

ORIGINAL RESEARCH

Open access

Multimodal Foundation Model for Predicting Care Delivery Delays Using Electronic Health Record Events, Operational Logs, Staff Assignments, Patient Messages, Facility Capacity Indicators, and Service Queue Data

Minh Tran^{1*}, Duc Pham¹

Abstract

Care delivery delays arise from asynchronous interactions among clinical decisions, operational constraints, staffing patterns, facility capacity, and patient communication. These delays are rarely represented within a single predictive framework that captures the hospital as a dynamic multimodal system. Existing delay prediction models often focus on one operational endpoint, such as discharge timing, transport coordination, or procedure scheduling. Such models may require extensive hand-engineered features and may not generalize across units, services, or delay types. This article proposes a conceptual multimodal foundation model for predicting care delivery delays using electronic health record events, operational logs, staff assignments, patient messages, facility capacity indicators, and service queue data. The goal is to describe a reusable model backbone that could support multiple downstream operational prediction tasks. The proposed model would use a transformer-based architecture pre-trained through self-supervised learning over heterogeneous temporal hospital data. Task-specific prediction heads could then be fine-tuned for medication delays, procedure delays, discharge delays, transport delays, and broader care progression bottlenecks. Conceptually, the model would learn a holistic representation of clinical workflow, operational pressure, staffing context, patient communication burden, and service demand. It would be expected to produce dynamic delay risk estimates that adapt to changing hospital conditions and tolerate incomplete modality availability. A multimodal foundation model could unify delay prediction across multiple operational domains. Such an approach may support proactive hospital management by transforming fragmented data streams into shared, contextualized representations of care delivery risk.

Keywords Hospital operations, Electronic health records, Self-supervised learning, Multimodal foundation model, Care delivery delays, Patient messages

*Correspondence:

Minh Tran

minh.tran@gmail.com

¹ Department of Smart Healthcare Systems, Faculty of Medicine, Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam

Introduction

Care delivery delays, including late medication administration, delayed procedures, postponed discharges, transport bottlenecks, and slow consult completion, are not isolated events but manifestations of a broader operational system under pressure. Discharge delays have been studied as outcomes that reflect patient complexity, service readiness, and

institutional workflow constraints, while patient transport and procedural scheduling studies show how downstream care progression depends on coordinated operational execution [1-3]. Medication and procedure timing are similarly shaped by workload, crowding, staffing, and queue accumulation, making delay prediction a temporal and multimodal problem rather than a simple classification exercise [4, 5]. A model intended for care delivery delay prediction must therefore reason over both clinical state and operational context.

Many existing predictive approaches are task-specific, designed for a single outcome such as discharge readiness, inpatient flow, case duration, or patient transport optimization. These models have demonstrated that machine learning can support hospital operational decisions, but they often depend on features tailored to one service line, one workflow, or one institutional dataset [6-8]. Even when such models are clinically useful, they may not capture the simultaneous effects of bed capacity, staff availability, patient communication, pending orders, and departmental backlogs. This creates a gap between the complexity of real hospital operations and the narrow scope of many predictive tools.

Foundation models offer a different modeling paradigm because they learn general-purpose representations through pre-training and can then be adapted to downstream prediction tasks. In healthcare, transformer-based models trained on structured electronic health record sequences have shown how contextualized event representations can support multiple clinical prediction tasks [9-11]. More recent work has argued that biomedical artificial intelligence increasingly requires multimodal systems capable of integrating structured data, text, imaging, and temporal signals, although the operational side of healthcare remains less developed than clinical diagnosis or prognosis [12, 13]. A multimodal operational foundation model could extend this logic from patient-centered clinical prediction to workflow-centered delay prediction.

The central thesis of this article is that a multimodal foundation model pre-trained on institutional care delivery data could provide a shared backbone for predicting diverse delay types. Instead of developing bespoke models for discharge delays, transport delays, medication delays, patient message escalation, and service queue congestion, one operational transformer could learn reusable temporal representations and support task-specific fine-tuning. Such a model would be expected to improve conceptual scalability, robustness to missing data, and cross-task transfer compared with isolated models built around single endpoints. The contribution is therefore model-oriented: a design framework for applying foundation-model principles to hospital operations without claiming experimental results.

Background

Types of care delivery delays

Care delivery delays can be organized into a taxonomy that includes medication delays, diagnostic delays, procedural delays, discharge delays, transport delays, and broader care progression delays. These categories are interdependent because a delayed laboratory result may postpone clinical decision-making, a transport bottleneck may delay imaging, and a discharge delay may constrain bed availability for incoming patients [3, 14, 15]. Prior discharge prediction studies illustrate how patient readiness, care coordination, and institutional capacity interact in determining whether a patient can leave the hospital at the intended time [1, 16]. A foundation model for delay prediction should therefore treat delay types as related operational outcomes rather than independent prediction problems.

Data modalities reflecting the hospital state

The hospital state is reflected in multiple data modalities, including timestamped electronic health record events, admission-discharge-transfer logs, bed management systems, staff rosters, patient portal messages, facility census, occupancy indicators, and service queues. Structured EHR events encode orders, results, diagnoses, and medications, while operational logs capture movement, bed cleaning, transfer requests, transport milestones, and departmental handoffs [7, 17]. Patient-generated messages add a communication modality that may reveal urgency, dissatisfaction, confusion, or unmet care needs before they appear in structured operational metrics [18, 19]. Capacity and queue data provide the system-level context needed to distinguish a patient-specific clinical delay from a hospital-wide bottleneck.

Foundation models in healthcare

Foundation models in healthcare build on large-scale pre-training, contextual representation learning, and adaptation to downstream clinical prediction tasks. Structured EHR transformers such as Med-BERT and BEHRT demonstrate how patient event histories can be embedded into reusable representations that support later prediction problems [9, 10]. Other work has extended language-model-style representation learning to electronic health records and generative transformer frameworks for disease outcome prediction, reinforcing the potential of pre-trained models for heterogeneous clinical sequences [11, 20]. For hospital operations, the analogous opportunity is to pre-train on workflow sequences rather than only disease trajectories.

Self-supervised learning for multimodal hospital data

Self-supervised learning is especially relevant for hospital operations because large volumes of timestamped events, logs, messages, and queue states are available even when explicit delay labels are incomplete or inconsistent. Masked event modeling could train the model to reconstruct hidden orders, transfers, transport milestones, or queue changes, while contrastive learning could align patient messages with contemporaneous operational states [11, 21]. Next-event and next-state prediction objectives could encourage the model to anticipate how clinical and operational sequences evolve, similar to temporal modeling strategies used in clinical time-series benchmarks [22]. These objectives would allow the model to learn useful operational structure before being fine-tuned for specific delay tasks.

Prior work on delay and operational risk prediction

Prior work on operational risk prediction shows the value of machine learning for discharge timing, inpatient flow, patient transport, procedural duration, and emergency department antibiotic delay, yet these studies usually address a bounded operational endpoint [4, 6, 8, 23, 24]. Discharge-focused models can support care coordination, while transport and procedural models can help manage movement and operating room workflows [3, 5, 16]. However, these approaches do not typically share a common representation across medication, diagnostic, transport, procedural, messaging, capacity, and discharge workflows. The absence of a unified pre-trained operational model motivates a foundation-model approach that treats delay prediction as a family of related tasks.

Model Development Overview

High-level framework

The proposed framework is a general-purpose operational transformer that would be pre-trained on longitudinal institutional data and then fine-tuned with task-specific heads for different care delivery delay outcomes. The backbone would learn contextual representations of patients, units, staff assignments, queues, capacity states, messages, and care events, drawing on the same pre-training principles used in large-scale EHR representation learning [9, 10, 20]. Downstream heads could formulate medication administration delay, delayed discharge, transport delay, or procedure delay as classification or regression tasks depending on operational need. This design emphasizes representation reuse rather than separate model construction for every hospital workflow.

Core input modalities

The core input modalities would include structured EHR events such as orders and results, operational logs such as bed assignments and transfer events, staff assignments such as shift rosters and unit coverage, patient portal messages represented through text and metadata, facility capacity indicators such as census and occupancy, and service queue states such as radiology backlog or pending consult volume. Prior multimodal healthcare modeling shows that combining structured EHR data with clinical text can improve contextual understanding, while patient message studies show that secure communications contain operationally meaningful information that can be classified and modeled [18, 21, 25].

Hospital flow and discharge studies further indicate that capacity and service state influence delay risk beyond individual patient characteristics [7, 16]. Together, these inputs would represent both the patient journey and the surrounding operational environment.

Design principles

The model should follow four design principles: modality-agnostic input encoding, robustness to missing data, shared representation across tasks, and efficient adaptation to new operational endpoints. A modality-agnostic design would allow orders, logs, messages, staff variables, and queue indicators to enter a shared latent space while retaining modality identity through learned embeddings [13, 26]. Robustness is essential because operational data feeds may be incomplete, delayed, or unavailable during real-time use, and a foundation model should continue to reason from the remaining signals [12, 22]. Efficient fine-tuning would make the backbone reusable for new delay definitions as hospital priorities evolve.

Multimodal Data Sources and Feature Engineering

Ehr events and operational logs

Electronic health record events and operational logs would be extracted as timestamped sequences that describe the progression of care. These could include order placement, medication verification, medication administration, result availability, consult requests, procedure scheduling, transfer orders, bed assignment, bed cleaning, transport request, and department arrival events. EHR foundation-model studies demonstrate that structured clinical events can be represented as temporal tokens, while hospital flow research shows that operational milestones can reveal bottlenecks in inpatient movement and discharge processes [7, 10, 17]. For the proposed model, each event would be encoded with patient context, unit context, event type, time information, and relevant operational metadata.

Staff assignments, patient messages, and capacity indicators

Staff assignments, patient messages, and facility capacity indicators would provide complementary signals about operational load and communication pressure. Time-varying staff counts, nurse-to-patient assignment patterns, and shift context could be combined with census, occupancy, bed status, and unit-level demand to represent the care environment in which clinical work occurs [1, 8]. Patient portal messages could be processed through natural language representations that capture urgency, topic, dissatisfaction, symptom escalation, or care coordination needs, building on studies of secure message classification and patient concern detection [18, 19, 25, 27]. These modalities would help the model identify delays that arise not only from clinical complexity but also from communication burden and resource strain.

Service queue data and temporal alignment

Service queue data would include pending radiology orders, laboratory workload, consult backlogs, transport queues, operating room readiness states, and other departmental demand indicators. Prior transport, procedure-duration, and operating room studies show that queue-like operational signals are central to predicting when a service can be completed and how delays propagate through hospital workflows [3, 5, 15, 24]. To make these signals usable by a transformer, heterogeneous streams would be aligned into patient-centered and unit-centered temporal sequences using event tokens, continuous-time encodings, and summary state tokens. This alignment would allow the model to jointly represent irregular clinical events and regularly updated operational capacity indicators.

Table 1 maps each proposed data modality to its operational meaning, representation strategy, and role in predicting different forms of care-delivery delay.

Table 1. Multimodal Input Architecture and Operational Representation Logic for the Proposed Foundation Model

Input modality	Primary operational meaning	Example raw signals	Recommended representation strategy	Delay mechanisms captured	Downstream prediction relevance
Electronic health record event sequences	Patient-level clinical progression and care-task state	Orders, medication verification, medication administration, laboratory results, consult requests, procedure scheduling, result availability	Event-token embeddings with timestamp, event type, patient context, unit context, and order-state metadata	Delays caused by pending clinical decisions, incomplete care steps, late results, or order accumulation	Medication delay, procedure delay, diagnostic delay, discharge delay, consult delay
Operational movement and workflow logs	Physical and administrative flow through the hospital	Admission-discharge-transfer logs, transfer orders, bed assignment, bed cleaning, transport request, department arrival	Operational event embeddings aligned to patient-centered and unit-centered timelines	Delays caused by transfer bottlenecks, bed turnover constraints, movement dependencies, or departmental handoffs	Transport delay, discharge delay, bed assignment delay, care progression bottleneck
Staff assignment and coverage data	Available labor capacity and workload context	Shift rosters, staff-to-patient ratios, unit coverage, role mix, assignment changes	Numeric and categorical staffing embeddings with shift-time and unit-pressure context	Delays caused by insufficient coverage, workload spikes, skill mismatch, or competing care demands	Medication delay, patient response delay, procedure readiness delay, escalation delay
Patient portal messages and communication metadata	Communication burden, unmet needs, urgency, and escalation risk	Secure messages, message topics, sentiment, urgency indicators, response timestamps, patient concerns	Text embeddings combined with message metadata and temporal proximity to clinical events	Delays caused by unresolved questions, care coordination gaps, dissatisfaction, symptom escalation, or administrative friction	Message escalation delay, follow-up coordination delay, discharge readiness delay, patient navigation bottleneck
Facility capacity indicators	System-level pressure surrounding the patient and unit	Census, occupancy, bed availability, unit crowding, isolation-room availability, pending admissions	Time-varying state tokens and continuous numerical projections	Delays caused by crowding, bed scarcity, downstream congestion, or system-wide demand imbalance	Discharge delay, admission-to-bed delay, transfer delay, care progression bottleneck

Service queue states	Departmental demand and backlog pressure	Pending radiology orders, laboratory workload, transport queue, consult backlog, operating room readiness, pharmacy queue	Queue-state embeddings with temporal summaries and service-specific backlog indicators	Delays caused by queue accumulation, service saturation, resource contention, or delayed downstream completion	Procedure delay, imaging delay, medication verification delay, transport delay, consult delay
Cross-modal temporal alignment layer	Shared temporal structure across patient, unit, and system states	Irregular clinical events plus regularly updated staffing, capacity, message, and queue feeds	Continuous-time encoding, modality-type embedding, patient-level timeline, unit-level context tokens	Interactions among clinical state, operational pressure, staffing, and communication burden	Enables cross-task transfer across all delay endpoints
Missing-modality handling	Realistic inference under incomplete live feeds	Absent messages, delayed staffing updates, unavailable queue feeds, inconsistent timestamps	Modality dropout during training, missingness indicators, feed-quality tokens, fallback inference paths	Degradation caused by missing or unreliable operational data streams	Robust prediction across units, services, and implementation environments

Foundation Model Architecture

Modality-specific input projection

Each input modality would be projected into a shared latent space before being processed by the foundation model. Structured orders, results, medications, transfers, and operational logs would use event embeddings; patient messages would use text embeddings and metadata embeddings; staff, capacity, and queue variables would use numerical projections with temporal context [13, 21]. The approach is consistent with multimodal biomedical artificial intelligence, where different data types are transformed into representations that can interact within a shared model space [13]. Modality-specific projection layers would preserve the distinct structure of each data stream while allowing cross-modal reasoning about delay risk.

Transformer encoder with cross-modal attention

The central architecture would use a transformer encoder that processes multimodal token sequences with modality-type embeddings, time encodings, and patient or unit context embeddings. Transformer-based EHR models have shown that attention mechanisms can represent longitudinal clinical histories, while graph-convolutional and multimodal transformer approaches illustrate how contextual relationships among events and modalities can be learned [10, 21, 26]. In the proposed operational model, cross-modal attention would allow the representation of a pending medication, a staffing change, a patient message, and a service backlog to influence one another. This structure is particularly important because many care delivery delays are caused by interactions among signals rather than by any single isolated feature.

Pre-training objectives

The architecture would be pre-trained using self-supervised objectives designed for hospital operations. Masked event prediction would ask the model to infer hidden clinical or operational events, cross-modal contrastive learning would align contemporaneous messages, queues, and EHR states, and next-state proxy prediction would encourage anticipation of near-future workflow changes [11, 20, 22]. These objectives extend the logic of pre-trained EHR models from disease prediction to operational representation learning [9, 11]. After pre-training, the same backbone could support fine-tuning for delay-specific tasks without requiring every downstream model to relearn the basic structure of hospital workflow.

Figure 1 illustrates the proposed multimodal foundation model architecture for transforming heterogeneous hospital data streams into a reusable operational representation that supports multiple care-delivery delay prediction tasks.

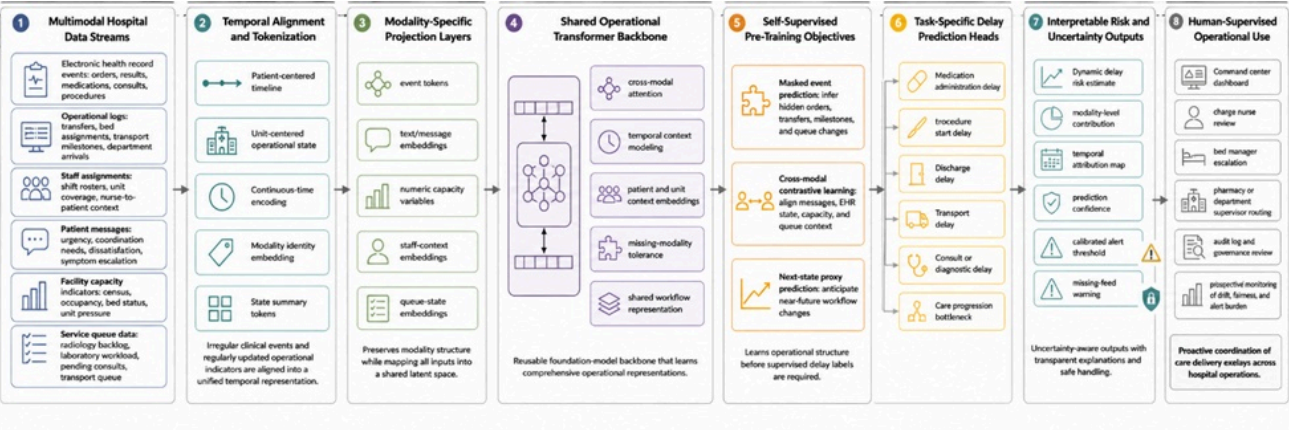


Figure 1. Multimodal Foundation Model Architecture for Predicting Care Delivery Delays across Hospital Operations

Pre-training and Fine-Tuning for Operational Delay Prediction

Pre-training on a large retrospective corpus

The model would be pre-trained on retrospective multimodal hospital sequences spanning clinical events, operational logs, patient communication, staff context, facility capacity, and service queues. The objective would not be to create a single static prediction model, but to learn reusable representations of how care states evolve over time under changing operational conditions. Prior work on EHR foundation models, multitask clinical time-series prediction, and multimodal inpatient length-of-stay modeling supports the premise that temporal pre-training can capture clinically meaningful patterns across structured and unstructured data streams [20, 22, 28]. For operational delay prediction, pre-training would emphasize reconstruction, temporal ordering, and cross-modal consistency rather than supervised delay labels alone.

Fine-tuning on specific delay tasks

For each downstream delay task, the pre-trained backbone would be paired with a lightweight prediction head designed for the relevant endpoint, such as medication administration delay, delayed discharge, transport delay, procedure start delay, or extended length of stay. The shared representation could be partially frozen or selectively adapted, allowing the model to retain general operational knowledge while learning task-specific relationships. This design follows the logic of earlier EHR pre-training and fine-tuning systems, while extending it to operational outcomes that have previously been modeled with narrower task-specific architectures [6, 9, 10, 29]. The goal would be to allow each delay task to benefit from broad institutional context rather than only the variables traditionally selected for that single endpoint.

Multi-task and continual learning capability

A multimodal operational foundation model should support multi-task learning because delay categories are related through shared hospital workflows. A delayed discharge can affect bed availability, bed availability can alter emergency department flow, transport queues can affect imaging completion, and imaging completion can influence clinical decision timing [3, 7, 16]. Multi-task training would therefore allow the backbone to learn common temporal structures while prediction heads specialize in particular operational endpoints. Continual learning would further allow hospitals to add new delay definitions over time, such as new message-triage delays or capacity-related escalation tasks, while preserving the model's core operational representation [12, 20].

Model Interpretability and Trust in Operational Predictions

Explaining delay predictions to operations managers

Operational users need explanations that connect model outputs to recognizable workflow conditions, not only abstract risk scores. Attention-based summaries, temporal attribution maps, and modality-level contribution displays could show whether a predicted delay is being driven by recent order activity, low staffing coverage, patient message urgency, bed occupancy, or service backlog [21, 25, 26]. For example, an explanation might indicate that a projected medication delay is associated with high unit census, increased patient messaging burden, and recent clustering of pending medication orders. Such explanations would be essential for charge nurses, bed managers, and departmental supervisors who must decide whether a prediction is actionable.

Uncertainty quantification and decision support

The model should communicate uncertainty alongside delay risk because operational decisions often involve tradeoffs among competing patients, units, and services. Confidence scores, prediction intervals, and calibrated alert thresholds could help distinguish routine workflow variation from high-impact delay risk requiring escalation. Prior work on hospital discharge prediction, patient transport optimization, and emergency department delay modeling suggests that operational forecasts are most useful when embedded into decision processes rather than treated as standalone outputs [1, 4, 15]. A foundation model should therefore support graded decision support, where high-confidence predictions trigger supervisory review and lower-confidence predictions remain visible as contextual signals.

Operational Deployment and Integration into Hospital Dashboards

Real-time inference architecture

A deployed version of the model would operate as a streaming inference service that consumes live EHR events, operational logs, staffing updates, patient messages, facility capacity feeds, and service queue states. Rather than relying on a single batch prediction, the system would update delay probabilities as new events arrive and as the operational context changes. Prior hospital flow, patient transport, and real-time tracking studies show that operational prediction becomes more actionable when aligned with live workflow data and system-state monitoring [3, 7, 23]. The inference architecture would therefore need reliable interfaces to clinical and operational systems, with careful monitoring for missing feeds, delayed updates, and temporal misalignment.

Integration with command center and charge nurse workflows

Predictions should be integrated into existing operational dashboards and role-specific workflows rather than introduced as an isolated artificial intelligence tool. A delay risk involving bed turnover might be routed to a bed manager, a medication delay risk to a charge nurse or pharmacy supervisor, and a procedure delay risk to a departmental coordinator. Studies of discharge prediction and inpatient flow highlight that predictive analytics can support multidisciplinary rounds and operational planning when outputs are embedded into the routines of the people who

coordinate care [6, 8, 16]. Explanation views, escalation pathways, and feedback mechanisms would help users interpret the model and correct predictions when the operational context is not fully captured.

Evaluation Strategy

Predictive performance across delay tasks

The evaluation strategy should compare the multimodal foundation model against task-specific baselines for each delay endpoint, while avoiding the assumption that a single metric fully captures operational value. Conceptually, the model should be assessed for medication delay classification, discharge delay prediction, transport delay forecasting, procedure timing estimation, and patient-message-related operational escalation using appropriate task-specific measures. Prior studies on discharge prediction, cardiovascular discharge usability, multimodal length-of-stay prediction, procedural case duration, and operating room time estimation provide examples of task-specific evaluation designs that could inform such comparisons [5, 14, 24, 28, 29]. The key question would be whether a shared pre-trained representation improves generalizability and operational usefulness relative to separately developed models.

Robustness to missing modalities and data drift

Robustness evaluation should examine whether the model remains useful when specific modalities are unavailable, delayed, sparse, or changing over time. Patient messages may be absent for some patients, staffing feeds may differ across units, operational logs may use inconsistent timestamps, and queue definitions may change as services redesign workflows [18, 19, 27]. Prior critiques of foundation models for EHRs emphasize that large pre-trained systems require careful validation, because representational power does not guarantee reliability under distribution shift or imperfect data capture [12]. Evaluation should therefore include missing-modality analysis, temporal drift monitoring, subgroup assessment, and prospective checks for changes in clinical and operational practice.

Prospective pilot and operational impact

A prospective pilot should begin in silent mode, where the model generates predictions without changing workflow, allowing investigators to assess calibration, alert burden, and alignment with real operational events. After sufficient review, the model could be introduced into selected command-center or charge-nurse workflows with careful governance, human oversight, and documentation of unintended consequences. Prior work on machine-learning-supported discharge processes, inpatient flow prediction, and transport optimization suggests that operational impact depends on whether predictions are timely, understandable, and connected to feasible interventions [6, 8, 15]. The pilot should therefore evaluate not only predictive behavior but also staff trust, alert fatigue, escalation appropriateness, and whether the system supports safer and more efficient care progression.

Table 2 provides an evaluation and deployment-readiness framework for determining whether the proposed model is predictive, interpretable, robust, fair, and operationally safe enough for prospective hospital use.

Table 2. Evaluation, Interpretability, and Deployment Readiness Framework for a Multimodal Operational Foundation Model

Evaluation domain	Core question	Recommended assessment approach	Model-specific indicators	Operational readiness interpretation	Governance risk if neglected
Cross-task predictive performance	Does the shared backbone improve prediction across multiple delay	Compare against separate baseline models for medication,	AUROC or AUPRC for classification tasks; MAE/RMSE for timing	Supports the claim that a reusable foundation model	A complex model may be adopted without evidence that shared

	types compared with task-specific models?	discharge, procedure, transport, consult, and care progression delay tasks	estimates; calibration slope; task-specific sensitivity at actionable thresholds	adds value beyond isolated predictive tools	representation improves operational usefulness
Transfer learning value	Does pre-training reduce the amount of labeled delay data needed for downstream tasks?	Fine-tune prediction heads using different labeled-data fractions and compare with models trained from scratch	Performance under low-label settings; convergence speed; representation reuse across units and services	Demonstrates whether the backbone is scalable for new delay definitions	Hospitals may invest in pre-training without meaningful benefit for emerging operational endpoints
Missing-modality robustness	Does the model remain useful when live feeds are absent, delayed, or sparse?	Conduct modality ablation, feed-delay simulation, missingness stress testing, and unit-level data-quality analysis	Performance drop by removed modality; missing-feed warning accuracy; stability under delayed queue, staffing, or message data	Indicates whether the model can function under realistic hospital data conditions	Predictions may fail silently when staffing feeds, message streams, or queue data are incomplete
Temporal drift and workflow change	Does performance remain stable as hospital workflows, staffing patterns, and queue definitions change?	Monitor calibration and error patterns across time periods, service changes, seasonal demand, and policy redesigns	Drift metrics, calibration decay, alert-volume change, endpoint-definition instability	Determines whether the model requires retraining, recalibration, or workflow-specific adaptation	A model may become unreliable after operational redesign, staffing change, or documentation changes
Interpretability for operational users	Can managers understand why a delay forecast was generated?	Evaluate modality-level contributions, temporal attribution summaries, attention-based explanations, and user-facing explanation displays	Contribution ranking by modality; explanation concordance with user judgment; explanation readability	Supports actionable use by charge nurses, bed managers, pharmacy supervisors, and command-center teams	Users may distrust or ignore predictions that cannot be linked to recognizable workflow causes
Uncertainty and alert calibration	Does the model distinguish high-confidence escalation risks from routine workflow variation?	Assess prediction intervals, confidence scores, threshold calibration, alert burden, and false escalation rates	Calibration curves; alert precision; alert-to-action ratio; uncertainty-stratified error	Allows graded decision support rather than indiscriminate alerting	Poorly calibrated alerts may create alarm fatigue, missed escalation, or inefficient resource allocation
Equity and workforce	Are predictions or escalations	Conduct subgroup assessment by unit,	Error disparity, alert-rate disparity,	Ensures that operational AI	The model may reinforce

fairness	unevenly distributed across patient groups, units, staff roles, or services?	service, patient demographics where appropriate, message availability, staffing context, and access patterns	missed-delay disparity, escalation burden by group or unit	does not worsen inequity or unfairly burden particular teams	documentation bias, staffing surveillance concerns, or unequal operational attention
Prospective silent-mode validation	Do predictions correspond to real operational delays before workflow intervention?	Run the model without changing workflow and compare predictions with observed delays, escalation events, and manager review	Real-time calibration; lead time before delay; silent-mode alert burden; agreement with operational review	Establishes safety before active deployment	Direct deployment without silent validation may disrupt care coordination or create unsafe reliance
Human-supervised deployment impact	Does the model improve coordination when embedded into real workflows?	Pilot role-specific routing for charge nurses, bed managers, pharmacy supervisors, transport coordinators, and command centers	Time-to-escalation, delay reduction, staff trust, overridden alert rate, documentation of action taken	Measures whether predictions are timely, interpretable, and connected to feasible interventions	A technically accurate model may fail if alerts are not integrated into accountable operational workflows
Governance, privacy, and auditability	Are sensitive operational and communication data used responsibly?	Review access controls, patient-message handling, staff-data governance, model audit trails, and secondary-use policies	Audit completeness; restricted-access compliance; documentation of model updates; governance review frequency	Confirms institutional readiness for responsible operational AI	Privacy concerns, workforce surveillance risks, and unclear accountability may undermine adoption

Limitations

Data integration and governance

The central limitation of the proposed approach is the difficulty of assembling a linked multimodal corpus across electronic health records, operational systems, staff assignment tools, patient messaging platforms, facility capacity feeds, and service queue databases. These systems often differ in ownership, timestamp conventions, data quality, update frequency, and governance rules, making integration a major institutional challenge. Patient messages and staffing data also raise privacy, labor, and security concerns because they may contain sensitive communication content or information about individual work patterns [18, 19, 27]. A responsible implementation would require strict access controls, de-identification where appropriate, auditing, stakeholder engagement, and explicit policies governing secondary use of operational data [12, 13].

Computational costs and inference latency

Foundation models can impose substantial computational demands during pre-training, fine-tuning, monitoring, and real-time inference. Even if a model is conceptually attractive, smaller hospitals may lack the hardware, cloud resources, data engineering infrastructure, or machine learning operations capacity required to deploy and maintain it safely. Prior discussions of large healthcare models and multimodal biomedical artificial intelligence caution that model scale must be balanced against reliability, transparency, implementation feasibility, and institutional readiness [12, 13]. Practical versions of the proposed system may therefore require compressed backbones, parameter-efficient fine-tuning, local caching, latency-aware serving, and simplified deployment options for resource-constrained environments.

Conclusion

A multimodal foundation model for predicting care delivery delays would treat hospital operations as a dynamic system rather than as a collection of isolated prediction tasks. By integrating electronic health record events, operational logs, staff assignments, patient messages, facility capacity indicators, and service queue data, the model could represent the clinical and operational context in which delays arise. This approach would shift delay prediction from narrow endpoint modeling toward a shared representation of care progression risk.

The key strength of the proposed model is its reusable pre-trained backbone. A single operational representation could support multiple downstream prediction heads, enabling fine-tuning for medication delays, procedure delays, discharge delays, transport delays, and other workflow bottlenecks. The architecture would also be designed for missing-modality robustness and interpretable outputs, allowing users to understand why a delay forecast was generated.

Important challenges remain before such a model could be responsibly used in hospital operations. Data integration would require coordination across clinical, operational, staffing, messaging, and capacity systems, and governance would need to address privacy, security, fairness, and workforce concerns. Computational requirements and prospective validation would also determine whether the model is feasible, trustworthy, and beneficial in real-world settings.

Future work should focus on collaborative development of open operational foundation models trained across multiple hospitals and evaluated under prospective conditions. Such work should emphasize transparent reporting, careful human oversight, and operational endpoints that matter to patients and staff. A foundation model for care delivery delays should ultimately be judged not by technical sophistication alone, but by whether it helps hospitals anticipate bottlenecks and coordinate care more effectively.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Jul 2025

Revised: 12 Sep 2025

Accepted: 06 Nov 2025

Published online: 25 February 2026

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

- Pahlevani M, Rajabi E, Taghavi M, VanBerkel P. Developing a decision support tool to predict delayed discharge from hospitals using machine learning. *BMC Health Serv Res.* 2025;25(1):56.
- Keller DS, Tantchou I, Flores-Gonzalez JR, Geisler DP. Predicting delayed discharge in a multimodal enhanced recovery pathway. *Am J Surg.* 2017;214(4):604–9.
- Kropp T, Faeghi S, Lennerts K. Evaluation of patient transport service in hospitals using process mining methods: Patients' perspective. *Int J Health Plann Manage.* 2023;38(2):430–56.
- Seo J, Park S, Cha WC, Kim T, Shin SY, Jung KY, et al. Predicting delays in antibiotic administration in the emergency department: a machine learning approach incorporating nursing workload and crowding factors. *Signa Vitae.* 2026;22(1).
- Kendale S, Bishara A, Burns M, Solomon S, Corriere M, Mathis M. Machine learning for the prediction of procedural case durations developed using a large multicenter database: Algorithm development and validation study. *JMIR AI.* 2023;2(1):e44909.
- Safavi KC, Khaniyev T, Copenhaver M, Seelen M, Zenteno Langle AC, Zanger J, et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Netw Open.* 2019;2(12):e1917221.
- Bertsimas D, Pauphilet J, Stevens J, Tandon M. Predicting inpatient flow at a major hospital using interpretable analytics. *Manuf Serv Oper Manag.* 2022;24(6):2809–24.
- 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).
- Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* 2021;4(1):86.
- Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for electronic health records. *Sci Rep.* 2020;10(1):7155.
- Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. *J Biomed Inform.* 2021;113:103637.
- Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med.* 2023;6(1):135.
- Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med.* 2022;28(9):1773–84.

Ghazalbash S, Zargoush M, Mowbray F, Papaioannou A. Examining the predictability and prognostication of multimorbidity among older delayed-discharge patients: A machine learning analytics. *Int J Med Inform.* 2021;156:104597.

Kropp T, Gao Y, Lennerts K. Data-based optimisation of intra-hospital patient transport capacity planning. *OR Spectr.* 2024;46:1–54.

Wei J, Zhou J, Zhang Z, Yuan K, Gu Q, Luk A, et al. Predicting individual patient and hospital-level discharge using machine learning. *Commun Med (Lond).* 2024;4(1):236.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1(1):18.

Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform.* 2017;105:110–20.

De A, Huang M, Feng T, Yue X, Yao L. Analyzing patient secure messages using a Fast Healthcare Interoperability Resources (FHIR)-based data model: Development and topic modeling study. *J Med Internet Res.* 2021;23(7):e26770.

Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun.* 2023;14(1):7857.

Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. *AMIA Annu Symp Proc.* 2022;2022:719–28.

Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data.* 2019;6(1):96.

Huang SW, Chiou SY, Chen RC, Sub-R-Pa C. Enhancing hospital efficiency and patient care: Real-time tracking and data-driven dispatch in patient transport. *Sensors.* 2024;24(12):4020.

Chu J, Hsieh CH, Shih YN, Wu CC, Singaravelan A, Hung LP, et al. Operating room usage time estimation with machine learning models. *Healthcare.* 2022;10(8):1518.

Ren Y, Wu Y, Fan JW, Khurana A, Fu S, Wu D, et al. Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. *J Am Med Inform Assoc.* 2024;31(8):1714–24.

Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. *Proc AAAI Conf Artif Intell.* 2020;34(1):606–13.

Liu S, Wright AP, McCoy AB, Huang SS, Steitz B, Wright A. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *J Am Med Inform Assoc.* 2025;32(6):1032–9.

Chen J, Wen Y, Pokojovy M, Tseng TL, McCaffrey P, Vo A, et al. Multi-modal learning for inpatient length of stay prediction. *Comput Biol Med.* 2024;171:108121.

Ahn I, Gwon H, Kang H, Kim Y, Seo H, Choi H, et al. Machine learning-based hospital discharge prediction for patients with cardiovascular diseases: Development and usability study. *JMIR Med Inform.* 2021;9(11):e32662.