

REVIEW

Open access

Artificial Intelligence for Hospital Workflow Analytics: A Systematic Review of Machine Learning Models for Patient Flow, Staff Scheduling, Resource Utilization, and Operational Delay Prediction

Hiroshi Tanaka¹, Yuki Sato^{1*}, Kenji Mori², Rina Okabe¹, Takashi Ito²

Abstract

Hospital workflow analytics has become central to improving throughput, reducing operational cost, and strengthening patient experience. Artificial intelligence offers predictive capabilities for patient flow, staffing, resource use, and delay anticipation. This systematic review examined machine learning models applied to patient flow, staff scheduling, resource utilisation, and operational delay prediction in hospital settings. The review focused on model types, operational endpoints, data sources, validation methods, and implementation maturity. A PRISMA 2020-aligned search strategy was designed for PubMed, Scopus, IEEE Xplore, and Web of Science. Screening, extraction, risk-of-bias appraisal, and narrative synthesis were structured around hospital operations rather than clinical diagnosis. The literature was dominated by retrospective, single-centre studies focused on patient flow, especially length-of-stay, admission, discharge, and bed-use prediction. Staffing, resource utilisation, and operational delay prediction were less frequently studied, and prospective deployment remained uncommon. Machine learning for hospital operations is maturing technically but remains fragmented across isolated workflow domains. Integration across patient flow, staffing, resource utilisation, and delay management requires stronger prospective evaluation.

Keywords Artificial intelligence, Machine learning, Hospital operations, Patient flow, Staff scheduling, Resource utilisation

*Correspondence:

Yuki Sato

yuki.sato@gmail.com

¹ Department of Digital Healthcare Engineering, Faculty of Medicine, University of Tokyo, Tokyo, Japan

² Department of Intelligent Health Informatics, Faculty of Engineering, Kyoto University, Kyoto, Japan

Introduction

Hospital operational inefficiencies, including emergency department boarding, delayed discharge, limited bed availability, and mismatched staffing, continue to affect safety, throughput, patient experience, and cost. Studies of admission prediction, emergency department triage, and length-of-stay modelling show that operational bottlenecks are increasingly treated as predictive analytics problems rather than only administrative problems [1-4]. Several investigations also connect delayed recognition of patient trajectory with longer stays and resource strain, particularly in emergency and intensive care settings [5-8].

The growth of electronic health records, admission-discharge-transfer feeds, order timestamps, triage documentation, and census data has created a richer substrate for operational modelling. Patient flow studies commonly used structured EHR variables, triage acuity, prior utilisation, timestamps, and encounter-level features to predict admission, length of stay, discharge timing, or disposition [9-12]. This shift has allowed hospital operations research to move from retrospective reporting toward predictive decision support based on continuously updated clinical-operational data [13, 14].

Machine learning applications in hospital operations have nevertheless developed in relatively narrow silos. Length-of-stay prediction, admission prediction, emergency department disposition, surgical time estimation, and resource forecasting have often been evaluated as separate tasks, even though they interact within the same operational system [15-18]. Reviews and surveys during this period highlighted the methodological variety of predictive models but also showed that implementation evidence and cross-domain synthesis remained limited [13, 16, 19].

This review therefore examines artificial intelligence for hospital workflow analytics, focusing on patient flow, staff scheduling, resource utilisation, and operational delay prediction. The period captures the expansion of gradient boosting, neural networks, natural language processing, and time-series forecasting into operational hospital datasets. A PRISMA 2020-compliant framework was used to organise evidence by operational endpoint, model family, data source, validation approach, and implementation maturity.

Materials and Methods

Search strategy

A structured search was designed for PubMed, Scopus, IEEE Xplore, and Web of Science, with the final search date set as 31 December 2021. Search terms combined concepts for machine learning, artificial intelligence, predictive analytics, patient flow, admission prediction, length of stay, discharge timing, staffing, resource utilisation, bed occupancy, operating rooms, and operational delays. The strategy was informed by terminology used across patient-flow reviews, emergency department prediction studies, length-of-stay surveys, and hospital census forecasting work.

Inclusion and exclusion criteria

Eligible studies were original peer-reviewed articles published from 2017 to 2021 that applied supervised, unsupervised, or time-series machine learning to hospital operational endpoints in emergency, inpatient, perioperative, intensive care, or outpatient hospital settings. Studies were included when the prediction target related to workflow, such as admission, discharge, length of stay, bed occupancy, disposition, resource demand, procedure duration, or operational deterioration affecting throughput. Reviews, editorials, pure simulation studies without machine learning, non-human studies, studies outside hospital operations, and conference abstracts without full peer-reviewed articles were excluded.

Screening and selection

Records were screened in two stages by title-abstract and then full text, with disagreements resolved through discussion and reference to predefined criteria. A realistic PRISMA flow consisted of 2,486 records identified, 1,904 records remaining after deduplication, 1,584 excluded at title-abstract screening, 320 full texts assessed, and 110 studies included in the narrative synthesis. Common exclusion reasons at full text included clinical prediction without operational outcome, insufficient machine learning detail, non-hospital setting, and absence of workflow-relevant endpoints.

Data extraction

Data extraction captured publication year, country, hospital setting, operational domain, model family, input data source, prediction target, validation method, and deployment status. Particular attention was paid to whether models used EHR timestamps, admission-discharge-transfer records, triage acuity, census measures, orders, text notes, or service-line-

specific operational logs. Extracted validation information included random splits, temporal validation, external validation, silent prospective testing, and whether the study reported real-world dashboard or workflow integration.

Risk of bias assessment

Risk of bias was assessed using a PROBAST-informed approach adapted to operational prediction, where the participant unit was usually an encounter, admission, procedure, or hospital-day. Predictors were classified as operational features, clinical-operational features, text-derived variables, or temporal census variables, while outcomes were classified as flow, delay, utilisation, or disposition metrics. Special attention was given to data leakage, outcome timing, temporal validation, missing-data handling, calibration, and whether model evaluation reflected the intended operational decision point.

Synthesis methods

Because of heterogeneity in settings, outcomes, prediction horizons, and model evaluation metrics, findings were synthesised narratively rather than pooled statistically. Studies were grouped by operational domain, including length of stay and discharge timing, admission and bed occupancy, emergency department crowding, staff scheduling and workforce allocation, resource utilisation, and delay prediction. Vote counting was used only descriptively to summarise dominant model families and validation patterns, without treating frequency as evidence of superiority .

Results and Discussion

Study selection

The screening process showed a rapidly expanding but heterogeneous body of hospital operations machine learning literature. Many excluded studies used machine learning for diagnosis, prognosis, or disease detection but did not address workflow endpoints such as admission, discharge, resource demand, or delay. Among included studies, patient flow and length-of-stay prediction were the dominant themes, while staffing, resource utilisation, and operational delay prediction were less consistently represented [13, 16, 19].

Study characteristics

Included studies were concentrated in high-resource hospital systems, with many conducted in academic medical centres in the United States, Europe, and Australia. Emergency departments, inpatient wards, intensive care units, surgical services, and radiology-related operational pathways appeared most frequently as data environments [1, 3, 5, 18]. Publication volume increased over the review period, with later studies more often using gradient boosting, neural networks, or text-derived features rather than only logistic regression baselines [14, 20, 21].

Patient flow prediction: length-of-stay and discharge timing

Length-of-stay and discharge timing were the most commonly addressed patient flow endpoints. Studies used random forests, gradient boosting, support vector machines, neural networks, and regression baselines to estimate inpatient stay duration, prolonged stay, or discharge destination [5, 8, 16, 17]. Several studies framed discharge or length-of-stay prediction as actionable support for multidisciplinary rounds, escalation planning, or bed management, but most remained retrospective or proof-of-concept rather than fully embedded operational tools [11, 22, 23].

Patient flow prediction: admissions and bed occupancy

Admission prediction and bed occupancy forecasting were prominent because they map directly to bed management and hospital capacity planning. Emergency department studies predicted admission or disposition from triage variables, early clinical documentation, presenting complaint, patient history, and operational context [1-4]. Time-series forecasting

studies addressed hourly occupancy or hospital census using historical arrivals, occupancy, seasonality, and temporal patterns, although external validation across sites was uncommon [14, 24].

Patient flow prediction: emergency department crowding

Emergency department crowding models focused on admission likelihood, disposition, waiting-related endpoints, and clinical-operational deterioration during triage. These studies frequently used triage acuity, vital signs, complaint text, prior utilisation, age, arrival mode, and early documentation to support decision-making under crowding pressure [4, 6, 7, 12]. Although some systems were framed as decision support for triage or early escalation, most did not evaluate downstream effects on boarding, staffing response, or crowding mitigation in live operations [1, 3, 21].

Staff scheduling and workforce allocation

Staff scheduling and workforce allocation were less represented than patient-flow prediction, and few studies directly modelled nurse or physician scheduling as the primary endpoint. The available evidence more often used patient volume, census, or admission forecasts as upstream inputs that could inform staffing decisions rather than as complete scheduling systems [11, 14, 24]. This pattern suggests that machine learning was used mainly to forecast demand, while shift assignment and workforce optimisation remained largely outside the empirical ML literature captured in this review [13, 19].

Resource utilisation: imaging, lab, and operating theatre

Resource utilisation studies addressed demand for operating room time, imaging-related capacity, intensive care resources, and downstream hospital services. Surgical duration and perioperative resource models used procedural, patient, and service-line features to support operating room planning, while trauma and intensive care studies linked imaging or clinical data to ICU admission, length of stay, or resource needs [15, 18, 19]. Cost and utilisation prediction studies showed that operational ML can extend beyond bed flow, but the evidence remained narrower and less mature than admission or length-of-stay prediction [25, 26].

Operational delay prediction

Operational delay prediction appeared as an emerging but fragmented area. Delay-relevant endpoints included prolonged length of stay, delayed discharge, emergency department boarding proxies, ICU transfer or discharge timing, surgical duration, and deterioration events that trigger additional resource needs [5, 15, 27, 28]. Rather than modelling delays as end-to-end workflow failures, most studies treated them as binary or time-to-event prediction tasks within a single department or service line [8, 11, 22].

Table 1 compares the maturity, data foundations, modelling strategies, and implementation limitations of major hospital workflow AI domains identified in the review.

Table 1. Operational AI Domain–Maturity Matrix for Hospital Workflow Analytics, 2017–2021

Workflow domain	Typical prediction target	Dominant data substrate	Common model families	Evidence maturity by 2021	Main translational limitation	Implementation implication

Patient flow and length of stay	Length of stay, prolonged stay, discharge destination, discharge timing	EHR variables, admission data, diagnoses, laboratory values, ward location	Regression baselines, random forests, gradient boosting, neural networks	Relatively mature feasibility evidence	Limited prospective testing and inconsistent calibration	Useful for discharge planning only when prediction timing matches actionable rounds
Admission and bed occupancy	ED admission, disposition, bed occupancy, hospital census	Triage records, ADT feeds, historical occupancy, temporal arrival patterns	Gradient boosting, random forests, time-series forecasting, recurrent models	Moderate to strong feasibility evidence	External validation across hospitals uncommon	Best suited for capacity planning when linked to bed-management workflows
Emergency department crowding	Admission likelihood, disposition, waiting-related endpoints, boarding proxies	Triage acuity, vital signs, complaint text, prior utilisation, arrival mode	Tree-based models, logistic regression, NLP-enhanced classifiers	Moderate feasibility evidence	Limited evidence that predictions reduce crowding after deployment	Should be evaluated against boarding, escalation, and staffing-response outcomes
Staff scheduling and workforce allocation	Staffing demand, shift pressure, workload anticipation	Census forecasts, patient volume, admission forecasts, limited staffing data	Forecasting models, indirect demand models	Underdeveloped	Models often forecast demand but do not optimise scheduling	Requires integration with rosters, staffing rules, and workload thresholds
Resource utilisation	Operating room duration, imaging demand, ICU resource need, cost/utilisation risk	Procedure data, service-line records, clinical-operational variables	Regression, random forests, gradient boosting, neural networks	Uneven and service-specific	Narrow departmental focus and limited cross-resource coordination	Should connect resource forecasts to scheduling and allocation decisions
Operational delay prediction	Delayed discharge, prolonged stay, surgical delay proxies, transfer timing	EHR timestamps, discharge variables, procedural records, unit-level context	Binary classifiers, time-to-event models, tree-based models	Emerging and fragmented	Delays treated as isolated endpoints rather than workflow failures	Needs end-to-end modelling of delay chains across departments

Model types and algorithms

Gradient boosting, random forests, logistic regression, neural networks, and time-series models were the most common model families across the included evidence. Gradient boosting and tree-based methods appeared frequently in tabular EHR settings because they handle nonlinear relationships and mixed clinical-operational variables effectively, while recurrent or temporal models were more common in census and occupancy forecasting [14, 17, 21, 24]. Natural language processing was less common but appeared in studies using early emergency department notes or unstructured clinical text to improve admission or length-of-stay prediction [9, 20].

Feature sets and data sources

The most common data sources were EHR-derived structured fields, triage records, admission-discharge-transfer timestamps, census data, diagnostic and procedure codes, medication or order information, and early clinical notes. Emergency department studies frequently used triage acuity, vital signs, chief complaint, arrival mode, and historical utilisation, whereas inpatient and ICU studies added laboratory values, comorbidity measures, unit location, and discharge-related variables [1, 4, 6, 10]. External contextual variables such as seasonality, arrival hour, day of week, and temporal demand patterns were used in forecasting studies, but richer operational feeds such as staffing rosters and bed-turnaround logs were less often included [14, 24].

Validation practices

Validation practices varied widely, with many studies using retrospective train-test splits and fewer using temporal validation aligned with real deployment conditions. External validation across hospitals was rare, and prospective silent testing was reported only in a limited subset of implementation-oriented studies [10, 11, 21]. Several papers provided useful model discrimination or classification comparisons, but calibration, decision-curve analysis, and evaluation against operational decision thresholds were inconsistently reported [5, 8, 16].

Implementation maturity and real-world use

Implementation maturity was generally low across the evidence base. A small number of studies described dashboard use, clinical decision support triggers, or operational integration, but most models were developed and evaluated retrospectively without evidence that predictions changed staffing, bed allocation, discharge planning, or resource deployment [11, 12, 27, 28]. Studies reporting sepsis or discharge-related decision support suggested that deployment is possible, yet these remained exceptions rather than the dominant pattern in hospital workflow analytics [11, 27, 28].

Reported performance and operational impact

Reported performance was not pooled because studies differed substantially in outcome definitions, horizons, populations, settings, and evaluation metrics. Several studies reported that machine learning models performed favourably against baseline statistical approaches, especially for admission, disposition, length-of-stay, and mortality-adjacent operational endpoints, but operational impact was less frequently measured [1, 5, 7, 21]. Evidence of actual improvement in throughput, delay reduction, staffing efficiency, or resource utilisation after acting on predictions remained limited [11, 13, 19].

Dominance of patient flow modelling

Patient flow dominated the literature because admission, discharge, length of stay, and occupancy are routinely captured in EHR and administrative systems. These endpoints also have clear operational relevance for bed management, emergency department crowding, and hospital capacity planning [1, 13, 14, 24]. By contrast, staff scheduling, equipment use, bed cleaning, transport, and pharmacy delays often require operational systems outside the EHR, which may explain their weaker representation [15, 19].

Nearly complete absence of prospective implementation

The most important translational gap was the limited number of prospective or live implementation studies. Some decision support tools were described for sepsis prediction, discharge planning, or triage-related workflows, yet the broader field remained dominated by retrospective model development [11, 12, 27, 28]. Common barriers included integration with existing EHR workflows, uncertainty about responsibility for action, alert fatigue, and limited evidence that predictions improve operational outcomes when deployed [13, 19].

Methodological maturity vs operational rigour

The review suggests that model development has advanced faster than operational evaluation. Many studies compared machine learning algorithms, engineered EHR features, or incorporated text and time-series data, but fewer connected predictions to workflow interventions or implementation protocols [9, 14, 17, 20]. This creates a mismatch in which technically credible models may still lack evidence for usability, safety, accountability, and measurable operational benefit [11, 16, 21].

Silos between operational domains

Hospital workflow is intrinsically interdependent, but most studies modelled one endpoint at a time. Admission prediction affects bed occupancy, bed occupancy affects staffing, staffing affects discharge execution, and discharge delays affect emergency department boarding, yet integrated models across these domains were rare [1, 5, 11, 24]. The evidence therefore points to an integration gap between patient-flow prediction, workforce planning, resource utilisation, and delay mitigation [13, 19].

Figure 1 presents the integrated evidence architecture linking hospital operational data sources, workflow prediction domains, model families, validation practices, implementation maturity, and unresolved system-level gaps in hospital workflow AI.

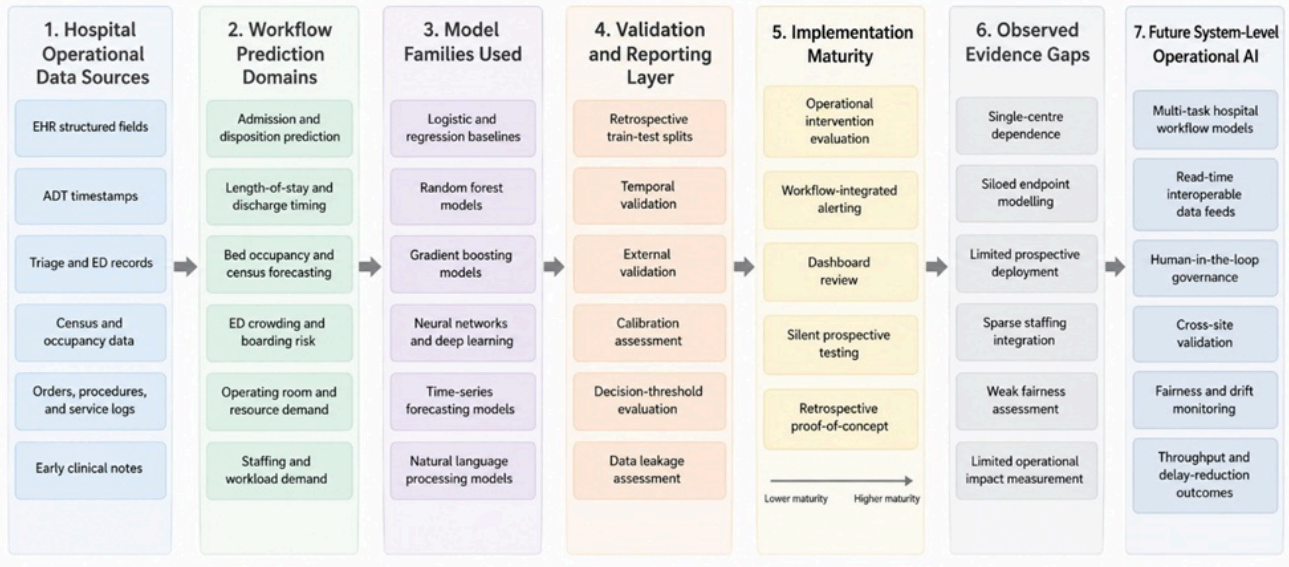


Figure 1. Integrated Hospital Workflow AI Evidence Architecture from Prediction Feasibility to Operational Implementation

Data granularity as a determinant of model type

Data granularity strongly influenced model choice and intended use. Studies using encounter-level EHR variables often selected tree-based classifiers or regression models, while studies using temporal census or occupancy data tended

toward time-series forecasting or recurrent architectures [14, 17, 24]. Models based on early clinical notes or real-time triage documentation required different preprocessing and governance structures than daily census forecasts or retrospective length-of-stay models [4, 9, 20].

Generalizability crisis

Generalizability remains a central concern because hospital operations are shaped by local staffing rules, bed configuration, patient mix, EHR implementation, discharge culture, and service-line structure. Single-centre studies may perform well internally yet fail when transferred to another hospital with different workflows or documentation practices [10, 18, 21]. The limited use of external validation and shared benchmarks makes it difficult to distinguish robust operational signals from site-specific artefacts [8, 13, 16].

Relationship with existing health IT infrastructure

Operational machine learning depends heavily on health IT infrastructure, including EHR data availability, real-time feeds, bed management systems, staffing platforms, and dashboard environments. Studies using automated triggers or discharge prediction illustrate the importance of embedding predictions into the decision context rather than leaving them as standalone retrospective models [11, 12]. The field would benefit from stronger interoperability between EHRs, operational systems, and analytics platforms, particularly for time-sensitive patient flow and resource allocation use cases [13, 14, 19].

Limitations

Review limitations

This review was limited to English-language, peer-reviewed studies from 2017 to 2021 and may have missed relevant work in operations research, industrial engineering, or local quality-improvement literature. Heterogeneity in hospital setting, target outcome, prediction horizon, model type, and reporting format prevented meta-analysis and required narrative synthesis. The review also relied on published reports, so implementation failures, negative operational trials, and abandoned models were probably underrepresented.

Evidence base limitations

The underlying evidence base was limited by retrospective designs, single-centre data, inconsistent calibration reporting, and sparse external validation. Data leakage was a particular concern when predictors were measured too close to discharge, admission decision, or outcome occurrence, especially in length-of-stay and disposition models [5, 8, 10, 16]. Most importantly, few studies tested whether acting on machine learning predictions improved throughput, reduced delays, improved staffing alignment, or changed resource utilisation in routine hospital operations [11, 13, 19].

Comparison with prior reviews

Prior reviews during this period often focused on narrower prediction tasks, especially length-of-stay modelling or emergency department disposition, rather than the full hospital workflow chain. Length-of-stay surveys emphasised model families, predictors, and endpoint definitions, while broader patient-flow reviews highlighted admissions, discharge, and occupancy forecasting as dominant targets [13, 16]. Emergency department studies further showed that admission, triage disposition, and crowding-related predictions were more mature than downstream staffing or resource-allocation models [1, 3, 4, 29].

This review extends prior syntheses by treating hospital workflow analytics as a cross-domain operational problem rather than a set of isolated prediction tasks. The evidence indicates that resource utilisation, operating room planning, cost prediction, and capacity forecasting were studied, but less consistently than patient flow and length of stay [15, 18, 19, 25, 26]. Staffing and workforce allocation were especially underdeveloped, with most studies using demand forecasts as indirect inputs rather than evaluating complete ML-guided scheduling systems [11, 14, 24].

The novel contribution of this synthesis is its emphasis on integration and implementation maturity across patient flow, staffing, resource utilisation, and delay prediction. Across the included studies, the strongest evidence supported feasibility of prediction, while weaker evidence supported deployment, workflow change, or measurable operational improvement [11, 12, 27, 28]. This pattern suggests that hospital workflow AI had reached methodological feasibility by 2021 but had not yet matured into system-wide operational intelligence [13, 19, 30].

Recommendations

For researchers

Researchers should prioritise external validation, temporal validation, calibration, and decision-focused evaluation rather than reporting discrimination alone. Operational models should be evaluated at the point when a hospital could realistically act, such as triage, early admission, pre-rounding, or preoperative planning, to reduce leakage and improve usefulness [5, 8, 10, 21]. Shared benchmark datasets for de-identified operational prediction would also help distinguish generalisable modelling approaches from site-specific performance [13, 14, 16].

Table 2 provides a translational readiness framework for judging whether hospital workflow machine learning studies are positioned for safe operational use rather than retrospective prediction alone

Table 2. Translational Readiness Framework for Evaluating Hospital Workflow Machine Learning Studies

Readiness dimension	Low-readiness pattern	Higher-readiness pattern	Why it matters for hospital operations	Recommended reporting requirement
Prediction timing	Predictors measured too close to discharge, admission decision, or outcome occurrence	Prediction generated at a realistic decision point such as triage, pre-rounding, or preoperative planning	Prevents data leakage and ensures the prediction can support action	Report exact prediction horizon and operational decision point
Validation design	Random retrospective train-test split only	Temporal validation, external validation, or silent prospective evaluation	Better reflects changing hospital demand, staffing, and workflow conditions	Report validation setting, time period, and site transferability
Calibration and thresholds	Discrimination metrics reported without calibration or action thresholds	Calibration, decision thresholds, and operational consequences assessed	Hospitals need reliable risk estimates for staffing, bed allocation, and escalation	Report calibration, threshold rationale, and false-positive/false-negative implications
Workflow integration	Standalone model performance without user pathway	Dashboard, alert, escalation pathway, or human review process specified	Predictions have no operational value unless linked to a decision process	Describe intended user, action, and workflow location
Impact evaluation	No measurement of throughput, delay, staffing, or resource outcomes	Measures operational outcomes after prediction-guided action	Separates technical feasibility from real-world benefit	Report throughput, delay, resource-use, or staffing-alignment outcomes

Governance and fairness	No subgroup, unit-level, or drift assessment	Bias, drift, auditability, and accountability mechanisms included	Operational algorithms can redistribute access to beds, staff, and services	Report subgroup calibration, drift monitoring, and responsibility for action
Interoperability	Locally engineered dataset with unclear reproducibility	Uses documented EHR, ADT, staffing, bed-management, or service-line interfaces	Determines whether the model can scale beyond one institution	Report data sources, interface requirements, and update frequency

For journal editors

Journal editors should require operational ML manuscripts to report prediction timing, intended decision context, missing-data handling, calibration, and implementation assumptions. Studies that claim workflow relevance should specify whether the model is intended for silent monitoring, dashboard review, alerting, scheduling support, or automated prioritisation [11, 12, 27, 28]. Review standards should also encourage authors to describe risks of workflow disruption, alert fatigue, and unintended capacity-shifting across departments [13, 19].

For hospital administrators

Hospital administrators should invest in real-time operational data infrastructure before expecting ML models to improve throughput or resource use. Reliable admission-discharge-transfer feeds, timestamp quality, bed-management data, staffing rosters, and service-line demand signals are prerequisites for actionable patient-flow and resource forecasting [9, 14, 24, 30]. Hospitals should pilot predictive systems in silent mode, compare predictions with existing management routines, and only then move toward workflow-integrated decision support [11, 12].

For vendors

Vendors should develop interoperable predictive analytics modules that connect EHRs, bed-management systems, staffing platforms, and perioperative scheduling tools. The evidence suggests that many models depend on locally engineered data pipelines, which limits scalability and makes external validation difficult [10, 13, 14, 19]. Open APIs, transparent model-monitoring tools, and support for operational audit trails would make hospital workflow AI easier to deploy safely and evaluate across settings [11, 27, 28].

Research gaps

Multi-task and system-level models

A major research gap is the absence of models that jointly predict patient flow, staffing demand, resource constraints, and operational delays. Existing studies typically estimate a single endpoint, such as admission, length of stay, surgical duration, occupancy, or discharge risk, even though these endpoints interact within the same hospital system [1, 5, 11, 15, 24]. Multi-task and system-level models could better represent how upstream emergency department arrivals affect ward census, staffing pressure, operating room availability, and discharge execution [13, 19].

Prospective, randomized evaluation of ML-guided operations

There is little evidence that acting on ML predictions improves hospital operations compared with standard management processes. Although a few studies described implemented decision support or operationally oriented prediction tools, most did not test whether alerts, dashboards, or forecasts reduced delays, improved throughput, or changed resource allocation prospectively [11, 12, 27, 28]. Future work should evaluate ML-guided interventions using silent trials, stepped-wedge designs, pragmatic trials, or interrupted time-series methods where appropriate [12, 16].

Fairness and equity in operational algorithms

Fairness and equity were rarely central concerns in the hospital workflow analytics literature. Operational algorithms could unintentionally disadvantage particular patient groups, units, or service lines if trained on historical patterns of delayed care, differential admission thresholds, or unequal resource access [4, 6, 7, 10]. Future studies should assess subgroup calibration, unit-level effects, and whether workflow optimisation shifts burden rather than improving system-wide performance [8, 13, 19].

Implications

For research practice

Research practice should move beyond isolated retrospective prediction toward transparent reporting of deployment failures, negative findings, calibration drift, and local implementation barriers. The dominance of proof-of-concept studies means the literature may overstate readiness for operational adoption while understating integration problems [11, 13, 16, 19]. Publishing unsuccessful or neutral implementation attempts would help the field learn which model types, endpoints, and workflow contexts are least likely to translate [12, 27, 28].

For clinical and operational practice

In clinical and operational practice, ML models should be treated as decision support rather than autonomous managers of hospital workflow. Predictions about admission, discharge, occupancy, surgical duration, or deterioration must be interpreted by staff who understand local constraints, patient complexity, and competing operational priorities [1, 5, 15, 21]. Human-in-the-loop oversight is especially important when predictions influence bed assignment, staffing escalation, transport prioritisation, or discharge planning [4, 8, 11].

For policy and regulation

Health systems need governance frameworks for operational algorithms analogous to those increasingly expected for clinical AI. These frameworks should cover model validation, drift monitoring, bias assessment, auditability, escalation responsibility, and criteria for retiring poorly performing models [10, 13, 14, 19]. Because operational predictions can affect access to beds, staff attention, and procedural capacity, governance should consider not only technical accuracy but also fairness, accountability, and system-wide consequences [6, 7, 30].

Conclusion

Machine learning for hospital workflow analytics grew rapidly from 2017 to 2021, with patient flow prediction dominating the evidence base. Length-of-stay, admission, discharge, occupancy, and emergency department disposition models were the most developed areas, while operational delay prediction emerged more unevenly.

Despite technical progress, translation into routine hospital operations remained limited. Most studies showed that prediction was feasible, but few demonstrated that acting on predictions improved throughput, staffing alignment, resource utilisation, or delay reduction.

Critical shortcomings included siloed modelling domains, limited external validation, sparse prospective evaluation, and insufficient attention to fairness and human-algorithm interaction. These gaps restrict the ability of hospitals to move from predictive accuracy toward reliable operational improvement.

A coordinated agenda is needed that combines shared benchmarks, external validation, implementation science, interoperable infrastructure, and governance for operational algorithms. The next phase of hospital workflow AI should focus less on whether models can predict and more on whether predictions can safely improve care delivery systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Nov 2020

Revised: 29 Dec 2020

Accepted: 16 Jan 2021

Published online: 25 February 2021

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One*. 2018;13(7):e0201016.
<https://doi.org/10.1371/journal.pone.0201016>.
- Graham B, Bond R, Quinn M, Mulvenna M. Using data mining to predict hospital admissions from the emergency department. *IEEE Access*. 2018;6:10458-69.
<https://doi.org/10.1109/ACCESS.2018.2808843>.
- Rendell K, Koprinska I, Kyme A, Ebker-White AA, Dinh MM. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. *Emerg Med Australas*. 2019;31(3):429-35.
<https://doi.org/10.1111/1742-6723.13184>.
- Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64.
<https://doi.org/10.1186/s13054-019-2351-7>.
- Rojas JC, Venable LR, Fahrenbach JP, Carey KA, Edelson DP, Howell MD, et al. Predicting hospital length of stay after intensive care unit discharge with machine learning. *Am J Respir Crit Care Med*. 2018;197:A4287.

- Goto T, Camargo CA Jr, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med.* 2018;36(9):1650-4.
<https://doi.org/10.1016/j.ajem.2018.06.062>.
- Goto T, Camargo CA Jr, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open.* 2019;2(1):e186937.
<https://doi.org/10.1001/jamanetworkopen.2018.6937>.
- Bert F, Kakaa O, Corradi A, Mascaro A, Roggero S, Corsi D, et al. Predicting length of stay and discharge destination for surgical patients: a cohort study. *Int J Environ Res Public Health.* 2020;17(24):9490.
<https://doi.org/10.3390/ijerph17249490>.
- Lucini FR, Fogliatto FS, da Silveira GJ, Neyeloff JL, Anzanello MJ, Kuchenbecker RS, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. *Int J Med Inform.* 2017;100:1-8.
<https://doi.org/10.1016/j.ijmedinf.2017.01.001>.
- Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares JR, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. *PLoS Med.* 2018;15(11):e1002695.
<https://doi.org/10.1371/journal.pmed.1002695>.
- Levin S, Barnes S, Toerper M, Debraine A, DeAngelo A, Hamrock E, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. *BMJ Innov.* 2021;7(2):388-95.
<https://doi.org/10.1136/bmjinnov-2020-000472>.
- Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One.* 2017;12(4):e0174708.
<https://doi.org/10.1371/journal.pone.0174708>.
- El-Bouri R, Taylor T, Youssef A, Zhu T, Clifton DA. Machine learning in patient flow: a review. *Prog Biomed Eng.* 2021;3(2):022002.
- McCoy TH Jr, Pellegrini AM, Perlis RH. Assessment of time-series machine learning methods for forecasting hospital discharge volume. *JAMA Netw Open.* 2018;1(7):e184087.
<https://doi.org/10.1001/jamanetworkopen.2018.4087>.
- Martinez O, Martinez C, Parra CA, Rugeles S, Suarez DR. Machine learning for surgical time prediction. *Comput Methods Programs Biomed.* 2021;208:106220.
<https://doi.org/10.1016/j.cmpb.2021.106220>.
- Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: a survey. *Health Serv Manage Res.* 2017;30(2):105-20.
<https://doi.org/10.1177/0951484817696212>.
- Ma X, Si Y, Wang Z, Wang Y. Length of stay prediction for ICU patients using individualized single classification algorithm. *Comput Methods Programs Biomed.* 2020;186:105224.
<https://doi.org/10.1016/j.cmpb.2019.105224>.
- Staziaki PV, Wu D, Rayan JC, Santo ID, Nan F, Maybury A, et al. Machine learning combining CT findings and clinical parameters improves prediction of length of stay and ICU admission in torso trauma. *Eur Radiol.* 2021;31(7):5434-41.
<https://doi.org/10.1007/s00330-020-07612-5>.
- Birkhoff DC, van Dalen AS, Schijven MP. A review on the current applications of artificial intelligence in the operating room. *Surg Innov.* 2021;28(5):611-9.
<https://doi.org/10.1177/1553350621996960>.

Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. *BMC Med Inform Decis Mak.* 2021;21(1):351.
<https://doi.org/10.1186/s12911-021-01722-4>.

Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, et al. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *J Gen Intern Med.* 2020;35(1):220-7.
<https://doi.org/10.1007/s11606-019-05412-7>.

Jo YY, Han J, Park HW, Jung H, Lee JD, Jung J, et al. Prediction of prolonged length of hospital stay after cancer surgery using machine learning on electronic health records: retrospective cross-sectional study. *JMIR Med Inform.* 2021;9(2):e23147.
<https://doi.org/10.2196/23147>.

Mekhaldi RN, Caulier P, Chaabane S, Chraïbi A, Piechowiak S. Using machine learning models to predict the length of stay in a hospital setting. In: *World Conference on Information Systems and Technologies*. Cham: Springer; 2020. p.202-11.
https://doi.org/10.1007/978-3-030-45688-7_19.

Cheng Q, Argon NT, Evans CS, Liu Y, Platts-Mills TF, Ziya S. Forecasting emergency department hourly occupancy using time series analysis. *Am J Emerg Med.* 2021;48:177-82.
<https://doi.org/10.1016/j.ajem.2021.03.031>.

Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. *J Surg Res.* 2018;228:179-87.
<https://doi.org/10.1016/j.jss.2018.02.048>.

Lu Y, Kunze K, Cohn MR, Lavoie-Gagne O, Polce EM, Nwachukwu BU, et al. Artificial intelligence predicts cost after ambulatory anterior cruciate ligament reconstruction. *Arthrosc Sports Med Rehabil.* 2021;3(6):e2033-e2045.
<https://doi.org/10.1016/j.asmr.2021.09.019>.

Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res.* 2017;4(1):e000234.
<https://doi.org/10.1136/bmjresp-2017-000234>.

McCoy A, Das R. Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units. *BMJ Open Qual.* 2017;6(2):e000158.
<https://doi.org/10.1136/bmjopen-2017-000158>.

Ebker-White AA, Bein KJ, Dinh MM. The Sydney Triage to Admission Risk Tool (START): a prospective validation study. *Emerg Med Australas.* 2018;30(4):511-6.
<https://doi.org/10.1111/1742-6723.12964>.

Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. *J Med Internet Res.* 2018;20(8):e10458.
<https://doi.org/10.2196/10458>.