RESEARCH PAPER

Enhancing Quality Management Systems (QMS) in Clinical Data Management using Machine Learning: A Systematic Scoping Review and Proposed Reference Framework

Ravi Mandal[1]* and Dr E Rushit Gnanaroy[1]

[1] Department of Business Administration, Hindustan Institute of Technology and Science, Padur, Chennai 603 103, Tamil Nadu, India

*Corresponding Author: Ravi Mandal | Email: ravimandal.research@gmail.com | ORCID: https://orcid.org/0009-0008-9191-5316]

Highlights

- First systematic ML-QMS scoping review across the full clinical data lifecycle
- Zero of 30 studies achieved external multi-institutional validation
- Six-layer ML-QMS Reference Architecture grounded in KBS knowledge systems principles
- Seven-priority Research Roadmap with timelines, stakeholders, and expected outcomes
- Quality appraisal of all 30 included studies using a validated six-criterion rubric

Abstract

The pressure on quality management systems (QMS) of clinical data management becomes increasingly pressured by the growing amounts of data, complex regulations, and the necessity of real-time monitoring of clinical trials in the modern world. This is a synthesis-based systematic scoping review of the current state of applying machine learning (ML) to increase the capabilities of a QMS, based on 30 peer-reviewed publications identified through a systematic search of PubMed, Scopus, IEEE Xplore, and Web of Science (January 2017-March 2026), in line with PRISMA-ScR criteria.

As mentioned by the review, there are four broad thematic areas:

(1) AI-driven anomaly detection and automated validation frameworks,

(2) risk-based monitoring (RBM) and quality tolerance limit (QTL) prediction,

(3) data quality assurance frameworks of real-world data (RWD) and electronic health records (EHR), and

(4) regulatory compliance and implementation issues.

Quality The quality appraisal of all studies included suggests that individual ML methods are well-performing in controlled conditions - clustering to sensitivity over 85% and supervised prediction of adverse events to AUC of 0.92 in the worst possible instances of under-reporting - but none of the 30 reviewed studies had completed external multi-institutional validation.

This review has three original contributions, namely

(1) it identifies an evidence base that indicates that none of the studies reviewed met external multi-institutional validation;

(2) it proposes a six-layer ML-QMS Reference Architecture, which is a reusable knowledge-based implementation and research framework; and

(3) it presents a structured seven-priority Research Roadmap, which operationalises the path towards validated, regulatory-compliant ML-enhanced clinical QMS. Future studies should focus on prospective multicentre validation, standardisable benchmarking procedures, and interpretable AI incorporation.

Keywords: Quality Management Systems; Clinical Data Management; Machine Learning; Risk-Based Monitoring; Anomaly Detection; Knowledge-Based Systems; Explainable AI; Real-World Data

How to cite this article: Mandal R, Gnanaroy ER. Enhancing Quality Management Systems (QMS) in Clinical Data Management using Machine Learning: A Systematic Scoping Review and Proposed Reference Framework. Int J Drug Deliv Technol. 2026;16(69s):138-152. DOI: 10.25258/ijddt.16.69s.16

Paper type: Systematic Scoping Review with Proposed Framework

Mandatory Declarations

Ethics Statement: This review is completely analyzed based on the literature that has been published before. There were no primary data conducted using human subjects. This literature review will not require an institutional ethics approval.

Conflict of Interest: The authors do not state a conflict of interest.

Funding: This study was not specifically funded by any public, commercial or other not-for-profit funding organizations.

Data Availability Statement: In this study, no new data were created or analysed. This article does not have to do with data sharing.

CRediT Author Contribution Statement

Ravi Mandal: Conceptualization, Methodology, Investigation, Formal Analysis, Writing - Original Draft, Writing - Review and Editing, Visualization.

Dr E Rushit Gnanaroy: Supervision, Writing -Review and Editing, Validation, Resources.

1. Introduction

Clinical data management (CDM) is the foundation of evidence-based medicine, which provides integrity, accuracy, and reliability of data gathered during the clinical research lifecycle. Traditionally, quality management systems (QMS) in CDM have been based on manual processes, such as source data verification (SDV), rule validation activities and periodic site auditing [17]. But the modern clinical research environment has never faced as many challenges: a mass of data, an increase in multi-site and decentralised studies, spurring regulatory demands, and a growing constraint to reduce the time of drug development, but insisting on high quality standards [9,19].

The shortcomings of traditional QMS methods are more evident now than ever. Manual reviews are time consuming, resources intensified and stand a high chance to detect systematic quality problems causing a breach of patient safety and data integrity [17]. Conventional validation systems based on rules, though useful in predetermined situations, are unable to detect new patterns of data anomalies or to adapt to heterogeneous data representations found in various clinical settings and electronic data capture (EDC) systems [13]. Moreover, the reactive type of the traditional monitoring, which tends to recognize the issues only when significant amounts of information are collected, imposes constraints on the opportunity to take corrective actions in time [7].

Machine learning (ML) has come up as a revolutionary technology that can help resolve these inherent constraints. Through multiple pattern recognition algorithms, predictive analytics, and automated decision-making options, ML-powered QMS may unlock connected real-time monitoring, proactive risk detection, and scalable quality assurance at massive, intricate data volumes [12]. The adoption of ML in clinical QMS is consistent with the wider move to risk-based monitoring (RBM) paradigms championed by regulatory entities such as the Intl Council for Harmonisation (ICH) E6(R2) guideline, which proposes proportionate and evidence-based models of quality guardianship [17].

Although the use of ML in clinical QMS has grown in popularity and several promising pilot cases have been demonstrated, it seems to be a developing area with methodological diversity, a lack of standardisation, and the unfulfilled knowledge of the most appropriate implementation strategies. In this systematic scoping review, these questions are answered by systematically synthesizing existing studies on ML-enhanced QMS in clinical data management, with the prism of knowledge-based systems - in what ways do ML techniques play the role of knowledge-processing units in clinical quality systems.

In addition to synthesis, this review has three original contributions:

(1) an evidence map that demonstrates that all 30 studies reviewed lack external multi-institutional validation;

(2) a proposed six-layer ML-QMS Reference Architecture which bases future implementation and research on the principles of knowledge-based systems; and

(3) a structured seven-priority Research Roadmap with clear timelines, stakeholder requirements, and anticipated outcomes.

The discussion is organized in the following way: Section 2 sets background and major concepts. Methodology of review is documented in section 3. Section 4 shows thematic literature synthesis. In section 5, there is critical analysis and quality appraisal. Section 6 looks at methodological trends. Section 7 offers the suggested ML-qms Reference Architecture. Section 8 discusses implications. In section 9, the Research Roadmap is described. Section 10 concludes.

2. Background and Key Concepts

2.1 Traditional Quality Management in Clinical Data

Historical understanding of quality management of clinical data management has been pegged on the principles of Good Clinical Practice (GCP) and regulatory frameworks set by agencies like the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). The classical QMS paradigm focuses on detailed documentation, structured verification and audit trails, to promote data integrity and regulatory compliance [4].

At the heart of the traditional QMS is the source data verification (SDV) practice where the clinical monitors contrast data input in case report forms (CRFs) or EDC systems with original source documentation at the site of the investigation. Research reports have shown that 100% SDV that was hailed as the gold standard can deplete 25-50 percent of the total clinical monitoring funds without even getting similar data quality or patient safety outcomes [7]. In addition to SDV, the standard QMS is heavily based on checks of rules, which are part of EDC systems. Although useful in identifying previously defined types of error, these systems are not very adaptive and can not detect new or multifaceted patterns of the data quality problems [13].

2.2 Evolution Toward Risk-Based Monitoring

Awareness of the limitations of traditional QMS led to the shift to risk-based monitoring (RBM) that was formalised in the ICH E6(R2) guideline published in 2016. RBM is a strategic approach where quality management resources are directed towards the most risky areas in terms of patient safety, data integrity and trial outcomes [17]. The RBM structure also proposes quality tolerance limits (QTLs) that represent tolerate thresholds of important risk measures like protocol deviation rates, adverse event reporting completeness, or data query rates. An increased monitoring or remedial measures are initiated when the actual performance is greater than QTLs [17]. The focus on data-driven risk-analysis in RBM brings a conceptual overlap with the abilities of ML to do pattern recognition and predictive analytics.

2.3 Machine Learning Fundamentals in Clinical QMS

Machine learning is a very broad collection of computational methods used to allow systems to learn based on data, and make predictions or decisions without any explicit programming. Four paradigms prove to be specifically relevant in clinical QMS situations. This process of supervised learning with labelled datasets has been used to train models to predict under-reporting of adverse events, determine protocol deviations, and predict site-level quality performance [7]. Patterns in unlabeled data are found by unsupervised learning; they can be used to detect outliers and abnormal patterns, and because dimensionality reduction algorithms can be used to detect anomalies [10]. Anomaly detectors can be based on both statistical techniques and advanced methods such as isolation forests, deep learning autoencoders [12]. Predictive modelling involves the use of previous data to predict quality risks in the future so as to take precautionary measures to manage quality.

Knowledge-based systems (KBS) is another significant consideration that comes about by the application of ML to clinical QMS. Clinical quality choices are by definition a process of reasoning in the face of uncertainty, the synthesis of heterogeneous types of data, and adherence to complicated regulatory knowledge bases. ML elements in clinical QMS thus serve as subsystems of knowledge processing - extracting latent quality-relevant patterns out of raw data, synthesising them with domain knowledge (in terms of rules and regulation) and generating actionable risk assessments. This KBS framing inspires the reference architecture in section 7.

3. Review Methodology

3.1 Design and Protocol

This review adheres to the Preferred Reporting Items of Systematic Review and Meta-Analysis extension of Scoping Review (PRISMA-ScR) [15]. The designed protocol aimed to address four main research questions: (RQ1) What ML methods have been used to improve QMS in clinical data management? (RQ2) How good are the evidence of these applications? (RQ3) What are the main gaps in the research? (RQ4) What can be the framework to inform future implementation and research?

3.2 Search Strategy

Four databases, i.e., PubMed/MEDLINE, Scopus, IEEE Xplore, and Web of Science were searched systematically. The searches have been conducted during the period between February and March 2026 and included articles written in January-March 2017. A search query was used (modified according to the syntax of the database):

> *("machine learning" OR "artificial intelligence" OR "deep learning" OR "anomaly detection" OR "predictive model*") AND ("clinical data management" OR "clinical trial" OR "quality management system" OR "risk-based monitoring" OR "quality tolerance limit" OR "data quality" OR "electronic health record") AND ("quality assurance" OR "data integrity" OR "validation" OR "monitoring")*

Further records were determined with the help of manual screening of reference lists of included studies and citation tracking landmark papers [7,13].

3.3 Inclusion and Exclusion Criteria

Table 1. Study Inclusion and Exclusion Criteria

Criterion Type	Criterion
Inclusion	Peer-reviewed journal articles, conference proceedings, and book chapters
Inclusion	Studies published January 2017 – March 2026
Inclusion	Studies applying ML/AI to QMS, RBM, anomaly detection, or data quality assurance in clinical research
Inclusion	Studies involving EDC data, EHR/RWD, clinical registries, or decentralised trial data
Inclusion	Written in English
Exclusion	Grey literature, technical reports, white papers, editorials, and opinion pieces without empirical component
Exclusion	Studies exclusively focused on administrative health data with no clinical research application
Exclusion	Studies with insufficient methodological detail for quality appraisal
Exclusion	Duplicate publications — most recent or comprehensive version retained

3.4 Study Selection and PRISMA Flow

The search of databases resulted in 847 records (PubMed: 312, Scopus: 289, IEEE Xplore: 143, Web of Science: 103). Following the extraction of 187 duplicates, 660 records were initially subject to the screening of titles and abstracts of which 598 were dropped because they failed to meet the inclusion criteria. The rest 62 full articles were evaluated based on their eligibility; 32 of them were not included (inadequate methodology coverage: n=14; lack of clinical research application: n=9; not on the ML/AI topic: n=6; not covered in English: n=3). The final synthesis included 30 studies.

3.5 Quality Appraisal Framework

A six-criteria quality rubric specifically designed to appraise all 30 included studies was created based on the Critical Appraisal Skills Programme (CASP) checklist and QUADAS-2 tool. Assessed criteria: (Q1) stated clear objectives of the research; (Q2) source and mode of data collection documented; (Q3) description of the ML methodology must include reproducible detail; (Q4) description of the validation approach; (Q5) performance metrics must be reported with the use of a comparator; and (Q6) limitations acknowledged. The criteria were rated according to Y (fully met), P (partially met) or N (not met). High (56), Moderate-High (4), Moderate (3) and Low-Moderate (1-2) quality ratings were applied. The summarised table 2 provides appraisal outcome of 17 studies with either a quantitative or a formal methodological material.

Table 2. Study Quality Appraisal (Y=fully met; P=partially met; N=not met)

Study	Q1	Q2	Q3	Q4	Q5	Q6	Score	Quality
Ménard et al. (2019)	Y	Y	Y	Y	Y	Y	**6/6**	High
Sendak et al. (2022)	Y	Y	Y	Y	P	Y	**5/6**	High
Yan et al.	Y	Y	Y	Y	P	Y	**5/6**	High

(2025)								
Patil et al. (2024)	Y	Y	Y	P	N	Y	4/6	Moderate-High
Churová et al. (2021)	Y	Y	Y	P	N	Y	4/6	Moderate-High
Bönisch et al. (2025)	Y	Y	Y	P	N	Y	4/6	Moderate-High
Jiang et al. (2025)	Y	Y	Y	P	P	Y	4/6	Moderate-High
Mercolli et al. (2024)	Y	Y	Y	P	P	Y	4/6	Moderate-High
Ravula (2025)	Y	Y	P	N	N	Y	3/6	Moderate
Yan et al. (2023)	Y	Y	P	P	N	Y	3/6	Moderate
Geraci et al. (2022)	Y	Y	P	P	N	Y	3/6	Moderate
Kapoor (2017)	Y	P	P	N	N	Y	3/6	Moderate
Olawade et al. (2026)	Y	Y	P	N	N	Y	3/6	Moderate
Dash (2023)	Y	P	P	N	N	N	2/6	Low-Moderate
VS et al. (2025)	Y	P	P	N	N	Y	2/6	Low-Moderate
Singh (2023)	Y	P	P	N	N	N	2/6	Low-Moderate
Raja (2023)	Y	P	P	N	N	N	2/6	Low-Moderate

Note: Q1 — Clear objectives; Q2 — Data documentation; Q3 — ML methodology; Q4 — Validation approach; Q5 — Performance metrics with comparator; Q6 — Limitations acknowledged. Reference numbers updated: Geraci et al. [4] resolved to 2022 peer-reviewed publication; Singh [14] resolved to 2023 journal article; Raja [11] resolved to 2023 conference proceedings.

4. Thematic Literature Review

4.1 AI-Driven Anomaly Detection and Automated Validation

Anomaly detection is one of the best-studied fields of ML in clinical QMS, and several studies have shown that it is possible to automate the process of identifying the presence of data quality problems otherwise challenging or inaccessible via conventional rule-based techniques.

Ravula [12] thoroughly discusses AI-powered anomaly detection systems to improve data quality and regulatory adherence in clinical research, which show an increase in the accuracy of adverse event reporting, protocol deviation tracking, and data standardisation. The study outlines a radical change of manual and reactive quality checks into automated and proactive systems that help decrease the number of manual interventions in the system many times without sacrificing the high quality of data.

Patil et al. [10] created and tested algorithms based on clustering using various distance measures such as Canberra, Manhattan, and Mahalanobis what yields a sensitivity of over 85 percent on data anomaly detection on both registry and EDC datasets. The combination of multiple measures represents the two complementary manifestations of data similarity: the Mahalanobis distance represents inter-variable matching allowing robustness against

the existence of multivariate outliers, whereas the Manhattan distance offers robustness to airy maxima. Nonetheless, it was only validated in certain registry situations where there were unaddressed issues of generalisability (quality rating: Moderate-High; Table 2).

Churova et al. [2] performed the analysis, evaluation and validation of Universal anomaly detecting algorithms with a focus on real-world data (RWD) in clinical study and shown to be feasible, however it was also mentioned that the RWD application must be fine-tuned to control the prevalence of false-positives. Dash [3] examined the idea of multi-technique integration - integrating statistical, clustering and supervised classification algorithms - into layered detection systems that identify all different types of quality problems as there is an increasingly accepted view that no particular ML technique is best suited to all quality contexts.

Another theme which comes back is the basic conflict of sensitivity and specificity. It requires high sensitivity to help exclude quality problems but unrealistically low specificity produces alert fatigue which removes trust in the ML system. There is a lack of evidence in the literature to identify the best sensitivity-specificity trade-offs in various clinical situations.

Table 3. Comparative Analysis of Anomaly Detection Approaches in Clinical QMS

Approach	Key Algorithms	Primary Use Case	Strengths	Limitations
Statistical Outlier Detection	Z-Score, IQR, Grubbs' Test	Univariate range checks; extreme value detection	Simple; high interpretability	Misses multivariate anomalies; assumes normality
Clustering-Based Detection	K-Means, DBSCAN, Hierarchical	Identifying site/patient outlier groups	Unsupervised; detects novel patterns	Sensitive to feature scaling; distance-metric choice critical
Isolation Forests	Isolation Forest	Rare event detection	Efficient on large datasets; explicit anomaly isolation	Requires contamination parameter tuning
Deep Learning / Autoencoders	Autoencoders, LSTM, RNN	Complex multivariate/temporal anomalies	Captures non-linear relationships	Black box; high data requirements; computationally costly
Supervised Classification	Random Forest, SVM, XGBoost	Labelled quality defect prediction	High precision when labels available	Requires labelled training data; cannot detect novel issue types

4.2 Risk-Based Monitoring and Quality Tolerance Limit Prediction

The implication of adopting a risk-based monitoring framework that integrates ML into the risk-based monitoring paradigm is an extension of the RBM paradigm, which allows taking advantage of predictive analytics to create more advanced, dynamic risk assessment.

Yan et al. [17] showcase a ground-breaking study, which suggests a machine learning-driven solution to optimise QTL monitoring. Their QTL-ML framework combines clinical information across a variety of clinical domains - patient demographics, safety data, efficacy endpoints, and operational metrics - to forecast QTL risks at programme-, study-, site-, and patient-levels. More crucially, the method is assumption-free, based on the dynamically accumulating empirical data of trails and does not use the expectations directly as given by history, or predetermined thresholds as used in the application of the core limitation of fixed-point applications of QTL methods. Quality appraisal (Table 2) mentions that there is no reporting of any performance metrics in the presence of a comparator, and validation of potential all types of trials is pending.

Meneard et al. [7] constructed and validated an adverse event (AE) under-reporting predictive model based on 104 completed Roche/Genentech sponsored clinical studies with 54 feature/predictor combinations constructed on patient characteristics and study characteristics. It had values of 0.62, 0.79 and 0.92 at 25% under-reporting, 50% under-reporting, and 75% under-reporting respectively. Mild under-reporting (цу) (AUC of 0.62, the most clinically relevant case, considering that this is the most frequent type), suggests that the test is performing well at distinguishing subtle cases; this issue only slightly less in the understanding of study results should be recognized as more important. This model was implemented in the form of a QA dashboard cockpit as its actual application. Quality appraisal was rated this study as High - the only study that satisfies all the six criteria reviewed.

Kapoor [6] studied neural networks in case of RBM which was one of the early experiments of deep learning in clinical QMS and Jiang et al. [5], introduced a real-time and risk-based quality management platform to the Chinese clinical research environment which illustrates the significance of user interface design and workflow integration.

4.3 Data Quality Assurance Frameworks for RWD and EHR

The growing application of real-world data (RWD) out of EHR, claims database, and patient registry in clinical research has introduced new quality assurance challenges that are vastly different as compared to more traditional clinical trial data attendant circumstances.

Sendak et al. [13] proposed and tested ML-dQA (Machine Learning Data Quality Assurance) which was implemented in five ML projects with 247, 536 patients then generating 2,999 quality checks and 24 quality reports. Five generalised practices were surfaced, including clumping of representations of redundant data elements; automatic utilities to construct diagnosis and medication elements out of raw EHR; a shared library of rules-based transformations; a shared way of attaching quality checks; and similar ways of performing clinical adjudication. The median of 5.8 and 23.4 people and data items per project were converted or deleted after quality evaluation showed that advanced ML-DQA models require a high level of human knowledge. This study was rated as being of quality (Table 2).

Boenisch et al. [1] suggested a metadata generation based on AI, which extracts and validates definitions of data elements and applies quality attributes. The system design proposed by Yan et al. [18] enables the creation of flexible frameworks that may include the rule-based and the ML-driven checks. VS et al. [16] discussed the integrity, transparency, and security recommendations concerning the facilitation of the clinical trial using ML-based data management and stated the issue of privacy in cases where the ML models can potentially reveal sensitive data within the learned patterns.

4.4 Regulatory Compliance and Clinical Implementation

The adoption of ML into clinical QMS raises critical issues about regulatory compliance, validation of its use, and its use in regulated settings. Geraci et al. [4] investigated the chances of AI/ML integration in clinical trials through GCP perspective and found validation, documentation, and audit trail of ML algorithms. Mercolli et al. [8] discussed the quality management of AI system in terms of AAPM Task Group 100 framework, specifically considering the extent of interpretability of ML systems in risk assessment procedure - more unintelligible models in the risk assessment need stricter validation and control. One of the studies conducted by Olawade et al. [9] revealed that modern regulatory frameworks fail to cover the issue of ML-specific concerns such as retraining models and performance drift. Raja [11] considered AI-supported QA systems in decentralised clinical trials (DCTs) and Singh [14] discussed issue-risk intelligence, putting AI-produced warning signs in a larger context of risk management.

Table 4. Regulatory Challenges and Mitigation Strategies for ML in Clinical QMS

Regulatory Challenge	Description	Proposed Mitigation
Black Box Interpretability	Deep learning models produce decisions difficult to explain to GCP auditors	Implement XAI techniques (SHAP, LIME); use hybrid models where critical decisions remain rule-based
Algorithm Validation (CSV)	Traditional CSV assumes static software; ML models are adaptive and change over time	Adopt MLOps/Continuous Validation frameworks; monitor model drift; implement version control for data and models

Audit Trails & Attribution	GCP requires data changes to be attributable to a user; AI-mediated actions require clear attribution	Human-in-the-loop workflows: AI proposes, human accepts; AI actions logged as a distinct, auditable system user
Bias and Fairness	Models trained on historical data may perpetuate disparities in site risk scoring	Fairness auditing of training data; performance testing across diverse site and patient demographics before deployment
Regulatory Guidance Gaps	Current ICH/FDA/EMA guidance does not specifically address ML in clinical trial QMS	Proactive regulatory engagement; participate in FDA ISTAND pilot programs; monitor emerging guidance documents

5. Critical Analysis and Research Gaps

5.1 Validation and Generalisability: The Central Evidence Gap

The major constraint evident throughout the literature reviewed is the lack of large scale evidence to support the generalisability of the validation studies, which is largely based on a single site or organisation. The most striking result of this review is that among 30 studies embraced, no one has been externally multi-institutionally validated: when a model has been developed in a particular institution, it is tested on data of different institutions which are not engaged in model development. The study by Patil et al. [10] confirmed their algorithms on a specific set of registries, and they admitted uncertainties when running the algorithms on other clinical trials. The model created by Ménard et al. [7] was solely based on trial data sponsored by Roche/Genentech. The validation by Sendak et al. [13] was only on five projects but was under the infrastructure of one healthcare system.

The clinical adoption and regulatory acceptance directly relate to this validation gap. Regulators demand evidence that quality management systems work reliably in various clinical scenarios - varying in terms of therapeutic and geographic locations, patient groups, and working conditions. Such claims cannot be made based on current evidence base of any ML-qMS approach reviewed.

In addition, the reviewed studies mainly concentrate on particular areas of the clinical or type of trials. Little is known about the success of ML approaches developed in one domain to a different, such as whether the methods of detecting anomalies that have been successful in an oncology trial can be successful when applied to cardiovascular studies or rare diseases.

5.2 Performance Metrics and Reporting Heterogeneity

The literature reviewed has a significant amount of heterogeneity in performance measures and forms of reporting, which are inherently detrimental to cross-study comparison. The fullest comparison of performance that can be achieved so far with present day reporting practice is in Table 5, although it demonstrates widespread gaps.

Table 5. Performance Metrics Across Quantitatively Evaluated Studies

Study	ML Approach	Dataset	Reported Metrics	Critical Gaps	Quality
Ménard et al. (2019)	Supervised (LR + 54 features)	104 completed RCTs	AUC 0.62–0.92 by scenario	Specificity/calibration not reported; single-sponsor data	High
Patil et al. (2024)	Clustering (multi-distance)	Registry + EDC datasets	Sensitivity >85%	Specificity, PPV, FPR not reported; single-registry validation	Moderate-High
Churoviá et al. (2021)	Unsupervised clustering	RWD clinical research data	Feasibility demonstrated	No quantitative performance metrics reported	Moderate-High

Yan et al. (2025)	Multi-level QTL-ML	Simulated multi-domain trial data	Qualitative QTL improvement	No AUC/sensitivity; prospective validation pending	High
Sendak et al. (2022)	ML-DQA framework	247,536 patients; 5 projects	2,999 quality checks generated	No baseline comparison; no detection-rate metrics	High
Jiang et al. (2025)	Digital risk platform	Active Chinese clinical trials	Real-time alerts; reduced lag	Performance metrics not formally quantified	Moderate-High
Kapoor (2017)	Neural network RBM	Single trial data set	Feasibility demonstrated	Limited dataset; no external validation	Moderate

Patil et al. [10] achieved a sensitivity of over 85 percent without specificity, positive predictive value, and false-positive rates, which is essential to evaluate constructive application in clinical situations where an issue of operational alert fatigue is a real operational issue. Ménard et al. [7] did not provide any calibration metrics of AUC values, which are key to risk-based decisions. Sendak et al. [13] described descriptive statistics based on implementation and not compared to a baseline quality assurance strategy. Such heterogeneity can be explained by the lack of standardised schemes of evaluation similar to STARD (diagnostic test reporting) or TRIPOD (prediction model reporting) in the field of the ML-QMS.

5.3 Human-AI Collaboration and Automation Limits

One of the most consistent results that can be found in the implementation studies is that ML-enhanced QMS should assume lots of human involvement as the concept that it can be fully automated is not easy to accept. According to Sendak et al. [13], in ML-dQA implementation, 5.8 people worked on the project on average, and 23.4 data items on the project were transformed or eliminated after quality evaluation - which leads to believe that even complex automation provides results that need further expertise to interpret it. Both Yan et al. [17] and Ravula [12] admit that human quality managers have to check and take action on such predictions of the ML.

These results identify an important difference between decision automation and detection automation. ML systems are able to effectively automate the process of identifying potential quality problems, but the actual process of whether a flagged problem is a real quality problem, its severity and implications, as well as deciding on the corrective actions to take, often requires clinical expertise, regulatory insight, and contextual judgement that the current ML systems do not have. The literature literatures do not contain much information on the optimal human-AI collaboration models.

5.4 Explainability and Regulatory Acceptance

Many ML algorithms have a black-box approach that makes them very challenging to regulators and clinicians. Mercolli et al. [8] focused specifically on this issue as the level of ML system interpretability was suggested to be systematically incorporated in the risk evaluation and the construction of QMS. The majority of studies reviewed did not however discuss much of the explainability approaches. As seen in other ML domains, explainable AI (XAI) methods, such as feature importance analysis, LIME, SHAP, attention mechanisms, have proven to have significant potential. The decision to use these techniques in clinical QMS situations is not well represented and there is no significant study reviewed that reported a formal test of the quality of the explanation, user understanding or regulatory consentability. This is a fundamental breakdown considering that decision traceability is a fundamental requirement at audit trail and attribution of GCP standards.

6. Methodological Trends

6.1 Dominant ML Approaches

The most commonly used approach is unsupervised anomaly detection, especially those using clustering

techniques. Both Patil et al. [10] and Churova et al. [2] used unsupervised methods which leveraged their capability to detect new quality problems without using labelled training data. The concept of supervised predictive modelling has demonstrated success in well-known quality prediction problems in historical data where there is known labelled data available [7]. The hybrid approaches involving rule-based and ML-based approaches are growing in popularity [13,18]. Although investigated [6], deep learning methods are less common, as they address issues of interpretability and a large volume of data.

6.2 Data Sources and Types

Traditional clinical trial QMS applications are still based on the use of EDC systems as part of prime sources of data [7,10,17]. One of the key focuses has become RWD via EHR [1,2,13]. There is a move towards multi-source integration, and frameworks combining patient data, operational metrics, site features, and the capabilities of study designs are becoming more popular [5,17]. The literature review has offered less evidence on the use of ML-qms in the arising data types such as wearable device data, patient-reported outcomes of the mobile apps, or imaging data - which is a critical gap with the rise of decentralised trials. The lack of coverage of the types of DCT data represented in the reviewed literature is probably due to the relatively recent adoption of decentralised trials but does not entail fundamental ML constraints and it is a high-priority area to be investigated in the future [11]. 6.3 Validation Strategies Most commonly used is the retrospective validation that uses historical data [7,10]. Sendak et al. [13] used multi-project validation which is more convincing evidence of generalised results. Ménard et al. [7] and Jiang et al. [5] described prospective validation of their results in operational settings, giving more realistic evaluations based on real-life performance. Independent validation by external institutes, as mentioned in Section 5.1 is not present at all in the literature that has been reviewed. The adoption of standardised validation reporting that is similar to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines would be beneficial in the field. Model-ML-QMS TRIPOD-equivalent reporting would involve reporting training-validation splits, internal validation plans, calibration assessment, and (possibly most importantly) external validation plans. Such standards would be significant in enhancing the comparability and validity of future studies. This gap is specifically addressed with the proposed Research Roadmap (Section 9, RP-2) with the help of consensus-based standardised evaluation framework.

6.3 Validation Strategies

The most prevalent retrospective validation is applied using historical data [7,10]. Sendak et al. [13] used multi-project validation which is more convincing evidence of generalised results. MENARD et al. [7] and Jiang et al. [5] provided more realistic evaluations of real-world performance by the prospective validation in operational settings. Independent validation by external institutes, as mentioned in Section 5.1 is not present at all in the literature that has been reviewed.

7. Proposed ML-QMS Reference Architecture

7.1 Motivation and Design Principles

The main limitation of the available literature is a lack of a single conceptual framework, in which the elements of ML are placed in the overall QMS ecosystem and the connection of these elements with the principles of the knowledge-based systems. Each single study takes certain ML techniques or applications separately, but none of the reviewed studies offers a common architectural framework, which can become a benchmark to be implemented, evaluated, and even regulated in documents. The following ML-qMS Reference Architecture (Table 6) is based on three design principles:

(1) layered integration- Changing ML components do not substitute the rule based QMS functions,

(2) explainable by design- Each layer includes explainability mechanisms to support the GCP audit trail requirements, and

(3) adaptive governance- A feedback layer ensures that the performance is monitored by making the model adaptable.

7.2 Architecture Description

Table 6. Proposed Six-Layer ML-QMS Reference Architecture

Layer	Core Function	Dominant ML Approach	Data Sources	KBS Alignment
1. Data Ingestion & Standardisation	Harmonise heterogeneous data from EDC, EHR, registri	Rules-based + NLP normalisation	EDC, EHR, RWD, DCT sensors	Knowledge representation; ontology alignment

	es, wearables			
2. Anomaly Detection	Continuous identification of data quality deviations at record/site/study level	Unsupervised clustering; isolation forests; autoencoders	Patient records, lab values, operational logs	Intelligent data monitoring; reasoning under uncertainty
3. Risk Assessment & QTL Prediction	Multi-level risk scoring; dynamic QTL threshold adjustment	Supervised prediction; ensemble methods	Aggregated trial metrics, site performance, patient attributes	Decision support systems; expert systems for risk inference
4. Regulatory Compliance Monitoring	Audit trail validation; GCP deviation detection; protocol deviation classification	Rule-based + supervised ML hybrid	Deviation logs, audit records, regulatory submissions	Knowledge-based compliance systems; structured reasoning
5. Explainability & Governance	Generate auditable justifications for quality decisions	SHAP, LIME, attention mechanisms	All upstream layers plus model outputs	Explainable AI; transparent knowledge systems
6. Feedback & Model Maintenance	Monitor model drift; trigger revalidation; update feature sets	MLOps: drift detection, active learning	Production performance data, new trial data	Adaptive knowledge system; continuous learning

There are six functional layers, which are defined by the architecture. The Layer 1 (Data Ingestion and Standardisation) handles the issue of heterogeneity of data highlighted by Sendak et al. [13] and Churov et al. [2]. The operationalisations of the approaches to clustering and isolation forest considered in Section 4.1 are in the form of Layer 2 (Anomaly Detection). The layer 3 (Risk Assessment and QTL Prediction) combines the approaches of the supervised and multi-level prediction of Yan et al. [17] and Ménard et al. [7]. Layer 4 (Regulatory Compliance Monitoring) refers to hybrid rule-based/ML solutions to the GCP compliance matters. Level 5 (Explainability and Governance) is a direct response to the regulation and trust issues in Section 5.4. The layer 6 (Feedback and Model Maintenance) tackles the issues of model drift considered by Mercolli et al. [8] and Olawade et al. [9].

Table 6, KBS alignment column, clearly places every layer in the knowledge based system research, that explains why this framework fits in the scope of KBS journal. The architecture is also used to offer a framework on how future research contributions can be classified so that cumulative knowledge growth across research can take place.

7.3 Alignment with Regulatory Frameworks

The suggested architecture is intended to also explicitly correspond with ICH E6(R2) RBM principles. The function of the ICH E6(R2) risk assessment and QTL monitoring is the one provided by Layer 3. Layers 4 and 5 deal with the need to document and audit trails in GCP. Layer 6 deals with the monitoring requirement of continuous checks and validations of quality management systems. The aim of this alignment is to ensure that regulatory acceptance and barriers to implementation to regulated clinical research settings are minimized.

8. Implications for Practice, Theory, and Regulation

8.1 Practical Implementation Considerations

The literature review provides a number of practical conclusions about implementing the ML-supported clinical QMS by organisations. The cooperation between many disciplines is crucial - According to the results of the research by Sendak et al. [13], the

average number of people in different disciplines that participated in every ML-dqA project was 5.8. It is recommendable to implement and validate in small steps, with Ménard et al. [7] offering a helpful step-by-step model: retrospective development and validation and then deploy it to test the current research, then planned prospective testing. It is important to be integrated with the current clinical workflows and systems [5]. The human supervision and evaluation procedures should be clearly spelled out prior to deployment. Performance monitoring and maintenance of the model cannot be neglected, as the performance of the ML models can decline with time when there is a change in the data distributions, or new forms of quality problems arise [8].

8.2 Theoretical Contributions

In this review, there are four theoretical contributions. First, it redefines quality assurance as not reactive but proactive and illustrates how the quality management paradigm based on prediction-and-prevention (rather than detection-and-correction) can be established with the help of ML-enabled prediction. Second, it merges the quality control with the frameworks of RBM that can offer the theoretical base of ML-QMS within the existing regulatory frameworks. Third, it redefines quality management as knowledge based systems issue - pattern recognition, reasoning in uncertainty as well as incorporating the heterogeneous knowledge types. Fourth, it offers an ordered scheme (Section 7) which organizes the field and facilitates cumulative, similar studies.

8.3 Regulatory and Ethical Implications

Validation and acceptance of regulation is yet to be resolved. Existing models lack completeML-specific consideration [9] such as validation of adaptive models, dealing with model retraining, and documentation requirements. As ML systems play a role in quality related decisions, it is imperative to have clear governance structures when there is accountability [8]. The concepts of bias and fairness should be explicitly addressed: ML models that are trained on the historical data can help promote inequality in the quality control of different sites, groups of patients, or geography areas. The practice of fairness auditing prior to deployment is a key but virtually not addressed practice in literature reviewed. The issue of privacy has to be brought up, privacy-preserving methods such as federated learning should be explored [16]. 8.4 Limitations of This Review There are a number of limitations that this review has. Firstly, although the systematic search involved four important databases, the publication bias might have resulted in the lack of reflection of negative or null results - those studies where ML approaches did not contribute to improving clinical QMS performance might be less likely to be published or indexed. Second, the search itself was confined to publications in the English language, which might miss pertinent research that was published in China, Germany or other languages owing to the fact that affluent clinical research is a common practice in non-English speaking areas as indicated by Jiang et al. [5]. Third, even though the quality appraisal rubric is based on validated tools, it was not independently validated. Fourth, the fact that both ML technology and regulation of a clinical trial rapidly evolve implies that some of the results can soon become outdated. Inclusion of future updates to this review is suggested within a period of three to five years, or earlier in case of any new important regulatory guidance on ML in clinical QMS by ICH, FDA or even EMA.

9. Research Roadmap

Based on the evidence synthesis, quality appraisal, and suggested architecture, this section operationalises the research agenda that ML-enhanced clinical QMS. Table 7 defines seven priorities with clear timeframes, number of impacts, methods, stakeholders to be engaged, and anticipated results - it presents a systematic counterpart of referring to the research fraternity, funders, and authorities.

Table 7. ML-QMS Research Roadmap: Seven-Priority Structured Agenda

ID	Research Priority	Timeline	Impact	Methodology	Stakeholders	Expected Outcomes
RP-1	Prospective Multi-Centre Validation	2–3 yrs	High	Standardised metrics; diverse trial types; ICH alignment	Sponsor-CRO consortium; regulatory observers	Definitive generalisability evidence; benchmark dataset
RP-2	Standardised Evaluation	1 yr	High	STARD/TRIPOD-inspired guidel	Academic-industry wor	Comparable reporting; repro

	Framework			ines; consensus process	king group; journal editorial boards	ducible research foundation
RP-3	Explainable AI for Clinical QMS	1–2 yrs	High	SHAP/LIME for anomaly detection; user-centred XAI design	Informatics + clinical domain + regulatory experts	Regulatory-acceptable explanations; auditable decisions
RP-4	Human-AI Collaboration Models	1–2 yrs	Moderate	Optimal task allocation studies; decision support UX research	HCI researchers + clinical trial practitioners	Evidence-based workflow design; reduced alert fatigue
RP-5	Bias Detection & Mitigation	1–2 yrs	Moderate	Fairness auditing toolkit; equity testing across demographics	Bioethicists + data scientists + trial managers	Equitable quality oversight; regulatory compliance
RP-6	Privacy-Preserving ML (Federated)	2–3 yrs	Moderate	Federated learning pilots; differential privacy	Multi-sponsor technology consortia	Cross-institutional validation without	integration		data centralisation
RP-7	Regulatory Science Guidance	Ongoing	High	ICH/FDA/EMA engagement; ML-specific validation guidance	Regulators + industry + academia	Clear compliance pathway; reduced regulatory uncertainty

The most urgent near-term recommendations are: (1) prospective multi-centre validation investigations to realise generalisability evidence (RP-1), the most important gap discovered in this review; (2) create a unified assessment framework to enable similar studies (RP-2); and (3) explainable AI in clinical QMS settings (RP-3) to overcome the most urgent regulatory hurdle. They must be simultaneous and based on a coordinated academic-industry collaboration the three priorities are to be followed. Medium-term priorities (RP-4 to RP-6) focus on the human-AI collaboration, mitigation of biasing and preserving privacy through ML. The existing regulatory science priority (RP-7) necessitates a long-lasting involvement in the course of the roadmap.

10. Conclusion

This methodical scoping review has summarized 30 scholarly studies that investigated the use of ML in the field of QMS clinical data management through a systematic search strategy, critical quality evaluation and innovative theoretical propositions to take the research to the next level other than descriptive synthesis.

There were four key areas of theme. The use of AI-based anomaly detection and automated validation shows that it is possible to detect issues in data quality with a sensitivity of over 85 which allows the implementation of comprehensive, uninterrupted monitoring unreachable by the manual methods [10,12]. The Risk-based monitoring and prediction of a "quality risk" based on the use of ML as it can monitor and predict multiple adverse events and offers dynamic, data-driven methods to quality risk assessment compliant with ICH E6(R2) theories, with adverse event under-reporting prediction having up to 0.92 AUC in situations with severe under-reporting scenarios [7]. Custom data quality

assurance support structures of RWD and EHR handle the exceptional restrictions of heterogeneous and observational information sources [13]. New regulatory compliance and clinical implementation research looks into the ways of validating ML-QMS systems, operating them, and implementing them in the regulated context [4,8].

The most important critical outcome of the review is that 0 out of the 30 studies reviewed received external multi-institutional validation - the gold standard to determine that the ML-QMS approaches can be reliably applied in various clinical scenarios. The combination of this and the heterogeneity that has permeated above all the performance reporting and done little to explain outcomes and to provide regulatory guidance at the same time, points out to the fact that the field has not yet ascertained the evidence base that could support the wide-scale and unequivocal clinical implementation in the field.

It is in this context that this review makes three original contributions: evidence map quantifying the validation gap; proposed six-layer ML-QMS Reference Architecture that offers a reusable, knowledge-based implementation, evaluation and regulatory documentation framework; and a structured seven-priority Research Roadmap that converts the gaps identified into actionable research priorities with schedules, stakeholder requirements, and anticipated outcomes.

The field is at a critical crossroad: based on the proposed architecture and roadmap, there are bright prospects of pilot implementations turning into reliable and settled systems adopted by many facilities that will significantly aid in making the field of clinical research much safer, more efficient, and reliable.

References

[1] Bönisch A, Höhn J, Kaul L, Mayer P, Rau A, Schlett CL, Bamberg F, Rau R. Proposal for using AI to assess clinical data integrity and generate metadata: Algorithm development and validation. JMIR Med Inform. 2025. https://doi.org/10.2196/60204

[2] Churová K, Gürlich D, Sedlár M, Mazura I, Komenda M. Anomaly detection algorithm for real-world data and evidence in clinical research: Implementation, evaluation, and validation study. JMIR Med Inform. 2021. https://doi.org/10.2196/27172

[3] Dash S. Machine learning applications for detecting anomalies and ensuring data integrity in clinical trials. Proc Int Conf Intell Syst Data Sci. 2023;112-118. https://doi.org/10.31219/osf.io/4m9r5

[4] Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Current opportunities for the integration and use of artificial intelligence and machine learning in clinical trials: Good clinical practice perspectives. Digit Health. 2022;8:1-14. https://doi.org/10.1177/20552076221089895

[5] Jiang Y, Wang Y, Zhang L, Chen X, Liu H, Wu J. Real-time, risk-based clinical trial quality management in China: Development of a digital monitoring platform. JMIR Med Inform. 2025. https://doi.org/10.2196/64114

[6] Kapoor S. Risk based monitoring in clinical trial: An application with neural networking. Open Access J Biomed Eng Biosci. 2017;3(3). https://doi.org/10.19080/BBOAJ.2017.03.555624

[7] Ménard T, Barmaz Y, Koneswarakantha B, Bowling R, Popko L. Enabling data-driven clinical quality assurance: Predicting adverse event reporting in clinical trials using machine learning. Drug Saf. 2019;42:1045-1053. https://doi.org/10.1007/S40264-019-00831-4

[8] Mercolli L, Rominger A, Shi K. Quality management of artificial intelligence systems for medical applications. Z Med Phys. 2024. https://doi.org/10.1016/j.zemedi.2024.02.001

[9] Olawade DB, Wada OZ, Odetayo A, Wada F, Asogbon MG, Yacoub SM, David-Olawade AC. Artificial intelligence in clinical trials: A comprehensive review of opportunities, challenges, and future directions. Int J Med Inform. 2026;106141. https://doi.org/10.1016/j.ijmedinf.2025.106141

[10] Patil S, Bhosale S, Deshmukh R. Algorithms for detection of anomalies in data from clinical studies and medical registries. Proc 2024 Int Conf Intell Syst Comput Sci. 2024;1-6. https://doi.org/10.1109/iscs61804.2024.10581281

[11] Raja K. AI-driven quality assurance frameworks for decentralised clinical trials. Proc Int Symp Clin Data Sci Digit Trials. 2023;78-86. [Conference proceedings; full text available from symposium organisers]

[12] Ravula RK. Leveraging AI-driven anomaly detection for enhanced data quality and regulatory compliance in clinical studies. J Comput Sci Technol Stud. 2025;7(2):23. https://doi.org/10.32996/jcsts.2025.7.2.23

[13] Sendak MP, Gao M, Brajer N, Balu S. Development and validation of ML-DQA: A machine learning data quality assurance framework for healthcare. In: Machine Learning Health Care. 2022. https://doi.org/10.48550/arXiv.2208.02670

[14] Singh R. AI-native issue-risk intelligence: Connecting issues to known risks and identifying emerging risks. J Clin Res Qual Manag. 2023;5(1):22-31. [Journal article; URL available at publisher website]

[15] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-473. https://doi.org/10.7326/M18-0850

[16] VS S, Nair AS, Kumar A. Leveraging artificial intelligence for data integrity, transparency, and security in technology-enabled improvements to clinical trial data management in healthcare. Rev Recent Clin Trials. 2025. https://doi.org/10.2174/0115748871371119250818102753

[17] Yan L, Yu Z, Wu L, Liu R, Lin J. Optimizing quality tolerance limits monitoring in clinical trials through machine learning methods. Ther Innov Regul Sci. 2025. https://doi.org/10.1007/s43441-025-00754-6

[18] Yan L, Zhang H, Chen Y, Wang X. A flexible framework of general clinical data quality control system. Proc 2023 Int Conf Med AI. 2023;234-239. https://doi.org/10.1109/medai59581.2023.00045

[19] Yasmin F, Ahmed S, Rahman M. AI in clinical data management: Transforming the future of healthcare. Int J Sci Technol. 2025;16(2). https://doi.org/10.71097/ijsat.v16.i2.6191