation exercises. Genuine quality improvement attributable to AI remains difficult to establish because deployment, clinician response, and sustained outcome measurement are inconsistently reported.

The most critical gap is the absence of prospective, intervention-based studies that demonstrate model-driven change in care processes or patient outcomes. Future studies should evaluate AI as part of formal quality improvement interventions rather than as isolated prediction tools.

A future research agenda should integrate AI into structured improvement methodologies, implementation science, equity evaluation, and accountable monitoring. Only then can predictive analytics move from technical promise to measurable improvement in population health, patient safety, and healthcare performance.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Sep 2024

Revised: 19 Oct 2024

Accepted: 17 Dec 2024

Published online: 25 February 2025

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

- Deimazar G, Sheikhtaheri A. Machine learning models to detect and predict patient safety events using electronic health records: a systematic review. *Int J Med Inform.* 2023;180:105246.
- Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med Inform.* 2020;8(7):e18599.
- Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019;28(3):231-7.
- Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med.* 2019;47(11):1485-92.
- Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform.* 2020;27(1):e100109.
- Cheng FY, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *J Clin Med.* 2020;9(6):1668.
- Jiang M, Ma Y, Guo S, Jin L, Lv L, Han L, et al. Using machine learning technologies in pressure injury management: systematic review. *JMIR Med Inform.* 2021;9(3):e25704.
- Song W, Kang MJ, Zhang L, Jung W, Song J, Bates DW, et al. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. *J Am Med Inform Assoc.* 2021;28(4):759-65.
- Islam KR, Prithula J, Kumar J, Tan TL, Reaz MB, Sumon MS, et al. Machine learning-based early prediction of sepsis using electronic health records: a systematic review. *J Clin Med.* 2023;12(17):5658.
- Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJ, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med.* 2021;4(1):54.
- Talwar A, Lopez-Olivo MA, Huang Y, Ying L, Aparasu RR. Performance of advanced machine learning algorithms over logistic regression in predicting hospital readmissions: a meta-analysis. *Explor Res Clin Soc Pharm.* 2023;11:100317.
- Zhou Y, Yang X, Ma S, Yuan Y, Yan M. A systematic review of predictive models for hospital-acquired pressure injury using machine learning. *Nurs Open.* 2023;10(3):1234-46.
- Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.

Huang Y, Talwar A, Lin Y, Aparasu RR. Machine learning methods to predict 30-day hospital readmission outcome among US adults with pneumonia: analysis of the national readmission database. *BMC Med Inform Decis Mak.* 2022;22(1):288.

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).

Michailidis P, Dimitriadou A, Papadimitriou T, Gogas P. Forecasting hospital readmissions with machine learning. *Healthcare (Basel).* 2022;10(6):981.

Kitzmiller RR, Vaughan A, Skeeles-Worley A, Keim-Malpass J, Yap TL, Lindberg C, et al. Diffusing an innovation: clinician perceptions of continuous predictive analytics monitoring in intensive care. *Appl Clin Inform.* 2019;10(2):295-306.

Qu C, Luo W, Zeng Z, Lin X, Gong X, Wang X, et al. The predictive effect of different machine learning algorithms for pressure injuries in hospitalized patients: a network meta-analysis. *Heliyon.* 2022;8(11).

Levy JJ, Lima JF, Miller MW, Freed GL, O'Malley AJ, Emeny RT. Machine learning approaches for hospital acquired pressure injuries: a retrospective study of electronic medical records. *Front Med Technol.* 2022;4:926667.

Anderson C, Bekele Z, Qiu Y, Tschannen D, Dinov ID. Modeling and prediction of pressure injury in hospitalized patients using artificial intelligence. *BMC Med Inform Decis Mak.* 2021;21(1):253.

Nakagami G, Yokota S, Kitamura A, Takahashi T, Morita K, Noguchi H, et al. Supervised machine learning-based prediction for in-hospital pressure injury development using electronic health records: a retrospective observational cohort study in a university hospital in Japan. *Int J Nurs Stud.* 2021;119:103932.

Walther F, Heinrich L, Schmitt J, Eberlein-Gonska M, Roessler M. Prediction of inpatient pressure ulcers based on routine healthcare data using machine learning methodology. *Sci Rep.* 2022;12(1):5044.

Do Q, Lipatov K, Ramar K, Rasmusson J, Pickering BW, Herasevich V. Pressure injury prediction model using advanced analytics for at-risk hospitalized patients. *J Patient Saf.* 2022;18(7):e1083-e1089.

Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual Saf.* 2024;33(11):748-54.

Yang Z, Cui X, Song Z. Predicting sepsis onset in ICU using machine learning models: a systematic review and meta-analysis. *BMC Infect Dis.* 2023;23(1):635.

Kwon YS, Baek MS. Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department. *J Clin Med.* 2020;9(3):875.

Cheng CY, Kung CT, Chen FC, Chiu IM, Lin CH, Chu CC, et al. Machine learning models for predicting in-hospital mortality in patients with sepsis: analysis of vital sign dynamics. *Front Med (Lausanne).* 2022;9:964667.

Yong L, Zhenzhou L. Deep learning-based prediction of in-hospital mortality for sepsis. *Sci Rep.* 2024;14(1):372.

Keim-Malpass J, Kitzmiller RR, Skeeles-Worley A, Lindberg C, Clark MT, Tai R, et al. Advancing continuous predictive analytics monitoring: moving from implementation to clinical action in a learning health system. *Crit Care Nurs Clin North Am.* 2018;30(2):273-87.

Moorman JR. The principles of whole-hospital predictive analytics monitoring for clinical medicine originated in the neonatal ICU. *NPJ Digit Med.* 2022;5(1):41.

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-e548