Contents lists available at IJMAHS
International Journal of Medical and Applied Health Science
journal homepage: IJMAHS
Volume 2, No. 1, 2026

Integrating Artificial Intelligence and Industrial Engineering for Smart Healthcare Management and Decision Support

Abbas Omidi [a]

[a] *Department of Engineering, Kharazmi university, Tehran, Iran, Head of IE & Strategic Planning, SPCO, Ezam Group, Tehran, Iran.*

ARTICLE INFO

Received: 2026/09/15
Revised: 2026/10/27
Accepted: 2026/11/12

Keywords:
Artificial intelligence; Industrial engineering; Smart healthcare; Healthcare operations management; Clinical decision support.

ABSTRACT

Healthcare organizations face challenges related to demand, limited bed capacity, workforce constraints, uncertain service times, and clinical priorities. Artificial intelligence (AI) can predict patient risk, arrivals, no-shows, length of stay, and resource requirements. At the same time, industrial engineering (IE) provides methods for scheduling, simulation, queueing, and optimization to translate predictions into decisions. This paper proposes an integrated AI–IE framework for healthcare management and decision support. The framework combines data integration, predictive analytics, mixed-integer optimization, simulation, and human oversight. The formulation coordinates beds, staffing, diagnostic capacity, and appointments while considering waiting, overtime, operating costs, capacity limits, clinical priority, skill compatibility, fairness, and uncertainty buffers. A 90-day scenario compares conventional planning, AI prediction alone, and integrated AI–IE decision support. Results show reductions in waiting time, overtime, cost per patient, and critical-priority delays, alongside lower bed utilization to preserve capacity slack. Because the results are synthetic, they do not establish clinical effectiveness. The framework offers a basis for prospective validation, careful governance, and operational improvement.

1. Introduction

Healthcare delivery has become a high-complexity service system. Hospitals coordinate emergency admissions, elective procedures, outpatient appointments, diagnostic services,

[a] Corresponding author email address: Abb.omidi@gmail.com (Abbas Omidi).
DOI: https://doi.org/10.22034/ijmahs.v2i1.242
Available online 2026/11/13
Licensee System Analytics. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).
3115-7378/BGSA Ltd.

inpatient beds, clinical staff, and expensive equipment. These activities are interdependent: a delayed test can postpone treatment; a delayed discharge can block an emergency admission; and a mismatch between staff skills and patient acuity can create inefficiency and safety risks. The resulting management problem is not simply a question of increasing capacity. It is a question of coordinating demand, capacity, time, uncertainty, and clinical priorities. Industrial engineering is particularly relevant because it studies the design, operation, and improvement of complex systems under constraints.

AI has expanded the analytical capabilities available to healthcare organizations. Machine-learning models can identify patterns in electronic health records, estimate no-show probabilities, predict appointment duration, classify risk, forecast demand, and estimate LOS. Natural-language processing can extract information from clinical notes, while computer vision can support selected imaging and monitoring tasks. MIMIC-IV illustrates the research potential of de-identified hospital data containing measurements, diagnoses, procedures, treatments, and clinical notes [1]. Nevertheless, a prediction is not an operational decision. A model estimating that a patient may be discharged tomorrow does not decide which bed to allocate today, how many nurses to roster, or how much reserve capacity to protect.

IE supplies the decision layer that complements AI. Healthcare operations research has examined appointment scheduling, operating-room planning, workforce rostering, patient flow, capacity allocation, and healthcare network design. Reviews of appointment scheduling classify problems by patient flow, preferences, random arrivals, service-time uncertainty, and solution approaches, including simulation, mathematical optimization, Markov models, and AI [2]. Research on healthcare planning and scheduling also emphasizes interdependence among hospital units, outpatient clinics, surgery, workforce, and wider networks [3]. These streams suggest that AI and IE should not be treated as alternatives. AI improves the information available to decision-makers; i.e., it converts that information into coordinated, feasible, and auditable plans.

Recent studies demonstrate the value of linking prediction and scheduling. Babayoff et al. [4] developed machine-learning models for outpatient no-shows and appointment length and connected their predictions to appointment-system decisions. Salah and Srinivas [5] proposed a predict-then-schedule approach for sequential clinical scheduling under no-show and service-time uncertainty. These studies move beyond standalone prediction. However, their operational scope

is narrower than a hospital-wide management system spanning beds, staff, appointments, diagnostics, and competing patient priorities.

LOS prediction is another important bridge between AI and operations. Forecasting inpatient LOS can support admissions planning and resource allocation, provided that estimates are adapted to the hospital context and do not replace clinical discharge authority [6]. Public EHR datasets such as MIMIC-IV enable reproducible methodological development, but external validity, institutional differences, missingness, and governance remain critical [1]. The practical value of a model therefore depends not only on discrimination or prediction error but also on the decisions made from its outputs.

Digital twins add a system-level perspective by connecting data streams, process models, simulation, and decision support. A healthcare digital twin may represent a patient pathway, hospital unit, or entire facility. A 2025 meta-review identifies fragmented applications, ethical concerns, and implementation barriers [7]. A 2026 review of hospital operations reports that many studies remain conceptual, simulation-based, or pilot-stage, with relatively few demonstrating operational implementation [8]. These findings reinforce the need for frameworks that connect prediction, optimization, workflow integration, and governance (See Figure A).

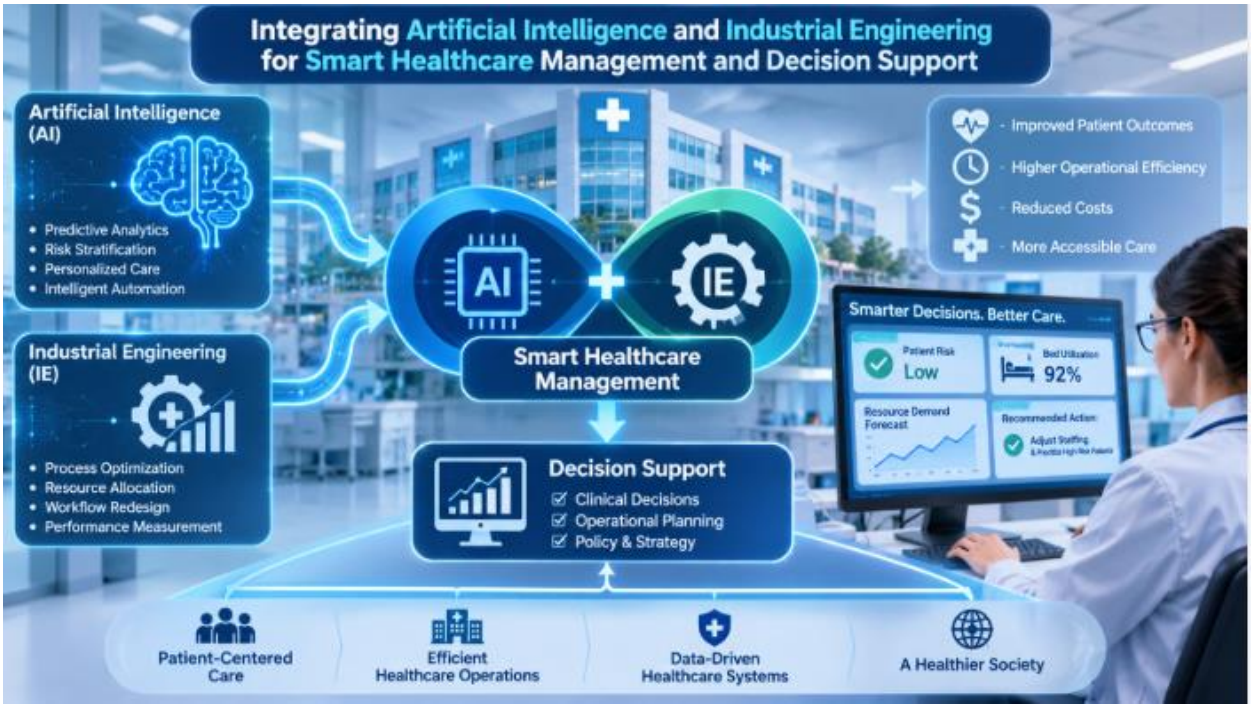


Figure A. Illustrative operational KPI comparison; each series retains its native unit.

The central research problem is therefore the gap between predictive intelligence and operational intelligence. Predictive intelligence estimates what may happen; operational intelligence selects what should be done, subject to clinical and organizational constraints. Accurate forecasts may have little value if they are not translated into coordinated schedules. Conversely, an optimizer can produce mathematically efficient but unsafe recommendations if inputs are biased, stale, poorly calibrated, or disconnected from clinical judgment. Integration requires a closed-loop design in which data quality, uncertainty, constraints, human review, and outcome monitoring are treated as parts of one system.

This paper has five objectives: (1) synthesize AI and IE contributions to smart healthcare management; (2) propose an integrated prediction-to-optimization architecture; (3) formulate a constrained multi-objective resource-allocation model; (4) demonstrate the evaluation logic through a reproducible synthetic experiment; and (5) identify governance and validation requirements. The contribution is a conceptual-methodological framework, a transparent mathematical formulation, and a scenario analysis. The numerical values are purely synthetic because no hospital dataset was provided; they should not be interpreted as empirical or clinical evidence.

The paper is organized as follows. Section 2 reviews relevant literature and identifies research gaps from 2021 to 2026. Section 3 presents the architecture and mathematical methodology. Section 4 reports the synthetic experiment. Section 5 discusses managerial and governance implications. Section 6 concludes and outlines future research.

2. Literature Review

2.1. AI as a predictive layer

AI in healthcare includes supervised and unsupervised learning, deep learning, reinforcement learning, natural-language processing, and hybrid methods. In management applications, supervised learning estimates outcomes such as appointment attendance, service duration, readmission risk, and LOS. Outputs can be translated into operational parameters: predicted appointment duration becomes a scheduling input; no-show probability informs controlled overbooking; LOS distributions inform bed occupancy forecasts; and acuity-related workload informs staffing. The essential design distinction is between a predicted probability and a decision. Reviews of AI implementation identify opportunities and challenges related to data quality, interoperability, explainability, workflow fit, privacy, bias, and organizational readiness [9]. These

issues are consequential because prediction errors can affect access, workload, and safety. Model evaluation should therefore include calibration, subgroup performance, temporal stability, sensitivity to missing data, and decision impact—not accuracy alone. A model may achieve high aggregate accuracy while systematically underestimating risk for a smaller patient group.

AI decision support also requires a defined human role. A prediction may flag a case for review, but decisions such as diagnosis, treatment, escalation, and discharge remain with qualified professionals. Recommendations should expose uncertainty, rationale, and an override mechanism. WHO guidance places ethics and human rights at the center of AI design, deployment, and use [10]. Responsible AI is therefore a design requirement, not a post-deployment checklist.

2.2. IE and healthcare operations

IE methods address the design and control of processes in which demand, capacity, and time are constrained. In healthcare, queueing theory explains congestion; discrete-event simulation represents patient pathways; mathematical programming allocates resources; forecasting supports capacity planning; lean methods identify waste; statistical process control monitors quality; and reliability engineering supports equipment and service continuity. These methods are complementary. Queueing models explain waiting mechanisms, simulation captures complex pathways, optimization selects allocations, and quality engineering monitors stability.

Appointment scheduling must balance patient waiting time, provider idle time, overtime, preferences, no-shows, and stochastic service durations. Ala and Chen [2] review more than 150 papers and classify appointment systems and solution methods. Youn et al. [3] place scheduling within broader healthcare planning, including hospitals, outpatient clinics, operating rooms, and networks. Their synthesis supports systems view: local optimization can merely transfer a queue to another department unless downstream capacity is taken into account.

Prediction-informed scheduling is an important development. Babayoff et al. [4] connected models of no-shows and appointment duration to an outpatient scheduling interface. Salah and Srinivas [5] integrated patient-specific uncertainty estimates with sequential scheduling policies. These studies imply that smart hospitals should use AI predictions as parameters in optimization models rather than as isolated dashboard indicators.

2.3. Bed management, LOS, and digital twins

Bed management is a multi-stage flow problem. Admissions depend on demand, bed type, infection-control requirements, staffing, expected discharge, and downstream services. LOS

prediction supports occupancy forecasting, but errors can create overbooking or unnecessary reserve capacity. Research on LOS prediction for hospital management emphasizes domain adaptation and operational integration [6]. A decision system should use distributions or prediction intervals rather than point estimates alone, and update forecasts as patient conditions and pathways evolve.

MIMIC-IV is an important research resource containing de-identified data from a tertiary academic medical center, including measurements, diagnoses, procedures, treatments, and clinical notes [1]. It supports retrospective analytics, subject to data-use requirements. Models trained at one institution may not transfer to another population or workflow without recalibration and external validation.

Digital twins extend the scope from individual predictions to dynamic process representations. They connect real-time data, process models, simulation, and scenario testing. A 2025 meta-review describes healthcare applications and implementation challenges [7]. A 2026 review focused on hospital operations reports varying levels of decision-support maturity and notes that operational deployment remains uncommon compared with conceptual and proof-of-concept studies [8]. A digital twin should therefore be more than a dashboard or a static simulation: it requires a maintained connection among the physical system, its digital representation, and decision-making.

2.4. Research synthesis and gap, 2021–2026

Table 1 provides a focused narrative synthesis of representative studies and guidance. It is not claimed to be a PRISMA systematic review. The 2021–2026 period captures the development of AI governance, prediction-informed scheduling, open EHR infrastructure, and healthcare digital twins.

Table 1: Research synthesis and gap, 2021–2026

Study	Focus	Contribution	Gap relevant here
WHO (2021) [10]	AI governance	Human rights, accountability, transparency, inclusiveness	Governance principles, not an operational allocation model
Ala & Chen (2022) [2]	Appointment scheduling	Broad review of settings and methods	No unified hospital-wide AI–IE architecture

Study	Focus	Contribution	Gap relevant here
Babayoff et al. (2023) [4]	Outpatient scheduling	Predict no-shows and duration; connect to scheduling	Focused outpatient scope
Salah & Srinivas (2022) [5]	Sequential scheduling	ML estimates inform scheduling under uncertainty	Not a multi-resource hospital model
Youn et al. (2022) [3]	Planning and scheduling	System-level research agenda	No single integrated operational model
Johnson et al. (2023) [1]	EHR data infrastructure	Reusable MIMIC-IV dataset	Dataset alone does not solve workflow integration
Wubineh et al. (2024) [9]	AI implementation	Reviews opportunities and challenges	Operational optimization remains separate
LOS prediction (2024) [6]	Bed management	ML-based LOS prediction for planning	Prediction-to-action remains challenging
Liu et al. (2025) [7]	Healthcare digital twins	Meta-review of applications and barriers	Fragmented applications and limited validation
Abd El-Mawla et al. (2026) [8]	Hospital operations twins	Reviews decision-support maturity	Few operational implementations

Five gaps emerge. First, prediction and prescription are frequently evaluated separately. A model may report AUC or MAE without showing whether its output changes resource allocation. Second, operational studies often focus on one service line—appointments, beds, or operating rooms—even though hospital performance depends on linked capacity pools. Third, uncertainty is not always propagated from prediction into optimization; point forecasts can create brittle schedules. Fourth, clinical priority and equity are often contextual considerations rather than explicit constraints. Fifth, implementation evidence remains limited: simulation and proof-of-concept studies are valuable but do not establish effectiveness in real-world clinical settings.

This study addresses these gaps by linking a predictive layer to multi-resource optimization and simulation-based monitoring. It represents clinical priority, staff skill, capacity, uncertainty buffers, and fairness. It distinguishes technical performance from operational outcomes and

clinical outcomes. The synthetic experiment demonstrates the analysis structure without claiming real-world effectiveness. Empirical research should validate the architecture prospectively using local hospital data and a pre-specified evaluation protocol.

3. Methodology

3.1. Research design and architecture

The study adopts a design-science and operations-research methodology. The unit of analysis is a hospital service system over a rolling planning horizon. The framework supports tactical capacity planning, daily planning, and near-real-time control. The system boundary includes arrivals, triage and acuity classification, appointment and admission queues, inpatient beds, staff teams, diagnostic capacity, and discharge forecasts. Diagnosis and treatment selection are outside the optimization model: the system supports operational coordination and does not replace professional judgment.

The architecture has five layers. The data layer integrates EHR events, appointments, rosters, bed status, diagnostic queues, equipment availability, and timestamps. The predictive layer estimates arrivals, no-shows, service duration, LOS, acuity-related workload, and discharge probabilities. The prescriptive layer converts estimates into allocation and scheduling decisions. The simulation and monitoring layer tests proposed plans under stochastic scenarios and tracks KPIs. The governance layer defines access control, auditability, model monitoring, human approval, escalation, and override procedures. The closed loop is data → prediction → optimization → simulation and approval → execution → monitoring → refreshed data.

Data preparation should include de-identification, duplicate resolution, timestamp harmonization, missingness profiling, unit standardization, and temporal validation. Leakage must be prevented by using only features available at the decision time. For example, a discharge summary written after discharge cannot be used to predict an earlier discharge. Training, validation, and test sets should be temporally separated; external validation should use a different site where possible. Subgroup analysis should examine clinically and legally appropriate groups and avoid unjustified use of sensitive attributes.

3.2. Predictive analytics

Let i index patients or patient groups, k index resources, and t index planning periods. The predictive layer estimates arrival probability, no-show probability, service duration, LOS distribution, and resource intensity. A generic supervised model estimates the outcome. y_i from

feature vector x_i as $\hat{y}_i = f\theta(x_i)$. Candidate model families include gradient-boosted trees, random forests, regularized regression, and neural networks. Selection should consider data volume, interpretability, calibration, and operational constraints. Temporal validation and operational utility should be prioritized over aggregate accuracy.

Prediction uncertainty is represented through quantiles or calibrated intervals. Expected lower and upper quantiles may accompany LOS. The optimizer can use a conservative quantile for critical capacity or scenario sets for stochastic planning. No-show estimates may inform controlled overbooking only where safety and access rules permit. Prediction outputs should be versioned and timestamped so that each optimization decision can be reconstructed from the information available at the time.

3.3. Sets, parameters, and decision variables

Let I denote the set of patients or patient groups (index i), K the set of resource types (index k), T the set of planning periods (index t), and S the set of staff skill categories (index s). The subset $I_H \subseteq I$ contains high-priority groups. For fairness assessment, I is partitioned into eligible subgroups $\{I_g\}_{g\in G}$, where G is the set of subgroups.

The parameters are the predicted demand d_{it}, the resource requirement a_{ik}, the clinical priority weight p_i, the waiting penalty w_i, the overtime penalty o_s, the nominal capacity C_{kt}, the staff availability H_{st}, the skill requirement coefficient g_{isk}, the unit cost c_{kt}, and the uncertainty buffer $b_{kt} \in [0,1)$. The target utilization band of the resource k is $[L_k, U_k]$, with $0 \le L_k < U_k \le 1 - b_{kt}$. The objective weights satisfy $\lambda_1, \dots, \lambda_4 \ge 0$ and $\sum_{r=1}^{4} \lambda_r = 1$; $\varepsilon_H \in [0,1]$ and $\delta \ge 0$ are the priority-protection and fairness tolerances.

The decision variables are the workload x_{ikt} assigned to resource k, the deferred workload u_{it}, the assigned staff hours h_{st}, the overtime OT_{st}, and, when indivisibility is required, the binary assignment indicators z_{ikt}. Two auxiliary variables, the resource usage y_{kt} and the band-deviation penalty M_{kt}, are defined in the model below. All symbols are summarized in the notation table below.

Symbol	Definition	Role
d_{it}	Predicted demand (workload) of group i in period t	Parameter

Symbol	Definition	Role
a_{ik}	Units of resource k required per unit of workload of group i	Parameter
p_i	Clinical priority weight of group i	Parameter
w_i	Waiting (deferral) penalty per unit of unmet workload of group i	Parameter
o_s	Overtime penalty per hour for skill category s	Parameter
C_{kt}	Nominal capacity of resource k in period t	Parameter
H_{st}	Regular available hours of skill category s in period t	Parameter
g_{isk}	Hours of skill s required per unit of resource-k usage by group i	Parameter
c_{kt}	Unit operating cost of resource k in period t	Parameter
b_{kt}	Uncertainty buffer (fraction of capacity held in reserve)	Parameter
L_k, U_k	Lower and upper bounds of the target utilization band of resource k	Parameter
λ_r	Objective weight of term r = 1,…,4; normalized so that $\sum_{r=1}^{4} \lambda_r = 1$	Parameter
ε_H	Maximum tolerated deferral fraction for high-priority demand	Parameter
δ	Maximum permitted disparity in normalized service rates	Parameter
x_{ikt}	Workload of group i assigned to resource k in period t	Decision variable
u_{it}	Deferred or unmet workload of group i in period t	Decision variable
h_{st}	Staff hours of skill s assigned in period t	Decision variable
OT_{st}	Overtime hours of skill s in period t	Decision variable
z_{ikt}	Binary assignment indicator (used when indivisibility is required)	Binary decision variable
y_{kt}	Usage of resource k in period t, defined in Eq. (1)	Auxiliary variable

Symbol	Definition	Role
M_{kt}	Deviation of usage of resource k from the target band in period t	Auxiliary variable

3.4. Multi-objective optimization

Resource usage is the workload assigned to resource k weighted by the resource requirement:

$$y_{kt} = \sum_{i \in I} a_{ik}\, x_{ikt} \qquad \forall\, k \in K,\ t \in T \tag{1}$$

The objective combines access (priority-weighted deferral), labor (overtime), operating cost, and capacity mismatch. Weights should be elicited with clinicians and managers, normalized, and tested. A lexicographic alternative can prioritize safety, then access, then cost. The weighted-sum formulation is:

$$\begin{aligned} \min Z = {} & \lambda_1 \sum_{i \in I} \sum_{t \in T} p_i\, w_i\, u_{it} + \lambda_2 \sum_{s \in S} \sum_{t \in T} o_s\, \mathrm{OT}_{st} \\ & + \lambda_3 \sum_{k \in K} \sum_{t \in T} c_{kt}\, y_{kt} + \lambda_4 \sum_{k \in K} \sum_{t \in T} M_{kt} \end{aligned} \tag{2}$$

The term M_{kt} penalizes deviation of resource usage from the target utilization band $[L_k, U_k]$. Penalizing both overload and underutilization avoids treating 100% occupancy as an ideal. Because the objective minimizes M_{kt} and each weight λ_r is nonnegative, the following inequalities make M_{kt} equal to the larger of the two band violations (or zero):

$$M_{kt} \geq y_{kt} - U_k C_{kt}, \qquad M_{kt} \geq L_k C_{kt} - y_{kt} \qquad \forall\, k \in K,\ t \in T \tag{3}$$

Demand balance ensures that, in each period, the workload of every group is either assigned to a resource or explicitly deferred:

$$\sum_{k \in K} x_{ikt} + u_{it} = d_{it} \qquad \forall\, i \in I,\ t \in T \tag{4}$$

Resource capacity, reduced by the uncertainty buffer, is constrained as follows:

$$y_{kt} \leq C_{kt}\,(1 - b_{kt}) \qquad \forall\, k \in K,\ t \in T \tag{5}$$

The buffer b_{kt} protects against forecast error, emergency arrivals, and disruption. Staff capacity and skill requirements are represented by the assigned-hours definition and the availability limit, where overtime covers any excess over regular hours:

$$h_{st} = \sum_{i \in I} \sum_{k \in K} g_{isk}\, a_{ik}\, x_{ikt} \qquad \forall\, s \in S,\ t \in T \tag{6}$$

$$h_{st} \le H_{st} + \mathrm{OT}_{st} \qquad \forall\, s \in S,\ t \in T \tag{7}$$

Nonnegativity of the continuous variables is required:

$$x_{ikt},\ u_{it},\ h_{st},\ \mathrm{OT}_{st},\ y_{kt},\ M_{kt} \ge 0 \tag{8}$$

For indivisible assignments, binary variables link assignment and workload:

$$x_{ikt} \le d_{it}\, z_{ikt}, \qquad z_{ikt} \in \{0,1\} \qquad \forall\, i \in I,\ k \in K,\ t \in T \tag{9}$$

Clinical priority is protected by limiting deferred high-priority demand in every period:

$$\sum_{i \in I_H} u_{it} \le \varepsilon_H \sum_{i \in I_H} d_{it} \qquad \forall\, t \in T \tag{10}$$

A simplified fairness constraint limits differences in normalized service rates R_g across eligible subgroups. Its use requires clinical and legal review. Because the denominators are constants, the constraint is linear in x_{ikt}:

$$R_g = \frac{\sum_{i \in I_g} \sum_{k \in K} \sum_{t \in T} x_{ikt}}{\sum_{i \in I_g} \sum_{t \in T} d_{it}}, \qquad \left|R_g - R_{g'}\right| \le \delta \qquad \forall\, g, g' \in G \tag{11}$$

Constraints (1)– (8), (10), and (11) are linear, so the model is a linear program; imposing (9) turns it into a mixed-integer linear program (MILP).

3.5. Uncertainty and rolling-horizon control

Demand and LOS are uncertain. A scenario-based extension samples ω from calibrated forecast distributions. First-stage decisions reserve capacity and establish staffing; recourse decisions adapt allocations after demand is observed:

$$\min_x \ c^\top x + \sum_{\omega \in \Omega} \pi_\omega\, Q(x, \omega) \tag{12}$$

A risk-averse extension penalizes conditional value at risk (CVaR) of operational loss:

$$CVaR_{\alpha}(L) = min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \sum_{\omega \in \Omega} \pi_{\omega} max(0,\, L_{\omega} - \eta) \right\} \quad (13)$$

The rolling-horizon controller solves at each decision epoch, implements near-term decisions, receives new operational data, updates predictions, and resolves. This is preferable to a static schedule in a dynamic hospital. A human supervisor reviews recommendations before execution, especially for high-acuity cases, unusual circumstances, and resource scarcity.

3.6. Evaluation and implementation

Predictive evaluation should include MAE or RMSE for continuous outcomes, Brier score and calibration for probabilities, discrimination measures where appropriate, and subgroup stability. Operational evaluation should include mean and percentile waiting times, left-without-being-seen rate, bed-blocking, utilization, overtime, cancellations, diagnostic turnaround, cost per completed episode, and critical-priority delay. Clinical safety outcomes require separate definitions and adjudication; operational proxies must not be presented as clinical benefit.

A stepped implementation is recommended. First, map processes and data lineage. Next, develop retrospective models and establish baseline KPIs. Then run the optimizer in shadow mode without changing operations. Conduct a controlled pilot with human approval and safety monitoring. Finally, evaluate prospectively against a pre-specified comparator, with independent review and incident response. Rollback criteria should be defined before deployment. Model drift, coding changes, staffing patterns, and seasonal demand should trigger review or recalibration.

4. Numerical Results

4.1. Experimental design and assumptions

The numerical experiment is a synthetic scenario analysis, not a study of actual patients. A 90-day workload stream is generated for low-, medium-, and high-acuity classes, with illustrative mean daily arrivals of 52, 46, and 22. LOS means are 1.2, 3.0, and 5.5 days, with standard deviations of 0.35, 0.80, and 1.20 days. Staff-hour requirements are 0.55, 1.25, and 2.40 hours per patient. A fixed random seed ensures reproducibility of the workload stream.

Three architectures are compared: conventional planning using historical averages and manual adjustments; AI prediction only, where forecasts are shown to managers but no integrated optimizer coordinates resources; and integrated AI–IE, where forecasts feed constrained allocation and scheduling with capacity buffers, priority protection, and human review. The KPI values are

explicit illustrative scenario assumptions. They are not estimated from a hospital, learned by a model, or generated by solving the full MILP. A real study must estimate outcomes from baseline data and evaluate the intervention prospectively.

4.2. Comparative performance

Table 2 reports the illustrative scenario values. The integrated scenario has the lowest waiting time, overtime, cost per patient, and critical-priority delay. Utilization is also lower. This is intentional: a hospital operating near full occupancy may have little slack for emergency arrivals, LOS variation, or staff shortages. The objective is safe flow and resilience, not maximal occupancy.

Table 2: Comparative performance

KPI	Conventional	AI prediction only	Integrated AI–IE
Mean waiting time (min)	74	62	41
Bed utilization (%)	91	89	84
Staff overtime (%)	18	15	9
Cost per patient (USD)	1,250	1,220	1,160
Critical-priority delay (%)	0.40	0.30	0.20

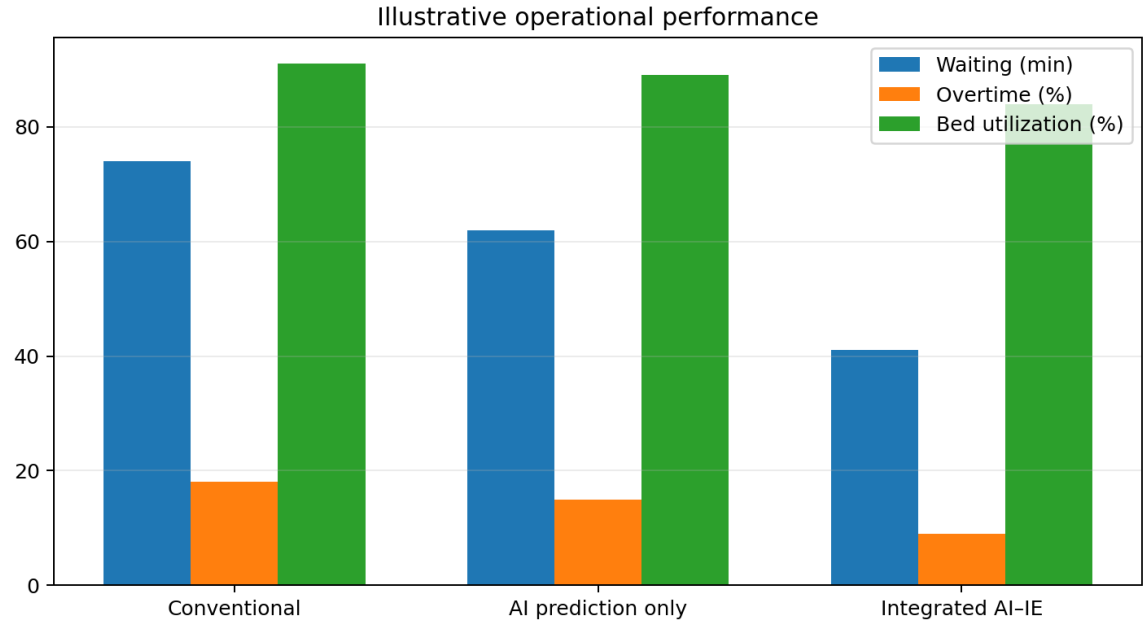


Figure 1. Illustrative operational KPI comparison; each series retains its native unit.

4.3. Relative improvement

Relative to conventional planning, the integrated scenario reduces waiting by (74−41)/74 = 44.6%, overtime by (18−9)/18 = 50.0%, cost per patient by (1,250−1,160)/1,250 = 7.2%, and critical-priority delay by (0.40−0.20)/0.40 = 50.0%. These are arithmetic comparisons between scenario assumptions. They are not statistically significant effects because no empirical sampling distribution or randomized comparison exists.

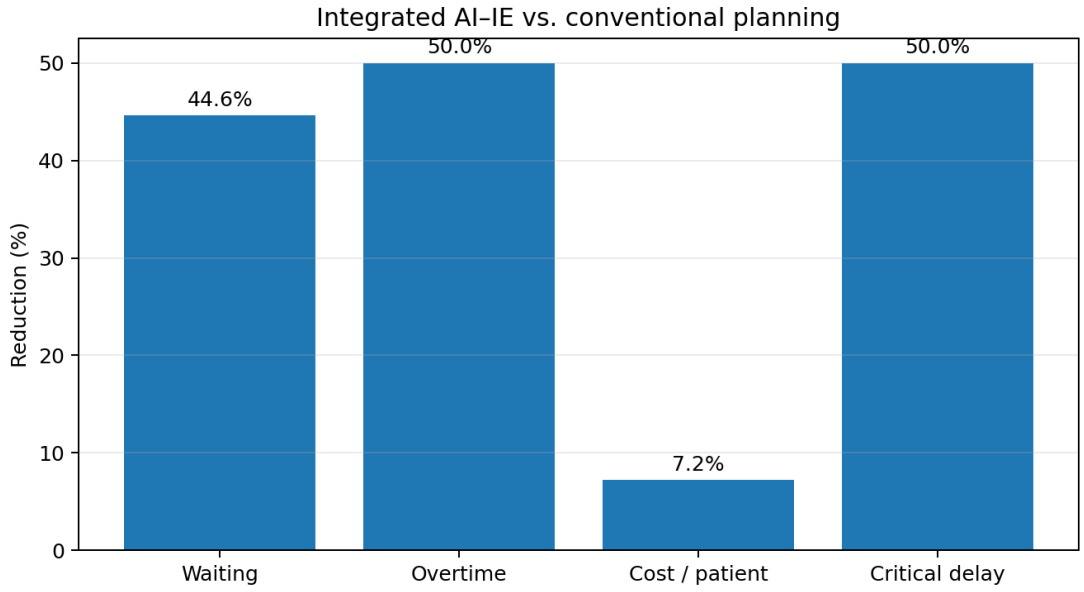


Figure 2. Relative reductions under the integrated AI–IE scenario.

The AI-only scenario reduces waiting time from 74 to 62 minutes, whereas the integrated scenario reduces it to 41 minutes. This illustrates the paper's central proposition: information creates operational value when it changes decisions. Prediction can improve awareness, but the benefit depends on whether the organization can reallocate staff, reserve beds, sequence appointments, and coordinate diagnostics. The 21-minute difference between AI-only and integrated planning is a scenario implication, not a separately estimated causal effect.

The assumed overtime reduction reflects better alignment between predicted workload and staffing. It does not imply minimizing staff without regard to fatigue, continuity, skill mix, shift rules, or safe staffing. Similarly, the decrease in utilization from 91% to 84% represents deliberate slack. A modest reserve can prevent emergency boarding and reduce cascading delays, provided the reserve is operationally and financially sustainable.

4.4. Sensitivity protocol

A real implementation should stress-test forecast error, demand volatility, capacity loss, and objective weights. Recommended tests include increasing arrivals by 10% and 20%; increasing LOS standard deviation by 25% and 50%; removing 5% and 10% of staffed beds; increasing staff absence; underpredicting high-acuity demand; and varying the cost-versus-waiting weight. Feasibility, priority violations, overtime spikes, and recommendation stability should be recorded. Robust or CVaR-based formulations are relevant when tail events dominate operational risk.

Stress dimension	Base	Stress levels	Main response
Daily arrivals	100% assumed mean	+10%, +20%	Waiting; deferred demand
LOS variability	Class-specific SD	+25%, +50%	Bed blocking; reserve
Staffed beds	100% available	−5%, −10%	Feasibility; critical delay
Staff hours	100% rostered	−5%, −10%	Overtime; skill coverage
High-acuity bias	Unbiased	−5%, −10% underprediction	Priority violations
Objective weights	Baseline λ	±20% cost/wait weights	Pareto stability

4.5. Reproducibility

The accompanying Python script generates the synthetic 90-day workload stream, writes the scenario KPI table, and creates the charts. The KPI assumptions are stated explicitly in the code. The script does not train a clinical prediction model or solve the full MILP; it reproduces the illustrative workload stream and scenario comparison. For an empirical extension, replace assumptions with de-identified hospital data, train and calibrate prediction models, implement the MILP in GAMS or Python/Pyomo, and compare policies through discrete-event simulation and prospective evaluation. A solver report should include feasibility, optimality gap, runtime, and sensitivity to parameter perturbations.

MIMIC-IV may support retrospective model development subject to its credentialing and data-use requirements [1]. The optimization layer can be implemented in GAMS, Pyomo, or a commercial

solver. Data transformations, model versions, random seeds, solver settings, and approval records should be archived.

5. Discussion

The framework treats AI and IE as complementary capabilities. AI estimates uncertain states from high-dimensional data; IE structures the decision problem, represents constraints, and selects allocations. Healthcare operations are not prediction problems alone. A forecast has value when it informs a decision that can be executed safely and fairly. The framework therefore requires organizations to define decision rights, constraints, and escalation pathways before selecting algorithms.

The numerical illustration also shows why utilization should not be the sole efficiency metric. High utilization can increase queues and reduce the ability to absorb uncertain demand. In queueing systems, waiting can rise sharply as utilization approaches capacity, particularly when arrivals and service times vary. A reserve can improve responsiveness and reduce downstream congestion. Lower utilization in the integrated scenario is consistent with a resilience-oriented policy, provided that reserve capacity is staffed and financially sustainable.

Clinical priority must be encoded carefully. An objective minimizing average waiting may disadvantage urgent patients unless their needs are explicitly represented. The proposed priority constraint limits deferral for high-priority demand. In practice, triage categories and escalation thresholds must be clinically governed; the model must never independently downgrade urgency. Recommendations should explain which constraints are binding and what trade-offs are involved.

Fairness is both a modeling and governance issue. Access disparities may arise from geographic, language, disability, socioeconomic, digital access, or historical factors. Excluding sensitive attributes does not eliminate bias because proxies may reproduce inequities. Hospitals should audit calibration and error rates across relevant groups, monitor waiting and cancellation outcomes, and investigate unjustified disparities. Fairness rules should be co-designed with clinicians, patient representatives, legal experts, and health-equity specialists.

Data governance is essential because integrated systems combine clinical and operational data. Access should follow least-privilege principles; data should be minimized; and audit logs should record access and model versions. WHO guidance [10] emphasizes accountability, transparency, inclusiveness, and human oversight. Practical controls include documented intended use, model

cards, data lineage, impact assessment, incident reporting, change management, and clinician override.

Digital twins may provide an environment for scenario testing, but they should be evaluated on whether they remain synchronized with the real process and improve decisions. Reviews from 2025 and 2026 [7,8] indicate that deployment and validation remain challenges. A hospital twin should expose data freshness, uncertainty, workflow state, and the distinction between simulated and observed quantities. It should not create false precision or become an opaque command center.

Limitations are substantial. The numerical values are synthetic and cannot support claims about actual hospital performance. The formulation is aggregate and omits detailed pathways, precedence constraints, specialty-specific resources, staff preferences, shift rules, infection-control restrictions, and emergency recourse. Objective weights and buffers are context-dependent. Predictive models can drift as populations and practices change. Operational improvements do not automatically imply better clinical outcomes. These limitations motivate prospective multi-site research.

6. Conclusion

This paper developed an integrated framework that combines AI and industrial engineering for smart healthcare management and decision support. AI estimates demand, no-shows, service duration, LOS, and workload. IE translates those estimates into actionable decisions on appointments, beds, staffing, diagnostics, and capacity buffers. Simulation and monitoring evaluate whether decisions remain robust as new information arrives, while human governance protects clinical authority and accountability.

The proposed formulation minimizes a weighted combination of waiting penalties, overtime, operating cost, and capacity mismatch, subject to demand balance, resource capacity, staffing, skill compatibility, clinical priority, and fairness constraints. A scenario-based extension incorporates uncertainty, while rolling-horizon control updates plans as the system changes. The framework supports operational decisions; it does not replace diagnosis, treatment selection, or discharge authority.

The synthetic scenario illustrates possible directions under the stated assumptions. Compared with conventional planning, the integrated scenario reduces waiting from 74 to 41 minutes (44.6%), overtime from 18% to 9% (50.0%), cost per patient from USD 1,250 to USD 1,160 (7.2%), and critical-priority delay from 0.40% to 0.20% (50.0%). Utilization decreases from 91% to 84%,

suggesting that planned slack may be preferable to continuous operation near full occupancy. These values are not empirical estimates, have not been statistically tested, and must not be interpreted as evidence of clinical benefit.

The managerial implication is that hospitals should invest not only in predictive models but also in the operational infrastructure required to use predictions: interoperable data, process ownership, transparent optimization, staffing and safety constraints, human review, and continuous auditing. Future research should evaluate integrated systems using real hospital data, temporal and external validation, discrete-event simulation, and prospective controlled implementation. Studies should report model accuracy alongside operational, equity, safety, and implementation outcomes.

Future work should extend the model to multi-hospital networks, operating room and diagnostic scheduling, home care coordination, emergency surge planning, and energy-aware hospital operations. Distributionally robust optimization, safe reinforcement learning, explainable machine learning, and digital twins are promising, but complexity should be justified by measurable decision value. A simpler interpretable system integrated into clinical workflows may be more valuable than a sophisticated system that staff cannot trust or use.

In conclusion, smart healthcare is a coordinated sociotechnical capability, not merely the adoption of AI software. The strongest path forward integrates AI-based situational awareness, IE-based allocation of scarce resources, and governance that aligns efficiency with safety, equity, resilience, and patient-centered care.

References

[1] Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, Article 1. https://doi.org/10.1038/s41597-022-01899-x

[2] Ala, A., & Chen, F. (2022). Appointment scheduling problem in complexity systems of the healthcare services: A comprehensive review. Journal of Healthcare Engineering, 2022, Article 5819813. https://doi.org/10.1155/2022/5819813

[3] Youn, S., Geismar, H. N., & Pinedo, M. (2022). Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. Production and Operations Management, 31(12). https://doi.org/10.1111/poms.13867

[4] Babayoff, O., et al. (2023). Improving hospital outpatient clinics appointment schedules by prediction models. Journal of Medical Systems, 47, Article 5. https://doi.org/10.1007/s10916-022-01902-3

[5] Salah, H., & Srinivas, S. (2022). Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling. Computers & Industrial Engineering, 169, Article 108270. https://doi.org/10.1016/j.cie.2022.108270

[6] Jain, R., Singh, M., Rao, A. R., & Garg, R. (2024). Predicting hospital length of stay using machine learning on a large open health dataset. BMC Health Services Research, 24(1), 860.

[7] Liu, et al. (2025). Advancing health care with digital twins: Meta-review of applications and implementation challenges. Journal of Medical Internet Research, 27, e69544. https://doi.org/10.2196/69544

[8] Abd El-Mawla, N., Shehata, M., & Elhosseini, M. A. (2026). Digital twins for hospital and healthcare operations: A systematic review of resource allocation, infection control, and workflow optimization. Bioengineering, 13(9), Article 1072. https://doi.org/10.3390/bioengineering13091072

[9] Wubineh, B. Z., Deriba, F. G., & Woldeyohannis, M. M. (2024). Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: A systematic literature review. Urologic Oncology: Seminars and Original Investigations, 42(3), 48–56. https://doi.org/10.1016/j.urolonc.2023.11.019

[10] World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. https://www.who.int/publications/i/item/9789240029200