

ORIGINAL RESEARCH

Open access

Temporal Graph Transformer for Predicting Interdepartmental Workflow Congestion Using Patient Movement Data, Service Requests, Staff Schedules, Unit-Level Dependency Networks, and Real-Time Capacity Signals

Elif Yilmaz^{1*}, Mehmet Demir¹, Ayse Kaya², Hasan Aydin¹

Abstract

Hospital congestion is an interconnected operational phenomenon in which local delays can cascade across emergency, inpatient, diagnostic, pharmacy, and consultative services. Prediction has often been organized around individual departments rather than the hospital as a coupled system. Current forecasting approaches do not sufficiently represent dynamic dependencies among units, including patient transfers, service requests, staffing availability, and real-time capacity. This limits their ability to anticipate how pressure in one department could propagate into another. This article proposes a temporal graph transformer for forecasting interdepartmental workflow congestion across a hospital network. The model is designed to learn from time-varying departmental relationships and produce forward-looking congestion estimates for operational decision support. The proposed framework represents hospital units as graph nodes annotated with capacity, queue, and staffing signals. Directed edges are weighted by patient movement and service request dependencies, and a temporal graph transformer processes graph sequences to estimate future congestion probabilities for each unit. Conceptually, the model would be expected to identify building pressure before it appears as visible delay. Such forecasts could support proactive coordination, including staff redeployment, transfer prioritization, and service sequencing. A graph-based temporal perspective could shift hospital congestion management from reactive firefighting toward coordinated, predictive flow control. The proposed model provides a conceptual foundation for future prospective evaluation in real-time operational settings.

Keywords Hospital operations, Graph neural networks, Patient flow, Temporal graph transformer, Workflow congestion, Capacity forecasting

*Correspondence:

Elif Yilmaz

elif.yilmaz@gmail.com

¹ Department of Healthcare Information Systems, Faculty of Medicine, Istanbul Technical University, Istanbul, Turkey

² Department of Clinical Informatics and AI Analytics, Faculty of Engineering, Middle East Technical University, Ankara, Turkey

Introduction

Interdepartmental workflow congestion imposes clinical and operational costs because delays rarely remain confined to a single unit. An emergency department boarding problem may reflect unavailable inpatient beds, delayed imaging interpretation, limited transport capacity, or downstream staffing constraints rather than only local emergency demand. Reviews of machine learning for patient flow emphasize that hospital admission and throughput forecasting should be understood as part of a broader operational system rather than as isolated prediction tasks [1, 2]. A model for congestion

prediction should therefore represent not only patient arrivals and queues but also the dependencies through which bottlenecks cascade across hospital services [3, 4].

Traditional bottleneck prediction has often focused on forecasting demand or admission risk at one point of care, especially the emergency department. Such models can support triage and admission decisions, but they may miss the network effects through which an apparent emergency department wait is actually produced by inpatient bed scarcity, diagnostic backlog, or delayed consult completion [5-7]. Graph-based representations are well suited to this setting because they can encode units as nodes and operational relationships as edges, allowing congestion to be treated as a system-level process rather than a sequence of disconnected local events [8, 9]. This distinction is central for hospital operations because a locally accurate model may still provide limited guidance when the true cause of delay lies upstream or downstream.

Temporal graph models have shown promise in other dynamic systems where flows, interactions, and dependencies evolve over time. Temporal graph networks, dynamic self-attention models, evolving graph convolutional networks, and temporal graph transformers have been developed to represent changing relationships and timestamped interactions in social, infrastructure, and transactional systems [10-17]. In transportation forecasting, spatio-temporal graph models have demonstrated how network topology and time-varying demand can be jointly modeled to anticipate congestion-like phenomena [18-23]. Hospital operations share important structural similarities with these domains, but the graph is defined by patient transfers, service dependencies, staffing constraints, and real-time capacity rather than road segments or social interactions.

The thesis of this article is that a temporal graph transformer could represent the hospital as a dynamic network and forecast how congestion may propagate across departments. In this framework, nodes correspond to units such as emergency care, imaging, laboratories, pharmacy, intensive care, wards, and procedural areas, while directed edges represent patient movement and service-request dependencies. Transformer-style temporal attention could allow the model to learn which recent events, upstream units, and changing capacities are most relevant to a predicted bottleneck. The result would be a conceptual model for predictive flow control, designed for prospective evaluation rather than for reporting retrospective experimental performance.

Background

Hospital congestion as a network phenomenon

Hospital congestion emerges when patient movement, service requests, staffing availability, and capacity constraints interact across units. A backlog in imaging may delay emergency department disposition, which in turn may increase boarding, reduce treatment-space availability, and slow subsequent admissions. Operations-focused studies have emphasized that healthcare delivery depends on coordination across departments, and data-driven network analysis can reveal how care processes are linked across organizational boundaries [3, 24]. This makes congestion a network phenomenon in which prediction should account for both local workload and interdepartmental dependency structure [4, 9].

Temporal graph neural networks

Temporal graph neural networks are designed for systems in which entities and relationships evolve over time, making them relevant for hospital workflows with timestamped transfers, orders, and capacity updates. Temporal graph networks encode event histories, dynamic embeddings, and interaction memory, while temporal attention and transformer approaches can represent changing dependencies across graph snapshots or event streams [10-13]. Evolving graph convolutional networks and temporal graph transformers extend this idea by allowing node representations and attention patterns to adapt as the graph changes [14-17]. These methods provide a conceptual basis for modeling how operational pressure accumulates, diffuses, and resolves across hospital units.

Patient movement and unit-level dependency data

Patient movement data can be derived from admission, discharge, and transfer events, location systems, transport logs, or other operational records that indicate where patients are located and when they move between units. These movements define one class of directed edges in a hospital graph, while order flows and consult requests define additional dependency edges between clinical and support services. Data-driven network analysis of hospital departments illustrates how care delivery can be represented through interdepartmental coordination patterns rather than only through individual encounters [24]. Graph-based healthcare representation learning further supports the idea that clinical and operational relationships can be encoded as structured dependencies for downstream prediction [8].

Service requests, staffing, and capacity signals

Service requests, staff rosters, occupancy feeds, and queue indicators provide operational context for interpreting the state of each hospital unit. A unit with moderate patient volume but low staffed capacity may be more vulnerable to congestion than a unit with higher volume and adequate staffing, so the model should include staffing and capacity signals as node attributes. Machine learning approaches to admission and patient flow prediction have shown the value of operational features, but these features are often used in tabular or unit-specific models rather than as part of a dynamic hospital graph [1, 2, 6, 7]. Integrating these streams into node-level features would allow the temporal graph transformer to learn how local capacity modifies the effect of upstream demand.

Prior work in hospital flow prediction

Prior work in hospital flow prediction includes machine learning models for emergency admission, simulation approaches for capacity planning, and operational analytics frameworks for managing healthcare delivery. Emergency department admission models demonstrate that machine learning could support early prediction, while simulation and hybrid modeling approaches show how patient flow interventions can be explored conceptually before deployment [2, 5-7, 25, 26]. However, many prior approaches do not explicitly encode evolving interdepartmental topology, which limits their ability to capture congestion propagation across units. Graph network techniques for emergency department patient flow begin to address this gap, but broader temporal graph models remain under-developed for whole-hospital operational forecasting [9].

Model Development Overview

High-level predictive framework

The proposed framework represents the hospital as a dynamic graph that is updated at regular operational intervals using recent movement, request, staffing, and capacity signals. At each interval, the temporal graph transformer processes the current graph together with prior graph states and produces a forecast of likely congestion for each department in the next operational horizon. This design draws from temporal graph networks and dynamic graph representation learning, where current predictions depend on the evolving history of node states and interactions [10-14]. In hospital operations, such a framework would be expected to support anticipatory coordination rather than merely describe congestion after it has already become visible [1, 3].

Core graph elements

The core graph elements include nodes, node features, directed edges, edge types, and temporal weights. Nodes represent individual departments or units, node features include current bed occupancy, pending service requests, staffed beds, queue length, and average patient length of stay, and edges are weighted by recent patient transfers or service dependencies with temporal decay. This mirrors graph-based learning approaches in which structured relationships are central to representation learning, while adapting traffic-forecasting logic to a hospital dependency network [8, 18-21]. The

graph should remain flexible enough to include heterogeneous edge types, because a transfer from the emergency department to an inpatient ward is operationally different from a laboratory order or imaging request.

Design principles

The model is designed to be real-time capable, sensitive to network topology, probabilistic in its outputs, and interpretable for operations managers. Real-time capability is needed because capacity and staffing conditions change throughout the day, while topology sensitivity is needed because a bottleneck's effect depends on the units connected to it. Probabilistic output is important because congestion forecasts should support risk-aware escalation rather than deterministic claims about future operations [2, 7]. Interpretability is also essential because operational leaders need to understand which upstream departments, service lines, or capacity constraints are contributing to a predicted bottleneck [27].

Data Sources and Graph Construction

Patient movement and service request data

Patient movement data would be extracted from admission, discharge, transfer, and location records, while service request data would be extracted from order timestamps for laboratory, imaging, pharmacy, transport, and consult workflows. These streams define directed edges between units, with edge weights reflecting recent movement frequency, service demand, or dependency strength. Prior work on healthcare operations management and patient flow emphasizes that operational prediction depends on timely, process-level information rather than only static patient characteristics [1, 3, 4]. In the proposed graph, movement and request records therefore function as the relational backbone through which congestion propagation can be learned.

Unit-level capacity and staff signals

Unit-level capacity and staff signals would serve as node attributes that describe whether each department can absorb incoming demand. Such attributes may include occupied beds, available staffed beds, active queues, scheduled staffing, shift-change indicators, and service-specific workload measures. Machine learning models for hospital admission and real-time aggregated prediction demonstrate the practical value of operational state variables, even when they are not embedded in a graph structure [2, 6, 7]. Incorporating these variables into graph nodes would allow the model to distinguish between high-volume units that are stable and lower-volume units that are vulnerable because of limited capacity.

Building the temporal dependency graph

The temporal dependency graph would be constructed by aggregating recent patient movements and service volumes over sliding operational windows, producing a weighted and directed graph that updates as hospital conditions change. Edges may decay over time so that recent transfers and service requests exert stronger influence than older interactions, which is consistent with dynamic graph learning approaches that emphasize evolving interaction histories [10-12]. This construction is also conceptually aligned with spatio-temporal traffic models, where current congestion predictions depend on both recent flow and network structure [18-23]. In the hospital setting, the resulting graph would represent a living map of operational dependency rather than a fixed organizational chart.

Table 1 defines how operational hospital data streams can be translated into graph components that preserve the distinction between local workload, interdepartmental dependency, temporal pressure, and data reliability.

Table 1. Graph Construction Logic for Modeling Interdepartmental Workflow Congestion

Graph Component	Operational Meaning in the Hospital Network	Example Data Elements	Temporal Representation	Congestion-Relevance Logic	Implementation Risk if Poorly Specified
Department node	A functional hospital unit capable of accumulating or transmitting workflow pressure	Emergency department, imaging, laboratory, pharmacy, ICU, inpatient ward, transport, procedural unit	Updated at each operational interval as a graph node with current state features	Identifies where congestion may emerge, absorb demand, or transmit downstream pressure	Overly broad nodes may hide bottlenecks; overly granular nodes may create unstable sparse graphs
Node capacity state	The unit's ability to absorb additional work at the current time	Occupied beds, staffed beds, physical beds, active treatment spaces, queue length, pending tasks	Time-varying node attributes refreshed from bed boards, staffing systems, and queue feeds	Distinguishes high demand from true congestion by comparing workload against available capacity	Missing staffed-capacity data may cause the model to underestimate operational strain
Node workload intensity	Current service burden within a department	Active orders, patient census, pending verifications, imaging queue, lab queue, consult queue	Aggregated within sliding windows and carried forward with staleness indicators when feeds are delayed	Captures local pressure before it becomes visible as delay or boarding	Inconsistent refresh rates may create misleading comparisons across units
Directed patient-movement edge	A flow relationship showing movement of patients from one department to another	ED-to-ward transfers, ICU step-down transfers, post-procedure recovery transfers, transport events	Weighted by recent movement frequency, direction, and temporal decay	Indicates how pressure can propagate from sending units to receiving units	Treating movement as undirected may obscure upstream and downstream congestion pathways
Directed service-request edge	A dependency created when one unit generates work for another	Imaging orders, laboratory requests, pharmacy verification requests, consult requests, transport requests	Weighted by request volume, urgency, pending status, and elapsed time	Represents hidden operational coupling even when patients do not physically move	Excluding non-transfer dependencies may miss diagnostic, pharmacy, and consult-driven bottlenecks
Edge type	The operational channel through which one	Transfer, imaging, laboratory,	Encoded as heterogeneous edge	Allows the model to learn that different workflow	Collapsing edge types may cause clinically distinct

	department affects another	pharmacy, consult, transport, procedural scheduling	categories or type-specific embeddings	dependencies have different congestion effects	dependencies to be treated as equivalent
Temporal edge weight	The recent strength of dependency between two units	Number of transfers in past hour, imaging requests over 30 minutes, pharmacy orders awaiting verification	Updated using sliding windows, exponential decay, or recency-weighted aggregation	Emphasizes recent pressure that is more likely to affect near-term congestion	Long windows may dilute acute surges; short windows may overreact to noise
Missingness and staleness marker	A signal that a data feed is incomplete, delayed, or stale	Time since last capacity update, missing staffing feed, delayed location event, unavailable queue value	Encoded as explicit node or edge features rather than silently imputed	Helps the model learn whether unreliable data streams are themselves associated with operational risk	Silent imputation can create false stability and reduce trust in real-time forecasts
Forecast target label	The future congestion state the model is trained or evaluated to predict	Congestion onset, queue threshold breach, boarding escalation, delayed transfer, service backlog	Defined over a pre-specified future horizon such as 30, 60, or 120 minutes	Anchors prediction to actionable operational windows rather than retrospective description	Poor target definition may produce statistically accurate but operationally unusable forecasts
Governance metadata	Context needed to audit graph construction and model behavior	Source system, timestamp quality, mapping rules, edge definition, feature refresh frequency	Stored with graph snapshots and prediction outputs	Enables monitoring of data quality, graph drift, and accountability during deployment	Weak governance may make forecasts difficult to validate, explain, or safely operationalize

Temporal Graph Transformer Architecture

Input representation

At each time step, the model receives a heterogeneous hospital graph composed of node features, edge weights, edge types, and temporal indicators. Node features describe unit-level operational status, while the adjacency structure encodes directed transfer and service-request dependencies among departments. Edge types may distinguish patient transfers from imaging orders, laboratory orders, pharmacy requests, transport dependencies, and consult relationships, allowing the model to treat different workflow channels differently. This input design builds on healthcare graph representation learning and dynamic graph models that use structured dependencies to produce time-sensitive embeddings [8-11].

Temporal graph encoding

The temporal graph encoder would use graph attention or transformer layers to compute node embeddings that capture both local dependencies and the broader hospital state. Across time, temporal attention would integrate a sequence of graph snapshots so that the model could learn whether a current congestion risk reflects immediate demand, repeated upstream pressure, or a delayed downstream blockage. Dynamic self-attention, temporal graph transformer, and evolving graph neural network methods provide conceptual precedents for modeling changing relationships with attention-based mechanisms [13-17]. Transportation forecasting models further support the value of combining spatial graph structure with temporal sequence learning when predicting congestion-like dynamics [19-22].

Congestion prediction head

For each node, the final temporal graph embedding would be passed to a prediction head that estimates congestion probability or ordinal congestion severity for the relevant future horizon. The output should be framed probabilistically so that operations teams can interpret forecasts as risk signals requiring judgment rather than as deterministic claims. This design is consistent with hospital admission and patient-flow prediction work, where model outputs are most useful when they support prioritization, escalation, and planning under uncertainty [2, 5-7]. Multi-step forecasting could also be considered conceptually, but it should be evaluated prospectively before being used to guide operational interventions.

Figure 1 illustrates the proposed temporal graph transformer architecture for converting hospital movement, service-request, staffing, dependency, and capacity signals into interpretable unit-level congestion forecasts for command-center decision support.

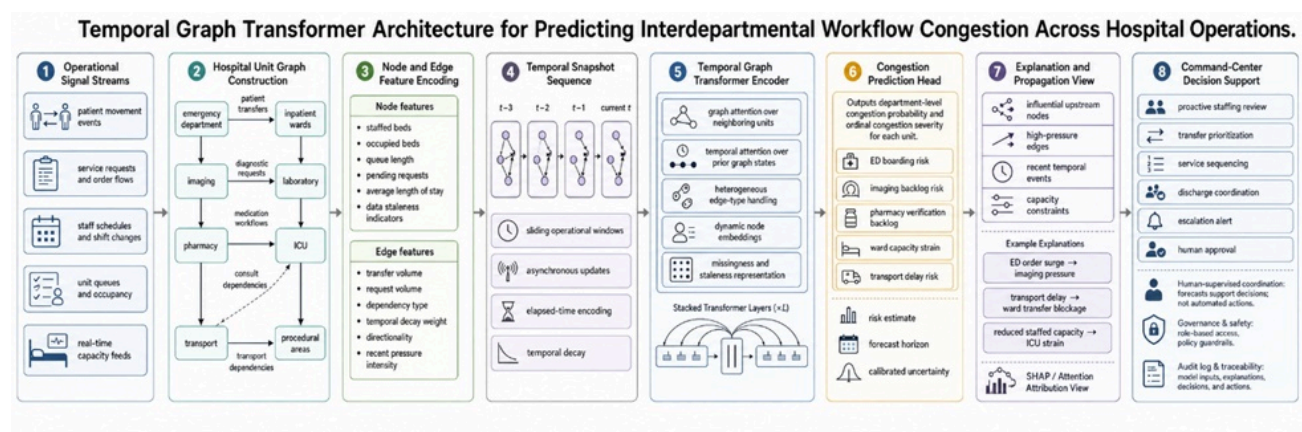


Figure 1. Temporal Graph Transformer Architecture for Predicting Interdepartmental Workflow Congestion Across Hospital Operations

Handling Missing Data and Asynchronous Updates

Missing or irregular feeds

Hospital operational feeds are often incomplete, delayed, or refreshed at different rates, so the model should represent missingness as an informative signal rather than only as a preprocessing defect. For a node with stale capacity data, a last-observation-carried-forward value could be paired with a staleness indicator so that the temporal graph transformer can learn whether uncertainty in the feed itself is associated with congestion risk. Temporal graph networks and dynamic embedding models provide a useful conceptual basis because they represent evolving histories rather than assuming that all observations arrive in synchronized tables [10-12]. In hospital operations, this approach would be expected to make the model more robust when bed boards, staff schedules, service queues, or location systems update unevenly [3, 4].

Asynchronous events and timestamp alignment

Patient moves, diagnostic orders, medication requests, consults, and staffing changes occur as irregular events rather than as neatly aligned observations. To use these signals in a temporal graph transformer, the events could be binned into graph snapshots while preserving timing information through positional encodings, elapsed-time features, or temporal decay weights. Temporal graph attention models and temporal graph transformers are relevant because they are designed to learn from changing interactions and time-aware dependencies rather than fixed static graphs [13, 15-17]. This is particularly important for hospital congestion, where a rapid cluster of imaging requests or transfers may carry different operational meaning than the same number of events distributed over a longer interval [2, 9].

Resilience to staff schedule changes

Staff schedule changes can alter the effective capacity of a department even when physical beds, rooms, or equipment remain unchanged. The proposed model would include shift-change indicators, scheduled staffing levels, and available staffed-capacity measures as node features so that abrupt staffing changes can be represented directly in the graph state. This design follows the broader logic of operational machine learning, where prediction should incorporate system context rather than only patient-level characteristics [1, 2, 6, 7]. A temporal graph model would be expected to adapt its congestion forecast when a unit's staffing state changes, while prospective evaluation should determine whether such adaptation improves operational usefulness.

Model Interpretability for Operational Decision-Makers

Attention-based explanation of congestion propagation

Attention-based explanation could help operations managers understand which upstream departments, service dependencies, or recent graph events contributed most strongly to a predicted congestion state. For example, a predicted imaging bottleneck might be traced to emergency department order volume, inpatient procedural demand, transport delays, or reduced staffed capacity, depending on which nodes and edges receive the greatest model attention. Explainability research in graph neural networks emphasizes that explanations should identify influential nodes, edges, and subgraphs rather than merely assign importance to isolated variables [27]. In this model, attention should therefore be treated as a decision-support aid that requires validation against clinical and operational expertise, not as a complete causal explanation.

From prediction to actionable mitigations

The purpose of interpretability is not only to explain a forecast but also to connect that forecast to feasible operational mitigations. If the model predicts congestion in a downstream ward and attributes the risk to emergency department boarding and delayed transport, the dashboard could suggest review of discharge readiness, transfer prioritization, or temporary transport escalation. Prior work on healthcare operations management and patient-flow simulation supports the idea that predictive models should be linked to operational policies and intervention testing rather than treated as isolated analytic products [3, 26]. Attention-guided root-cause analysis would therefore be expected to help staff distinguish between local overload and network-driven congestion that requires coordinated action.

Integration Into Command Center Dashboards

Real-time congestion map

A command center dashboard could display the hospital as a live dependency graph, with departments represented as nodes and predicted flow disruptions represented as directed edges. Node color or intensity could represent forecast congestion severity, while edge thickness could represent recent patient-transfer or service-request pressure. Health information system research emphasizes that patient-flow management depends on integrating operational information into usable, timely displays for decision-makers [4]. A graph-based map would extend this idea by showing not only which

unit is predicted to become congested, but also how that congestion may be connected to upstream and downstream departments [9, 24].

Alerting and escalation

When predicted congestion crosses a locally defined threshold, the model output could trigger an alert to bed management, service-line leadership, or the hospital command center. Such alerts should include an explanation view showing the main contributing nodes, edges, and recent temporal signals so that staff can judge whether escalation is warranted. Real-time aggregated prediction work in hospital admission demonstrates the potential value of early warning signals when they are embedded into operational workflows rather than left as standalone scores [2]. The alerting layer should therefore be evaluated for usability, calibration, and workflow fit before being used to guide time-sensitive operational decisions [27].

Evaluation strategy

Table 2 outlines a deployment-readiness evaluation framework that separates predictive performance, graph-specific added value, temporal robustness, interpretability, operational usefulness, and governance safeguards.

Table 2. Evaluation and Deployment Readiness Framework for a Temporal Graph Transformer in Hospital Command-Center Operations

Evaluation Domain	Core Question	Recommended Assessment Approach	Comparator or Reference Standard	Decision-Relevance for Hospital Operations	Deployment Readiness Criterion
Node-level predictive accuracy	Can the model forecast congestion for each department before visible delay occurs?	Evaluate discrimination, calibration, sensitivity to congestion onset, and forecast horizon performance by unit	Unit-specific time-series models, tabular machine learning models, historical baseline rates	Determines whether forecasts are reliable enough for department-specific monitoring	Model performance is stable across high-volume and low-volume units, not only the emergency department
System-level congestion propagation	Does the graph model capture how pressure moves across departments?	Evaluate prediction of downstream congestion after upstream surges, edge-pressure events, and transfer blockages	Non-graph models using only local unit histories	Tests whether explicit graph structure adds value beyond local workload trends	Graph model improves prediction during cascade-prone episodes such as ED boarding, imaging backlog, or ward saturation
Added value of edge structure	Do patient-movement and service-request edges improve prediction?	Conduct ablation studies removing edge weights, edge direction, edge types, or temporal decay	Full temporal graph transformer versus reduced graph or non-graph variants	Shows whether the relational representation is necessary rather than decorative	Removing graph components meaningfully degrades calibration, early warning, or propagation detection

Temporal robustness	Does the model remain valid across changing demand and staffing patterns?	Perform temporal validation across seasons, weekdays, weekends, holidays, shift changes, and policy changes	Earlier-period training versus later-period testing	Ensures the model does not fail when hospital operations shift over time	Forecasts remain calibrated during demand surges and non-routine staffing conditions
External and local transferability	Can the architecture be adapted across hospitals with different topologies?	Test local graph reconstruction, recalibration, and site-specific validation before operational use	Original-hospital model versus locally adapted model	Prevents unsafe transfer of topology-sensitive assumptions between institutions	Each new hospital validates its own node definitions, edge meanings, and workflow dependencies
Missing-data resilience	Does the model behave safely when feeds are stale or incomplete?	Stress-test delayed capacity feeds, missing staff schedules, asynchronous order feeds, and location-data gaps	Complete-feed performance versus degraded-feed scenarios	Determines whether real-time deployment remains trustworthy under imperfect data conditions	Forecasts degrade transparently, uncertainty increases appropriately, and fallback rules are triggered
Interpretability quality	Can operational leaders understand why congestion is predicted?	Review influential nodes, edges, temporal events, and capacity features with command-center staff	Expert review of known congestion episodes and operational narratives	Links predictions to plausible mitigation actions rather than opaque risk scores	Explanations identify credible upstream contributors and do not overstate causal certainty
Alert usability and workflow fit	Are alerts actionable without worsening alert fatigue?	Conduct silent prospective monitoring, usability testing, alert-threshold tuning, and command-center review	Current command-center escalation practices	Determines whether the forecast can be integrated into real operational routines	Alerts are timely, interpretable, threshold-calibrated, and tied to feasible mitigation pathways
Operational impact	Does the model improve coordination without creating unintended bottlenecks?	Use simulation first, then governed pilots measuring delay, queue burden, escalation timing, and staff workload	Standard operations, simulation-informed scenarios, phased pilot controls	Evaluates whether predictions improve flow control rather than only model metrics	Evidence shows improved coordination or earlier mitigation without harmful redistribution of congestion
Governance and safety monitoring	Can the system be audited, monitored, and controlled during use?	Track model drift, graph drift, data-quality failures, alert overrides, human decisions, and audit logs	Predefined governance thresholds and escalation rules	Supports accountability for high-stakes operational decision support	Human review, auditability, fallback procedures, and periodic recalibration are in place before live deployment

Prediction accuracy

Prediction accuracy should be evaluated at both node and system levels, comparing the temporal graph transformer against non-graph baselines such as unit-specific time-series models, tabular machine learning models, and simulation-informed predictors. Candidate metrics could include discrimination for congestion onset, calibration of predicted congestion probabilities, and error in predicted queue or occupancy states, though these should be selected before prospective evaluation. Prior patient-flow and admission-prediction studies provide useful methodological precedents for evaluating whether a model can anticipate operational strain, while traffic-forecasting graph models offer examples of comparing spatio-temporal architectures against simpler baselines [2, 5-7, 18-23]. The central question would be whether explicit graph structure adds operationally meaningful predictive value beyond local historical trends.

Temporal and external validation

Temporal validation should assess whether the model trained on earlier operational history remains useful in later periods with different demand patterns, staffing conditions, and service utilization. External validation should examine whether a model developed in one hospital can be adapted to another hospital with a different physical layout, departmental structure, and workflow topology. General surveys of graph neural networks for traffic forecasting and dynamic graph learning show that topology, temporal distribution, and domain shift can strongly affect model behavior, making validation across time and settings essential [12, 23, 28]. In healthcare, prospective validation is especially important because operational data streams and care processes can change as policies, staffing models, and information systems evolve [1, 4].

Operational impact

Operational impact should be evaluated conceptually through simulation and eventually through carefully governed prospective pilots, not through unverified claims of improved throughput. Simulation could examine how predicted congestion alerts might influence patient routing, service prioritization, discharge coordination, and staffing escalation under realistic workflow constraints. Hybrid patient-flow simulation and machine learning research supports the idea that model-triggered interventions should be tested in simulated environments before live deployment [25, 26]. The ultimate evaluation should ask whether the temporal graph transformer improves coordination, reduces avoidable delays, and supports staff decision-making without creating alert fatigue or unintended bottlenecks elsewhere [3, 4].

Limitations

Data integration complexity

Building a consistent real-time graph across electronic health records, bed-management systems, order-entry platforms, staffing systems, and location feeds is technically demanding. Differences in timestamps, identifiers, update frequencies, and documentation practices may introduce noise that affects node features, edge weights, and temporal alignment. Health information system reviews show that patient-flow management depends heavily on the quality and integration of operational data, which may vary across institutions [4]. The proposed model should therefore be implemented with data-quality monitoring, governance procedures, and fallback rules for degraded feeds [3].

Generalizability across hospital topologies

A graph model trained on one hospital may not transfer directly to another because the meaning of edges depends on local layout, staffing models, service configuration, and clinical workflow. For example, an imaging dependency, intensive care transfer pathway, or pharmacy bottleneck may have different operational implications across hospitals even when the same labels are used. Dynamic graph and spatio-temporal forecasting studies indicate that learned topology-sensitive representations can be powerful but may also be vulnerable to distribution shift when the network changes [14, 23, 28].

Any deployment beyond the originating hospital should therefore include recalibration, local graph reconstruction, and prospective validation before operational use.

Conclusion

A temporal graph transformer offers a conceptual framework for predicting interdepartmental workflow congestion by representing the hospital as a dynamic operational network. In this framework, units become nodes, patient movements and service requests become directed dependencies, and capacity and staffing signals define the evolving state of each department. The model is intended to forecast where congestion could emerge and how it may propagate across the hospital.

The main strength of this approach is its explicit representation of network structure. Rather than treating an emergency department delay, imaging backlog, or ward capacity problem as an isolated event, the model can represent how each unit contributes to system-wide flow. Its temporal design also allows recent operational history to influence prediction, while its attention mechanisms can support interpretable review by command center staff.

Important challenges remain before such a model could be used safely in live operations. Data integration across multiple hospital systems is complex, and graph construction choices could influence which dependencies are learned. Transferability across hospitals also requires caution because each institution has its own physical layout, staffing model, and workflow topology.

Future work should focus on carefully governed pilot studies in large, multi-unit hospitals. These pilots should begin with retrospective feasibility assessment, proceed to silent prospective monitoring, and only then consider controlled operational integration. The goal should be to move from conceptual proof of principle toward a live decision-support tool that improves coordination without replacing professional judgment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Jan 2026

Revised: 06 Mar 2026

Accepted: 16 Apr 2026

Published online: 20 July 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

- 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.
- King Z, Farrington J, Utley M, Kung E, Elkhodair S, Harris S, et al. Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digit Med.* 2022;5(1):104.
- Pianykh OS, Guitron S, Parke D, Zhang C, Pandharipande P, Brink J, et al. Improving healthcare operations management with machine learning. *Nat Mach Intell.* 2020;2(5):266-73.
- Nguyen Q, Wybrow M, Burstein F, Taylor D, Enticott J. Understanding the impacts of health information systems on patient flow management: A systematic review across several decades of research. *PLoS One.* 2022;17(9):e0274493.
- Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One.* 2018;13(7):e0201016.
- De Hond A, Raven W, Schinkelshoek L, Gaakeer M, Ter Avest E, Sir O, et al. Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope? *Int J Med Inform.* 2021;152:104496.
- Patel D, Cheetirala SN, Raut G, Tamegue J, Kia A, Glicksberg B, et al. Predicting adult hospital admission from emergency department using machine learning: An inclusive gradient boosting model. *J Clin Med.* 2022;11(23):6888.
- Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based attention model for healthcare representation learning. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM; 2017. p. 787-95.
- Reychav I, McHaney R, Babbar S, Weragalaarachchi K, Azaizah N, Nevet A. Graph network techniques to model and analyze emergency department patient flow. *Mathematics.* 2022;10(9):1526.
- Rossi E, Chamberlain B, Frasca F, Eynard D, Monti F, Bronstein M. Temporal graph networks for deep learning on dynamic graphs. *arXiv.* 2020;arXiv:2006.10637.
- Kumar S, Zhang X, Leskovec J. Predicting dynamic embedding trajectory in temporal interaction networks. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* New York: ACM; 2019. p. 1269-78.
- Xu D, Ruan C, Korpeoglu E, Kumar S, Achan K. Inductive representation learning on temporal graphs. *arXiv.* 2020;arXiv:2002.07962.
- Sankar A, Wu Y, Gou L, Zhang W, Yang H. DySAT: Deep neural representation learning on dynamic graphs via self-attention networks. In: *Proceedings of the 13th ACM International Conference on Web Search and Data Mining.* New York: ACM; 2020. p. 519-27.
- Paraja A, Domeniconi G, Chen J, Ma T, Suzumura T, Kanezashi H, et al. EvolveGCN: Evolving graph convolutional networks for dynamic graphs. *Proc AAAI Conf Artif Intell.* 2020;34(4):5363-70.
- Wang Z, Li Q, Yu D, Han X. Temporal graph transformer for dynamic network. In: *International Conference on Artificial Neural Networks.* Cham: Springer; 2022. p. 694-705.

- Zhong Y, Huang C. A dynamic graph representation learning based on temporal graph transformer. *Alex Eng J.* 2023;63:359-69.
- Chen H, Jiao P, Du M, Guo X, Zhao Z, Jin D, et al. TGFormer: Towards temporal graph transformer with auto-correlation mechanism. *Pattern Recognit.* 2026;170:112053.
- Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv.* 2017;arXiv:1709.04875.
- Guo S, Lin Y, Feng N, Song C, Wan H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proc AAAI Conf Artif Intell.* 2019;33(1):922-9.
- Wu Z, Pan S, Long G, Jiang J, Zhang C. Graph WaveNet for deep spatial-temporal graph modeling. *arXiv.* 2019;arXiv:1906.00121.
- Zheng C, Fan X, Wang C, Qi J. GMAN: A graph multi-attention network for traffic prediction. *Proc AAAI Conf Artif Intell.* 2020;34(1):1234-41.
- Zhang X, Huang C, Xu Y, Xia L, Dai P, Bo L, et al. Traffic flow forecasting with spatial-temporal graph diffusion network. *Proc AAAI Conf Artif Intell.* 2021;35(17):15008-15.
- Jiang W, Luo J. Graph neural network for traffic forecasting: A survey. *Expert Syst Appl.* 2022;207:117921.
- Abuhay TM, Nigatie YG, Metsker OG, Yakovlev AN, Kovalchuk SV. Investigating coordination of hospital departments in delivering healthcare for acute coronary syndrome patients using data-driven network analysis. In: *International Conference on Computational Science.* Cham: Springer; 2020. p. 430-40.
- Nas S, Koyuncu M. Emergency department capacity planning: A recurrent neural network and simulation approach. *Comput Math Methods Med.* 2019;2019:4359719.
- Abuhay TM, Robinson S, Mamuye A, Kovalchuk SV. Machine learning integrated patient flow simulation: Why and how? *J Simul.* 2023;17(5):580-93.
- Yuan H, Yu H, Gui S, Ji S. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(5):5782-99.
- Li B, Li Q, Zhou X, Deng M, Ling H. Dynamic Graph Transformer with Spatio-Temporal Attention for Streamflow Forecasting. *Hydrology.* 2025;12(12):322.