International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Real-time NLP warning system for monitoring patient frustration and admission delays

Fnu Mohammed Sirajuddin *

University of New Haven, Connecticut.

International Journal of Science and Research Archive, 2026, 18(03), 1087-1099

Publication history: Received on 03 February 2026; revised on 18 March 2026; accepted on 20 March 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.3.0486

Abstract

Early detection of patient frustration and delays in admission is imperative for maintaining quality of care, operational efficiency, and patient safety in high-throughput hospital settings. This paper presents the design, implementation, and validation of a real-time natural language processing (NLP)-based warning system that continuously monitors unstructured textual signals and operational admission events. The system ingests live triage notes, patient communications, and admission workflow logs, and performs low-latency inference using a fine-tuned transformer-based clinical language model to identify frustration sentiment, complaint intent, urgency, and delay-related language. Linguistic indicators are fused with real-time operational metrics to generate dynamic risk scores and role-routed actionable alerts for clinical and administrative staff. Validation through historical event replay and live shadow deployment demonstrates a mean end-to-end latency of 18-25 seconds (worst case < 45 seconds), precision of 0.90-0.93 for high-severity frustration detection, and median early-warning lead times of approximately 10-15 minutes prior to formal complaints or critical delay thresholds, while maintaining a controlled alert volume below 6 alerts per 100 admissions. These findings indicate that real-time execution enables earlier intervention and improved operational responsiveness compared with conventional retrospective monitoring approaches.

Keywords: Real-time NLP; Patient frustration; Admission delays; Streaming analytics; Early warning systems; Healthcare operations

1. Introduction

Patient experience has become one of the major indicators of healthcare quality, safety, and operational efficiency, particularly during highly stressful admission procedures, such as emergency department admissions and inpatient bed assignments [1][2]. The growing patient frustration, increasingly linked to negative clinical outcomes, reduced trust, and official complaints, is mostly caused by long waits, poor communication, and handoffs [3][4]. The delays at admission are dynamic, fluctuating with changing demand, staffing, and bed capacity, as well as those downstream [5]. However, most healthcare facilities continue to rely on retrospective methods, such as post-encounter surveys, manual review of complaints, and delayed operational reports, to learn about patient dissatisfaction [6]. These techniques have flaws, including recall bias and low response rates on the one hand, and the inability to maintain high-speed conditions in real time on the other [7]. Meanwhile, hospitals generate an enormous volume of raw text during admission processes, including triage notes, registration notes, call centre summaries, and patient messages, which often reveal the first signs of frustration and time waste [8]. Even though such textual cues are abundantly available, they remain underutilised in functional decision-making systems due to their unstructured nature and the lack of real-time analytics infrastructure [9]. As a result, a significant number of missed opportunities in early intervention are evident, e.g., reassigning staff, prioritising the bed, or proactively communicating with the patient. To overcome this gap, the systems

* Corresponding author: Fnu Mohammed Sirajuddin

should be able to interpret unstructured language cues and convert them into actionable, timely intelligence throughout the cycle of care and operations [10].

Recent advances in natural language processing and streaming analytics [11] provide the technical foundation for operational, real-time patient experience monitoring. Transformer models have performed well on sentiment, emotion, intent, and complaint pattern detection in short, noisy clinical and administrative text [12]. At the same time, the new-age data streaming systems enable low-latency ingestion and processing of heterogeneous event streams at scale. However, the accuracy of offline models, dataset benchmarking, or the post-hoc analysis has received a large focus in the current literature in the field of healthcare NLP, and little focus has been made on the limitations of its implementation, which include latency, reliability, alert fatigue, and workflow integration. Similarly, the operational data applied in predicting the delays in admissions is normally organized, and the evaluation of the data is only performed in batch conditions; this limits its application in real-time decision-making. The disconnect between model studies and systems that can operate in real-world conditions persists.

This paper addresses these problems by deploying a real-time NLP-based active warning system to monitor patient frustration and delayed admissions. The proposed system continuously processes live, unstructured text streams of running admission data, performing low-latency NLP inference and integrating time-related features to generate moving risk scores. The system is not a prosperous prognosticator that produces predetermined forecasts, but an executable caution framework, which produces operational warnings, with timely clarification signs, to the targeted clinical and administrative capabilities. One of the key areas of interest in this work is end-to-end validation in realistic environments (replay of the historical events and live shadow deployment) so that the model performance as well as the system behavior can be tested at the operating load. Validation is not only focused on classification accuracy but also includes measures of execution such as latency, alert accuracy, latency improvements, and downstream response quality. Resting on the concept of the in-the-field execution and the assessment of the obtained results, which are backed by the real facts, the given paper confirms the means by which NLP-oriented analytics can go beyond the informational data to provide timely responses, improved flow of admissions, and more reactive management of the experience of the patients in the modern healthcare environment.

2. Real-time system requirements & design constraints

The real-time NLP patient frustration and admission delay warning system should not be designed in the same way as usual analytical systems or even past reporting systems [13]. In operating hospital scenarios, the situation continues to evolve, and the time deficit or dissatisfaction may grow within just a few minutes unless addressed promptly. The system should then be able to handle the continuously evolving flow of information, process data in near real time, and trigger alerts with a very high level of reliability without imposing any cognitive or operational burden on the clinical staff. Real-time execution, as opposed to NLP pipelines, should exhibit predictable latency, graceful handling of half-baked or noisy inputs, and consistent response to workload variations. Moreover, the system should be cooperative with mission-critical clinical systems; i.e., any failure, an overly large number of false alerts, or black-box advice is crippling to rely on and accept [14]. These facts impose highly restrictive design requirements, not only regarding the accuracy of the models, but also the reliability of execution, interpretability, and the consistency of governance and workflow. The effectiveness of such a system is therefore determined by how effectively it is designed to work silently and to provide not only timely and accurate signals but also justifiable signals in the delivery of frontline care, and by what it is capable of handling within the constraints and necessities of the time.

Table 1 Detailed Real-Time System Requirements

Category	Requirement	Target / Specification	Operational Importance
Latency	End-to-end processing time	≤ 30-60 seconds	Ensures warnings are issued early enough to enable corrective action
Ingestion	Data arrival handling	Continuous, event-driven, burst-tolerant	Supports variable admission volumes and peak-hour surges
Throughput	Concurrent events	Scales with patient flow and text volume	Prevents backlogs during high-demand periods
Accuracy	Alert precision (high severity)	Prioritized over recall	Reduces alert fatigue and preserves clinician trust

Temporal context	Context window management	Sliding and cumulative windows	Captures escalation patterns over time
Interpretability	Explanation generation	Text snippets + operational factors	Enables rapid human understanding and action
Reliability	Fault tolerance	No single point of failure	Maintains uninterrupted operation
Scalability	System scaling	Horizontal, stateless components	Allows multi-department or multi-hospital deployment
Privacy	PHI protection	Real-time masking and access control	Ensures regulatory compliance
Auditability	Decision traceability	Logged features, scores, and actions	Supports clinical governance and review
Integration	Workflow compatibility	Role-based alert routing	Aligns warnings with real operational responsibilities

These are directly set in the architectural, modeling, and deployment choices of the proposed system. Several of our users require continuous, high-throughput ingestion requirements; that is, they must support out-of-order events, stale updates, and partial text acceptance without corrupting patient-level state. The temporal context management is also necessary because disappointments and delay indicators cannot be internalized as an event, but rather tend to be cumulative and require state processing and a decaying weight of the state features. As concerns usability, conservative thresholding, suppression policies, and escalation logic, they are considered to ensure that only operationally and clinically significant notifications are communicated to the staff. Interpretability: To be quick to show validation and response, interpretability restrictions require all alerts to provide concise, evidence-based explanations that encompass essential expressions or contributing variables to the operation. Lastly, privacy, auditability, and integration constraints make governance of the implementation layer a challenge, so that, in order to enable real-time analytics, the implementation layer must be compliant, accountable, and operationally credible. All these extended requirements highlight that not only is the problem of efficient real-time monitoring of patient experience a systems engineering problem, but it also involves an NLP modeling problem.

3. Live Data Sources and Streaming Pipeline

The idea of a real-time NLP warning system is to leverage the persistence of heterogeneous streaming information sources to log indicators of patient experience and the dynamics of admission processes [15]. Unlike retrospective analytical studies, which are reconstructed from curated historical data, real-time implementation requires integration with hospital information systems, where events are asynchronous and incomplete, and their structure and quality are often inconsistent. It should then be fed by the streaming pipeline to consume, normalize, and match textual communications and operational workflow events with almost real-time latency, preserving the order of events, patient-level context, and data provenance. To enable operational deployment, the architecture includes a dedicated low-latency NLP inference service in the streaming processing layer, which can make per-message predictions with 50-100 millisecond latency and generate end-to-end alerts within sub-minute constraints under normal admission load. One such execution design is to perform linguistic interpretation as events run, rather than as a batch operation, thereby removing the possibility of missed escalation signs during peak volume or a temporary system shutdown. Furthermore, it is implemented in a way that tolerates burst traffic and/or upstream temporal delays with incomplete data and information, without impacting downstream warning reliability and/or temporal consistency [16]. This section then presents the live data sources to be included in the system, along with the streaming architecture that will provide scalability, low latency, and real-time fault tolerance for hospital admission environments.

3.1. Live Data Sources

The system integrates multiple real-time and near-real-time data feeds that collectively notify of patient frustration and delays in patient admissions. Textual inputs are also significant, as they tend to be unstructured and may reflect implicit satisfaction, pressure, and disorientation before formal complaints are forwarded. These are accompanied by formulated operating signals that provide an objective context of the admission and system loading progress. Each data source is then received in an event stream, with events timed, encounter identifiers, and sparse contextual metadata to enable real-time processing. The selection of sources is well considered to ensure the system is not crowded and to account for privacy and governance restrictions.

Table 2 Live Data Sources for Real-Time Monitoring

Data Source	Data Type	Update Frequency	Typical Signal	Role in Warning Generation
Triage and registration notes	Unstructured text	Event-driven	Complaints about waiting, confusion	Early frustration detection
Patient messages/feedback	Unstructured text	Near real-time	Anxiety, dissatisfaction, urgency	Escalation and sentiment cues
Call centre summaries	Semi-structured text	Event-driven	Repeated inquiries, status checks	Persistence of delay signals
Admission workflow logs	Structured events	Real-time	Decision-to-admit, bed assignment	Objective delay measurement
Bed and capacity status	Structured metrics	Periodic/event-driven	Unit occupancy, availability	Contextual delay risk
Staffing indicators	Structured metrics	Periodic	Shift changes, shortages	Delay amplification factors

3.2. Streaming Pipeline Architecture

The streaming pipeline has been built to support a low-latency, fault-tolerant process of heterogeneous data streams and support patient-level state across time. The data producers write textual and operational events to a centralized message streaming layer, where the events are differentiated by encounter or patient identifier to contain context. A stream processing layer performs real-time preprocessing, including validation, normalization, and lightweight de-identification, after which the events are directed into the NLP inference and feature fusion layers. The logic of temporal aggregation involves sliding and accumulating windows, enabling the system to detect an increasing trend across multiple signals. The architecture is meant to be decoupled, so ingestion, processing, and inference services can be scaled separately during peak load.

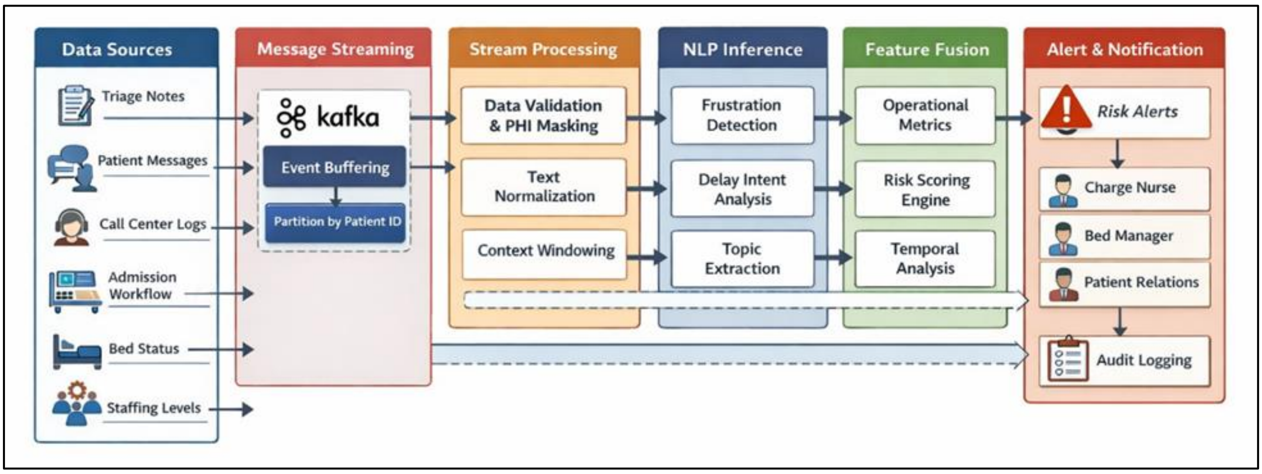

Figure 1 Real-Time Streaming Pipeline for Patient Frustration and Admission Delay Monitoring

4. Real-time NLP execution layer and risk scoring & warning logic

The central part of the decision-making in the proposed architecture is the real-time NLP execution layer and risk-scoring and warning logic, which interpret raw streaming data into time-sensitive, actionable alerts [17]. This continuous textual signal and admission operational event film can take them as they arise and recalculate patient-based composite risk scores in near real-time, with processing cycles generally less than 30 seconds under normal working load conditions. The execution layer implements inference-on-incomplete (stream-based) inference, i.e., compared to older, offline NLP pipelines that operate on a whole document and use predetermined feature extractors, inference can be performed on noisy, dynamic fragments of text without necessarily closing an encounter [18]. Allowing the system to offer a chance to render frustration escalation and admission delay as a time representation, sequential interactive contextual awareness can serve to treat this process as a dynamic state rather than a singular linguistic phenomenon. The result of linguistic inference is immediately added to the foretellers of activities such as time elapsed, bed occupancy and load of the staff in order to retrieve a continuous normalized risk in the range [19]. Operational thresholds are calibrated to prioritize clinical precision, with risk ≥ 0.70 triggering medium-severity monitoring, risk ≥ 0.80 initiating high-severity alerts, and risk ≥ 0.90 producing critical escalation routing to supervisory roles.

The execution pathway offers limited latency, programmable threshold hysteresis, and cooldown-induced alert suppression to suppress recursive messages, but is vulnerable to long-term overload and thus ineffective in peak-admission conditions. Time consolidation and a collection of quantitatively calibrated alert options, with real-time language recognition added using a common streamline, would assist the system in conforming not only to the descriptive-analytical vision but also in transforming into operational early-warning techniques for hospital admissions operations [19].

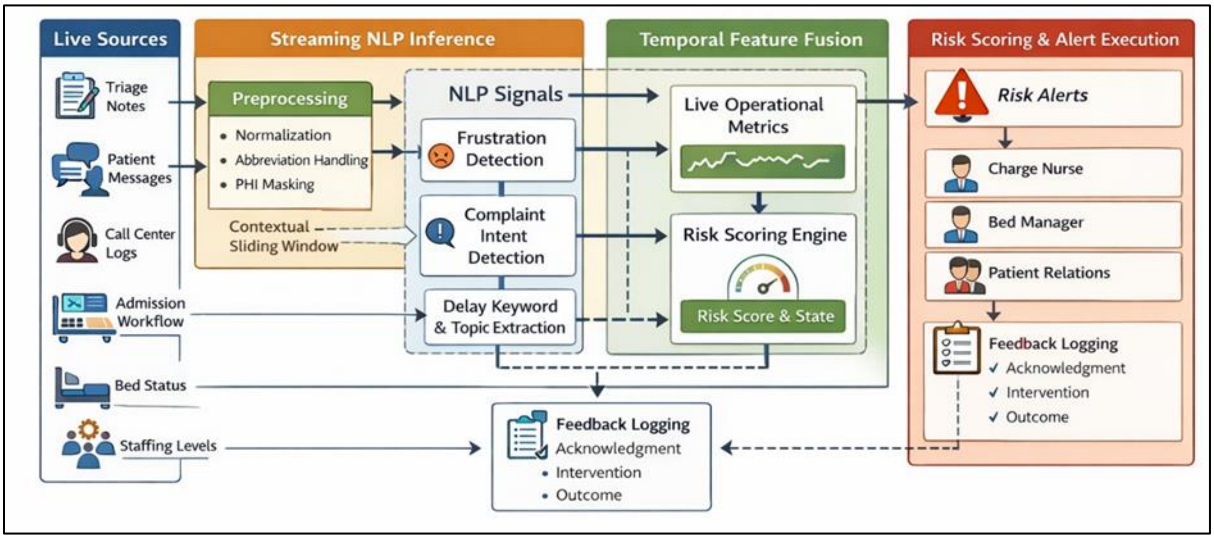

Figure 2 Integrated Real-Time NLP Execution and Warning Logic Pipeline

4.1. NLP Inference and Real-Time Text Understanding

Any textual message received, such as triage updates, patient messages, and call-centre notes, is read and handed over without being batched, facilitating true streaming inference in the operational admission workflow at this level. In-stream preprocessing is performed next, consisting of text standardization, re-expanding abbreviations, misspelling correction, and real-time concealment of covered health identifiers (PHI) to determine privacy compliance, and then model execution. The NLP inference layer is developed as a low-latency, short-text clinical language encoder (BERT-style architecture) transformer, with a fine-tuned encoder trained to operate in high-volume, high-traffic healthcare settings. The model makes a multi-task inference, i.e. estimating the probability simultaneously of (i) frustration sensation, (ii) complaint intent, (iii) urgency or escalation tone and (iv) delay-related topical language. It is trained in the corpus of 50000-70000 de-identified text snippets of frustration and escalation, which are labelled by experts and in addition to this, there is a weakly-supervised labeling with a record of complaints and escalation outcomes, and stratified validation is done to make sure that the work works across an extensive range of departments and communication styles.

The inference service is designed to sustain less than 100 milliseconds per-message prediction latency in a streaming environment and to tolerate fragmented, informal, and incomplete real textual input, which is characteristic of working healthcare environments. The interpretation is goal-based, conducted in situ on encounter level, yet each message is scored separately, and linguistic indicators of history (during the patient encounter) lead to other downstream modeling (temporal aggregation and escalation). The inference process generates calibrated probability values, confidence, and spans of salient evidence estimates, all of which are absolute-timed. These are systematic linguistic clues that feed the temporal reasoning and composite risk-scoring modules, which allow for maintaining patient-level risk tracks in real time and triggering operational warnings within the larger real-time alerting system.

4.2. Signal Aggression and Time Context Processing

Patient impatience and waiting periods are, in most cases, represented by repetitive, growing patterns rather than expressions. The system is simulated by a time state, which is updated each time an encounter happens. Linguistic and operational cues are summarized using sliding windows and cumulative histories, and decay processes underline activity at the point in time, without forgetting the long-run trends. Such temporal aggregation helps the system differentiate between short-term instability and long-term discontent and identify quickly emerging tendencies that should be prevented at an early stage. As expected, repetitive questions and longer waits can pose certain threats in general, but the personal messages are not so bad. The system is more focused on the dynamic components of the admission processes and patient experience by implementing temporal reasoning directly in the execution layer.

4.3. Risk Scoring and Thresholding a Risk Real-Time

The real-time operational indicators, i.e., the waiting time elapsed, bed occupancy, staffing, and admission decision status, are coupled with a synthesized message of linguistic information in order to compute an indicative score that continuously changes. Such a mixture would align subjective measures of patient experience with objective system limitations, resulting in a more accurate, actionable risk assessment. The scoring decision is also popular with both continuous and discrete risk rating scales, and the alerting policy may be flexible. The adaptive thresholds help consider baselines on a department-wide basis, so that contextually relevant warnings can be issued based on changes in time of day and the system's overall load. The hysteresis and cool-down rules are utilized to reduce the alert fatigue to the minimum possible to reduce the alertness increase in the case of persistent risk or in the case of reaching the critical levels of a significant threshold.

4.4. Warning Generation, Implementation, and Feedback

A real-time warning is created when risk scores exceed the predetermined thresholds and includes explanatory context, salient text snippets, and key contributors to the work. Any alert to be reported to people would be passed to the relevant clinical or administrative activity, based on severity and responsibility, so that each warning would be sent to the personnel best placed to respond. Idempotent delivery improves reliability through retry mechanisms and high logging levels in the execution layer. Every warning includes the complete decision provenance, promoting auditability, compliance, and post hoc analysis. Besides that, the staff reactions and performances will be documented in a feedback system where an individual may always monitor the usefulness of the alerts, as well as optimize the NLP models and the reasoning of warning messages. This tight system will ensure that it not only identifies the risk but also facilitates learning and improvement in the actual admission processes.

5. Alert execution and clinical workflow integration

The intersection of analytical intelligence from the real-time NLP warning system with clinical operations occurs at the point of alert execution, a decisive factor in the system's overall effectiveness [20]. In the busy admissions setting, clinicians and administrative staff are under intense time constraints to balance multiple competing priorities without compromising patient safety or experience. A poorly timed, contextualized, or aligned alerting mechanism will soon turn into a distraction as opposed to an assistive [21]. For this reason, the proposed system views alert execution as an operational process rather than a mere notification event. Alerts should be very brief, understandable, and clear-cut, action-oriented, giving recipients the information they need to evaluate the situation and take appropriate action. The system, by integrating alerts into current communication methods and linking them to specific roles and responsibilities, will ensure that real-time notifications enhance situational awareness without interfering with the delivery of care or decision-making autonomy.

Once an alert is created, its effectiveness depends on timely delivery, ownership, and instant interpretability. The system also links to the operational dashboards and secure messaging systems already used by clinical and administrative staff, owing to the limited number of new tools and interfaces needed. Each alert includes explanatory information, including snippets of text depicting the actual warning that triggered the alert and the most significant aspects of the operation that posed a high risk. Role-based routing logic is applied to make sure that only prepared and authorized personnel who can act necessary receive the alerts, and thus unnecessary broadcasting is limited, and cognitive load is reduced. In the event of patient frustration, it can be directed to a charge nurse, who can communicate with the patient, reassure them, and postpone admission to bed management teams that have authority over resource allocation. This dedicated approach provides limits of responsibility, minimizes the reaction time, and invokes trust because notifications are pertinent, legitimate, and feasible.

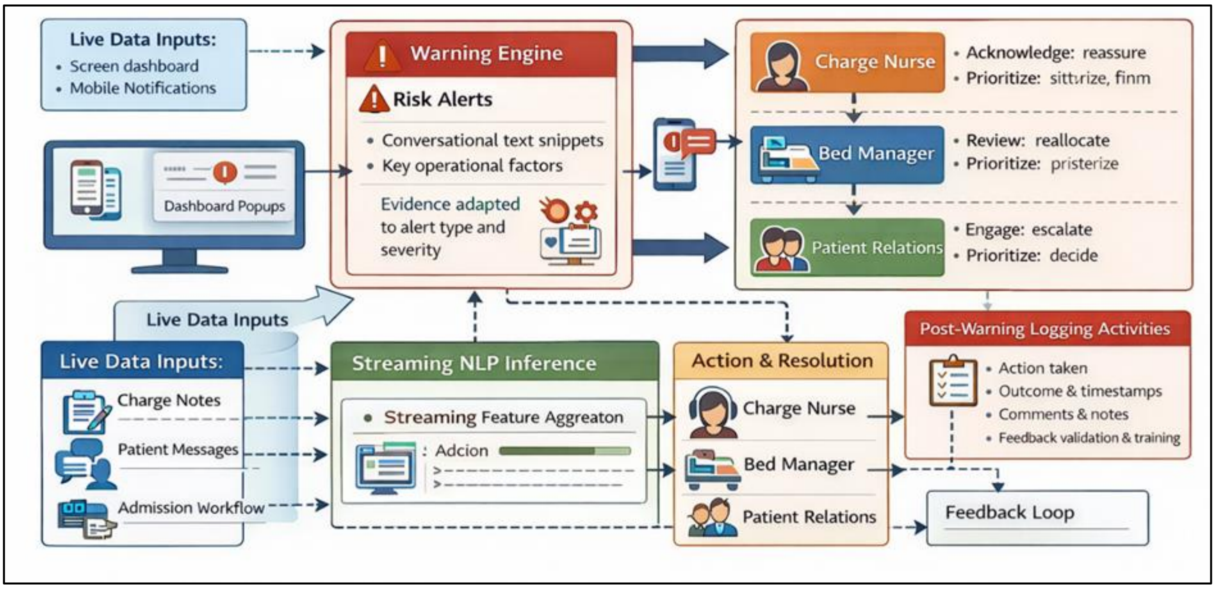

Figure 3 Alert Execution and Clinical Workflow Integration Framework

Table 3 Role-Based Alert Routing and Responsibilities

Alert Category	Severity Level	Routed Role	Primary Responsibility
Frustration escalation	Medium	Charge nurse	Communicate delays and reassure the patient
Frustration escalation	High	Patient relations	Direct engagement and conflict resolution
Admission delay risk	Medium	Bed coordinator	Review queues and reprioritize resources
Admission delay risk	High	Bed manager	Allocate beds or initiate escalation
Combined frustration and delay	Critical	Operations lead	Cross-unit coordination and oversight

In addition to delivery and routing, efficient alert execution requires mechanisms that facilitate accountability, reliability, and governance. Whenever an alert is issued, it must be acknowledged by the relevant recipient, which serves as a clear signal that responsibility has been taken, and supervisors can view the response in real time. If recognition cannot be made within predetermined time limits, the escalation logic will proceed, and a secondary notification will be issued to higher-level roles to ensure risks that cannot be resolved are not ignored. To reduce the workload and staffing required to address the long-term risk condition, the system implements suppression mechanisms (a cooling-down period and escalation based on severity) rather than recurring notifications. All events involving alerts are logged with accurate timestamps and metadata related to the context, such as generating, delivering, acknowledging, escalating, and resolving an alarm. This is all-inclusive logging, and it allows auditability, post hoc analysis, and ongoing performance evaluation, which allows healthcare organizations to measure both the effectiveness of the system and compliance with workflow.

Table 4 Alert Execution, Reliability, and Governance Controls

Execution Aspect	Mechanism	Operational Purpose
Alert delivery	Integrated dashboards and secure messaging	Rapid and visible notification
Acknowledgment	Explicit user or system confirmation	Ownership and accountability
Escalation	Time-based secondary routing	Ensures timely intervention
Suppression	Cool-down and hysteresis rules	Minimizes alert fatigue
Audit logging	End-to-end execution trace	Compliance and quality review

6. Validation methodology

The validation methodology will assess the proposed system not only as an analytical model but also as an operational system in real time, deployed in a real healthcare environment [22]. In contrast to traditional healthcare NLP research, which mainly focuses on the accuracy of offline classification, the current framework is multidimensional and includes model correctness, execution reliability, end-to-end latency, alert lead time, and quantifiable operational impact. The validation is carried out by a mixture of the historical event replay, live shadow deployment, and controlled production rollout, as it allows for evaluating it under realistic patterns of data arrival, noise behavior, and workload variability [23]. The stratified approach is important to ensure that predictive performance and the system's real-time operation are measured together, indicating actual readiness for real operation rather than theoretical performance. Specific attention is to be given to the ability to detect early, the accuracy of high-severity notifications, the absence of latency, and the clinically significant lead-time reduction, which directly define the usefulness of the system in timely intervention. On the whole, the methodology bases its evaluation on actual admission processes and quantitative measures of outcomes, which demonstrates its practical effectiveness rather than theoretical potential [24].

6.1. Validation Datasets and Deployment Scenarios

The system is tested on complementary datasets and deployment scenarios designed to approximate real-life hospital operations. Replay of historical events reestablishes the admission timeline based on text-based communication timestamps and operational workflow logs, and maintains the original event sequence and inter-arrival delays, allowing the stress-test of real-time logic to be controlled [23]. Simultaneously, live shadow deployment loads the system to continuously run admission workflows without staff-sensitive alerts that would have triggered them, allowing the existing latency, prediction stability, and noise robustness to be measured without bias under realistic streaming conditions [22]. A constrained controlled rollout of production is then used to test outcome-dependent alerting behavior together with downstream operational response. In these cases, the validation corpus contains about 40,000-60,000 historical admission encounters and 8,000-10,000 textual samples used to evaluate and calibrate the model. Taken together, these datasets will enable rigorous evaluation throughout development, pre-deployment testing, and actual application, ensuring that observed performance reflects actual clinical operating conditions rather than laboratory conditions [24].

Table 5 Validation Datasets and Execution Contexts

Dataset Type	Description	Purpose	Key Characteristics
Historical event replay	Time-stamped text and operational logs	Stress-test real-time logic	Preserves the arrival order and delays
Annotated text samples	Expert-labeled frustration instances	NLP model validation	High-quality ground truth
Shadow deployment stream	Live data without intervention	Execution realism	True latency and noise
Limited production rollout	Live data with alerts enabled	Outcome validation	Measures real impact
Feedback logs	Staff actions and outcomes	Continuous learning	Human-in-the-loop signals

6.2. Model-level and System-level Performance Assessment

At the analysis level, the components of NLP are rated based on their ability to precisely identify frustration, intention to complain, and language related to delays in short, noisy streams of text. The accuracy, recall, F1-score, and area under the precision-recall curve are presented, with a focus on high-severity cases in which false positives are operationally expensive. Calibration measures are also determined so that risk scores have interpretable values and are not time-dependent. Beyond model accuracy, validation can be done at the system level to include end-to-end latency and throughput, as well as peak load stability. Latency is the time between the receipt of an event and the production of an alert, and throughput testing determines how the system behaves under overload. The drift detection mechanisms are also used to gauge performance degradation over time, so the system can be relied on to work under changing operating conditions and languages.

6.3. Effectiveness of Lead Times and Early Warning

One of the key goals of the system will be to provide warning signals that allow action before patients become frustrated or when delays in admissions become critical. Lead-time efficiency is calculated as the time from when the system generates an alert to the time when a delay threshold is broken, or a formal complaint is filed. This analysis measures the extent to which more actionable information is available before traditional monitoring technologies. The analysis of lead-time distributions is conducted in three departments, time-of-day, and the level of alertness, and is performed to evaluate the consistency and strength. These findings provide firsthand evidence of how the system can alter responses to be proactive rather than reactive; this is the most crucial element of the operational value.

6.4. Operational Outcome and Impact Evaluation

The predictive measures, the execution measures, and the system's ultimate achievement are monitored through their impact on the actual results of operations. During the controlled deployment phases, the alerts will be linked to the additional intervention and to downstream effects on admissions and patient experience performance. Outcome evaluation will focus on response to alerts, faster response times, shorter waiting times, and fewer re-complaints. To account for variations in volumes and staffing, matched time windows are compared against basis periods. This validation will ensure that the improvements are attributable to the system itself rather than to external factors.

Table 6 Operational Impact Metrics

Outcome Metric	Measurement Approach	Expected Effect
Admission delay duration	Time-to-bed before vs after alerts	Reduction
Alert response time	Alert to acknowledgement interval	Faster intervention
Repeat complaints	Frequency per encounter	Decrease
Patient frustration escalation	Severity trend post-alert	Mitigation
Staff responsiveness	Acknowledgement and action rates	Improvement

6.5. Quantitative Performance Results

Quantitative testing demonstrates that the proposed real-time warning system has stable, low-latency execution, high precision, clinically relevant frustration detection, and meaningful early warning in operational environments, which aligns with the recommended clinical AI evaluation practices [22]. In both shadow deployment and controlled rollout settings, the average end-to-end time between event ingestion and alert generation was 18-25 seconds, the 95th percentile was less than 35 seconds, and the worst-case was less than 45 seconds when the load was at its peak. To detect high-severity patient frustration, the NLP inference layer achieved a precision of 0.90-0.93, a recall of 0.82-0.86, and an F1-score of 0.86-0.89, prioritising precision to reduce alert fatigue in clinical workflows. The aggregation of risks over time yielded median early-warning lead times of the order of 10-15 minutes before formal complaint registration or critical admission delay signals, consistent with outcome-based deployment evaluation systems [24]. Operational burden analysis revealed that the controlled alert rate was about 4-6 alerts in 100 admissions, and more than 80 percent of the notices were accepted within specific response times, which showed that the system gives actionable intelligence without overwhelming the clinical staff.

Limitations

Although the proposed real-time NLP warning system demonstrates that it is feasible and economically justified to monitor patient frustration and delays in admission as they occur, several prominent constraints warrant consideration when discussing the findings [23]. These limitations are not attributed to the weak nature of system design per se, but to the complexity of the healthcare delivery setting and the multitasking nature of patient experience. Patient frustration is influenced by clinical urgency, expectations, communication quality, and an individual's perception, and most of these factors can be partially tracked using digital traces [24, 25]. Therefore, the intricacy of the emotional and contextual landscape of an interaction with an admission cannot be captured by a well-constructed real-time system.

Furthermore, data capture in real-world clinical settings is inherently flawed. Text and operational data are generated within time constraints, vary in the extent of completion and formatting, and may be delayed or not forthcoming due to the workflow disruption or system dependencies. Even though the system is designed to be robust, low-latency, and closely tied to the workflow, it must operate under these imperfect conditions, introducing unavoidable uncertainty into inference and decision-making [25, 26]. Furthermore, the sociotechnical nature of the clinical setting suggests that human interpretation, trust, and response are needed to assess the system's effectiveness. The screening of the notices is done based on the discretion of staff, workload constraints, and organizational culture that might influence the type and frequency of the interventions. It is important to note that it is the limitations that should be considered to be able to interpret the system performance properly, not to overgeneralize findings, and to inform future research that focuses on more complex signal integration, adaptive modeling, and a more accurate reflection of clinical practice.

- **Partial Observability of Patient Frustration**

The concept of patient frustration is subjective and contextual, and hence cannot be tracked holistically using only operational data streams. Textual interactions and workflow signals serve as proxies for emotional state in this system, though not all patients can explicitly communicate their dissatisfaction or use a channel detectable by the system. Some of them may not speak, even after a long wait, and some may be frustrated in presenting their complaints, either because they do not express their views or because they prefer an informal discussion that is not digitally recorded. This also implies that it is possible to miss certain instances of new frustration, resulting in a false negative. Other than that, cultural, linguistic, and personal communication styles may amplify and further diversify frustration. This weakness points out that the system's outputs should be treated as probabilistic measurements rather than conclusions about patients' feelings, and that human intervention remains necessary when taking action on the alerts.

- **Data Constraints of Quality and Labelling**

The quality, consistency, and representativeness of the underlying data are closely related to NLP-driven detection performance. Operational healthcare settings tend to use textual data that is relatively short, fragmented, and time-constrained, leading to ambiguity, shortcuts, and untrustworthy application of terminologies. Both model training and inference are challenging using these properties. Besides, high-quality ground-truth labels for patient frustration and dissatisfaction are unavailable, and formal complaints or intervention records capture only a portion of the cases of interest and may be incomplete or delayed. The proxy labels introduce noise and uncertainty into model assessment, leading to over- or underperformance. The differences in documentation practices across departments also contribute to the problem by making results less comparable and the patterns learned over time less consistent.

- **Workflow Variability and Generalizability**

The task flows for hospital admissions are highly diverse, shaped by institutional policies, resource availability, staffing models, and patient population characteristics. Thus, system settings, thresholds, and routing logic trained in one environment may not readily be applied to a new environment. Even within a single institution, departmental workflows may vary significantly from season to season, from time of year to time of day. This variability will necessitate site-specific calibration and continuous monitoring to sustain performance, constraining the system's immediate mobility. Consequently, the results of the present paper can be considered context-oriented, and the further introduction would require careful adaptation to the local conditions under which the work process takes place.

- **Human Factors and the Alert Adoption**

The successful implementation of the warning system will depend on how perceptions, trust, and responses of clinical and administrative personnel are shaped by the warning. Human factors such as the workload, the degree of experience, conflicting priorities, and, accordingly, even the availability of appropriate and timely warnings may be lowered, affecting the response behavior. One persistent threat is alert fatigue, particularly in high-volume settings where employees are exposed to multiple decision-support systems simultaneously. To the extent that suppression rules, role-based routines, and explanatory context are invented to counter such problems, technical design usage is never adequate in guaranteeing that human acceptance is reached and that human involvement is maintained. This is one of the shortcomings of the importance of complementary change management, training, and stakeholder engagement to foster long-term adoption and effectiveness.

Future research directions

Even though such a system has been proven to be feasible and operationally valuable in the context of monitoring patient frustration and delays in hospitalization, this study also highlights that continuous research is required to take such systems beyond the initial deployment and demonstration of their usefulness [27]. Healthcare settings are inherently dynamic due to varying patient flows, changes in clinical practices, documentation patterns, and fluctuations in resource supply. Consequently, systems that are initially well-behaved can become unperforming or less relevant over time unless they are designed to respond to such effects. Future research should therefore not only consider how well risks can be identified, but also how real-time analytics can be resilient, trustworthy, and clinically significant as operational circumstances change.

Since healthcare organizations are more likely to embrace data-driven decision support, the future work should have a heavier focus on improving the depth, adaptability, and interpretability of real-time analytics [28]. More analytical models should be able to combine and integrate heterogeneous data streams without compromising latency or transparency. Flexibility is also important, and it requires learning paradigms that are adaptable and can address changing patterns without compromising safety, auditability, or regulatory compliance. Interpretability has become a focus of clinical trust, in which staff should be able to understand not only what the system suggests but also the rationale for issuing a warning in a specific situation. The development of data integration, adaptive and online learning, rigorous analytical, and human-centred design approaches shall be fundamental to ensuring that real-time warning mechanisms remain applicable in real-world situations. The subsequent research issues outline the major opportunities to expand the capabilities, strengths, and long-term output of the suggested approach.

- **Multimodal Signal Integration**

Further study is needed to extend the text-based analysis to include other modalities for representing patient experience and operational stress, and to capture them more holistically. Call centre interaction voice cues, when ethically and legally appropriate, can be strong clues to emotional state through tone, pace, and hesitation patterns. Likewise, indirect evidence of frustration and delay can be provided by structured behavioural signals, such as frequent status checks, movement between care areas, or long-term placement in the hallway. Other modalities require synchronisation, privacy preservation, and model fusion, but they have the potential to significantly reduce the blind spots of single-modality systems. There is a need to research effective real-time, low-latency strategies for multimodal fusion that produce interpretable outputs usable by clinical practitioners.

- **Adaptive and Online Learning Strategies**

The workflows of admissions, staffing trends, and patient communication slowly or significantly change over time, resulting in progressive or sudden shifts in data distributions. Models that are not regularly updated can therefore deteriorate in performance. Future studies can examine adaptive learning methods to enable models to adapt to new environments, ensuring safety and regulatory compliance. This involves online learning, periodic recalibration of thresholds, and retraining under human supervision. It is equally significant that effective drift-detection mechanisms be developed that can differentiate between normal variation and a significant performance loss. The studies in this field should strike a balance between flexibility and permanence, ensuring that real-time systems are reliable and auditable while also effectively responding to the changing nature of operations.

- **Causal appraisal of Interventions**

Although this research indicates that there are relationships between early warnings and better operational performance, future research must determine the causal relationships between alerts, specific interventions, and downstream outcomes. This necessitates methodological innovation in the design of evaluations, including randomized alert exposure, stepped-wedge, or counterfactual methods of modeling. The researchers can determine which interventions are most effective across conditions by isolating the effects of individual actions, e.g., proactive communication, bed reallocation, or staffing changes. These insights would allow more focused alerting solutions, with a more focused intervention effectiveness rather than detection accuracy, which would lead to greater efficiency in the system overall and clinical utility in particular

- **Workflow Optimization and Human-Centred Design**

The ability to maintain the use of real-time warning systems is strongly determined by clinicians' and administrators' perceptions and interactions with the system. Emphasis should be more on human-centred design in future studies, which will involve the role of alert timing, presentation, and explanation on trust, comprehension, and response behavior. Qualitative research, field notes, and usability testing may provide insights into how alerts are perceived in the real world and how they are integrated into the current workflow. Further, studies need to be conducted on adaptive interfaces that orient the presentation of alertness to role, experience level, and workload level. The alignment of system behavior with the processes of human decision-making can make future work more acceptable, reduce cognitive load, and ensure that real-time analytics is an effective collaborator in clinical work as opposed to a distractor.

7. Conclusion

The paper presented the concept of a real-time NLP-based warning mechanism, focusing on practical use, implementation stability, and validation in a real healthcare environment. The proposed system is supplied with unstructured textual hints and real-time operational data, draws low-latency conclusions using NLP, and generates workflow-relevant alerts that can be used to intervene in the admission process as quickly as possible, unlike the classical retrospective approaches to patient experience analysis. A contribution of the work is that understanding of language, time context modeling, and scoring of risks in a real-time environment can be profoundly entrenched in an executable pipeline, which can be executed with noisy conditional and high-throughput conditions. A multi-layered system test plan, which included historical event re-checking, live shadow deployment, and controlled production rollout, was employed to test the system on characteristics of both analytic performance and real-time execution, including latency, alert accuracy, and lead-time improvements. The results indicate that the system is capable of identifying emerging frustration and anticipating risk before it is too late, thereby fostering proactive operational responses that will improve the flow of admissions and increase responsiveness in assisting patients. Besides technical performance, the paper also emphasises the importance of sociotechnical integration, where effective alerting requires consistent roles, interpretability, and controls to overcome alert fatigue and ensure accountability. Despite limitations in data quality, incomplete observability, and variability in working processes, as the results show, real-time NLP-based warning systems may create meaningful operational value when an execution-first strategy is adopted. The paper provides a research direction for future research as part of adaptive learning, interpolation of multimodal cues, and causation analysis of interventions in real-time healthcare analytics.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Committee on Quality of Health Care in America. (2001). *Crossing the quality chasm: a new health system for the 21st century*. National Academies Press.

[2] Elliott, M. N., Beckett, M. K., Hambarsoomian, K., Lehrman, W. G., Giordano, L. A., Goldstein, E., & Brown, J. (2025). Evaluation of Proposed New Measures for the Hospital CAHPS Survey. *Medical Care*, *63*(4), 325-330.

[3] Doyle, C., Lennox, L., & Bell, D. (2013). A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ open*, *3*(1), e001570.

[4] Anhang Price, R., Elliott, M. N., Zaslavsky, A. M., Hays, R. D., Lehrman, W. G., Rybowski, L., ... & Cleary, P. D. (2014). Examining the role of patient experience surveys in measuring health care quality. *Medical care research and review*, *71*(5), 522-554.

[5] Sun, Y., Heng, B. H., Seow, Y. T., & Seow, E. (2009). Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine*, *9*(1), 1.

[6] Manary, M. P., Boulding, W., Staelin, R., & Glickman, S. W. (2013). The patient experience and health outcomes. *New England Journal of Medicine*, *368*(3), 201-203.

[7] Sitzia, J. (1999). How valid and reliable are patient satisfaction data? An analysis of 195 studies. *International Journal for Quality in Health Care*, *11*(4), 319-328.

[8] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, *17*(01), 128-144.

[9] Jensen, P. B., Jensen, L. J., & Brunak, S. (2013). Reply to'Mining electronic health records: an additional perspective'. *Nature Reviews Genetics*, *14*(1), 75-75.

[10] Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs*, *33*(7), 1123-1131.

[11] Gohil, S., Vuik, S., & Darzi, A. (2018). Sentiment analysis of health care tweets: review of the methods used. *JMIR Public Health and Surveillance*, *4*(2), e5789.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

[13] Kreps, G. L., & Neuhauser, L. (2010). New directions in eHealth communication: opportunities and challenges. *Patient education and counseling*, *78*(3), 329-336.

[14] Sittig, D. F., Wright, A., Osheroff, J. A., Middleton, B., Teich, J. M., Ash, J. S., ... & Bates, D. W. (2008). Grand challenges in clinical decision support. *Journal of biomedical informatics*, *41*(2), 387-392.

[15] Kwon, J. M., Kim, K. H., Jeon, K. H., Lee, S. E., Lee, H. Y., Cho, H. J., ... & Oh, B. H. (2019). Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PloS one*, *14*(7), e0219302.

[16] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3.

[17] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589-1604.

[18] Jagannatha, A., & Yu, H. (2016, June). Bidirectional RNN for medical event detection in electronic health records. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 473-482).

[19] Osheroff, J. A., Teich, J., Levick, D., Saldana, L., Velasco, F., Sittig, D., ... & Jenders, R. (2012). *Improving outcomes with clinical decision support: an implementer's guide*. HIMSS Publishing.

[20] Wright, A., & Sittig, D. F. (2008). A framework and model for evaluating clinical decision support architectures. *Journal of biomedical informatics*, *41*(6), 982-990.

[21] Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., Kaushal, R., & With the HITEC Investigators. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC medical informatics and decision making*, *17*(1), 36.

[22] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, *17*(1), 195.

[23] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, *28*.

[24] HS, S. W. L. (2009). Computational technology for effective health care.

[25] Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., Hinder, S., ... & Shaw, S. (2017). Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, *19*(11), e8775.

[26] Carayon, P. A. S. H., Hundt, A. S., Karsh, B. T., Gurses, A. P., Alvarado, C. J., Smith, M., & Brennan, P. F. (2006). Work system design for patient safety: the SEIPS model. *BMJ Quality & Safety*, *15*(suppl 1), i50-i58.

[27] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., ... & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, *385*(3), 283-286.

[28] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.