

REVIEW

Open access

Natural Language Processing for Healthcare Administration from 2017 to 2024: A Review of Models for Clinical Documentation, Coding Support, Incident Reporting, Patient Communication, and Workflow Automation

Thomas Andersen^{1*}, Lars Nielsen¹, Mette Sørensen²

Abstract

Healthcare administration generates large volumes of textual data, including clinical notes, billing documentation, incident narratives, portal messages, and operational records. These data create administrative burden but also provide opportunities for natural language processing to support documentation, coding, communication, safety review, and workflow automation. This systematic review examined natural language processing models applied to healthcare administration from 2017 to 2024. The review focused on five domains: clinical documentation, coding support, incident reporting, patient communication, and workflow automation. A PRISMA 2020-compliant review was conducted using structured searches of PubMed, Scopus, IEEE Xplore, and Web of Science for studies published between 2017 and 2024. Eligible studies were screened by two reviewers, extracted using a structured template, and synthesized narratively by administrative domain, model type, evaluation approach, and implementation maturity. Transformer-based models were increasingly prominent across the reviewed literature, particularly in documentation support, automated coding, and clinical text classification. Clinical documentation and coding support had the strongest evidence base, while incident reporting, patient communication, and workflow automation remained less mature and less frequently evaluated in real-world settings. Natural language processing for healthcare administration is technically promising, but most applications remain at the retrospective or proof-of-concept stage. Prospective validation, workflow integration, governance, and human oversight are needed before widespread operational adoption.

Keywords Natural language processing, Clinical documentation, Healthcare administration, Patient communication, Automated coding, Incident reporting

*Correspondence:

Thomas Andersen

thomas.andersen@gmail.com

¹ Department of Health Data Analytics and Systems, Faculty of Medicine, University of Copenhagen, Copenhagen, Denmark

² Department of Intelligent Clinical Informatics, Faculty of Engineering, Technical University of Denmark, Lyngby, Denmark

Introduction

Healthcare administration depends heavily on unstructured and semi-structured text, including progress notes, discharge summaries, coding justifications, safety narratives, portal messages, and operational reports. These textual sources are essential for care continuity, billing, quality monitoring, and institutional accountability, yet they also contribute to clinician and staff workload. Reviews of clinical natural language processing have shown that unstructured documentation

contains valuable operational and clinical information that is difficult to use at scale without automated methods [1, 2]. The administrative relevance of these data has grown as healthcare organizations seek tools that can reduce documentation burden while preserving accuracy, traceability, and compliance [3, 4].

Natural language processing has been proposed as a way to transform healthcare administrative work by extracting, structuring, summarizing, and generating text across multiple operational domains. Clinical documentation tools include digital scribes, clinical conversation summarizers, and note-structuring systems, while coding support models map clinical text to billing and diagnostic codes [3-6]. Incident reporting systems use NLP to classify safety narratives and identify adverse events, whereas patient communication systems apply language models to portal messages, triage requests, and summarize concerns [7-10]. Workflow automation is emerging through systems that assist medical necessity justification, prior authorization preparation, and administrative report generation [11, 12].

Despite this expanding literature, prior reviews have often focused on clinical NLP in general, specific disease areas, or technical model families rather than the administrative pipeline as a whole. Foundational reviews have synthesized clinical note processing, chronic disease NLP, and biomedical embeddings, but they do not fully integrate documentation, billing, incident reporting, patient communication, and workflow automation as linked administrative functions [1, 2, 13, 14]. This gap matters because administrative NLP tools are increasingly evaluated not only by classification accuracy or text similarity but also by their capacity to change workflow burden, compliance risk, staff workload, and communication quality. A domain-spanning review is therefore needed to distinguish mature applications from early-stage automation claims.

The objective of this systematic review was to synthesize peer-reviewed evidence from 2017 to 2024 on NLP models for healthcare administration across five domains: clinical documentation, coding support, incident reporting, patient communication, and workflow automation. The review followed principles and used a narrative synthesis because the included studies differed substantially in data sources, tasks, model families, evaluation metrics, and deployment settings. The synthesis emphasizes systematic-review language and avoids experimental claims beyond the reviewed evidence. Particular attention is given to implementation maturity, governance needs, human oversight, and the gap between retrospective evaluation and operational utility.

Materials and Methods

Search strategy

A structured search strategy was designed to identify peer-reviewed studies published from January 1, 2017, to December 31, 2024, on NLP applications in healthcare administration. Searches were conducted in PubMed, Scopus, IEEE Xplore, and Web of Science using combinations of terms related to “natural language processing,” “clinical documentation,” “automated coding,” “incident reporting,” “patient communication,” “portal messages,” “prior authorization,” “workflow automation,” “transformer,” “BERT,” “GPT,” and “healthcare administration.” The search strategy was informed by PRISMA 2020 and PRISMA-S principles, with attention to reproducibility, database coverage, and transparent reporting of search concepts.

Inclusion and exclusion criteria

Studies were eligible if they reported NLP methods applied to a healthcare administrative task, including clinical documentation generation or structuring, coding support, incident reporting, patient communication, or workflow automation. Eligible articles included original research, methodological studies, and systematic or scoping reviews when they directly informed administrative NLP domains, were published in English, and appeared between 2017 and 2024. Studies were excluded if they focused exclusively on biological text mining, radiology-only diagnosis without administrative relevance, non-healthcare customer service, or purely technical language modeling without a healthcare

administrative use case. The eligibility framework was aligned with prior clinical NLP reviews but narrowed toward administrative functions such as documentation support, coding, safety-event processing, and message triage.

Screening and selection

Records were screened in two stages by title and abstract followed by full-text assessment, with disagreements resolved through discussion and consensus. The review process identified 2,100 records, removed 420 duplicates, screened 1,680 records by title and abstract, assessed 260 full-text reports, and included 85 studies in the qualitative synthesis. Common exclusion reasons at full-text review included absence of an NLP method, lack of administrative relevance, conference abstracts without sufficient methodological detail, non-English publication, and studies outside the 2017–2024 time window. The PRISMA flow diagram should report these numbers and distinguish included evidence across documentation, coding, incident reporting, patient communication, and workflow automation.

Figure 1 presents the PRISMA 2020 study selection process for the systematic review of NLP applications in healthcare administration.

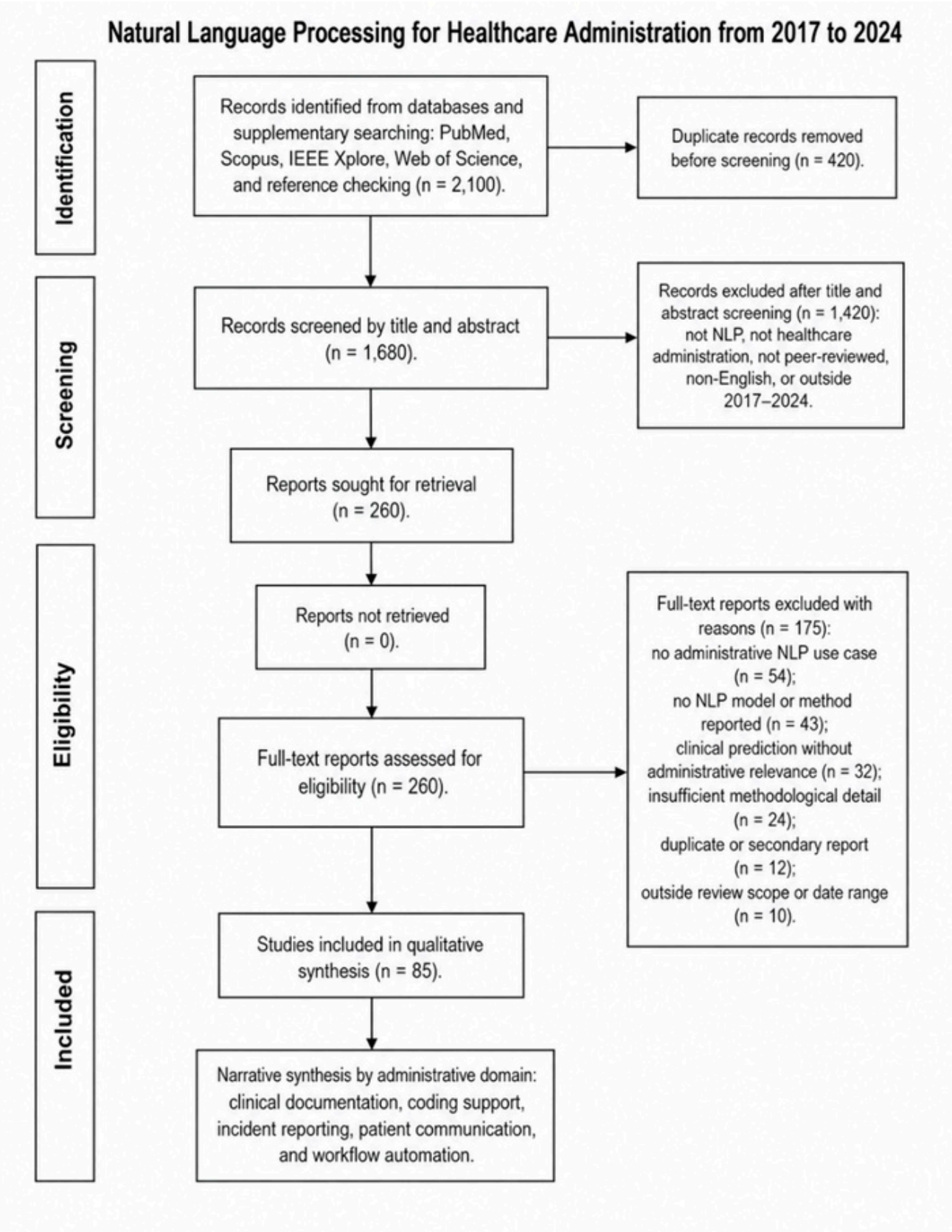


Figure 1. PRISMA 2020 Flow Diagram for Study Selection in the Systematic Review of Natural Language Processing for Healthcare Administration

Data extraction

A structured extraction form captured publication year, country or region, healthcare setting, administrative domain, text data source, NLP task, model family, evaluation metrics, validation approach, and deployment status. For coding studies, extracted variables included diagnostic or billing code targets, label structure, explainability approach, and whether the model addressed ICD-10, diagnosis-related groups, or broader coding workflow. For documentation and communication studies, extracted variables included whether the system summarized encounters, structured patient information, supported digital scribe functions, or processed portal messages. For incident reporting and workflow automation studies, extraction focused on adverse-event detection, contributing-factor classification, safety narrative analysis, medical necessity support, and administrative report generation.

Risk of bias assessment

Risk of bias and methodological quality were assessed using an adapted framework informed by prediction-model appraisal and qualitative evidence assessment. The appraisal considered data representativeness, annotation quality, internal and external validation, handling of class imbalance, reporting transparency, and whether performance claims were linked to realistic administrative workflows. Particular attention was given to whether studies used curated datasets that may not reflect noisy operational settings, because administrative NLP often depends on heterogeneous documentation practices, local coding policies, and site-specific workflows. Studies reporting only retrospective evaluation without prospective or user-centered validation were treated as having limited implementation evidence, even when their technical methods were well described.

Synthesis methods

A narrative synthesis was conducted because the included studies varied widely in administrative domain, task definition, model type, dataset, validation design, and outcome reporting. Studies were grouped into five domains: clinical documentation, coding support, incident reporting, patient communication, and workflow automation, with cross-cutting attention to transformer-based architectures and implementation maturity. Frequency counts were used descriptively to summarize model families and deployment status, but no meta-analysis was performed because metrics such as F1 score, ROUGE, BERTScore, accuracy, and coding precision were not consistently reported across comparable tasks. The synthesis drew on established clinical NLP reviews and task-specific studies to interpret how administrative NLP is evolving from extraction and classification toward generation, summarization, and operational decision support.

Results and Discussion

Study selection

The PRISMA selection process resulted in 85 included studies from 2,100 identified records after duplicate removal, title and abstract screening, and full-text assessment. Exclusions most often occurred because studies addressed clinical prediction without NLP, biomedical text mining without administrative relevance, or documentation systems without model evaluation. The final evidence base included studies on digital scribes and documentation summarization, automated coding, incident-report classification, patient-message analysis, and early workflow automation. This distribution was consistent with prior evidence showing that clinical NLP is well developed for information extraction but less consistently evaluated for administrative implementation [1, 2, 4, 7].

Study characteristics

The included literature showed increasing publication activity from 2017 to 2024, with a visible shift from rule-based and conventional machine-learning approaches toward deep learning and transformer-based models. Most studies were conducted in hospital, academic medical center, or EHR-linked settings, while fewer addressed ambulatory administration, revenue-cycle workflows, or cross-institutional deployment. Several studies used retrospective clinical

notes, coding records, safety reports, or patient portal messages, and relatively few described prospective operational testing. The evidence base therefore reflected strong technical development but uneven administrative generalizability across settings and health systems [4, 9, 10, 13, 15].

Table 1 organizes the reviewed evidence by administrative NLP application domain, healthcare setting, data source, model family, target outcome, validation approach, practical use, and implementation readiness.

Table 1. Evidence Taxonomy and Application Matrix for Natural Language Processing in Healthcare Administration

Application domain	Healthcare setting	Main data sources	AI/model families	Target outcomes	Validation approach	Practical use	Implementation readiness
Clinical documentation: summarization and note generation	Hospitals, outpatient clinics, academic medical centers, EHR-linked documentation workflows	Clinical conversations, progress notes, discharge summaries, patient records, clinician-entered narrative text	Digital scribe architectures, clinical summarization models, transformer-based language models, GPT/T5-style generation, patient-information summarization systems	Reduced documentation burden, structured notes, improved continuity of care, condensed patient summaries, more usable clinical narratives	Mostly retrospective or prototype evaluation; some human review of generated or summarized text; limited prospective workflow testing	Assist clinicians by drafting or summarizing documentation while preserving human verification	Medium
Clinical documentation: quality and optimization	Post-acute care, hospital documentation systems, outpatient documentation review, administrative quality monitoring	Unstructured notes, copied-forward text, semi-structured templates, discharge documentation, documentation completeness indicators	Rule-based NLP, information extraction, clinical note structuring models, classical ML, clinical text representation models	Identification of missing documentation, narrative structuring, improved completeness, better downstream usability for coding and quality reporting	Retrospective note analysis and qualitative assessment; limited direct evaluation of documentation burden or staff workload	Support documentation improvement, quality monitoring, and preparation of cleaner text for coding or operational review	Medium
Coding support: ICD-10, diagnosis-related groups, and billing code prediction	Hospital coding departments, revenue-cycle operations, diagnosis-related group assignment, EHR-linked	Clinical notes, discharge summaries, diagnosis fields, coding records, billing documentation, ICD code labels	CNN models, hierarchical attention networks, label-wise attention, retrieval-and-reranking models,	Suggested ICD codes, improved coding support, coder prioritization, coding consistency, explainable	Mostly retrospective multi-label classification using benchmark or institutional datasets; limited	Assist human coders with candidate codes, documentation review, and audit preparation	High

	coding workflows		transformer-based classifiers, explainable coding models	code assignment	prospective production evaluation		
Coding support: medical necessity and compliance	Revenue-cycle management, payer documentation, prior authorization support, billing compliance review	Medical necessity narratives, clinical justification text, payer forms, authorization documentation, diagnostic evidence	LLM-based extraction, multi-agent administrative automation, retrieval-supported generation, supervised coding support systems	Structured medical necessity justification, denial-risk support, evidence extraction, compliance documentation, prior authorization preparation	Early-stage proof-of-concept and retrospective evaluation; limited real-world administrative outcome testing	Assist staff in compiling evidence and drafting justification while maintaining human approval	
Incident reporting: event detection and classification	Patient safety offices, hospital risk management, quality improvement departments, safety surveillance systems	Free-text incident reports, adverse-event narratives, patient safety reports, event descriptions, surveillance text	Supervised classifiers, probabilistic language models, rule-based extraction, NLP classification pipelines, transformer-supported text classification	Adverse-event detection, safety report categorization, incident signal identification, structured safety analytics	Retrospective classification and extraction studies; limited linkage to real-time safety intervention	Support safety teams by organizing narrative reports and identifying recurring safety-event categories	
Incident reporting: root-cause and contributing-factor analysis	Quality improvement teams, patient safety committees, risk management units, incident review boards	Root-cause analysis documents, contributing-factor narratives, free-text safety event reports, action-plan notes	NLP topic classification, contributing-factor classifiers, clinical text mining, supervised and semi-supervised models	Identification of contributing factors, thematic safety analysis, prioritization of quality-improvement actions, structured RCA review	Retrospective narrative analysis; limited prospective testing of whether extracted themes lead to implemented action	Assist reviewers in identifying recurrent contributory patterns and organizing RCA documentation	
Patient communication: message triage and summarization	Patient portals, ambulatory care, primary care, specialty clinics,	Patient portal messages, patient requests, free-text concerns,	Pretrained language models, message classifiers,	Message routing, primary-concern identification,	Retrospective corpus analysis and model validation;	Support triage teams by routing messages, summarizing	

	centralized message pools	administrative messages, clinician-patient communication threads	concern-detection models, summarization systems, triage NLP pipelines	reduced inbox burden, patient-request summarization, escalation support	limited prospective testing of staff workload, access, or patient safety outcomes	concerns, and identifying administrative or clinical urgency
Patient communication: chatbots and conversational agents	Patient access centers, appointment services, administrative helpdesks, portal support, outpatient communication systems	Patient questions, appointment requests, FAQ content, portal messages, administrative instructions	Conversational AI, chatbot systems, GPT-style dialogue models, retrieval-supported response generation, intent classification	Appointment support, FAQ response, patient-friendly explanation, administrative navigation, communication efficiency	Often conceptual, prototype-based, or indirectly evaluated; limited rigorous real-world testing in administrative settings	Provide first-line administrative communication with escalation to human staff
Workflow automation: prior authorization and administrative report generation	Prior authorization teams, utilization management, operational reporting, administrative decision-support units	Prior authorization forms, payer documentation, operational reports, medical necessity statements, EHR-derived text	LLM-based administrative automation, multi-agent systems, extraction-generation pipelines, retrieval-augmented methods	Faster preparation of authorization materials, structured evidence extraction, report drafting, administrative task prioritization	Early proof-of-concept studies; sparse prospective evidence on turnaround time, denial rates, or staffing burden	Assist administrative staff with evidence compilation, drafting, and task organization
Cross-domain administrative NLP	Health-system operations, integrated EHR and revenue-cycle environments, enterprise analytics platforms	Combined documentation, coding, safety narratives, portal messages, operational reports, administrative logs	Modular NLP pipelines, transformer models, clinical language models, potential RAG and LLM orchestration, hybrid human-AI systems	Coordinated workflow support, cross-domain administrative intelligence, integrated documentation-coding-safety-communication signals	Rarely evaluated; mostly inferred from separate task-specific evidence	Future health-system infrastructure for linking documentation, coding, safety, communication, and administrative automation

Clinical documentation NLP: summarization and note generation

Clinical documentation NLP included digital scribe concepts, encounter summarization, clinical conversation processing, and patient-information summarization. Several studies and reviews described digital scribes as tools intended to reduce documentation burden by capturing clinical encounters and transforming them into structured or semi-structured notes [3-

5]. Scoping evidence on patient-information summarization further indicated growing interest in condensing fragmented clinical records into usable summaries for clinicians and administrative workflows [6]. However, the reviewed evidence also showed that many documentation systems remained conceptual, retrospective, or limited by evaluation methods that did not fully measure real-world documentation burden.

Clinical documentation NLP: quality and optimization

Documentation-quality applications used NLP to structure unstructured notes, identify missing information, and improve the usability of clinical documentation for downstream administrative processes. Reviews of clinical note NLP emphasized the importance of standardizing unstructured information, especially when documentation supports coding, quality measurement, and continuity of care [1, 2, 16]. Several studies reported that clinical documentation systems must manage copy-forward text, inconsistent templates, narrative redundancy, and site-specific documentation norms, although these issues were not always evaluated as primary outcomes. The evidence suggests that documentation optimization remains an important but methodologically challenging subdomain because text quality is both a clinical and administrative construct.

Coding support NLP: ICD-10 and CPT prediction

Coding support was one of the most technically mature areas of administrative NLP, with studies applying convolutional networks, hierarchical attention, transformer-based methods, and retrieval-based strategies to assign diagnosis or billing codes from clinical text. Automated coding studies addressed ICD code prediction, ICD-10 assignment, diagnosis-related group support, and supervised coding systems in hospital environments [15, 17-22]. Several studies reported that model architectures increasingly incorporate label hierarchies, explainable attention mechanisms, or retrieval and reranking to address the complexity of multi-label coding [17-20]. Nevertheless, most evidence remained retrospective, and the administrative implications of coding errors, payer requirements, and compliance review were not consistently evaluated.

Coding support NLP: medical necessity and compliance

A smaller subset of studies addressed medical necessity, billing justification, and administrative compliance rather than code prediction alone. Medical necessity automation was represented by emerging systems that used language models or multi-agent approaches to prepare or justify administrative documentation for coverage decisions [11]. Related coding studies suggested that extracting evidence from clinical text may help support billing review, diagnosis-related grouping, and denial prevention, but these functions were rarely evaluated as end-to-end administrative workflows [15, 22]. The evidence therefore indicates that compliance-oriented NLP is promising but still less mature than retrospective ICD coding models.

Incident reporting NLP: event detection and classification

Incident reporting NLP focused on extracting adverse events, classifying safety narratives, and identifying signals from free-text patient safety reports. A systematic review of incident reporting and adverse-event analysis found that NLP was frequently used for classification tasks, but study designs and reporting quality varied across settings [7]. Subsequent studies applied supervised learning, probabilistic language models, and NLP-based surveillance methods to categorize safety events and detect incidents from textual data [8, 23, 24]. The reviewed evidence suggests that incident reporting NLP can structure large volumes of safety narratives, but it remains more commonly retrospective than embedded in real-time safety improvement workflows.

Incident reporting NLP: root-cause and contributing factor analysis

Root-cause and contributing-factor analysis was addressed by studies using NLP to identify themes, contributing conditions, and action-relevant categories from incident narratives. A natural language processing approach to patient safety event reports categorized contributing factors, supporting the administrative task of turning unstructured safety

reports into analyzable categories [8]. Additional work on critical incident reports demonstrated that NLP can support structured analysis of safety narratives, although the evidence base remains smaller than that for coding or documentation [25]. Across studies, common barriers included local terminology, sparse labels, inconsistent reporting practices, and the challenge of linking extracted factors to actual quality-improvement action.

Patient communication NLP: message triage and summarization

Patient communication NLP included portal message analysis, primary-concern identification, triage support, and summarization of patient-generated text. One study used NLP to analyze patient messages at corpus scale, while another used pretrained language models to uncover patient primary concerns in portal messages [9, 10]. These studies indicate that patient communication data contain operationally important signals related to urgency, routing, administrative requests, and care coordination. However, the reviewed evidence also suggests that automated message handling requires careful evaluation because errors can affect access, patient trust, and staff workload.

Patient communication NLP: chatbots and conversational agents

Conversational agents and chatbot-related systems were discussed as tools for administrative communication, appointment support, frequently asked questions, and patient-facing explanation. Although some reviewed studies centered on patient messages rather than deployed chatbots, the broader literature on digital scribes and administrative automation indicates increasing interest in generative and conversational language interfaces [3, 12]. Patient communication systems differ from internal administrative NLP because outputs are visible to patients and may influence comprehension, expectations, and perceived empathy. As a result, systematic evaluation should include not only task completion but also safety, tone, escalation behavior, and human oversight [9, 10].

Workflow automation NLP: prior authorization and report generation

Workflow automation was the least mature but rapidly emerging administrative NLP domain in the reviewed evidence. Studies addressing medical necessity justification and administrative task automation suggested that large language models may assist with preparing structured documentation, extracting required evidence, and generating administrative reports [11, 12]. These applications are closely related to coding support because both depend on connecting clinical text to payer, compliance, and operational requirements [15, 22]. However, the available studies provided limited evidence on whether NLP systems reduce turnaround time, denial rates, staff burden, or administrative cost in prospective healthcare settings.

NLP model types and architectures

Model architectures shifted substantially during the review period, moving from feature-based and neural classification systems toward pretrained biomedical and clinical language models. BioBERT, ClinicalBERT, and domain-specific language model pretraining studies provided foundations for transformer-based clinical NLP, while transfer-learning evaluations showed the relevance of pretrained models across biomedical and clinical tasks [26–29]. Embedding surveys and comparative studies further demonstrated that representation learning became central to NLP model development in healthcare text processing [13, 14]. In administrative NLP, these architectures supported coding, documentation, patient-message classification, and emerging generative workflows, although deployment evidence remained uneven.

Evaluation and real-world deployment

Evaluation metrics varied by task and included classification measures, ranking measures, text similarity measures, and qualitative review of generated outputs. Coding studies commonly used multi-label classification metrics, documentation studies often considered summarization or note-quality measures, and incident reporting studies emphasized classification and extraction outcomes [6, 7, 15, 17–22]. Few studies reported prospective operational evaluation, user-centered outcomes, or direct measures of administrative burden reduction. This gap was especially important for digital

scribes, patient-message tools, and workflow automation systems, where usability, trust, accountability, and workflow fit may be as important as model output quality [3-5, 9, 10, 12].

Figure 2 summarizes the evidence-to-implementation pathway linking administrative text sources, NLP model families, healthcare application domains, target outcomes, validation maturity, governance risks, and future research priorities.

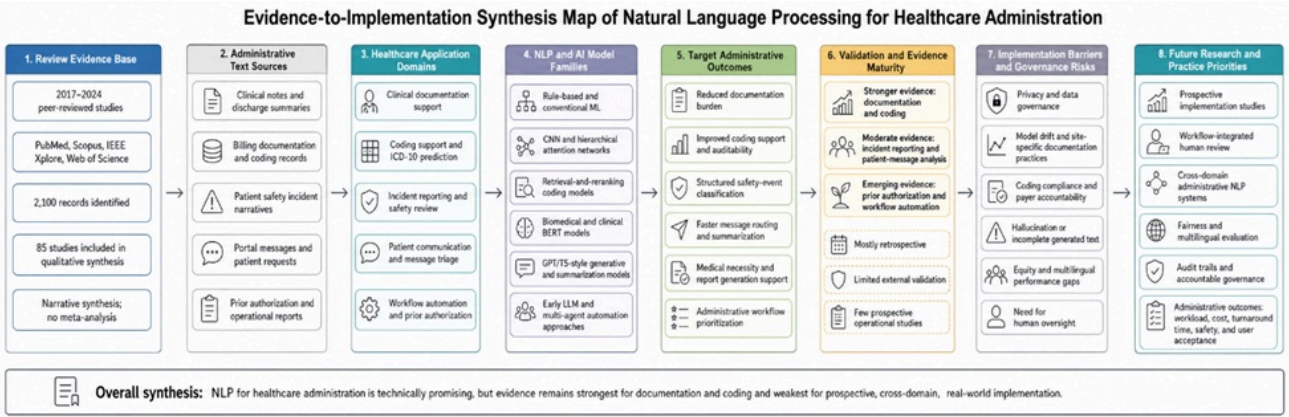


Figure 2. Evidence-to-Implementation Synthesis Map of Natural Language Processing for Healthcare Administration

Clinical documentation leads in maturity and volume

Clinical documentation emerged as one of the most visible and operationally relevant areas of administrative NLP. Digital scribe research, documentation summarization, and clinical note structuring reflect direct attempts to reduce documentation burden and improve the usability of clinical text [3-6, 16]. The evidence suggests that documentation NLP has advanced beyond simple extraction toward summarization and generation, but validation remains uneven across settings. Even when tools are conceptually mature, studies frequently emphasize the need for careful integration into clinical documentation workflows and clinician review [4, 5].

Automated coding is reliable in retrospect, uncertain in production

Automated coding studies showed substantial methodological development, including hierarchical attention, convolutional neural networks, retrieval and reranking, explainable coding models, and ICD-10 assignment systems [15, 17-22]. These studies indicate that coding support is technically advanced compared with many other administrative NLP domains. However, retrospective coding benchmarks do not fully capture production challenges such as payer-specific rules, audit risk, changing documentation standards, and institutional coding practices. The evidence therefore supports cautious interpretation: automated coding may assist human coders, but the reviewed literature does not establish fully autonomous coding in routine healthcare administration.

Incident reporting NLP remains retrospective and siloed

Incident reporting NLP has demonstrated value for classifying safety reports, identifying adverse events, and extracting contributing factors from narrative text. The included studies show that safety narratives can be processed using NLP methods to support categorization, surveillance, and incident analysis [7, 8, 23-25]. Nevertheless, most evidence remains retrospective and focused on organizing existing reports rather than enabling real-time safety response. A key limitation is that incident reporting models are rarely connected to operational quality-improvement systems, action tracking, or feedback mechanisms that would demonstrate measurable administrative impact.

Patient communication NLP balances efficiency and empathy

Patient communication NLP addresses a growing administrative challenge as portal messages and digital communication channels increase staff workload. Studies using NLP and pretrained language models to identify patient concerns suggest that automated triage and summarization can help organize message volume and support routing [9, 10]. However, patient-facing or patient-adjacent NLP requires more than technical classification because communication quality, tone, escalation, and trust are central to safe implementation. The reviewed evidence indicates that message triage systems should be evaluated with both operational metrics and patient-centered safeguards.

Workflow automation is the newest frontier

Workflow automation, including prior authorization support, medical necessity justification, and administrative report generation, represents a newer frontier for healthcare NLP. Emerging studies describe large language model-based systems and multi-agent approaches for administrative task automation, but the evidence base remains small and early-stage [11, 12]. These systems may eventually connect documentation, coding, payer requirements, and operational reporting into a more integrated administrative pipeline. At present, however, the reviewed literature provides limited evidence on prospective effectiveness, safety, compliance, or economic outcomes.

Cross-domain integration is missing

Most reviewed models were designed for a single administrative task rather than integrated administrative workflows. Documentation models rarely linked outputs directly to coding support, incident reporting models rarely connected with documentation-quality tools, and patient-message systems were generally separate from broader workflow automation [3-6, 9, 10, 15, 16]. This separation limits the ability of NLP to address administrative burden as a system-level problem. Cross-domain models could theoretically connect clinical documentation, billing evidence, safety signals, patient communication, and workflow routing, but the reviewed literature shows that such integration remains underdeveloped.

Implementation and trust barriers

Implementation barriers included privacy constraints, governance uncertainty, workflow disruption, model drift, lack of external validation, and limited transparency of generated or classified outputs. Studies of clinical documentation and digital scribes emphasized usability and clinician trust, while coding and incident reporting studies highlighted explainability, auditability, and local adaptation [3-5, 8, 17]. Transformer-based and large language model systems add further concerns because generated text may be fluent but still incomplete, inaccurate, or misaligned with administrative requirements [12, 26-29]. The evidence supports human oversight as a central requirement for administrative NLP, especially in coding, safety, and patient communication contexts.

Limitations

Review limitations

This review was limited to English-language peer-reviewed publications from 2017 to 2024 and may have missed relevant gray literature, vendor evaluations, internal health-system reports, or non-English studies. Heterogeneity in task definitions, data sources, model architectures, and evaluation metrics prevented meta-analysis and required narrative synthesis. The included studies also varied in whether they addressed healthcare administration directly or contributed foundational methods relevant to administrative NLP. These limitations are consistent with prior reviews of clinical NLP, which have noted variability in reporting standards, validation methods, and generalizability across institutions [1, 2, 4, 13, 14].

Comparison with prior reviews

Prior reviews of healthcare NLP have generally focused on broad clinical text processing, biomedical embeddings, chronic disease note extraction, or documentation-specific applications rather than healthcare administration as an integrated operational domain. Reviews of clinical notes and unstructured information extraction have shown that NLP

can standardize and structure clinical text, but their primary emphasis was often on clinical information capture rather than administrative use across coding, incident review, communication, and automation [1, 2]. Similarly, reviews of embeddings and clinical NLP model representations provided important technical foundations but did not organize the evidence around administrative workflow needs [13, 14]. This systematic review differs by treating administrative text as a cross-cutting operational resource rather than as a single clinical data source.

This review uniquely links clinical documentation, coding support, incident reporting, patient communication, and workflow automation as parts of a broader administrative pipeline. Documentation systems may generate or structure text that later supports coding, coding systems may depend on documentation completeness, incident reporting systems may reveal safety and quality problems, and patient communication systems may generate administrative routing demands [3-6, 9, 10, 15, 16]. Prior reviews of digital scribes and patient-information summarization have emphasized documentation burden, while incident reporting and coding studies have usually remained task-specific [4, 6-8, 15, 17-22]. By synthesizing these domains together, the present review highlights how administrative NLP could move from isolated model development toward coordinated health-system operations.

A central distinction from prior reviews is the emphasis on the gap between technical performance and administrative utility. Several studies reported technically advanced models for ICD coding, clinical text representation, incident classification, and message analysis, but fewer evaluated whether these systems improved throughput, reduced burden, prevented errors, or supported accountable administrative decisions [7, 9, 10, 13-15, 17-22, 26-29]. This distinction is particularly important for generative and transformer-based systems, where fluent output may not guarantee compliance, accuracy, or workflow safety [12, 26-29]. The reviewed evidence therefore suggests that future reviews should assess implementation maturity and operational value alongside conventional NLP metrics.

Table 2 summarizes the major implementation gaps, governance risks, future research priorities, and practical implications for deploying NLP systems in healthcare administration.

Table 2. Implementation Gaps, Governance Risks, and Future Research Agenda for Administrative Natural Language Processing in Healthcare

Recurring limitation or gap	Practical consequence for healthcare administration	Safety or governance concern	Recommended future research direction	Implementation implication
Heavy reliance on retrospective datasets	Models may perform well on curated historical text but fail in noisy, real-time administrative workflows	Retrospective performance may overestimate operational reliability	Conduct prospective, pragmatic implementation studies in live documentation, coding, safety, and communication workflows	Deploy tools gradually with monitoring, staff feedback, and rollback procedures
Limited external validation across institutions	Models may not generalize across hospitals, specialties, EHR systems, payer policies, or documentation cultures	Site-specific drift may produce inaccurate codes, summaries, classifications, or patient-message routing	Require multi-site validation and subgroup analysis across settings, specialties, and administrative practices	Avoid broad deployment based on single-center evidence

Heterogeneous metrics and task definitions	Results are difficult to compare across documentation, coding, incident reporting, communication, and automation studies	Inconsistent reporting may obscure clinically or administratively meaningful failure modes	Develop domain-specific reporting standards for administrative NLP outcomes	Journals and health systems should require standardized reporting of task, user, workflow, and intended use
Weak measurement of administrative outcomes	Studies often report model metrics rather than workload, cost, turnaround time, denial reduction, or staff burden	Tools may add hidden review burden despite appearing technically successful	Measure operational outcomes such as time saved, documentation burden, coder workload, message backlog, and authorization turnaround	Procurement and adoption decisions should depend on administrative outcome evidence
Limited human-centered evaluation	Models may be misaligned with coder, clinician, safety-team, or administrative staff workflows	Poor usability can reduce trust, increase workarounds, and produce unsafe overreliance	Use mixed-methods evaluation with end users, including workflow observation, interviews, and usability testing	Administrative NLP should be co-designed with frontline users and governance teams
Incomplete auditability of NLP outputs	Generated notes, suggested codes, or authorization text may be difficult to trace back to source evidence	Lack of source attribution can create legal, billing, and compliance risk	Build audit trails linking model outputs to source text, user edits, confidence indicators, and final decisions	Human reviewers must be able to verify, contest, and correct NLP outputs
Hallucination and incomplete generated text	Generative models may produce fluent but inaccurate documentation, summaries, patient responses, or administrative reports	Errors may affect reimbursement, care continuity, patient understanding, or safety escalation	Evaluate factual consistency, omission risk, source grounding, and human correction rates	Use generative systems only with source-grounding and mandatory human review in high-risk tasks
Coding compliance and payer accountability	Automated coding suggestions may conflict with documentation rules, payer policies, or audit expectations	Incorrect codes can produce reimbursement errors, compliance exposure, and inequitable billing consequences	Study coder-in-the-loop systems, denial outcomes, audit discrepancy rates, and payer-policy sensitivity	Coding NLP should remain assistive unless validated under compliance and audit conditions
Limited linkage between incident-report NLP and quality improvement action	Safety-report classifiers may organize narratives without changing safety response or prevention	Extracted signals may not translate into accountability or implemented corrective actions	Evaluate whether NLP outputs improve event review timeliness, RCA completeness, action tracking, and safety learning	Incident-report NLP should be integrated with safety governance and action-management systems

Risk of missed urgency in patient-message triage	Misrouted or incorrectly summarized messages may delay care, access, or administrative response	Patient safety, trust, equity, and service quality may be affected	Test message-triage systems prospectively with escalation audits and patient-centered outcome measures	Patient communication NLP needs conservative escalation thresholds and human oversight
Limited evidence for workflow automation	Prior authorization and report-generation tools are promising but not yet supported by strong operational evidence	Automation may create compliance, privacy, and accountability concerns if adopted prematurely	Measure authorization turnaround, denial rates, staff burden, documentation quality, and error correction	Workflow automation should begin with narrow, supervised pilots and clear accountability
Fragmentation across administrative domains	Documentation, coding, communication, safety, and workflow automation are usually studied separately	Siloed tools may duplicate work, create inconsistent records, or miss cross-domain signals	Develop interoperable NLP systems that link documentation quality, coding evidence, safety signals, and communication demands	Health systems should plan administrative NLP as infrastructure, not as isolated point solutions
Privacy and data governance constraints	Administrative NLP often requires access to sensitive clinical, billing, safety, and patient-message text	Unauthorized exposure or secondary use of administrative text may create legal and ethical risk	Evaluate privacy-preserving NLP, secure deployment, access controls, and data minimization strategies	Governance boards should review data flows, retention, monitoring, and vendor access before deployment
Fairness and multilingual under-evaluation	Models may perform unevenly across language groups, health literacy levels, patient populations, and institutional contexts	Automated triage, coding, or documentation may worsen inequities if errors cluster in underrepresented groups	Conduct fairness, subgroup, and multilingual evaluation for coding, messaging, documentation, and authorization tasks	Administrative NLP should include equity monitoring and multilingual support before scale-up
Lack of clear responsibility for final decisions	Staff may be uncertain whether model output, vendor logic, clinician review, or administrative approval is authoritative	Accountability gaps can arise in documentation, billing, safety review, and patient communication	Define responsibility frameworks for human review, override, correction, and escalation	Implementation policies should specify who reviews outputs, who approves final actions, and how errors are handled

Research gaps

Prospective evaluation of administrative NLP

A major research gap is the lack of controlled prospective evidence showing that administrative NLP reduces workload, cost, delay, or error in routine healthcare operations. Many coding, documentation, incident reporting, and communication

studies evaluated models retrospectively on existing datasets, while relatively few examined real-time use by coders, clinicians, safety teams, or administrative staff [3-7, 9, 10, 15, 17-22]. Workflow automation studies were especially early-stage and did not yet establish measurable effects on prior authorization turnaround, medical necessity review, report completion, or administrative staffing burden [11, 12]. Future studies should include pragmatic designs that measure implementation outcomes under real operational constraints.

Integration across administrative domains

Another gap is the absence of integrated NLP systems that connect documentation improvement, coding support, incident reporting, patient communication, and workflow automation. Current studies typically address one task at a time, such as ICD prediction, portal-message concern detection, or safety-report classification [7, 9, 10, 15, 17-22]. Yet administrative work is interconnected because documentation quality affects coding, coding affects billing and compliance, patient messages affect workload, and incident narratives affect quality governance [3-6, 8, 15, 16]. Research should therefore explore modular but interoperable NLP systems that preserve accountability while allowing administrative signals to move across health-system functions.

Fairness and multilingual administrative NLP

Fairness and multilingual support remain underdeveloped in administrative NLP. Patient-message triage, automated coding, and documentation generation may perform differently across language groups, health literacy levels, clinical specialties, and institutions, but these subgroup risks were not consistently evaluated in the reviewed literature [9, 10, 15, 17, 21, 22]. Transformer-based clinical language models provide powerful representations, yet they may reproduce biases embedded in training data, documentation habits, or institutional coding practices [26-29]. Future research should evaluate fairness in administrative outcomes, including message escalation, billing classification, medical necessity support, and access to timely administrative services.

Implications

For research practice

For research practice, the findings imply that administrative NLP should move from benchmark-centered evaluation toward implementation science. Studies should combine model metrics with workflow outcomes, qualitative user feedback, failure-mode analysis, and site-to-site generalizability testing [3-8, 10, 15, 17-22]. Reporting should make clear whether an NLP system is intended to extract information, classify administrative text, summarize records, generate documentation, or support a human decision. This distinction is essential because the risks of a wrong code, an omitted safety signal, a misleading patient response, or an inaccurate generated note are operationally different [3, 8, 10, 11, 15, 17].

For administrative practice

For administrative practice, current NLP tools should be treated as assistive systems rather than replacements for professional judgment. Coding support models may help prioritize review or suggest likely codes, but coders and compliance teams must remain responsible for final billing decisions [15, 17-22]. Incident reporting and patient-message tools may help organize high-volume text streams, but safety teams and clinical staff should review escalations and ambiguous outputs [7-10, 23-25]. Documentation and workflow automation tools should similarly include human verification, because generated administrative text can influence care continuity, reimbursement, legal records, and patient trust [3-6, 11, 12, 16].

For policy

For policy, governance of administrative NLP must address accuracy, privacy, auditability, fairness, and financial accountability. Automated coding, medical necessity support, and documentation generation may affect reimbursement and compliance, making transparent evidence trails and review responsibility essential [11, 12, 15, 22]. Patient

communication and incident reporting tools also require safeguards because errors can affect access, safety recognition, and institutional response to harm [7-10, 23-25]. Policymakers should therefore encourage standards that require validation in operational settings, monitoring for drift, documentation of intended use, and mechanisms for contesting or correcting NLP-generated administrative outputs [4, 5, 26-29].

Conclusion

Natural language processing has made significant progress in automating and augmenting healthcare administrative tasks, with clinical documentation and automated coding leading the field. These areas have the strongest technical evidence and the clearest connection to everyday administrative burden. However, technical maturity does not automatically translate into safe or effective operational deployment. Human oversight remains necessary wherever NLP outputs influence documentation, billing, safety review, or patient-facing communication.

Incident reporting, patient communication, and workflow automation are rapidly emerging but remain less mature and less consistently evaluated. These domains show strong potential because they address large volumes of free-text administrative data that are difficult to manage manually. At the same time, their risks are substantial because misclassification, poor escalation, or inaccurate generated text can affect safety, access, and trust. The evidence therefore supports cautious, staged implementation rather than broad automation.

The critical shortcoming across the field is the limited availability of prospective, real-world evaluation. Most studies remain retrospective, task-specific, and focused on model outputs rather than administrative outcomes. Future research should measure workload, cost, turnaround time, quality, fairness, user acceptance, and downstream consequences. Integrated evaluation is especially important as transformer-based and generative systems become more capable and more widely available.

A focused agenda on implementation science, fairness, and cross-domain integration is needed to realize the potential of NLP in healthcare administration. The next phase of research should connect technical development with governance, workflow design, and accountable human review. Administrative NLP should be evaluated as a health-system intervention, not merely as a text-processing task. With careful validation and oversight, NLP can support more efficient, transparent, and responsive healthcare administration.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Published online: 25 February 2025

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

- Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform.* 2019;7(2):e12239.
- Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform.* 2017;73:14-29.
- Coiera E, Kocaballi AB, Halamka J, Laranjo L. The digital scribe. *NPJ Digit Med.* 2018;1(1):58.
- Van Buchem MM, Boosman H, Bauer MP, Kant IM, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med.* 2021;4(1):57.
- Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med.* 2019;2(1):114.
- Keszthelyi D, Gaudet-Blavignac C, Bjelogrić M, Lovis C. Patient information summarization in clinical settings: scoping review. *JMIR Med Inform.* 2023;11(1):e44639.
- Young IJ, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform.* 2019;132:103971.
- Tabaie A, Sengupta S, Pruitt ZM, Fong A. A natural language processing approach to categorise contributing factors from patient safety event reports. *BMJ Health Care Inform.* 2023;30(1):e100731.
- Tafti Ahmadi P. Probing patient messages enhanced by natural language processing: a top-down message corpus analysis. *Health Data Sci.* 2021;2021:9816723.
- Ren Y, Wu Y, Fan JW, Khurana A, Fu S, Wu D, et al. Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. *J Am Med Inform Assoc.* 2024;31(8):1714-24.
- Pandey HG, Amod A, Kumar S. Advancing healthcare automation: multi-agent system for medical necessity justification. In: *Proc 23rd Workshop Biomed Nat Lang Process*; 2024. p. 39-49.
- Gebreab SA, Salah K, Jayaraman R, ur Rehman MH, Ellaham S. LLM-based framework for administrative task automation in healthcare. In: *12th Int Symp Digit Forensics Secur (ISDFS)*; 2024. p. 1-7.
- Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform.* 2018;87:12-20.
- Kalyan KS, Sangeetha S. SECNLP: A survey of embeddings in clinical natural language processing. *J Biomed Inform.* 2020;101:103323.

Dai HJ, Wang CK, Chen CC, Liou CS, Lu AT, Lai CH, et al. Evaluating a natural language processing-driven, AI-assisted International Classification of Diseases, 10th Revision, Clinical Modification, coding system for diagnosis related groups in a real hospital environment: algorithm development and validation study. *J Med Internet Res.* 2024;26:e58278.

Scharp D, Hobensack M, Davoudi A, Topaz M. Natural language processing applied to clinical documentation in post-acute care settings: a scoping review. *J Am Med Dir Assoc.* 2024;25(1):69-83.

Dong H, Suárez-Paniagua V, Whiteley W, Wu H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J Biomed Inform.* 2021;116:103728.

Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. *Proc AAAI Conf Artif Intell.* 2020;34(5):8180-7.

Niu K, Wu Y, Li Y, Li M. Retrieve and rerank for automated ICD coding via contrastive learning. *J Biomed Inform.* 2023;143:104396.

Hu S, Teng F, Huang L, Yan J, Zhang H. An explainable CNN approach for medical codes prediction from clinical text. *BMC Med Inform Decis Mak.* 2021;21(Suppl 9):256.

Sun Y, Sang L, Wu D, He S, Chen Y, Duan H, et al. Enhanced ICD-10 code assignment of clinical texts: a summarization-based approach. *Artif Intell Med.* 2024;156:102967.

Chen PF, He TL, Lin SC, Chu YC, Kuo CT, Lai F, et al. Training a deep contextualized language model for International Classification of Diseases, 10th Revision classification via federated learning: model development and validation study. *JMIR Med Inform.* 2022;10(11):e41342.

Ozonoff A, Milliren CE, Fournier K, Welcher J, Landschaft A, Samnaliev M, et al. Electronic surveillance of patient safety events using natural language processing. *Health Informatics J.* 2022;28(4):14604582221132429.

Walsh CG, Wilimitis D, Chen Q, Wright A, Kolli J, Robinson K, et al. Scalable incident detection via natural language processing and probabilistic language models. *Sci Rep.* 2024;14(1):23429.

Denecke K, Paula H. Analysis of critical incident reports using natural language processing. In: *dHealth 2024*; 2024. p. 1-6.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234-40.

Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: *Proc 2nd Clin Nat Lang Process Workshop*; 2019. p. 72-8.

Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proc 18th BioNLP Workshop Shared Task*; 2019. p. 58-65.

Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc.* 2021;3(1):1-23