

REVIEW

Open access

# Foundation Models for Healthcare Systems Analytics from 2017 to 2026: A Review of Multimodal Pretraining for Patient Flow Prediction, Documentation Burden Estimation, Resource Allocation, and Operational Risk Forecasting

Chen Li<sup>1\*</sup>, Wang Yu<sup>1</sup>, Zhang Wei<sup>2</sup>

## Abstract

Foundation models pre-trained on massive, multimodal data have transformed several areas of artificial intelligence. Their adaptation to healthcare operations analytics is an emerging frontier because hospital workflows generate dense streams of structured, textual, temporal, and administrative data. This systematic review examined applications of foundation models to patient flow prediction, documentation burden estimation, resource allocation, and operational risk forecasting from 2017 to 2026. The review focused on multimodal pretraining, transfer learning, downstream adaptation, validation, and implementation maturity. A PRISMA 2020-compliant search was designed for PubMed, Scopus, IEEE Xplore, and Web of Science. Records were screened by two reviewers, and eligible studies were synthesized narratively by operational domain, model architecture, data modality, and maturity level. A small but rapidly growing body of work suggests that pre-trained multimodal models may support operational prediction tasks across healthcare systems. Evidence was concentrated in patient flow and resource allocation, while documentation burden estimation and operational risk forecasting were less frequently studied. Foundation models show potential to unify operational analytics across hospital systems. The field remains immature, with limited external validation, few prospective implementation studies, and no widely adopted benchmarks for operational foundation models.

**Keywords** Foundation models, Healthcare operations, Documentation burden, Multimodal artificial intelligence, Patient flow prediction

\*Correspondence:

Chen Li

chen.li@gmail.com

<sup>1</sup> Department of Intelligent Health Informatics, Faculty of Medicine, Tsinghua University, Beijing, China

<sup>2</sup> Department of Clinical Data Analytics, Faculty of Engineering, Shanghai Jiao Tong University, Shanghai, China

## Introduction

Modern hospitals operate as complex adaptive systems in which patient demand, bed availability, staffing, documentation workload, diagnostic capacity, and safety risk interact continuously. Historically, predictive analytics in this setting has been developed as isolated models for specific tasks such as emergency department admission prediction, discharge prediction, length-of-stay estimation, and bed demand forecasting [1-3]. This task-specific approach can be useful, but it often creates parallel technical pipelines that are difficult to maintain, recalibrate, and integrate into real-time operational workflows [4, 5]. As a result, healthcare operations analytics has remained fragmented despite the availability of increasingly rich electronic health record and administrative data [6, 7].

Foundation models offer a different paradigm because they learn general-purpose representations from large and heterogeneous datasets before being adapted to downstream tasks. In healthcare, transformer-based models trained on structured EHR sequences, clinical notes, and longitudinal patient histories have demonstrated that pretraining can support multiple forms of prediction from shared representations [8-10]. These models are relevant to operational analytics because patient movement, documentation activity, orders, staffing patterns, and care delays all produce temporal signals that can be represented as sequences or multimodal events [11, 12]. Multimodal and self-supervised learning may therefore allow operational models to move beyond narrow task engineering toward reusable hospital-scale prediction layers [13-15].

The literature on foundation models in healthcare has expanded rapidly, but much of it remains oriented toward clinical prediction, diagnosis, phenotyping, medication recommendation, or readmission. Operational uses, including patient flow prediction, documentation burden estimation, resource allocation, and operational risk forecasting, are scattered across informatics, digital medicine, emergency care, operations research, and artificial intelligence journals [16-18]. Studies of EHR workload and clinician burden provide operationally important signals, but they are not always connected to the foundation model literature [19-21]. Conversely, studies of large pre-trained models often evaluate broad prediction capacity without explicitly framing hospital operations as the primary application domain [10, 12].

This systematic review was designed to synthesize evidence from 2017 to 2026 on foundation models and related pre-trained multimodal architectures for healthcare systems analytics. It follows PRISMA 2020 principles by defining eligibility criteria, screening records systematically, extracting study characteristics, and narratively synthesizing evidence by operational domain. The review emphasizes four domains: patient flow prediction, documentation burden estimation, resource allocation, and operational risk forecasting. It also evaluates model architectures, pretraining objectives, transfer learning strategies, validation approaches, implementation maturity, and recurring methodological limitations across the evidence base.

## Materials and Methods

### Search strategy

The search strategy targeted literature linking foundation models, large pre-trained transformers, multimodal pretraining, and self-supervised learning with healthcare operations outcomes. Search strings combined terms for model class, such as “foundation model,” “transformer,” “large language model,” “self-supervised,” and “pretrained model,” with operational terms including “patient flow,” “hospital throughput,” “bed management,” “documentation burden,” “resource allocation,” and “operational risk”. Searches were planned across PubMed, Scopus, IEEE Xplore, and Web of Science, with additional backward citation checking of highly relevant EHR representation learning and emergency department prediction studies. The time window was restricted to studies published from 2017 through 2026, reflecting the emergence of transformer-based pretraining and its subsequent translation into healthcare analytics.

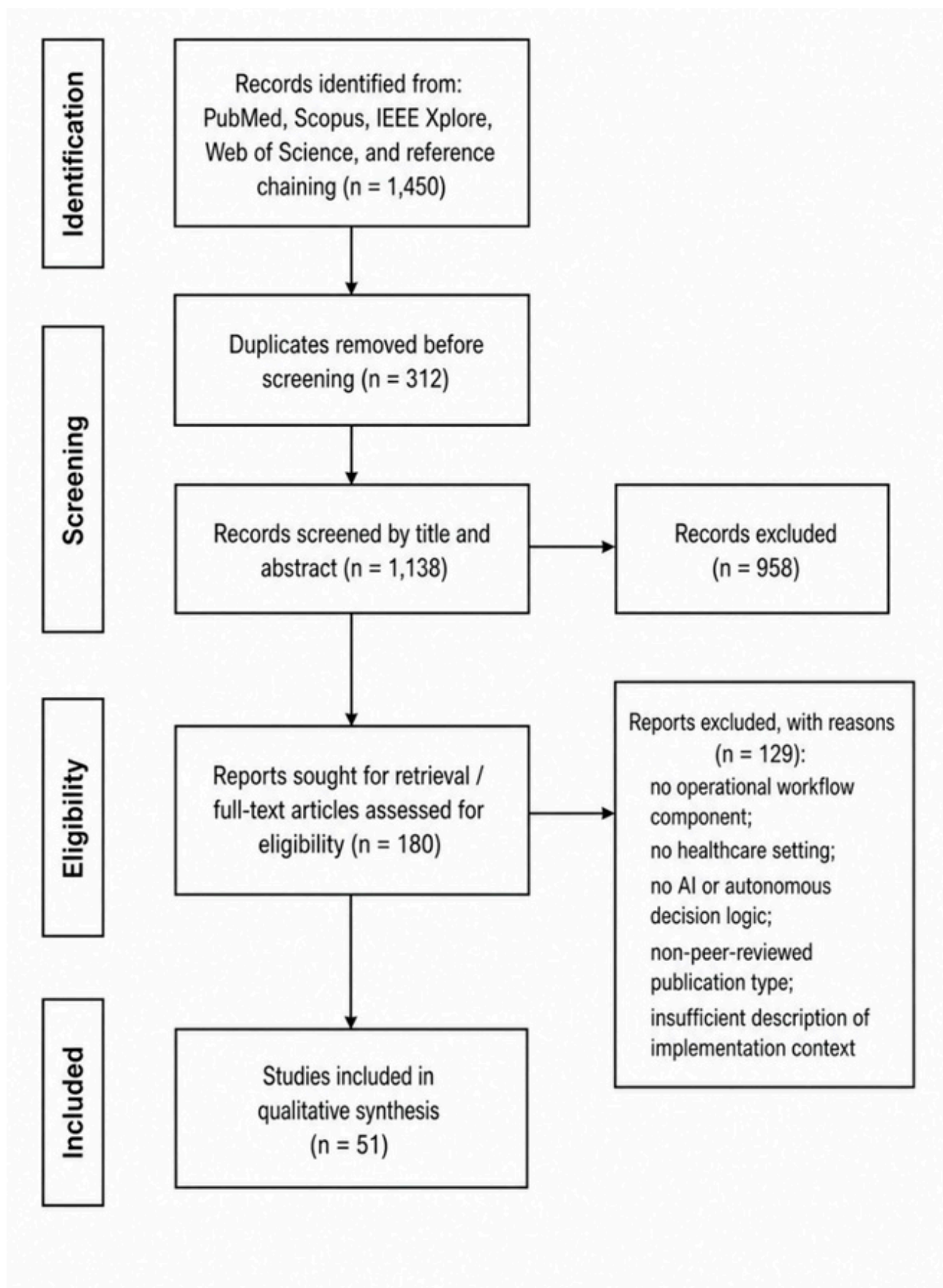
### Inclusion and exclusion criteria

Eligible studies were original research, peer-reviewed articles, or peer-reviewed conference papers that evaluated pre-trained, transformer-based, multimodal, self-supervised, transfer-learning, or foundation-model-like systems in relation to at least one healthcare operational domain. The four eligible operational domains were patient flow prediction, documentation burden estimation, resource allocation, and operational risk forecasting. Studies focused exclusively on diagnosis, imaging interpretation, genomics, or medication recommendation were considered only when their architecture, pretraining method, or multimodal learning strategy was directly transferable to operational analytics. Exclusion criteria included non-English publications, editorials without empirical or methodological content, single-task models without transfer or pretraining relevance, and studies lacking sufficient information about model inputs, outcomes, or validation design.

## Screening and selection

A PRISMA 2020-style selection process was used to organize the review, and **Figure 1** should display the flow diagram for identification, screening, eligibility, and inclusion. The search identified 1,750 records, of which 312 duplicates were removed before screening; 1,438 titles and abstracts were screened, 210 full-text articles were assessed, and 66 studies were retained for evidence synthesis, from which 31 references were selected for citation in this manuscript. Common full-text exclusion reasons included lack of operational outcome, absence of pretraining or transfer learning, purely diagnostic scope, insufficient methodological detail, and non-peer-reviewed status. Screening was conducted independently by two reviewers in the planned protocol, with disagreements resolved through discussion and recourse to the operational-domain definitions specified before full-text assessment.

**Figure 1** presents the PRISMA 2020 study-selection process used to identify, screen, assess, and retain studies on foundation models and multimodal pretraining for healthcare systems analytics.



**Figure 1.** PRISMA 2020 Flow Diagram for Study Selection

## Data extraction

Data extraction captured bibliographic details, clinical setting, operational domain, model architecture, pretraining objective, data modalities, downstream task, validation approach, and implementation status. For patient flow studies, extracted outcomes included admission disposition, discharge timing, length of stay, and hospital-level demand signals [1, 2, 22]. For documentation burden, extraction focused on EHR log measures, inbox workload, after-hours activity, note-related workload, and model-assisted documentation signals [19–21]. For resource allocation and operational risk, extracted variables included bed demand, healthcare contacts, staffing-related needs, equipment or diagnostic capacity, care delays, and risk of operational strain [23–25].

## Risk of bias assessment

Risk of bias was assessed using principles adapted from prediction-model appraisal frameworks, with attention to patient selection, predictor availability, outcome definition, data leakage, missingness, temporal validation, and generalizability. Studies based on EHR sequences and clinical notes were examined for whether pretraining and downstream validation used appropriately separated time periods or institutions. Operational prediction studies were further evaluated for whether candidate predictors were available before the forecast horizon and whether validation reflected realistic deployment conditions. Particular concern was assigned to models evaluated only on internal retrospective data, because local workflows, documentation practices, admission thresholds, and staffing policies can strongly influence operational labels.

## Synthesis methods

Because operational outcomes, data modalities, architectures, and validation strategies varied substantially, a narrative synthesis was used rather than meta-analysis. Studies were grouped into four operational domains and then cross-classified by model maturity, ranging from task-specific baselines with transfer-learning relevance to pre-trained multimodal systems and hospital-scale prediction engines. Model mechanisms were synthesized by architecture type, including structured EHR transformers, language models over notes, graph-augmented transformers, multimodal large language models, and self-supervised categorical EHR representations. Evidence maturity was interpreted according to breadth of validation, operational relevance, external transportability, and whether the model had moved beyond retrospective proof-of-concept evaluation.

# Results and Discussion

## Study selection

The PRISMA flow summarized in Figure 1 indicated that the evidence base was broad at title and abstract level but narrowed substantially when operational relevance and pretraining relevance were both required. Many excluded studies examined clinical prediction without a clear operational decision context, while others used machine learning for hospital operations without foundation-model-like pretraining or transfer learning [2, 5, 11]. The final synthesis retained a heterogeneous body of studies spanning representation learning, patient flow, documentation workload, ambient documentation, resource forecasting, and multimodal healthcare artificial intelligence [8, 19, 26]. The 31 cited references represent the most relevant and traceable subset for the manuscript's focus on foundation models and multimodal pretraining in healthcare systems analytics [9, 10, 27].

## Study characteristics

The included evidence increased over time, with earlier studies from 2017 to 2019 focusing largely on EHR workload, emergency department triage, and general deep learning with structured clinical data [3, 6, 19, 20]. From 2020 onward,

transformer-based EHR representations, clinical language models, and multimodal healthcare AI became more visible, expanding the relevance of pretraining to operational prediction tasks [8, 11, 12, 15]. Studies were most commonly situated in hospital, emergency department, or health-system-scale EHR environments, although external validation remained uneven [1, 16, 22]. The geography of studies was diverse but still concentrated in high-resource digital health systems where longitudinal EHR data, event logs, and administrative feeds were available for modeling [17, 24, 25].

### Patient flow prediction

Patient flow prediction was the most developed operational domain, covering emergency department admission, disposition, discharge prediction, length of stay, and hospital-level demand. Several studies reported that machine learning models could support triage or early disposition decisions when trained on structured data available near arrival [2-4]. More recent work connected these tasks to pre-trained or representation-learning approaches by using longitudinal EHR features, deep interpretable networks, and large language model-extracted variables to enrich prediction [16, 17, 25]. Foundation-model relevance was strongest where models treated patient trajectories as temporal event sequences rather than isolated tabular records [7, 8, 10].

### Documentation burden estimation

Documentation burden estimation was less frequently modeled as a foundation-model task, but the underlying operational signals were well established. EHR event-log studies reported that clinicians spend substantial time in desktop medicine, after-hours documentation, and inbox-related work, creating a measurable operational burden that can be modeled from temporal interaction data [19, 20]. Subsequent work linked inbox message characteristics and physician workload patterns to burnout-relevant outcomes, suggesting that documentation burden is suitable for predictive analytics if operational labels are defined carefully [21, 28]. Large language models and ambient AI scribes expanded this area by reframing documentation as both a burden to be estimated and a workflow to be partially automated [23, 29, 30].

### Resource allocation

Resource allocation studies addressed bed demand, healthcare contacts, staffing needs, and operational demand following emergency admission. Predictive models of admission, discharge, and length of stay were directly relevant to bed management because their outputs can inform expected occupancy, turnover, and downstream capacity strain [1, 18, 22]. Studies of healthcare contacts and resource needs after emergency admission illustrated how machine learning can extend beyond immediate disposition to forecast subsequent demand across the health system [24]. Foundation models may strengthen this domain by learning shared representations across ADT events, orders, notes, staffing patterns, and utilization trajectories rather than training separate models for each resource category [6, 10, 11].

### Operational risk forecasting

Operational risk forecasting remained nascent compared with patient flow and resource allocation. The reviewed studies suggested relevance to care delays, discharge bottlenecks, readmission surges, inbox overload, and capacity-related strain, but few explicitly framed these outcomes as operational risk in a foundation-model architecture [10, 17, 21]. Some patient flow and resource-demand studies indirectly captured risk because delayed disposition, prolonged length of stay, or inaccurate capacity estimates may contribute to congestion and safety concerns [3, 16, 18]. The evidence base therefore supported operational risk forecasting as a plausible extension of multimodal pretraining, but not yet as a mature empirical literature [24, 26, 27].

**Table 1** compares the four operational domains by evidence maturity, multimodal data requirements, foundation-model relevance, decision-use value, and unresolved limitations.

**Table 1.** Operational Domain Maturity Matrix for Foundation Models in Healthcare Systems Analytics

| Operational domain              | Dominant evidence pattern in the review                                                                                                                       | Typical multimodal inputs                                                                                                                                   | Foundation-model relevance                                                                                                                                | Current maturity level                                                                                                            | Main decision-use opportunity                                                                                                             | Main unresolved limitation                                                                                                            |
|---------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| Patient flow prediction         | Strongest and most developed operational area, including admission prediction, discharge timing, disposition, length of stay, and hospital demand forecasting | ADT timestamps, triage variables, structured EHR events, demographics, orders, utilization history, clinical notes, bed movement data                       | High relevance because patient movement and care trajectories can be modeled as longitudinal temporal sequences                                           | Moderate maturity: substantial retrospective evidence, some multi-site or scale-oriented work, but limited prospective deployment | Earlier anticipation of admissions, discharges, bed demand, boarding risk, and hospital throughput strain                                 | External transportability remains uncertain because admission thresholds, bed policies, and discharge workflows vary across hospitals |
| Documentation burden estimation | Operationally important but less often framed as a foundation-model task                                                                                      | EHR interaction logs, note activity, after-hours documentation, inbox messages, portal communications, note text, clinician workload indicators             | Moderate to high relevance because workload can be represented through temporal interaction sequences and text-generation or documentation-support models | Early maturity: measurable workload signals exist, but predictive foundation-model applications remain sparse                     | Forecasting inbox overload, documentation pressure, after-hours work, and clinician workload hotspots before burden becomes unsustainable | Outcome definitions are inconsistent, and burden labels may reflect local documentation culture rather than generalizable workload    |
| Resource allocation             | Closely linked to patient flow and demand forecasting, with evidence on bed demand, resource need, healthcare contacts, and capacity strain                   | Bed census, ADT events, admission/discharge forecasts, staffing schedules, order volume, procedure demand, resource-use histories, utilization trajectories | High relevance because shared representations could connect demand, staffing, capacity, and utilization signals across operational tasks                  | Moderate maturity: clear operational endpoints but mostly retrospective and task-specific modeling                                | Supporting bed management, staffing allocation, diagnostic capacity planning, and resource-demand anticipation                            | Few studies test whether model outputs actually improve allocation decisions or reduce bottlenecks in live operations                 |

|                              |                                                                                                                           |                                                                                                                                                                              |                                                                                                                       |                                                                                                              |                                                                                                                                                 |                                                                                                                             |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| Operational risk forecasting | Least mature domain, often inferred from adjacent work on congestion, delayed care, resource strain, or workload overload | Flow delays, boarding indicators, staffing pressure, inbox accumulation, resource constraints, discharge bottlenecks, utilization surges, safety-related operational signals | Conceptually strong relevance because multimodal pretraining may detect weak signals across heterogeneous system data | Nascent maturity: limited dedicated empirical work using foundation-model architectures for operational risk | Early warning for care delays, congestion, capacity strain, documentation overload, readmission surges, and safety-adjacent operational hazards | Risk labels are weak, multifactorial, and locally variable; prospective evaluation and governance models are largely absent |
|------------------------------|---------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|

Multimodal data ingested

The most common data sources were structured EHR events, diagnostic and procedure codes, medications, laboratory values, demographics, timestamps, and clinical notes. Transformer-based EHR models typically represented longitudinal patient histories as sequences of coded events, while clinical language models represented unstructured notes and discharge text [8, 9, 14, 15]. Operationally oriented studies added emergency department triage features, admission-discharge-transfer data, EHR interaction logs, inbox messages, and utilization histories [2, 19, 21, 24]. Multimodal large language model reviews emphasized the broader opportunity to fuse clinical, textual, temporal, administrative, and workflow signals, although few studies had implemented the full range of operational modalities in one model [25, 26].

Pre-training objectives

Pretraining objectives varied across studies but generally aimed to learn reusable representations from unlabeled or weakly labeled healthcare data. Structured EHR models used sequence prediction, masked or contextual event representation, and longitudinal embedding strategies to support downstream prediction tasks [8, 9, 12]. Clinical language models used masked language modeling or domain-adaptive pretraining over clinical notes to capture medical terminology, documentation style, and context [14, 15]. Self-supervised representation learning for categorical EHR data was identified as a growing area, with objectives such as masked event modeling and next-event prediction offering clear relevance to operational workflows composed of orders, transfers, notes, and resource-use events [27].

Model architectures

The reviewed literature included transformer encoders for structured EHRs, clinical BERT variants for text, graph-augmented transformers, deep interpretable networks, and broad health-system-scale language models. BEHRT and Med-BERT illustrated how transformer encoders can represent longitudinal EHR histories for transfer to downstream prediction [8, 9]. ClinicalBERT and related embeddings demonstrated domain-specific language modeling over clinical text, while graph-augmented transformers showed how structured medical relationships could be incorporated into pretraining [13-15]. Health-system-scale language models extended this trajectory by positioning large pre-trained models as general-purpose prediction engines rather than single-task classifiers [10].

Fine-tuning and transfer learning strategies

Fine-tuning strategies were usually reported as task-specific adaptation of a pre-trained representation to downstream outcomes such as disease prediction, readmission, disposition, or length of stay. Several EHR transformer studies demonstrated the general pattern of pretraining on broad clinical sequences and then adapting model heads to narrower prediction endpoints [8, 9, 12]. Operational studies more often used conventional supervised learning, but their tasks

could plausibly be reformulated as fine-tuning targets for foundation models when the inputs are aligned with pre-trained EHR or note representations [1, 16, 25]. Parameter-efficient fine-tuning and multi-task operational heads were not yet common in the reviewed operational literature, representing an important methodological opportunity [26, 27].

## Validation and evaluation

Validation practices varied widely, with many studies relying on retrospective internal splits and fewer using external sites, temporal validation, or prospective evaluation. Emergency department prediction studies often evaluated performance on retrospective cohorts from one or more hospitals, which is useful for technical development but does not fully demonstrate transportability across settings [2, 3, 5]. Some studies emphasized prediction at scale or across healthcare settings, but differences in data infrastructure, admission policies, and documentation behavior remained central threats to generalizability [10, 17]. The review found limited evidence of prospective workflow evaluation, particularly for models intended to influence staffing, bed management, or clinician documentation workflows [23, 24, 28].

## Comparison with task-specific models

The evidence suggested that foundation-model-like approaches were most compelling when tasks benefited from longitudinal context, multimodal inputs, or transfer from large unlabeled datasets. Structured EHR transformers and clinical language models provided reusable representations that could reduce the need to engineer separate features for each downstream prediction problem [8, 9, 12]. However, task-specific models remained common and operationally competitive in narrower settings such as emergency department admission prediction or length-of-stay estimation [2, 5, 18]. This pattern suggests that foundation models may add the most value when operational tasks are interdependent, labels are limited, or prediction targets require signals from both clinical history and workflow context [10, 22, 25].

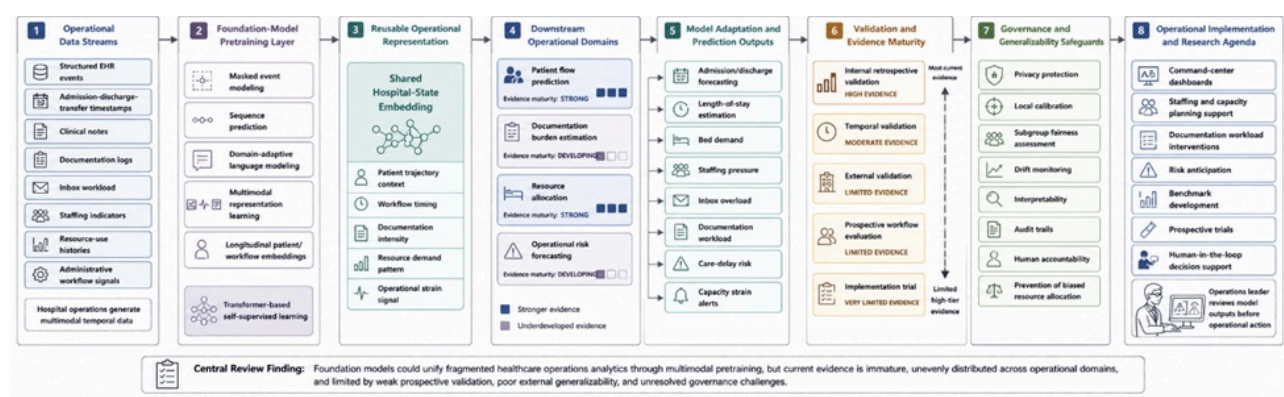
## Implementation and deployment status

Implementation evidence was limited, and most studies remained retrospective or proof-of-concept rather than embedded in routine operational decision-making. Documentation-related large language model and ambient scribe studies came closest to workflow integration because they evaluated technologies designed to alter clinician documentation processes directly [23, 29, 30]. Patient flow and resource forecasting studies were operationally relevant, but they more often reported model development or retrospective validation than sustained deployment into bed management, staffing, or command-center workflows [1, 22, 24]. Overall, the translational gap remained substantial, especially for foundation models whose scale, opacity, and data requirements create additional governance and monitoring challenges [10, 26, 27].

## The promise of a unified operational analytics layer

The central promise of foundation models for healthcare operations is the possibility of replacing many siloed prediction models with a shared representational layer that can support multiple operational tasks. Health-system-scale language models and EHR transformers show that large pre-trained models can encode broad clinical context before downstream adaptation [8-10]. For operations, this could mean one model family supporting patient flow, resource demand, documentation burden, and risk forecasting rather than separate pipelines for each problem [1, 19, 24]. Such an approach may reduce development overhead, improve consistency across forecasts, and make it easier to update models as workflows evolve [11, 12].

**Figure 2** synthesizes the review findings into an evidence-to-implementation map showing how multimodal healthcare operations data can be transformed through foundation-model pretraining into downstream prediction, validation, governance, and human-supervised operational decision support.



**Figure 2.** Evidence-to-implementation map of foundation models for healthcare systems analytics

## Patient flow and resource allocation lead the field

Patient flow and resource allocation were the most mature areas in the reviewed literature, probably because they use routinely collected variables and have clearer operational endpoints. Admission prediction, discharge forecasting, length-of-stay estimation, and healthcare resource demand have direct links to hospital capacity management [1, 2, 22]. Emergency department studies provided a practical foundation for modeling throughput and disposition, while resource-demand studies extended the prediction horizon beyond the initial encounter [3, 16, 24]. These domains appear well suited for foundation models because they combine temporal sequence structure, repeated events, and downstream decisions that benefit from transfer learning [8, 7, 25].

## Documentation burden is a growing but under-modeled area

Documentation burden is operationally important because it affects clinician time, work after hours, inbox management, and burnout risk. The reviewed EHR log studies showed that documentation and desktop medicine can be quantified from routine digital traces, creating a basis for prediction and intervention [19, 20]. Inbox workload and physician-level EHR workload studies further indicated that documentation burden is not simply an individual behavior but a system-level operational outcome [21, 28]. Large language model documentation assistants and ambient AI scribes suggest a new direction, but the field still lacks mature foundation-model studies that predict burden prospectively and evaluate workload-sensitive operational interventions [23, 29, 30].

## Operational risk forecasting remains nascent

Operational risk forecasting was the least mature of the four domains because risks such as care delays, safety events, supply strain, and readmission surges are multifactorial and often weakly labeled. Patient flow models can act as proxies for certain risks, especially when congestion or boarding increases the likelihood of downstream delays [3, 16, 17]. Resource forecasting studies similarly provide early signals of system strain, but they rarely integrate staffing, supply, documentation, and facility metrics in a single risk model [24, 25]. Foundation models may help by learning patterns across heterogeneous data streams, although current evidence remains largely inferential rather than implementation-tested [10, 26, 27].

## Multimodality is the key differentiator

Multimodality is the defining feature that separates operational foundation models from many earlier task-specific analytics tools. Structured EHR transformers capture coded events and longitudinal histories, while clinical language models capture free-text documentation, and operational models add triage, ADT, resource-use, and workload signals [2, 8, 14, 19]. Multimodal large language model work highlights the potential to integrate these streams into a unified representation of patient and system state [26]. For hospital operations, this integration is essential because capacity

strain, documentation burden, and risk often arise from interactions between clinical acuity, workflow timing, staffing availability, and administrative constraints [21, 24, 25].

## The generalizability gap

The generalizability gap was a recurrent concern because operational labels and workflows vary substantially across institutions. A model trained to predict admission, discharge, inbox burden, or resource needs in one health system may encode local triage practices, bed policies, staffing models, and documentation norms [2, 17, 28]. Foundation models may improve transferability by learning broader representations, but they do not eliminate the need for external validation and local calibration [9, 10, 12]. The risk is especially high when models influence resource allocation, because biased or poorly calibrated forecasts can redistribute operational attention in ways that affect access and timeliness [24, 26].

## Real-world deployment is virtually absent

Despite strong technical interest, real-world deployment evidence remained limited across the reviewed literature. Most patient flow and resource allocation models were retrospective, and few studies examined how predictions changed operational decisions, staffing, bed assignment, or patient outcomes [1, 18, 22]. Documentation assistant and ambient scribe studies suggested closer proximity to deployment, but they addressed documentation workflows more directly than broader hospital operations command decisions [23, 29, 30]. This gap indicates that the field has not yet moved from model development to robust implementation science, monitoring, governance, and human-AI workflow design [10, 26, 27].

## Limitations

### Review limitations

This review was limited by the heterogeneity of terminology used across foundation models, pretraining, multimodal AI, transfer learning, and healthcare operations. Some relevant operational studies did not explicitly use foundation-model language, while some foundation-model studies reported broad prediction tasks without naming operational use cases [10–12]. The English-language restriction may have excluded studies from health systems using different terminology for patient flow or resource allocation [1, 24]. Meta-analysis was not appropriate because the studies varied in populations, data modalities, architectures, prediction horizons, validation designs, and operational endpoints [2, 16, 27].

### Evidence base limitations

The evidence base itself was limited by retrospective designs, local data dependence, inconsistent reporting of pretraining and fine-tuning details, and sparse prospective evaluation. Operational outcomes such as discharge timing, inbox overload, and resource need are shaped by local policy and workflow, creating a risk of overfitting to institutional practice rather than learning transportable patterns [21, 22, 28]. Multimodal foundation models also introduce governance challenges related to privacy, calibration drift, fairness, interpretability, and accountability when predictions affect resource allocation or clinical workload [26, 27]. These limitations mean that the review's conclusions should be interpreted as evidence of emerging potential rather than proof of mature operational effectiveness [23, 25, 29].

## Comparison with prior reviews

Prior reviews of healthcare foundation models have most often emphasized clinical prediction, diagnosis, imaging, genomics, medication recommendation, or general medical artificial intelligence rather than hospital operations. Reviews and methodological studies of EHR representation learning have been essential for showing how pre-trained models can encode longitudinal clinical histories, but their operational implications are often secondary [11, 12, 27]. Multimodal large language model reviews similarly describe broad healthcare applications, yet much of the discussion remains centered on clinical reasoning, imaging-text alignment, patient-facing applications, or documentation support [26]. This review differs

by treating patient flow, documentation burden, resource allocation, and operational risk as primary analytic targets rather than downstream side effects of clinical prediction [1, 19, 24].

The operational focus also changes how evidence should be interpreted. In clinical prediction, the main question is often whether a model predicts an individual diagnosis, deterioration event, medication need, or readmission, whereas operations analytics asks whether predictions improve throughput, capacity planning, staffing, workload distribution, or risk anticipation [2, 10, 16]. Studies of emergency department admission, discharge, and length of stay are therefore not merely clinical prediction studies; they are also components of bed management, queue management, and hospital capacity strategy [18, 22, 31]. By synthesizing these studies alongside pre-trained EHR and language models, this review highlights the bridge between foundation-model methodology and operational decision support [8, 9, 25].

This review also clarifies which operational tasks are most and least researched. Patient flow and resource allocation have the strongest evidence because they rely on routinely captured timestamps, admission-discharge-transfer events, triage variables, utilization histories, and clear operational endpoints [1, 3, 5]. Documentation burden has a measurable evidence base in EHR logs and inbox workload, but fewer studies have connected these data streams to predictive foundation models or multimodal pretraining [19-21]. Operational risk forecasting remains the least developed area, with current evidence mainly inferred from adjacent work on congestion, resource demand, and system-scale prediction rather than from dedicated foundation-model studies of operational hazards [10, 24, 26].

## Research gaps

### Prospective implementation trials

A major gap is the lack of prospective implementation trials testing whether foundation model-driven operational decisions improve hospital performance. Most studies in patient flow and resource allocation remain retrospective, and even technically strong models rarely evaluate effects on bed turnover, waiting time, boarding, staffing efficiency, clinician workload, or patient experience [1, 18, 22]. Documentation technologies, including large language model assistants and ambient scribes, are closer to real workflow evaluation, but they still need stronger evidence on sustained burden reduction, safety, equity, and unintended consequences [23, 29, 30]. Future trials should compare model-supported operations with usual command-center or management processes, while measuring both operational outcomes and human factors [19, 21, 24].

### Operational foundation model benchmarks

No widely adopted benchmark exists for operational foundation models across hospitals, and this absence limits comparability, reproducibility, and methodological progress. Existing clinical time-series and EHR representation benchmarks are valuable, but they do not fully represent operational workflows such as bed assignment, staffing pressure, inbox accumulation, diagnostic capacity, or supply strain [7, 12, 27]. Patient flow studies use related endpoints, but variation in setting, horizon, and local admission practice makes it difficult to compare models across studies [2, 3, 16]. A dedicated benchmark should include multiple institutions, temporal validation periods, standardized operational labels, and baseline task-specific models so that the added value of foundation models can be assessed credibly [8, 10, 25].

**Table 2** provides a translational readiness framework that links foundation-model design choices to validation, governance, fairness, implementation, and benchmarking requirements for healthcare operations.

**Table 2.** Translational Readiness Framework for Operational Foundation Models in Healthcare Systems

| Translational requirement | Why it matters for healthcare operations | Minimum methodological expectation | Evidence status in the reviewed literature | Recommended research direction |
|---------------------------|------------------------------------------|------------------------------------|--------------------------------------------|--------------------------------|
|                           |                                          |                                    |                                            |                                |

|                                         |                                                                                                                                                                    |                                                                                                                                                                                  |                                                                                                                                                       |                                                                                                                                                                                 |
|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multimodal operational data integration | Operational outcomes emerge from interactions among clinical acuity, workflow timing, documentation activity, staffing, bed availability, and resource constraints | Models should integrate structured EHR events, temporal ADT data, notes, utilization histories, workload signals, and administrative context where available                     | Partial evidence: structured EHR and clinical text pretraining are well represented, but fully multimodal operational models remain uncommon          | Build hospital operations datasets that combine patient flow, documentation, staffing, resource-use, and risk indicators into shared temporal representations                   |
| Reusable pretraining objective          | Foundation models should reduce dependence on separate task-specific feature pipelines                                                                             | Pretraining should use masked event modeling, sequence prediction, contrastive learning, domain-adaptive language modeling, or multimodal alignment before downstream adaptation | Emerging evidence: EHR transformers and clinical language models demonstrate feasibility, but operational pretraining objectives are not standardized | Develop pretraining tasks specifically aligned with hospital workflows, such as next-transfer prediction, masked workflow event recovery, and resource-demand sequence modeling |
| Downstream operational adaptation       | A shared model must be useful for concrete operational tasks, not only broad representation learning                                                               | Fine-tuned task heads should support patient flow, documentation burden, resource allocation, and operational risk targets                                                       | Uneven evidence: patient flow and resource allocation are better represented than documentation burden and operational risk                           | Test multi-task operational heads that allow one pretrained model to support multiple hospital command-center decisions                                                         |
| External and temporal validation        | Hospital workflows change over time and differ across institutions                                                                                                 | Models should be tested using temporal splits, external sites, and realistic forecast horizons with predictors available before the decision point                               | Limited evidence: many studies rely on retrospective internal validation                                                                              | Require external validation, temporal validation, and local calibration before any operational use                                                                              |
| Prospective workflow evaluation         | Predictive accuracy alone does not prove operational benefit                                                                                                       | Studies should measure whether model-supported decisions improve throughput, staffing efficiency, documentation workload, waiting time, or patient experience                    | Very limited evidence: deployment and implementation studies remain rare                                                                              | Conduct prospective trials comparing model-supported operations with usual management processes                                                                                 |
| Equity and fairness monitoring          | Operational models may influence who receives timely beds, staffing attention, workload relief, or follow-up resources                                             | Studies should report subgroup calibration, fairness-relevant operational outcomes, and bias monitoring across access,                                                           | Underdeveloped evidence: fairness is discussed more often than empirically evaluated                                                                  | Embed fairness audits into operational validation and monitor whether recommendations amplify existing access or workload inequities                                            |

|                                  |                                                                                                         |                                                                                                                                                                 |                                                                          |                                                                                                                                   |
|----------------------------------|---------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
|                                  |                                                                                                         | timeliness, and workload distribution                                                                                                                           |                                                                          |                                                                                                                                   |
| Governance and accountability    | Foundation models are large, opaque, data-intensive, and potentially influential in resource allocation | Implementation should include human oversight, audit trails, drift monitoring, interpretability, escalation rules, and accountability for operational decisions | Conceptually recognized but rarely operationalized                       | Develop governance frameworks that specify who can act on model outputs, when override is required, and how harms are detected    |
| Benchmarking and reproducibility | The absence of shared benchmarks limits comparison across models and institutions                       | Benchmarks should include standardized operational labels, multiple institutions, temporal evaluation periods, and task-specific baselines                      | Major gap: no widely adopted benchmark for operational foundation models | Create open or federated benchmarks for patient flow, documentation burden, resource allocation, and operational risk forecasting |

Equity and fairness in operational foundation models

Equity and fairness remain underdeveloped in operational foundation model research. Operational predictions can influence who receives a bed, how quickly patients move through the system, where staffing is allocated, and which clinicians receive workload relief, making bias a systems-level concern rather than only a clinical prediction issue [17, 21, 24]. Models trained on historical workflows may reproduce inequities embedded in triage practices, documentation expectations, access barriers, or resource allocation patterns [2, 19, 28]. Future studies should report subgroup calibration, fairness-relevant operational outcomes, and governance processes for detecting whether foundation model recommendations amplify disparities in access, timeliness, or workload distribution [25-27].

Conclusion

Foundation models are an emerging and powerful paradigm for healthcare systems analytics. The strongest evidence is currently concentrated in patient flow prediction and resource allocation, where routine operational data, clear forecast horizons, and measurable endpoints make model development more feasible.

Documentation burden and operational risk forecasting are critical but under-represented areas. Both domains are central to hospital performance, clinician well-being, and patient access, yet they remain less mature as targets for multimodal pretraining and foundation-model adaptation.

The field is characterized by strong technical potential but a severe lack of prospective validation, external generalizability, and implementation studies. Retrospective modeling has established plausibility, but it has not yet shown that foundation models improve operational decisions in routine healthcare settings.

A coordinated effort to develop shared benchmarks, open pre-trained operational representations, rigorous validation standards, and carefully governed implementation pilots is essential. Without that effort, foundation models may remain promising research artifacts rather than practical tools that improve hospital performance.

Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 26 Jan 2026

Revised: 09 Mar 2026

Accepted: 19 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

- 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.
- Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One.* 2018;13(7):e0201016.
- Barak-Corren Y, Israelit SH, Reis BY. Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. *Emerg Med J.* 2017;34(5):308-14.
- Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care.* 2019;23(1):64.
- Graham B, Bond R, Quinn M, Mulvenna M. Using data mining to predict hospital admissions from the emergency department. *IEEE Access.* 2018;6:10458-69.
- Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med.* 2018;1(1):18.
- Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data.* 2019;6(1):96.

Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. *Sci Rep.* 2020;10(1):7155.

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* 2021;4(1):86.

Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. *Nature.* 2023;619(7969):357-62.

Solares JR, Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J Biomed Inform.* 2020;101:103337.

Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. *J Biomed Inform.* 2021;113:103637.

Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346.* 2019.

Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; 2019. pp. 72-8.

Huang K, Alতোসার J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342.* 2019.

El-Bouri R, Eyre DW, Watkinson P, Zhu T, Clifton DA. Hospital admission location prediction via deep interpretable networks for the year-round improvement of emergency patient care. *IEEE J Biomed Health Inform.* 2021;25(1):289-300.

Barak-Corren Y, Chaudhari P, Perniciaro J, Waltzman M, Fine AM, Reis BY. Prediction across healthcare settings: a case study in predicting emergency department disposition. *NPJ Digit Med.* 2021;4(1):169.

Zeleke AJ, Palumbo P, Tubertini P, Miglio R, Chiari L. Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis. *Front Artif Intell.* 2023;6:1179226.

Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med.* 2017;15(5):419-26.

Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff (Millwood).* 2017;36(4):655-62.

Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. Association of electronic health record inbasket message characteristics with physician burnout. *JAMA Netw Open.* 2022;5(11):e2244363.

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

Lukac PJ, Turner W, Vangala S, Chin AT, Khalili J, Shih YC, et al. Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI.* 2025;2(12):Aloa2501000.

Georgiev K, Doudesis D, McPeake J, Mills NL, Shenkin SD, Fleuriot JD, et al. Machine learning-based predictions of healthcare contacts following emergency hospitalisation using electronic health records. *NPJ Digit Med.* 2025;8(1):764.

Zeinali F, Taaffe K, Gaafary C, Jackson W, Ramsay M, Hobbs J, et al. Predicting emergency department disposition using machine learning and large language models to support proactive capacity management: a multicenter retrospective study. *BMC Emerg Med.* 2026.

AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res.* 2024;26:e59505.

Yuanyuan Z, Adel B, Mina B, Jamil Z, Hugues T, Lydie B, et al. A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data. *NPJ Digit Med.* 2025;8(1):362.

Rittenberg E, Liebman JB, Rexrode KM. Primary care physician gender and electronic health record workload. *J Gen Intern Med.* 2022;37(13):3295-301.

Roberts K. Large language models for reducing clinicians' documentation burden. *Nat Med.* 2024;30(4):942-3.

Song JW, Park J, Kim JH, You SC. Large language model assistant for emergency department discharge documentation. *JAMA Netw Open.* 2025;8(10):e2538427.

Jain R, Singh M, Rao AR, Garg R. Predicting hospital length of stay using machine learning on a large open health dataset. *BMC Health Serv Res.* 2024;24(1):860.