MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 10, Issue 2, July-December 2025

Predictive Analytics and Intelligent Data Integration for Improving Decision-Making in Healthcare Management

Dr. Manavendra Pal[1], Nirmala Venkatesh[2], Dhruvendra Rathore[3]
Assistant Professor[1], Associate Professor[2], Research scholar[3]
Department of Artificial Intelligence and Data Science
Baselios Mathews II College of Engineering, Kerala
E-mail: *manavendra.pal98@outlook.com[1]*

ABSTRACT

Healthcare organizations increasingly generate heterogeneous data through electronic health records (EHRs), laboratory systems, pharmacy transactions, claims, scheduling platforms, medical devices, and patient-facing applications. Yet data fragmentation, inconsistent terminology, missingness, temporal misalignment, and delayed access frequently prevent managers from converting information into timely decisions. This paper examines an integrated framework that combines standards-based data integration with predictive analytics to support capacity planning, readmission-risk management, workforce allocation, and proactive care coordination. A simulation-based experimental design was developed using 60,000 synthetic patient encounters and two analytical pathways: an EHR-only baseline and an integrated longitudinal dataset incorporating utilization, pharmacy, claims, and appointment features. Three models were compared: regularized logistic regression using the baseline feature set, regularized logistic regression using integrated features, and histogram-based gradient boosting using the integrated dataset. The simulation demonstrated that integration improved AUROC from 0.659 to 0.686 for logistic regression and improved F1 from 0.310 to 0.346 at an operational threshold of 0.20. A complementary latency simulation indicated a reduction in median decision-data assembly time from approximately 183 seconds to 133 seconds when standardized and cached integration services were used. The findings suggest that predictive performance and managerial responsiveness depend not only on model choice but also on the quality, interoperability, provenance, and timeliness of the data layer. The paper proposes a closed-loop architecture in which data quality monitoring, prediction, managerial action, and outcome feedback are treated as one governance process. The results are simulation findings rather than evidence from a real hospital and are

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

therefore intended to demonstrate the engineering and analytical value of the proposed architecture rather than claim clinical effectiveness.

KEYWORDS: *Predictive analytics, healthcare management, electronic health records, data integration, FHIR, OMOP, machine learning, clinical decision support, interoperability, data quality*

INTRODUCTION

Healthcare management is increasingly characterized by decisions that must be made across multiple time horizons. At the operational level, managers allocate beds, staff, operating-room capacity, diagnostic resources, and discharge services. At the tactical level, they monitor readmissions, patient flow, utilization, quality indicators, and service bottlenecks. At the strategic level, they evaluate population risk, technology investment, workforce requirements, and the financial consequences of changing patterns of demand. These decisions are data-intensive, but the underlying data are rarely organized as a single coherent information resource. EHR systems capture clinical encounters, laboratory values, diagnoses, medications, and observations; claims systems emphasize reimbursement and utilization; pharmacy systems capture dispensing behavior; appointment systems describe access and missed visits; and connected devices may generate high-frequency physiological measurements. Predictive analytics can estimate future events from these data, but prediction quality is constrained by the quality and completeness of the information supplied to the model [1], [2].

The technical problem is therefore broader than selecting a sophisticated machine-learning algorithm. A hospital can deploy a high-performing classifier and still obtain weak managerial value if relevant information arrives late, patient identities are not reconciled, coding systems are inconsistent, or missingness is interpreted incorrectly. Reviews of EHR prediction studies have documented limited use of longitudinal information, incomplete treatment of bias and missingness, and relatively weak external validation [2]. More recent implementation evidence similarly shows that information-technology infrastructure, access to knowledge, organizational culture, and perceived complexity influence whether machine-learning systems become useful in practice [3]. Interoperability standards provide an important foundation. HL7 FHIR is designed for structured exchange of healthcare information, while the OMOP Common Data Model supports transformation of heterogeneous observational databases into a standardized structure and vocabulary suitable for analytics [4], [5]. These standards do not automatically solve data-quality problems, but they create explicit interfaces for provenance, terminology mapping, and reusable analytical workflows.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

This study addresses the problem by treating data integration and predictive analytics as a coupled engineering system. The central research question is: how much can a standards-oriented, longitudinal integration layer improve predictive performance and decision responsiveness compared with an EHR-only analytical workflow? The work focuses on healthcare-management use cases rather than autonomous diagnosis. The intended outputs are risk stratification, workload prioritization, capacity alerts, and resource-allocation signals that remain subject to human review. To make the contribution reproducible without exposing patient data, a synthetic simulation was constructed with realistic data-generation assumptions, explicit missingness, and measurable model and latency outcomes. The paper therefore contributes an analytical framework, a simulation protocol, and a set of design principles for implementing predictive management systems responsibly.

LITERATURE REVIEW

Predictive analytics in healthcare has progressed from manually engineered statistical scores to models capable of learning from high-dimensional longitudinal EHR data. Rajkomar et al. demonstrated that deep-learning approaches can learn from large-scale raw EHR representations and predict multiple clinical outcomes across centers [6]. This work is significant because it illustrates the value of preserving temporal information rather than reducing a patient record to a small static feature vector. However, the ability to learn from large EHR streams does not remove the need for governance, interoperability, and validation.

Hospital readmission prediction is a representative management application. A scoping review of machine-learning studies reported frequent use of tree-based methods, neural networks, regularized logistic regression, and support-vector machines, while also noting substantial variability in predictive performance [7]. A separate systematic review found that logistic regression remained dominant and that external validation across sites was uncommon [8]. These findings support a pragmatic position: model complexity can improve discrimination, but transparent models remain attractive when decision-makers need understandable risk factors and stable operational thresholds.

Longitudinal modeling creates additional methodological issues. A systematic methodological review of AI prediction models using longitudinal EHR data found substantial heterogeneity in follow-up periods, feature construction, validation strategies, and reporting quality; approximately one-third of the reviewed studies were judged to have high risk of bias [9]. The implication for management systems is important: predictive accuracy should be accompanied by transparent definitions of the prediction window, index time, available information, missingness treatment, and validation population. TRIPOD+AI now provides updated reporting guidance for prediction models using regression or machine-learning methods [10].

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Data integration is equally consequential. The OMOP CDM explicitly separates source data, standardized data, derived analytical data, and metadata, enabling disparate databases to be transformed into a common representation [5]. FHIR, meanwhile, provides a standardized mechanism for exchanging healthcare information and is designed to make digital records structured and machine-processable [4]. In practical architectures, FHIR can support near-real-time exchange while an OMOP-like analytical layer can support longitudinal modeling, cohort construction, and reproducible research. The two approaches are therefore complementary rather than mutually exclusive.

Data quality is a persistent constraint. Recent reviews emphasize completeness, conformance, plausibility, correctness, recency, fairness, and stability as relevant dimensions of EHR data quality [11]. Missingness is particularly difficult because absence of a value may indicate workflow behavior rather than absence of the clinical condition itself. Contemporary methodological work consequently recommends evaluating missingness patterns as part of model development and deployment rather than treating imputation as a purely technical preprocessing step [12]. For healthcare management, this distinction matters because a missing laboratory test can reflect either low clinical risk or limited access to testing, and these situations imply different managerial actions.

RESEARCH GAP

The literature establishes three capabilities separately: predictive models can estimate clinical or operational risk; interoperability standards can move and standardize information; and data-quality frameworks can assess EHR reliability. The research gap is the limited integration of these capabilities into a single management-oriented architecture with measurable links between data harmonization, predictive performance, and decision latency. Many published models are evaluated as isolated algorithms rather than as components of an operational information pipeline [2], [9]. Furthermore, model performance is often reported without quantifying how quickly the required data can be assembled, whether information from different sources is temporally aligned, or how data-quality monitoring affects managerial confidence. The present study addresses this gap by evaluating an end-to-end conceptual pipeline and explicitly measuring both prediction metrics and simulated data-assembly latency.

OBJECTIVES

- To design a standards-oriented architecture for integrating EHR, claims, pharmacy, laboratory, scheduling, and utilization data into a longitudinal management dataset.
- To compare predictive performance between an EHR-only baseline and an integrated feature set using regularized logistic regression and gradient boosting.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- To quantify the operational effect of standardized and cached data services on simulated decision-data assembly latency.
- To identify governance mechanisms for data quality, model monitoring, interpretability, privacy, and human oversight.
- To derive practical implementation principles for hospital managers seeking predictive decision support without replacing professional judgment.

METHODOLOGY

Study Design and Simulation Environment

A simulation-based analytical study was selected because real hospital data were not supplied and because the objective was to isolate the contribution of data integration. A synthetic cohort of 60,000 encounters was generated using a fixed random seed (42). The outcome represented a binary management event defined as high near-term utilization risk, operationalized as an event probability generated from age, comorbidity burden, prior admissions, emergency visits, length of stay, laboratory instability, medication count, pharmacy activity, claims activity, missed appointments, and a continuous care-gap indicator. The simulated prevalence was approximately 15%, representing a sufficiently imbalanced classification problem for evaluating precision-recall behavior.

Two data configurations were created. The baseline configuration contained seven EHR-derived variables: age, comorbidity count, prior admissions, emergency visits, length of stay, laboratory instability, and medication count. The integrated configuration added pharmacy utilization, claims utilization, appointment history, and a care-gap feature, yielding eleven predictors. Twelve percent of integrated feature cells were randomly masked to represent incomplete multi-source capture. Median imputation with missingness indicators was used for integrated logistic regression, while histogram-based gradient boosting was allowed to process missing values directly. The train-test split was stratified at 75:25.

Integration and Standardization Workflow

The proposed integration layer follows a source-to-feature sequence: ingestion, patient/entity matching, terminology normalization, temporal alignment, quality validation, feature construction, and model serving. FHIR-compatible resources are treated as the exchange interface for event-level information, while an OMOP-compatible analytical representation is used conceptually for standardized longitudinal features [4], [5]. Each transformed record should retain source provenance, timestamp, transformation status, and a data-quality flag. Duplicate encounters are removed through deterministic and probabilistic matching rules, while terminology mapping translates source-specific codes into standardized concepts.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Temporal alignment is performed using an index time so that the model receives only information available before the prediction point.

Predictive Models and Evaluation

Three predictive configurations were evaluated. Model A was regularized logistic regression using the EHR-only baseline. Model B was regularized logistic regression using the integrated feature set. Model C was histogram-based gradient boosting using the same integrated data. Logistic regression was selected as an interpretable baseline, while gradient boosting represented a nonlinear model capable of capturing interactions without requiring a deep-learning infrastructure. Discrimination was measured using AUROC and average precision. Calibration-oriented performance was assessed using the Brier score. Because the simulated outcome was imbalanced, operational classification was evaluated at a threshold of 0.20, chosen to represent a management workflow in which recall is prioritized over a default 0.50 threshold. Precision, recall, and F1 score were then calculated. This reporting approach is consistent with the broader emphasis on transparent prediction-model reporting advocated by TRIPOD+AI [10].

Operational Latency Simulation

A separate 1,000-run latency simulation compared two data-assembly pathways. The native pathway represented sequential retrieval from disconnected systems and was modeled with a log-normal distribution centered on approximately 180 seconds. The standardized pathway represented a harmonized integration service plus a feature cache, with component distributions centered on approximately 95 seconds and 35 seconds, respectively. Median end-to-end latency was compared, and percentage reduction was calculated as (baseline median − integrated median) / baseline median × 100. This experiment is illustrative: the distributions are engineering assumptions rather than measurements from a production hospital.

Figure 1 Proposed standards-oriented architecture linking multisource healthcare data, intelligent integration, predictive models, and managerial decision support.

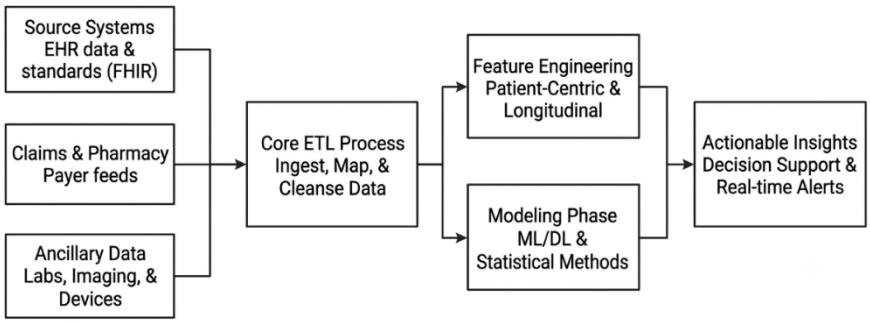


Figure 1. Intelligent data integration and predictive analytics architecture

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

RESULTS AND FINDINGS

The simulation produced a 14.98% positive-event prevalence. The EHR-only logistic regression achieved an AUROC of 0.659 and average precision of 0.260. Adding pharmacy, claims, appointment, and care-gap information increased AUROC to 0.686 and average precision to 0.292. The integrated gradient-boosting model achieved an AUROC of 0.677 and average precision of 0.278. Thus, the largest improvement in this scenario came from expanding and harmonizing the information base rather than simply increasing algorithmic complexity.

At the operational threshold of 0.20, integrated logistic regression achieved precision 0.287 and recall 0.436, compared with 0.260 and 0.382 for the EHR-only model. The resulting F1 score increased from 0.310 to 0.346. Gradient boosting achieved precision 0.277, recall 0.409, and F1 0.330. These results indicate that the integrated logistic model provided the best balance in the simulated decision setting even though the gradient-boosting model was more flexible. The result reinforces the importance of selecting a model according to the decision objective rather than assuming that a more complex algorithm will always be superior.

Table 1: Predictive performance in the synthetic simulation

Model	Feature set	AUROC	Average Precision	Brier Score	F1 @ 0.20
EHR-only LR	7 EHR variables	0.659	0.260	0.122	0.310
Integrated LR	11 integrated variables	0.686	0.292	0.119	0.346
Integrated HGB	11 integrated variables	0.677	0.278	0.120	0.330

Figure 2 shows the relative separation in AUROC and F1. The AUROC improvement of the integrated logistic model over the baseline was 0.027, corresponding to a relative increase of approximately 4.1%. Although modest, such an improvement can be operationally meaningful when a model is used to rank patients for limited interventions such as care-management calls, discharge planning, or utilization review. The simulation should not be interpreted as demonstrating a clinical benefit; it demonstrates how additional longitudinal information can change ranking performance under controlled assumptions.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

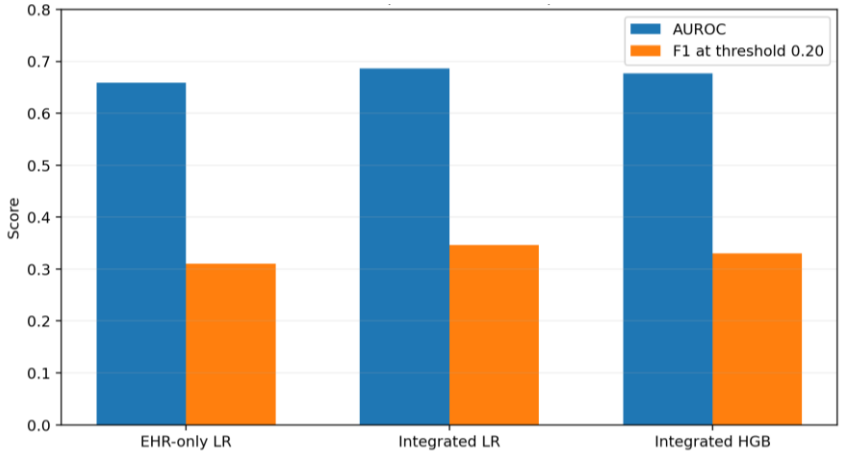


Figure 2. Simulated comparison of predictive performance across analytical configurations

The latency simulation produced a median of 183.0 seconds for the disconnected/native pathway and 132.9 seconds for the standardized-plus-cache pathway, representing a 27.4% reduction in decision-data assembly time. This result is technically important because managerial usefulness is partly determined by whether a prediction is available while a decision can still be changed. A risk score delivered after discharge planning has finished is less useful than an equally accurate score delivered during the planning window.

Table 2: Summary of simulated analytical and operational changes

Measure	EHR-only / Native	Integrated / Standardized	Change
Median decision-data assembly	183.0 s	132.9 s	−27.4%
AUROC (logistic regression)	0.659	0.686	+0.027
Average precision	0.260	0.292	+0.032
F1 at threshold 0.20	0.310	0.346	+0.036

DISCUSSION

The principal finding is that predictive analytics and intelligent data integration should be designed as one decision system. The integrated logistic model outperformed the EHR-only baseline despite using the same model family. This suggests that the information boundary around the model may be as influential as the optimization algorithm itself. Pharmacy activity, claims utilization, appointment behavior, and care-gap signals add dimensions that are not always visible in a single clinical encounter. This observation is consistent with evidence that EHR prediction studies have historically underused

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

longitudinal information [2], and with research showing that machine-learning models can benefit from representations of the full patient record [6].

The result also illustrates why management systems should distinguish predictive discrimination from decision utility. An AUROC increase of 0.027 does not automatically mean that clinical outcomes will improve. The actual value depends on the intervention attached to the score, the capacity of the organization to act on alerts, the threshold selected, and the cost of false positives and false negatives. For example, a high-recall threshold may be reasonable for a low-cost telephone follow-up program but inappropriate for an intervention that consumes scarce specialist capacity. Thresholds should therefore be selected using explicit utility functions, workload constraints, and stakeholder review rather than model statistics alone.

The modest advantage of integrated logistic regression over integrated gradient boosting is also instructive. Healthcare management often requires explanations that can be translated into operational actions. A linear model can show how recent utilization, comorbidity, or care gaps contribute to risk in a relatively direct manner. Gradient boosting can represent nonlinear interactions but requires additional explanation mechanisms and monitoring. The choice should therefore reflect the governance burden, not only predictive accuracy. Transparent reporting guidelines such as TRIPOD+AI reinforce this principle by requiring clearer descriptions of data sources, modeling methods, validation, and reporting practices [10].

Interoperability standards can reduce the technical cost of integration, but they do not guarantee semantic correctness. FHIR establishes a common exchange framework [4], while OMOP establishes a standardized analytical structure and vocabulary [5]. A practical hospital architecture can use FHIR at the operational exchange boundary and an OMOP-compatible warehouse or lakehouse for analytical harmonization. This division of responsibilities can preserve near-real-time access while supporting reproducible longitudinal analysis. The architecture should also preserve source values and transformation lineage so that managers and analysts can trace an alert back to the underlying information.

Missingness requires particular attention. In the simulation, 12% of integrated feature cells were masked. The integrated logistic model used missingness indicators in addition to median imputation, while gradient boosting handled missing values natively. This design reflects the reality that missingness can carry workflow information. A test not ordered, a medication not dispensed, or an appointment not scheduled may itself be informative, but the meaning can differ by department, patient population, and access pathway. Recent methodological literature emphasizes that missing-data treatment should be evaluated as part of prediction development and implementation rather than treated

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

as an invisible preprocessing step [12].

From a governance perspective, the proposed system should be managed through four controls: data quality, model quality, workflow quality, and accountability. Data-quality controls monitor completeness, conformance, plausibility, timeliness, and provenance [11]. Model-quality controls monitor discrimination, calibration, drift, subgroup performance, and threshold behavior. Workflow controls monitor alert volume, response time, override frequency, and whether predictions actually trigger the intended intervention. Accountability controls define who can approve model updates, who investigates harmful alerts, and how users can challenge or override recommendations. Implementation evidence indicates that infrastructure and organizational factors are major determinants of whether ML becomes embedded in healthcare practice [3].

Figure 3 presents the resulting closed-loop management model. Unlike a one-directional analytics pipeline, the closed-loop design treats outcomes and user responses as feedback. If a prediction repeatedly fails for a specific department or population, the system should trigger investigation of data quality, feature definitions, threshold policy, or model drift. This makes monitoring a continuous management function rather than an afterthought.

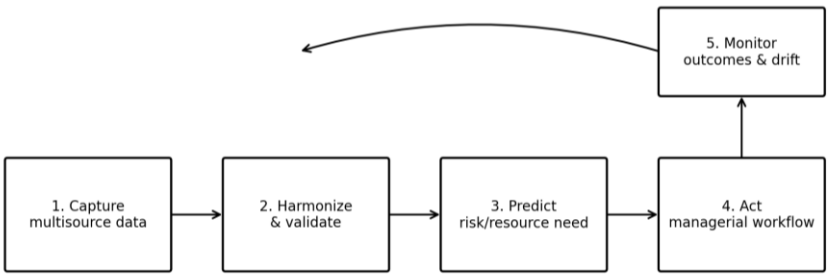


Figure 3. Closed-loop decision management with monitoring and feedback

Table 3: Governance-oriented architecture for predictive healthcare management

Layer	Primary function	Key controls	Managerial output
Data layer	Capture and reconcile sources	Identity, terminology, provenance, completeness	Trusted longitudinal record
Analytics layer	Construct features and predictions	Validation, calibration, drift, fairness	Risk and demand scores

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Workflow layer	Translate scores into action	Thresholds, escalation, workload limits	Prioritized interventions
Governance layer	Monitor and authorize use	Audit, privacy, review, accountability	Safe and sustainable operation

PRACTICAL APPLICATIONS

The framework is applicable to several management scenarios. In bed and capacity management, integrated utilization histories can support forecasts of admission pressure and discharge readiness. In readmission management, pharmacy and appointment information can complement clinical variables when identifying patients who may require follow-up. In workforce planning, predicted service demand can be linked to staffing rosters and historical throughput. In population health management, longitudinal claims and encounter histories can help prioritize outreach. In procurement, aggregated forecasts can support planning for high-use medicines and consumables. In each case, the prediction should be positioned as a prioritization signal rather than an automatic decision.

CONCLUSION

Predictive analytics can improve healthcare-management decision support, but its effectiveness depends on the information infrastructure surrounding the model. This study proposed and simulated a standards-oriented architecture that combines FHIR-compatible exchange, OMOP-compatible analytical standardization, longitudinal feature construction, predictive modeling, and closed-loop governance. In a 60,000-encounter synthetic experiment, adding integrated utilization and care-coordination features improved logistic-regression AUROC from 0.659 to 0.686 and F1 from 0.310 to 0.346 at an operational threshold of 0.20. A separate latency simulation showed a 27.4% reduction in median data-assembly time. These findings support the proposition that intelligent integration can contribute to both analytical quality and managerial responsiveness. The most important engineering lesson is that data pipelines, models, workflows, and governance should be designed together. The most important scientific lesson is that simulation results must not be confused with evidence of clinical effectiveness. Real-world deployment requires external validation, prospective evaluation, monitoring, and stakeholder oversight before predictions are used to influence patient-facing or resource-allocation decisions.

LIMITATIONS

- The study used synthetic data rather than a real hospital dataset; therefore, the numerical results demonstrate methodological behavior under controlled assumptions and cannot be generalized as clinical performance.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- The outcome-generation mechanism was specified by the researchers, which may not reproduce the complex causal and documentation processes present in real EHRs.
- The simulation used a single train-test split and did not perform temporal or external validation across hospitals.
- Only three model configurations were compared. Deep learning, survival analysis, temporal transformers, and causal decision models were outside the scope.
- The latency distributions were engineering assumptions and were not benchmarked against a production integration platform.
- Fairness, privacy risk, cybersecurity, and subgroup calibration were discussed as governance requirements but were not experimentally quantified.

FUTURE SCOPE

Future work should evaluate the architecture using multi-institutional longitudinal datasets and temporal external validation. Research should compare FHIR event streams with batch OMOP pipelines, quantify terminology-mapping error, and investigate how integration latency changes as data volume and source count increase. Prospective studies should evaluate whether alerts actually change managerial behavior and whether those changes improve outcomes, cost, equity, or staff workload. Model-development studies should incorporate calibration, decision-curve analysis, subgroup fairness, and drift detection. Further work can also investigate federated learning and privacy-preserving analytics so that institutions can collaborate without centralizing identifiable records. Finally, digital-twin and simulation environments could be used to test staffing and capacity policies before deployment, enabling healthcare managers to compare alternative interventions under uncertainty.

REFERENCES

1. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1:18.
2. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2017;24(1):198-208.
3. Implementation of Machine Learning Applications in Health Care Organizations: Systematic Review of Empirical Studies. JMIR Medical Informatics. 2024.
4. HL7 International. FHIR Release 5.0.0: Fast Healthcare Interoperability Resources. 2023.
5. Observational Health Data Sciences and Informatics (OHDSI). OMOP Common Data Model. 2026.
6. Huang Y, Talwar A, Chatterjee S, Aparasu RR. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC Medical Research Methodology.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved

2021;21:96.

7. Teo K, et al. Current Trends in Readmission Prediction: An Overview of Approaches. Arabian Journal for Science and Engineering. 2021.
8. Predictive models for hospital readmission risk: a systematic review of methods. BMC Medical Research Methodology. 2018.
9. Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review. Journal of the American Medical Informatics Association. 2023.
10. Cohen JF, Bossuyt PMM, Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
11. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. Journal of Medical Internet Research. 2024.
12. Lin S, Groenwold RHH, Mehta HB, et al. Addressing Missingness in Predictive Models That Use Electronic Health Record Data. Annals of Internal Medicine. 2025;178(10):1451-1463.

Page 95-107 © MANTECH PUBLICATIONS 2026. All Rights Reserved