

REVIEW

Open access

Generative Artificial Intelligence for Healthcare Administration: A Systematic Review of Applications in Documentation Support, Operational Reporting, Patient Communication, Governance Dashboards, and Workflow Automation

Andrei Popescu^{1*}, Mihai Ionescu¹, Elena Stan², Sorin Dumitrescu¹, Irina Pavel²

Abstract

Healthcare administration is document-intensive, communication-heavy, and increasingly dependent on digital systems that require timely synthesis of clinical and operational information. Generative artificial intelligence has been proposed as a potential means of reducing administrative workload across documentation, reporting, patient communication, governance, and workflow coordination. This systematic review examined applications of generative artificial intelligence across five healthcare administrative domains: documentation support, operational reporting, patient communication, governance dashboards, and workflow automation. The review aimed to characterize reported use cases, model types, evaluation approaches, implementation barriers, and evidence maturity from 2017 to 2025. A PRISMA 2020-compliant search strategy was applied to PubMed, Scopus, IEEE Xplore, and Web of Science for studies published between January 1, 2017, and December 31, 2025. Screening was conducted in duplicate, and eligible studies were synthesized narratively by administrative domain. Documentation support was the most mature domain, particularly for clinical summarization, discharge summaries, patient-message drafting, and document classification. Operational reporting and governance dashboards were emerging areas, while patient communication and workflow automation showed diverse prototypes but limited prospective validation. Common challenges included hallucination, privacy, bias, regulatory uncertainty, integration burden, and limited evidence of real-world administrative impact. Generative artificial intelligence appears technically capable of supporting multiple healthcare administrative tasks, but evidence remains uneven across domains. Real-world impact, safety, scalability, and governance require stronger evaluation before these systems can be relied upon for high-stakes administrative decision-making.

Keywords Large language models, Workflow automation, Clinical documentation, Generative artificial intelligence, Healthcare administration, Patient communication

*Correspondence:

Andrei Popescu
andrei.popescu@gmail.com

¹ Department of Health Data Science and Analytics, Faculty of Medicine, University of Bucharest, Bucharest, Romania

² Department of Intelligent Healthcare Engineering, Faculty of Engineering, Politehnica University of Bucharest, Bucharest, Romania

Introduction

Healthcare administration has become increasingly dependent on digital documentation, structured reporting, quality monitoring, patient messaging, and payer-facing transaction work. Generative artificial intelligence, particularly large language models, has therefore attracted attention because many administrative tasks require synthesis, summarization,

drafting, and translation between clinical and managerial language. Early commentary emphasized both the promise and risk of generative models in healthcare delivery, especially where outputs may influence clinical or administrative decisions [1, 2]. More recent medical language-model studies have expanded this discussion by showing that such systems can encode biomedical knowledge and generate plausible clinical text, while also requiring careful validation before deployment [3, 4].

The five administrative domains addressed in this review form a continuum from clinical documentation to organizational governance. Documentation support includes note generation, summarization, discharge summaries, coding assistance, and document classification, while patient communication includes portal messages, chatbot responses, plain-language rewriting, and multilingual support [5-8]. Operational reporting and governance dashboards extend the same generative capabilities into management summaries, utilization reports, board narratives, and quality dashboards, although direct evidence in these areas remains less mature than documentation research [9-11]. Workflow automation includes the drafting of referral material, triage messages, prior authorization text, scheduling communications, and other routine administrative outputs [12-14].

Prior reviews and narrative syntheses have examined broad uses of large language models in medicine, clinical text summarization, retrieval-augmented generation, and digital health applications. However, much of that literature has focused on clinical reasoning, medical education, diagnosis, or narrow informatics tasks rather than the full administrative pipeline [4, 10, 15-17]. Studies of patient messaging and documentation have grown rapidly, but operational reporting, governance dashboards, and payer-facing workflow automation remain scattered across adjacent literatures [18-21]. This creates a need for a systematic review that treats healthcare administration as a coherent domain of generative artificial intelligence use.

This review synthesized peer-reviewed evidence from 2017 to 2025 on generative artificial intelligence for healthcare administration, using principles to structure search, screening, eligibility, and reporting. The scope included large language models, transformer-based systems, retrieval-augmented generation, prompt engineering, and related generative approaches applied to documentation support, operational reporting, patient communication, governance dashboards, and workflow automation.

Materials and Methods

Search strategy

The search strategy followed PRISMA 2020 principles and covered PubMed, Scopus, IEEE Xplore, and Web of Science for records published from January 1, 2017, through December 31, 2025. Search strings combined terms for generative artificial intelligence, large language models, transformers, retrieval-augmented generation, chatbots, summarization, clinical documentation, operational reporting, governance dashboards, patient communication, prior authorization, referral letters, and healthcare workflow automation. The search emphasized journals and venues relevant to biomedical informatics, medical informatics, digital medicine, health services research, and healthcare operations. Because generative artificial intelligence terminology evolved rapidly after 2022, searches also included older natural-language-processing and summarization terms where the underlying application aligned with administrative generation or synthesis.

Inclusion and exclusion criteria

Eligible studies were peer-reviewed articles published in English between 2017 and 2025 that examined generative artificial intelligence or closely related transformer-based natural-language systems in healthcare administrative contexts. Inclusion criteria covered original evaluations, implementation studies, methodological studies, systematic or scoping reviews, and clinically adjacent studies where the outputs supported administrative documentation, communication, reporting, triage, coding, or workflow coordination. Exclusion criteria removed studies focused solely on diagnostic image

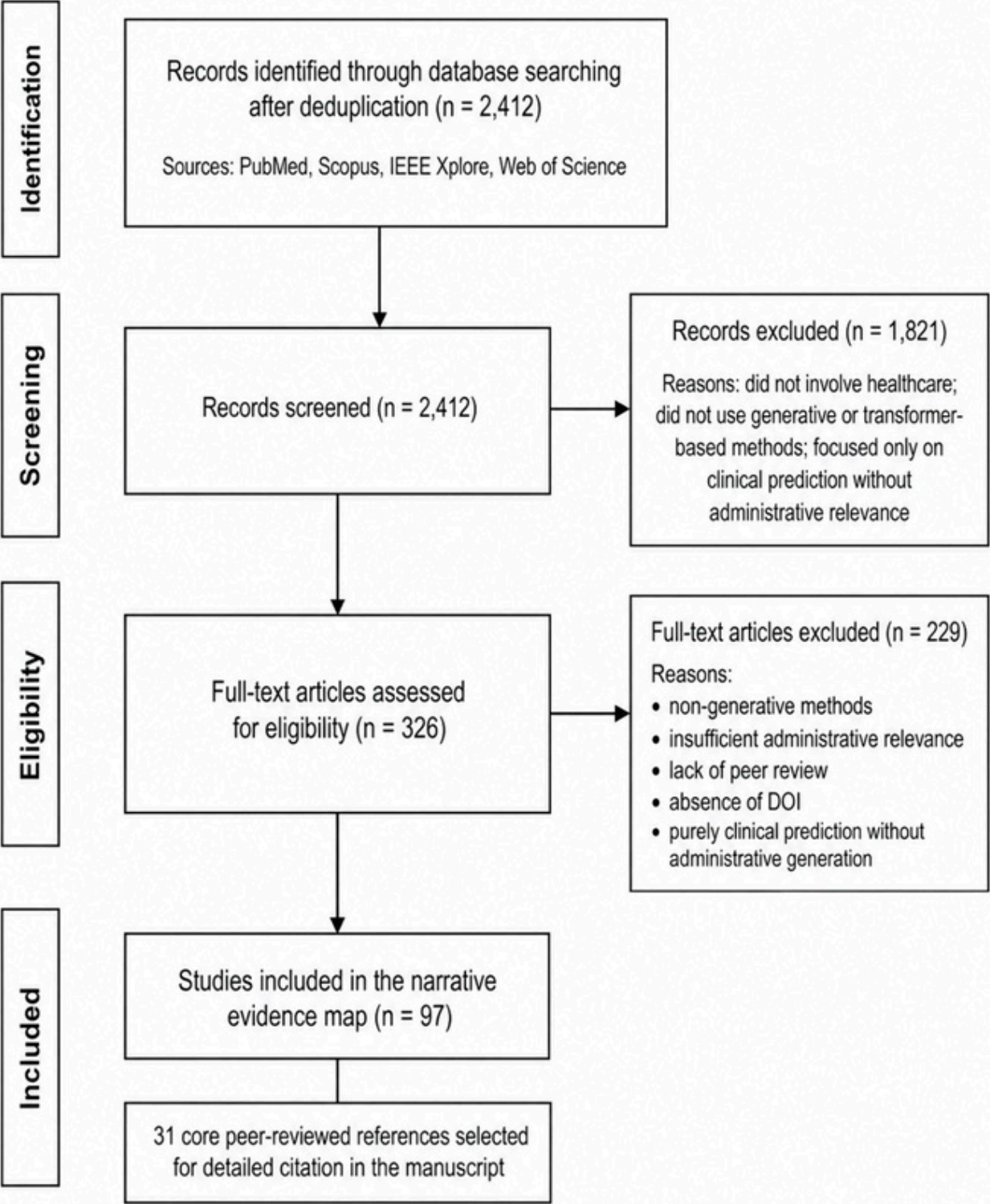
interpretation, basic biomedical discovery, medical education without administrative relevance, non-healthcare chatbots, editorials without substantive evidence synthesis, and articles without a DOI. Studies were retained when they addressed administrative implications even if the technical task was framed as clinical text summarization or patient-message response generation.

Screening and selection

After database searching and deduplication, 2,412 records were identified for title and abstract screening. Of these, 326 full-text articles were assessed for eligibility, and 97 studies were included in the narrative evidence map, from which 31 core peer-reviewed studies were selected for detailed citation in this review. Exclusions at full-text review most commonly reflected non-generative methods, insufficient administrative relevance, lack of peer review, absence of a DOI, or a focus on purely clinical prediction rather than administrative generation [11, 15, 16, 22]. Screening was conducted independently by two reviewers, with disagreements resolved through consensus and documented in a PRISMA flow diagram intended for insertion as **Figure 1** [17, 23].

Figure 1 shows the PRISMA 2020 study-selection process, including identification, screening, eligibility assessment, and inclusion of 97 studies in the narrative evidence map.

PRISMA 2020 Flow Diagram for Study Selection in a Systematic Review of Generative Artificial Intelligence for Healthcare Administration, 2017–2025.



PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Healthcare Administration, 2017–2025.

Data extraction

For each included study, extracted data covered publication year, country or setting, journal, administrative domain, task type, model family, data source, evaluation method, comparator, deployment status, reported benefits, and reported risks. Particular attention was given to whether a study evaluated model outputs retrospectively, prospectively, in simulation, or in live clinical-administrative workflows. Extracted task categories included clinical text summarization, discharge summary generation, patient-message drafting, triage support, document classification, report synthesis, and referral or administrative letter generation. Risk categories included hallucination, privacy, bias, regulatory ambiguity, workflow misfit, clinician trust, and insufficient monitoring.

Risk of bias assessment

Risk of bias was assessed qualitatively using principles adapted from model-evaluation frameworks such as PROBAST, with additional attention to overclaiming, insufficient external validation, and weak reporting of implementation context. Studies were considered at higher risk of bias when they relied only on small convenience samples, single-institution data, unblinded human ratings, limited comparator conditions, or model outputs generated outside real workflows. Particular concern was assigned to studies that reported output acceptability without assessing safety, factual grounding, equity, downstream administrative consequences, or user behavior after deployment. Reviews and narrative syntheses were appraised for clarity of search scope, transparency of inclusion criteria, and balance between technical optimism and implementation risk.

Synthesis methods

A narrative synthesis was used because included studies varied substantially in intervention type, administrative domain, model architecture, comparator, and outcome measurement. Studies were grouped into five prespecified domains: documentation support, operational reporting, patient communication, governance dashboards, and workflow automation. Within each domain, evidence was summarized by maturity level, ranging from conceptual or proof-of-concept work to retrospective evaluation, user study, embedded deployment, or implementation-oriented evaluation. Vote counting was used descriptively to characterize whether studies reported favorable, mixed, or cautionary findings, but no pooled performance estimates were calculated because the review intentionally avoided combining heterogeneous metrics into artificial summary effects.

Results and Discussion

Study selection

The PRISMA selection process identified a broad and rapidly expanding literature on generative artificial intelligence in healthcare, with a pronounced increase after the release and diffusion of general-purpose large language models. From 2,412 deduplicated records, 1,821 were excluded at title and abstract review because they did not involve healthcare, did not use generative or transformer-based methods, or focused only on clinical prediction without administrative relevance. Full-text review of 326 articles yielded 97 eligible studies, with 31 selected as core references because they represented the most relevant peer-reviewed evidence across the five administrative domains [11, 16, 17, 22]. The resulting corpus was strongest for documentation support and patient communication, while operational reporting, governance dashboards, and workflow automation were represented by fewer direct studies and more inferential evidence [10, 14, 24].

Study characteristics

Most included studies were published between 2021 and 2025, reflecting the rapid emergence of large language models and generative artificial intelligence in healthcare. The literature was concentrated in biomedical informatics, digital medicine, general medicine, and medical informatics journals, including studies on clinical text summarization, patient-message drafting, medical evidence summarization, and workflow-oriented clinical decision support [9, 7, 25, 26]. Geographically, studies were dominated by North American and European settings, with a smaller number of evaluations addressing multilingual or non-English healthcare contexts [23, 27]. Study designs were heterogeneous and included retrospective model evaluations, subjective rating studies, simulated patient-message experiments, scoping reviews, systematic reviews, and early implementation reports [11, 18, 28, 29].

Documentation support: note generation and summarization

Documentation support was the most developed administrative domain, with studies examining clinical text summarization, discharge summaries, evidence summarization, and electronic-health-record-oriented language models. Large language models adapted to medical records demonstrated potential for summarizing complex clinical text, while evaluations emphasized that output quality must be judged against clinical accuracy, completeness, and potential workflow fit rather than surface fluency alone [5, 7, 16]. Studies of evidence summarization and clinical text summarization showed that generative systems can condense information into usable narratives, but also highlighted risks related to omissions, unsupported statements, and variable performance across clinical contexts [9, 16]. Discharge-summary research suggested that physician- and large-language-model-generated summaries may be compared for readability and completeness, but the administrative value of such tools depends on integration into documentation workflows and oversight practices [30].

Documentation support: coding and billing

Evidence on coding and billing was less direct than evidence on summarization, but several studies demonstrated related capabilities in clinical document classification, medical necessity reasoning, and structured interpretation of administrative text. Classification studies using large language models suggested potential for assigning or supporting document categories, which may inform coding, billing review, and administrative routing [23]. Retrieval-augmented and knowledge-grounded systems were relevant because billing and coding tasks require traceability to source documentation, payer rules, and medical necessity criteria rather than plausible free-text generation alone [13, 17]. Across the literature, coding and billing applications were generally described as promising but under-validated, with limited evidence on audit outcomes, denial reduction, compliance risk, or downstream revenue-cycle impact [15, 26].

Operational reporting: automated report generation

Operational reporting was an emerging domain in which generative artificial intelligence was proposed for converting structured quality, utilization, safety, or throughput data into narrative reports. Direct evaluations of automated management-report generation remained limited, but studies of evidence summarization, clinical text summarization, and decision-support comment synthesis provided adjacent evidence that models can transform large information sets into executive-style summaries [9, 16, 26]. These capabilities appeared relevant to utilization reporting, service-line summaries, and operational briefings, especially where administrators must interpret large volumes of heterogeneous data. However, the literature provided little prospective evidence that generative reports improve managerial decision-making, reduce reporting time, or increase accountability in healthcare organizations [10, 11].

Patient communication: chatbots and message drafting

Patient communication was one of the fastest-growing application areas, especially for portal message drafting, chatbot responses, and clinician-facing response suggestions. Comparative studies reported that artificial-intelligence-generated answers to patient questions could be rated favorably for empathy or quality, but the review found that such findings required caution because public-question datasets and simulated workflows may not capture real patient safety risks [6, 8, 28]. Early electronic-health-record message studies described implementation lessons for drafting patient responses,

including the need for clinician review, careful prompt design, and integration into existing inbox workflows [18, 21]. Other studies examined the effect of large language models in responding to patient messages and suggested that communication quality, appropriateness, and workflow impact must be evaluated together rather than independently [19, 20].

Patient communication: language translation and simplification

Generative artificial intelligence was also studied or discussed as a tool for simplifying medical language and supporting patient-facing communication. Reviews and broad syntheses noted that large language models can convert technical clinical language into more accessible explanations, but they can also introduce inaccuracies, omit caveats, or generate advice that exceeds the intended communication scope [4, 11, 15]. Patient-message studies implied that language tone, readability, and empathy are central evaluation dimensions, particularly when models draft replies that clinicians may accept with limited editing [8, 19, 24]. Evidence for multilingual support was thinner, although studies in non-English clinical documentation contexts suggested that local language, regulatory, and workflow factors may significantly affect deployment feasibility [27].

Governance dashboards: automated board reports

Governance dashboard applications were discussed primarily as an extension of summarization, reporting, and decision-support synthesis rather than as a mature standalone evidence base. Studies on medical evidence summarization, decision-support comment summarization, and broad generative artificial intelligence in healthcare indicated that models can compress information into narratives suitable for governance review, quality oversight, or compliance monitoring [9, 11, 26]. However, board reports and governance dashboards have higher accountability requirements than ordinary summaries because they may influence resource allocation, regulatory reporting, and institutional risk management. The reviewed literature therefore supported feasibility in principle but did not provide strong empirical evidence that automated board-report generation is safe, auditable, or superior to human-prepared governance materials [2, 10, 15].

Workflow automation: prior authorization and referral letters

Workflow automation evidence was most visible in studies involving triage, referral, message classification, and clinical decision support rather than direct prior authorization trials. Large language model workflows for triage, referral, and diagnosis suggested that generative systems may help structure or draft administrative communications when tasks are repetitive and rule-bound [14]. Patient portal triage studies similarly indicated that models may help identify urgency, route messages, or draft next-step communications, although the safety consequences of misclassification remain substantial [12, 13]. For prior authorization and referral letters, the literature supported early promise but showed limited deployment evidence on payer acceptance, turnaround time, denial prevention, or staff workload reduction [15, 17].

Workflow automation: other administrative tasks

Other workflow automation tasks included scheduling communications, consent-related text, staff messaging, administrative reminders, and the conversion of clinical information into operationally useful language. Studies on patient message guidance and response drafting showed how generative systems could help users formulate more complete requests or help staff prepare efficient replies [24, 31]. In-basket message studies also suggested that administrative communication workflows may benefit from drafting support, but only when outputs are reviewed and adapted by clinicians or trained staff [8, 20]. Overall, workflow automation remained more advanced for text drafting and routing than for autonomous task completion, reflecting persistent concerns about safety, accountability, and system integration [18, 29].

Model types and architectures

The reviewed literature was dominated by large language models, including general-purpose commercial systems, adapted medical language models, and retrieval-augmented approaches. Electronic-health-record language models and medically adapted large language models illustrated the trend toward domain adaptation for clinical documentation and administrative text [5, 7]. Prompt engineering emerged as a practical method for shaping patient-message outputs and improving consistency without full model retraining [21]. Retrieval-augmented generation was increasingly emphasized as a way to connect generated text to source documents or knowledge bases, especially for tasks requiring factual grounding, triage justification, or compliance-sensitive administrative output [13, 17].

Evaluation approaches

Evaluation approaches were heterogeneous and often surface-level, including expert ratings, preference comparisons, readability assessments, summarization metrics, classification accuracy, and qualitative judgments of usefulness. Studies of patient-message responses frequently used physician or reviewer ratings of quality, empathy, correctness, or safety, while summarization studies often examined completeness, factual consistency, and clinical usefulness [6, 7, 9, 19]. Some workflow-oriented studies incorporated task-specific assessments such as triage appropriateness, document classification, or decision-support comment summarization [12, 23, 26]. However, few studies measured administrative endpoints such as turnaround time, staff burden, reporting reliability, denial rates, governance actionability, cost, or sustained user adoption [22, 29].

Implementation and deployment barriers

Common barriers across domains included hallucination, privacy risk, bias, lack of explainability, uncertainty over regulatory status, weak integration with electronic health records, and limited user trust. Ethical analyses and narrative reviews repeatedly warned that generative models may produce fluent but unsupported content, which is especially concerning when administrative documents influence patient access, billing, compliance, or organizational decisions [1, 2, 15]. Implementation reports on patient-message drafting emphasized the practical importance of workflow design, human oversight, prompt governance, monitoring, and clear responsibility for final communication [18, 20, 29]. The evidence base therefore suggested that generative artificial intelligence should be treated as an assistive administrative tool rather than an autonomous administrative agent [4, 11, 17].

Table 1 presents a domain-level evidence maturity matrix showing that documentation support and patient communication are the most developed areas, whereas operational reporting, governance dashboards, and workflow automation remain under-validated for real-world administrative impact.

Table 1. Administrative Domain–Evidence Maturity Matrix for Generative AI in Healthcare Administration

Administrative domain	Dominant generative AI tasks identified in the review	Relative evidence maturity	Main analytical contribution to healthcare administration	Key unresolved evidence problem	Implementation interpretation
Documentation support	Clinical text summarization, discharge summaries, note drafting, document classification, coding-adjacent interpretation	Highest maturity among reviewed domains	Shows the clearest alignment between language-model capability and administrative documentation burden	Evidence remains concentrated in retrospective, simulated, or output-quality evaluations rather than sustained embedded workflow studies	Most appropriate near-term use case, but only with human review, factual checking, and integration into documentation workflows

Patient communication	Portal-message drafting, chatbot responses, plain-language rewriting, empathy-oriented response generation, message triage	Moderate and rapidly expanding maturity	Demonstrates potential to improve response drafting, tone, readability, and clinician inbox support	Simulated patient-question datasets and reviewer ratings may not capture real safety, personalization, equity, or liability risks	Useful as assisted drafting, not autonomous patient communication
Operational reporting	Utilization summaries, quality narratives, safety reports, executive briefings, service-line reporting	Emerging maturity	Extends summarization capability from clinical text to managerial and operational interpretation	Limited direct evidence that generated reports improve decision-making, timeliness, accountability, or organizational performance	Promising but should remain source-grounded, auditable, and reviewed by operational leaders
Governance dashboards	Board narratives, quality committee summaries, compliance dashboards, risk and safety oversight reports	Low direct maturity	Highlights a high-value but high-accountability administrative use case	Very limited empirical evidence on safety, auditability, regulatory adequacy, or board-level decision consequences	Requires the strongest governance safeguards because errors may affect institutional oversight and resource allocation
Workflow automation	Referral letters, prior authorization drafts, triage messages, scheduling communications, administrative reminders	Uneven maturity	Shows potential to reduce repetitive transactional workload by drafting and routing administrative text	Sparse evidence on payer acceptance, denial reduction, turnaround time, workload transfer, or autonomous task safety	Best framed as workflow assistance with staff approval rather than independent task completion

Figure 2 synthesizes the review findings into an evidence-to-implementation map linking administrative domains, generative AI functions, model patterns, evidence maturity, evaluation gaps, governance risks, and implementation implications.

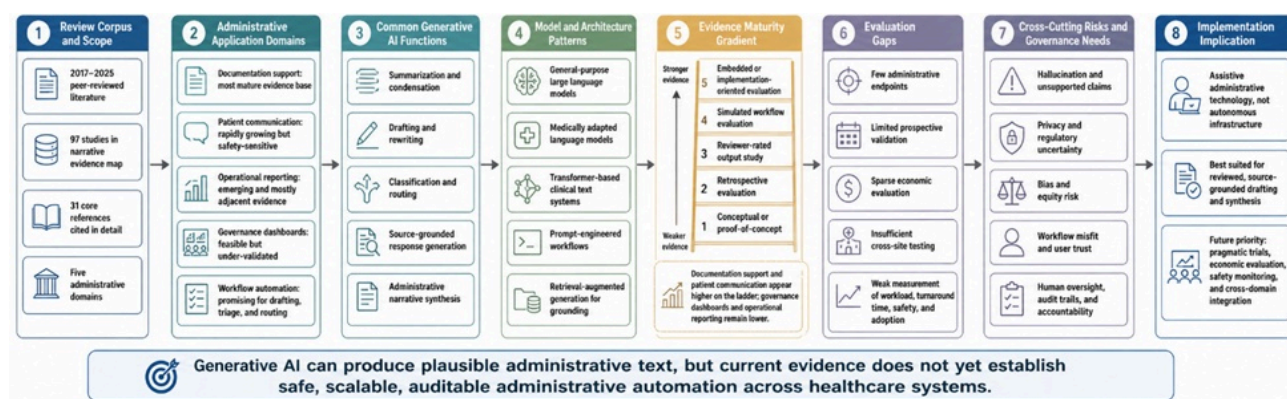


Figure 2. Evidence-to-Implementation Map of Generative Artificial Intelligence for Healthcare Administration

Documentation support is the most mature domain

Documentation support emerged as the most mature application area because it aligns closely with the core strengths of generative artificial intelligence: summarizing, drafting, rephrasing, and structuring language. Studies on electronic health record language models, clinical text summarization, discharge summaries, and adapted large language models provided the clearest evidence that generative systems can support documentation-related tasks [5, 7, 16, 30]. Nevertheless, most evidence remained retrospective or simulation-based, and relatively few studies tested documentation tools as embedded workflow interventions. The review therefore found that documentation support is advanced relative to other administrative domains, but not yet supported by enough prospective implementation evidence to justify autonomous use [23, 27].

Operational reporting and governance dashboards: emerging with potential

Operational reporting and governance dashboards appeared promising but underdeveloped in the peer-reviewed literature. Adjacent evidence from summarization, evidence synthesis, and decision-support comment analysis suggested that large language models can generate concise narratives from complex information streams [9, 16, 26]. These capabilities could plausibly support quality dashboards, utilization summaries, compliance reports, and board-facing narratives, particularly in organizations with fragmented reporting infrastructure. However, direct studies of automated governance reporting were scarce, and the reviewed literature did not establish whether such tools improve oversight, accountability, or executive decision-making [10, 11].

Patient communication balances efficiency and empathy

Patient communication studies showed that generative systems can draft responses that are often perceived as clear, empathetic, and useful, but they also highlighted the difficulty of balancing efficiency with safety and personalization. Comparative studies of chatbot and physician responses suggested potential communication benefits, while electronic-health-record message studies emphasized that real-world deployment requires clinician review and careful monitoring [6, 8, 18]. Prompt-engineering and implementation studies further showed that output quality depends on workflow context, instruction design, and the way clinicians interact with suggestions [20, 21]. Across the literature, generative patient communication was best understood as assisted drafting rather than replacement of professional judgment [19, 28].

Workflow automation reduces transactional burden

Workflow automation may reduce transactional administrative burden by drafting repetitive communications, triaging incoming messages, and supporting referral or authorization-related documentation. Studies on patient-message guidance, triage, referral workflows, and knowledge-grounded message classification illustrated how generative models could help structure administrative work that currently consumes staff and clinician time [12-14, 31]. Such tasks are attractive candidates for automation because they often follow recognizable patterns, require synthesis of existing information, and involve high message volume. However, the evidence remained thin for payer-facing prior authorization, scheduling automation, and autonomous completion of administrative forms, where consequences of errors may include delays, denials, or inequitable access [15, 17].

Hallucination and factual grounding: the central technical challenge

Hallucination was the central technical and governance challenge across all five administrative domains. Generative systems may produce outputs that appear fluent and authoritative while misrepresenting source information, omitting necessary qualifiers, or inventing details, which is hazardous in documentation, billing, compliance, and patient communication [2, 4, 15]. Retrieval-augmented generation was repeatedly proposed as one mitigation strategy because it

can link outputs to source documents or knowledge bases, but it does not eliminate the need for validation and monitoring [13, 17]. The review found that factual grounding must be treated not only as a technical model feature but also as an organizational safety requirement involving audit trails, human review, and accountability [1, 10].

Evaluation is surface-level

Many studies evaluated generated outputs using reviewer ratings, text-quality measures, or short-term preference judgments rather than administrative outcomes. Patient-message studies frequently measured empathy, accuracy, or usefulness, while summarization studies often assessed completeness or factuality, but few connected these measures to workload, cost, turnaround time, patient access, or organizational performance [6, 7, 9, 20]. Implementation-oriented work began to address real workflow effects, yet the evidence base remained dominated by small-scale or early-stage evaluations [18, 29]. As a result, the literature supported cautious optimism about output quality but offered limited evidence on whether generative artificial intelligence improves healthcare administration as a system [11, 22].

Missing: prospective, implementation-science work

A major gap was the limited amount of prospective, implementation-science research evaluating generative artificial intelligence as an embedded administrative workflow tool. Several studies moved toward real-world message drafting or implementation lessons, but most did not test long-term adoption, safety monitoring, economic effects, or cross-site scalability [18, 19, 29]. Reviews and scoping studies consistently emphasized the need for stronger evaluation frameworks, better reporting, and clearer attention to governance and human oversight [11, 16, 22]. Future work should therefore move from output-level testing to pragmatic studies that measure whether generative systems actually reduce administrative burden without compromising safety, equity, or accountability [17, 24].

Table 2 proposes an evaluation and governance framework showing that safe implementation requires evidence beyond text quality, including factual grounding, administrative outcomes, equity monitoring, human oversight, workflow integration, and economic evaluation.

Table 2. Governance and Evaluation Framework for Translating Generative AI from Administrative Text Generation to Safe Healthcare Implementation

Evaluation or governance dimension	Why it matters for healthcare administration	Minimum evidence needed before routine use	Higher-standard evidence needed before scaled deployment	Domains where this is especially critical
Factual grounding	Administrative outputs may influence documentation, billing, referrals, reports, compliance, and governance decisions	Source-linked outputs, citation to institutional documents, review of unsupported claims and omissions	Prospective monitoring of factual errors, source mismatch, and downstream administrative consequences	Documentation support, governance dashboards, operational reporting, prior authorization
Human oversight	Generated content may appear authoritative even when incomplete, biased, or wrong	Clear role assignment for review, editing, approval, and final responsibility	Measured reviewer behavior, override rates, automation bias, and accountability outcomes	Patient communication, board reporting, workflow automation

Administrative outcome measurement	Output fluency does not prove operational value	Task-specific outcomes such as completion time, staff burden, turnaround time, message volume, or report timeliness	Pragmatic trials comparing AI-assisted workflows with usual administrative processes	Documentation support, operational reporting, workflow automation
Safety and hallucination monitoring	False or unsupported administrative statements can create access, compliance, billing, or governance harms	Error taxonomy covering false statements, omissions, inappropriate advice, and fabricated details	Continuous safety surveillance, escalation pathways, and audit-ready error reporting	Patient communication, governance dashboards, referral letters, prior authorization
Equity and bias assessment	Administrative automation may affect access, responsiveness, routing, documentation quality, or payer-facing decisions unevenly	Subgroup review of output quality, tone, routing, and completeness	Prospective equity monitoring across language, demographic, insurance, access, and clinical-complexity groups	Patient communication, workflow automation, documentation support
Workflow integration	Tools may increase burden if they create extra review, editing, or system-switching work	Usability testing and fit with existing electronic health record or administrative platforms	Longitudinal adoption, workload transfer analysis, and staff experience evaluation	All five domains
Privacy and access control	Administrative workflows often combine clinical, operational, financial, and governance-sensitive information	Defined data access rules, de-identification where appropriate, and secure processing environment	Role-based access, provenance tracking, audit logs, and institutional governance review	Governance dashboards, operational reporting, documentation support
Economic evaluation	Administrative AI is often justified by efficiency and cost-reduction claims	Basic resource-use estimates and staff-time analysis	Full economic evaluation including cost, productivity, error burden, maintenance, monitoring, and implementation costs	Documentation support, workflow automation, operational reporting
Cross-domain interoperability	Administrative work often moves from documentation to communication, reporting, governance, and workflow action	Clear boundaries between data inputs, generated outputs, and downstream use	Tested cross-domain architecture with provenance, access control, and governance across systems	Integrated administrative platforms, governance dashboards, workflow automation

Limitations

Review limitations

This review was limited by English-language inclusion, heterogeneity in task definitions, and the rapid evolution of generative artificial intelligence terminology between 2017 and 2025. Because studies differed substantially in model type, task, setting, comparator, and outcome measures, meta-analysis was not appropriate and the synthesis remained narrative [11, 16, 22]. The administrative focus also required interpretive judgment, particularly for studies of clinical text summarization or patient communication that had both clinical and administrative relevance [7-9]. Publication bias may have favored positive or technically novel findings, while unsuccessful deployments and negative administrative outcomes were likely underreported [15, 29].

Evidence base limitations

The evidence base itself was limited by a high proportion of early-stage evaluations, simulated tasks, single-center studies, and output-quality assessments without prospective administrative endpoints. Many studies demonstrated that generative systems could draft, summarize, classify, or respond, but fewer evaluated whether these outputs improved efficiency, reduced costs, prevented errors, or strengthened governance in real healthcare organizations [18, 20, 26, 28]. Vendor-linked deployment environments, changing model versions, and incomplete reporting of prompts or safety controls also limited reproducibility and external validity [17, 21, 27]. Consequently, the review found that claims about healthcare administrative transformation remain plausible but not yet proven across the five domains [2, 4, 14].

Comparison with prior reviews

Prior reviews of large language models and generative artificial intelligence in medicine have often focused on broad clinical use, clinical knowledge encoding, biomedical reasoning, or clinical text summarization rather than healthcare administration as an integrated organizational function. Reviews of clinical text summarization and medical language models provided important evidence on documentation support, but they generally did not extend systematically to operational reporting, governance dashboards, or payer-facing workflow automation [4, 10, 15, 16]. Similarly, retrieval-augmented generation reviews clarified methods for grounding biomedical outputs, yet their primary emphasis was usually technical capability and clinical development rather than administrative implementation outcomes [17]. This review therefore differs by treating documentation, communication, reporting, governance, and workflow automation as interrelated administrative domains rather than isolated informatics tasks.

This review also extends prior work by including the latest peer-reviewed studies on patient-message drafting, in-basket response generation, portal-message triage, clinical document classification, and discharge-summary generation. Studies of patient communication have shown rapid movement from general chatbot comparisons toward embedded electronic-health-record workflows, prompt engineering, and early implementation lessons [6, 8, 18-21]. Documentation-focused studies likewise demonstrate movement from general clinical summarization toward adapted large language models, multilingual documentation settings, document classification, and discharge-summary generation [7, 23, 27, 28]. By integrating these streams with adjacent evidence on decision-support summarization and workflow-oriented triage, this review captures the administrative pipeline more comprehensively than prior task-specific reviews [12-14, 26].

A common theme across prior reviews and the present synthesis is that generative artificial intelligence is limited less by the ability to produce fluent text than by the difficulty of ensuring safe, grounded, auditable, and workflow-compatible outputs. Narrative and scoping reviews repeatedly identified hallucination, bias, privacy, regulation, and human oversight as cross-cutting barriers in medical applications [1, 2, 11, 15, 22]. This review found that the same concerns are intensified in healthcare administration because generated documents may influence billing, referrals, access, reporting, compliance, and governance decisions. Consequently, the administrative use of generative artificial intelligence requires evaluation frameworks that move beyond text quality toward accountability, reproducibility, and organizational impact [17, 24, 29].

Research gaps

Prospective implementation and economic studies

The most important evidence gap is the lack of prospective studies evaluating generative artificial intelligence against manual or conventional workflows using administrative endpoints. Although several studies assessed patient-message drafting, clinical summarization, discharge summaries, and triage support, few measured sustained changes in staffing burden, cost, turnaround time, access delays, documentation completeness, or organizational productivity [7, 12, 19, 30]. Economic studies are especially scarce, even though administrative automation is often justified by claims of efficiency and cost reduction [11, 22, 24]. Future research should use pragmatic designs that compare generative artificial intelligence-assisted workflows with usual administrative processes while monitoring safety, equity, and unintended workload transfer [18, 29].

Safety and hallucination mitigation in administrative contexts

Few studies systematically examined hallucination and factual-error consequences in administrative contexts such as coding, billing, referral letters, operational reports, or governance dashboards. Existing research and reviews described hallucination as a general risk, but administrative harms may differ from clinical harms because errors can cause claim denials, delayed authorizations, inaccurate compliance summaries, misleading board reports, or inappropriate patient routing [2, 10, 15]. Retrieval-augmented generation, prompt engineering, and domain adaptation may reduce some risks, yet the evidence does not show that these strategies are sufficient without human review and audit trails [13, 17, 21]. Future studies should report false statements, unsupported claims, omissions, source mismatch, and downstream administrative consequences as primary safety outcomes [9, 23, 26].

Cross-domain integration

No mature evidence demonstrated a unified generative architecture spanning documentation support, operational reporting, patient communication, governance dashboards, and workflow automation. Most studies addressed one task at a time, such as summarization, portal-message response, document classification, discharge-summary generation, or triage support [8, 14, 16, 20, 30]. Yet healthcare administration often requires information generated in one domain to feed another, such as clinical documentation informing referral letters, patient communication, quality reporting, and governance review [5, 26, 31]. Cross-domain systems will therefore require stronger interoperability, provenance tracking, role-based access control, and governance mechanisms than single-task tools [1, 11, 17].

Conclusion

Generative artificial intelligence for healthcare administration is advancing rapidly, with documentation support leading the field. The strongest evidence concerns summarization, discharge documentation, patient-message drafting, and document classification, where the language-centered nature of the task aligns closely with the capabilities of current models. Even in these relatively mature areas, however, the evidence base remains concentrated in retrospective, simulated, or early implementation settings.

Operational reporting, governance dashboards, patient communication, and workflow automation are gaining traction but remain unevenly evaluated in real healthcare environments. These domains may offer substantial administrative value because they involve repetitive synthesis, communication, and reporting work. Yet the literature does not yet establish that generative artificial intelligence improves organizational performance, reduces costs, or safely scales across diverse institutions.

Critical challenges apply across all five domains, including hallucination, weak factual grounding, privacy risk, limited interoperability, unclear accountability, and insufficient prospective evaluation. The central question is no longer whether generative artificial intelligence can produce plausible administrative text, but whether it can produce reliable, auditable, context-appropriate outputs within real healthcare systems. Human oversight remains essential wherever generated content may affect patients, staff, payers, regulators, or organizational governance.

A rigorous program of pragmatic trials, standardized administrative outcome measures, safety-first deployment, and transparent governance is needed before generative artificial intelligence can be trusted to run the administrative backbone of healthcare. Future research should prioritize real-world implementation, economic evaluation, equity monitoring, and cross-domain integration. Until that evidence is available, generative artificial intelligence should be treated as a promising assistive technology rather than an autonomous administrative infrastructure.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jul 2025

Revised: 17 Aug 2025

Accepted: 24 Sep 2025

Published online: 25 February 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med.* 2021;4(1):93.
- Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine.* 2023;90:104512.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172-80.
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930-40.

Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. *NPJ Digit Med.* 2022;5(1):194.

Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589-96.

Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* 2024;30(4):1134-42.

Liu S, McCoy AB, Wright AP, Carew B, Jenkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. *J Am Med Inform Assoc.* 2024;31(6):1367-79.

Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med.* 2023;6(1):158.

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond).* 2023;3(1):141.

Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Afrash MR. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. *Int J Med Inform.* 2024;188:105474.

Pham JH, Thongprayoon C, Miao J, Suppadungsuk S, Koirala P, Craici IM, et al. Large language model triaging of simulated nephrology patient inbox messages. *Front Artif Intell.* 2024;7:1452469.

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

Gaber F, Shaik M, Allega F, Bilecz AJ, Busch F, Goon K, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *NPJ Digit Med.* 2025;8(1):263.

Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med.* 2024;177(2):210-20.

Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, et al. Scientific evidence for clinical text summarization using large language models: scoping review. *J Med Internet Res.* 2025;27:e68998.

Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inform Assoc.* 2025;32(4):605-15.

Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. *JAMIA Open.* 2024;7(2):ooae028.

Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health.* 2024;6(6):e379-e381.

Small WR, Wiesenfeld B, Brandfield-Harvey B, Jonassen Z, Mandal S, Stevens ER, et al. Large language model-based responses to patients' in-basket messages. *JAMA Netw Open.* 2024;7(7):e2422399.

Yan S, Knapp W, Leong A, Kadkhodazadeh S, Das S, Jones VG, et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. *J Am Med Inform Assoc.* 2024;31(10):2263-70.

Hu D, Guo Y, Zhou Y, Flores L, Zheng K. A systematic review of early evidence on generative AI for drafting responses to patient messages. *NPJ Health Syst.* 2025;2(1):27.

Mustafa A, Naseem U, Azghadi MR. Large language models vs human for classifying clinical documents. *Int J Med Inform.* 2025;195:105800.

Subramanian CR, Yang DA, Khanna R. Enhancing health care communication with large language models—the role, challenges, and future directions. *JAMA Netw Open.* 2024;7(3):e240347.

Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc.* 2023;30(7):1237-45.

Liu S, McCoy AB, Wright AP, Nelson SD, Huang SS, Ahmad HB, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. *J Am Med Inform Assoc.* 2024;31(6):1388-96.

Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C. Viability of open large language models for clinical documentation in German health care: real-world model evaluation study. *JMIR Med Inform.* 2024;12:e59617.

Reynolds K, Nadelman D, Durgin J, Ansah-Addo S, Cole D, Fayne R, et al. Comparing the quality of ChatGPT- and physician-generated responses to patients' dermatology questions in the electronic medical record. *Clin Exp Dermatol.* 2024;49(7):715-8.

Bootsma-Robroeks CM, van Workum JD, Schuit SC, Hoekman A, Mehri T, Doornberg JN, et al. AI-generated draft replies to patient messages: exploring effects of implementation. *Front Digit Health.* 2025;7:1588143.

Williams CY, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P, et al. Physician- and large language model-generated hospital discharge summaries. *JAMA Intern Med.* 2025;185(7):818-25.

Liu S, Wright AP, McCoy AB, Huang SS, Jenkins JZ, Peterson JF, et al. Using large language model to guide patients to create efficient and comprehensive clinical care message. *J Am Med Inform Assoc.* 2024;31(8):1665-70.