
| RESEARCH ARTICLE

Operational and Financial Evidence for Generative AI in United States Healthcare Administrative Workflows

Clapher Ankur Gomes¹, Mahfuz Islam Khan Javed²✉, Tanvir Ahammad³, Sourav Mandal⁴, Sakib Hassan⁵, Mahfujul Karim Sejan⁶, Prudvi Saisaran Ponduru⁷, Md Masum Bhuiyan⁸ and Md Luman Jamali⁹

¹Department of Information Technology, Washington University of Science and Technology, Alexandria, VA, United States

²Department of Information Technology, Washington University of Science and Technology, Alexandria, VA, United States

³Department of Business Administration, University of The Potomac, Washington D.C., United States

⁴Department of Computer Science, St Francis College, Brooklyn, New York, United States

⁵Department of Information Technology, Washington University of Science and Technology, Alexandria, VA, United States

⁶Department of Business Administration, Washington University of Science and Technology, Alexandria, VA, United States

⁷Department of Computer Science, University of the Potomac, Washington D.C., United States

⁸Department of Information Technology, Washington University of Science and Technology, Alexandria, VA, United States

⁹Department of Business Administration, Trine University, Arizona, United States

Corresponding Author: Mahfuz Islam Khan Javed, **E-mail:** mahfuzislamkhanjaved@gmail.com

| ABSTRACT

Generative artificial intelligence (GenAI) is increasingly used for clinical documentation and administrative work, but reported benefits often mix time savings, capacity, billing proxies, and realized financial value. This structured narrative review synthesized peer-reviewed evidence on operational and financial outcomes in United States healthcare administrative and hybrid workflows and developed a governance-integrated evaluation framework. PubMed, PubMed Central, publisher records, backward citation searching, and DOI or PMID verification were used to identify empirical studies and relevant syntheses. Fifteen primary studies and one systematic review met the use-case criteria; ten primary studies evaluated ambient documentation, including two randomized trials. Ambient tools generally produced modest reductions in documentation time or work burden, whereas visit volume and productivity effects were inconsistent. One controlled study estimated 1.81 additional relative value units and 0.80 additional encounters per physician-week but did not measure complete implementation costs or net collected cash. Patient-message tools showed utilization near 20% and no consistent time savings. Summarization and coding studies identified clinically important omissions, inaccuracies, and poor agreement with human coding. Evidence of net return on investment remained sparse. Adoption decisions should pass sequential gates for technical validity, workflow fit, operational effectiveness, financial conversion, and equity and sustainability, with total cost and human review measured before return on investment is claimed.

| KEYWORDS

Generative artificial intelligence, machine learning, healthcare administration, ambient documentation, operational efficiency, economic evaluation, return on investment, artificial intelligence governance

| ARTICLE INFORMATION

ACCEPTED: 14 October 2025

PUBLISHED: 30 November 2025

DOI: 10.32996/jcsts.2025.4.4.10

1. Introduction

Administrative work consumes substantial clinician and staff time across billing, insurance, documentation, and electronic health record activities. [1, 2, 3] This burden has made documentation, messaging, coding, billing, prior authorization, scheduling, and related processes prominent targets for automation.

Copyright: © 2025 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (<https://creativecommons.org/licenses/by/4.0/>). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

Generative artificial intelligence systems can transform speech, free text, and structured records into draft notes, messages, summaries, codes, and workflow recommendations. Adoption is already widespread. A national hospital survey estimated that 31.5% of nonfederal acute-care hospitals used an EHR-integrated large language model in 2024, while another 24.7% planned adoption within one year. [4] Adoption, however, is an exposure measure. It does not establish effective use, safer work, lower cost, or financial return.

The evaluation challenge is a sequence of conversions. A technically plausible output must be accepted, reviewed, integrated, and used in a real workflow. Any reduction in task time must create net capacity after review and rework. Capacity must then be converted into improved access, avoided expenditure, incremental collected revenue, or another verifiable benefit. Those benefits must exceed license, integration, training, security, governance, monitoring, human-review, incident, and exit costs. Most published studies observe only part of this chain.

Regulatory and governance requirements also shape implementation. The NIST Artificial Intelligence Risk Management Framework and its generative AI profile provide voluntary lifecycle controls, federal health information technology rules impose transparency requirements for specified decision-support interventions, HIPAA duties continue when cloud providers handle protected health information, and federal prior-authorization requirements are changing payer workflows. [25, 26, 28, 29, 30] These sources define obligations and context; they are not evidence that a particular GenAI product improves outcomes.

This review addresses two questions. First, what operational and financial outcomes have been measured for GenAI-enabled healthcare administrative and hybrid workflows? Second, what evaluation structure can prevent technical, operational, and financial measures from being treated as interchangeable? The review separates observed empirical findings from policy context and proposes a staged framework for local decision making.

1.1 Conceptual Foundation

GenAI implementation is treated as a sociotechnical intervention in which the model, interface, data, workflow, users, organizational policies, external rules, and monitoring process jointly determine results. [5] PRISM and RE-AIM add organizational fit, reach, adoption, implementation, maintenance, and sustainability to the assessment. [6, 7]

The proposed pathway begins with the use case, model, vendor, data, and integration. Eligibility, activation, task-level use, and use intensity determine exposure. Draft quality, review time, substitution, and rework are mechanisms. Time, turnaround, backlog, throughput, quality, safety, and access are operational outcomes. Only verified operational changes can be translated into avoided expense, incremental contribution margin, and net benefit. Organizational resources, payer mix, interoperability, workforce capacity, governance maturity, language, and patient mix may modify each stage.

2. Materials and Methods

2.1 Review Design

This study used a structured narrative review and framework-development design because the literature spans heterogeneous interventions, outcomes, study designs, policy documents, and implementation frameworks. Reporting followed the core narrative-review principles of explaining the review's importance and aims, documenting source identification, supporting claims with references, distinguishing evidence levels, and reporting relevant endpoint data. [8]

2.2 Scope and Eligibility

The review focused on GenAI-enabled administrative and hybrid clinical-administrative workflows in United States healthcare delivery. Eligible empirical studies evaluated ambient documentation, patient-message drafting, summarization, coding, hospital adoption, or related operational workflows and reported adoption, time, throughput, productivity, financial, safety, quality, equity, or workforce outcomes. Systematic reviews were eligible when they synthesized the same use cases. Purely clinical diagnosis or treatment studies without an administrative workflow were excluded. Vendor marketing claims and product demonstrations without independent outcome evaluation were excluded from effectiveness conclusions.

2.3. Information Sources and Search

Peer-reviewed sources were identified through PubMed, PubMed Central, accessible journal publisher records, backward citation searching, and targeted identifier verification. Search concepts combined generative artificial intelligence, large language model, ambient scribe, AI-generated draft, and agentic workflow with healthcare, hospital, clinician, administrative, documentation, message, coding, billing, prior authorization, scheduling, revenue cycle, time, productivity, safety, cost, and return on investment. DOI, PMID, publication year, journal, volume, issue, and article identifiers were checked against publisher or bibliographic metadata. Official NIST, HHS, CMS, health information technology, and GAO records were used only for policy, governance, or federal context.

2.4. Selection and Data Extraction

A single-reviewer process screened titles, abstracts, and accessible full text for relevance. For each included empirical source, the extraction recorded publication identity, setting, design, sample, intervention or use case, comparator, exposure or adoption

definition, operational outcomes, financial measures, safety and quality findings, principal limitations, and funding or conflict information when reported. Financial measures were classified as workload or time, capacity, billing proxy, avoided cost, realized cash, contribution margin, or net benefit. This hierarchy prevented time saved, RVUs, gross charges, or avoided hiring from being labeled as cash savings without evidence of conversion.

2.5. Evidence Appraisal and Synthesis

Evidence strength was assessed by design features rather than a pooled score. Randomization, a concurrent or credible counterfactual, objective telemetry, sample size, follow-up, exposure measurement, outcome definition, adjustment for selection, missing data, product specificity, and sponsor involvement were considered. Findings were synthesized narratively because interventions, denominators, study designs, and endpoints were not sufficiently comparable for meta-analysis. Framework development integrated the empirical pathway with sociotechnical and implementation models, NIST risk-management functions, and economic-reporting principles. [5, 6, 7, 24, 25, 26]

3. Results

3.1. Evidence Base

The use-case synthesis contained 15 primary empirical studies and one systematic review. Ten primary studies evaluated ambient documentation, two evaluated patient-message drafting, one evaluated hospital-course summarization, one benchmarked general-purpose large language models for coding, and one described national hospital adoption. The patient-message systematic review synthesized 23 earlier studies. Designs included two randomized trials, multisite and single-site controlled observational studies, interrupted time series, pre-post quality-improvement studies, audit-log analyses, a national survey, and a benchmark study. Exact database-screening counts were not available and are therefore not reported.

3.2. Adoption and Organizational Context

The national hospital study estimated current EHR-integrated GenAI use at 31.5% in 2024, with 24.7% planning adoption within one year. Independent hospitals, critical-access hospitals, and hospitals with high Medicaid discharge shares were more likely to lag. [4] The study measured adoption category, not task-level use, implementation quality, or outcomes. Adoption rates therefore describe diffusion and organizational readiness rather than effectiveness.

3.3. Ambient Documentation

Ambient documentation had the most mature operational evidence. In a longitudinal multisite cohort of 8,581 clinicians, access to an AI scribe was associated with 13.4 fewer total EHR minutes and 16.0 fewer documentation minutes per eight scheduled patient hours, along with 0.49 additional visits per week. After-hours EHR time did not change significantly. [9]

The two randomized evaluations produced product- and outcome-specific findings. One trial involving 238 physicians found a 9.5% reduction in time in note for one product but not the other, with actual use in approximately one-third of eligible visits. [10] A stepped-wedge pragmatic trial involving 66 practitioners reported lower work exhaustion and 0.36 fewer note hours per day; its after-hours result was sensitive to outlier handling. [11]

Controlled observational studies generally supported modest efficiency gains while showing heterogeneity. A propensity-balanced study reported 8.5% lower total EHR time and 15.9% lower note time per appointment, without significant changes in appointment volume or after-hours documentation. [12] An interrupted time-series analysis found an immediate reduction of 0.26 minutes per note, a gradual reduction in after-hours documentation, no change in appointments per day, and an immediate RVU increase without a sustained RVU trend. [13] In emergency departments, ambient AI was associated with 1.6 fewer adjusted documentation minutes per note, compared with 3.3 fewer minutes for human scribes, and no wRVU difference. [14]

Short pre-post studies also reported favorable signals. A 45-physician pilot found lower documentation time but substantial variation in use and outcome magnitude. [15] A 46-clinician study reported time in notes falling from 10.3 to 8.2 minutes per appointment and same-day closure increasing from 66.2% to 72.4%; the adjusted after-hours reduction was 16.9%. [16] Across six systems, self-reported burnout decreased from 51.9% to 38.8% after 30 days, but participation was voluntary and only 60.3% of enrolled clinicians completed both surveys. [17]

3.4. Financial Evidence for Ambient Documentation

The clearest financial-productivity study included more than 1.2 million ambulatory encounters and used a difference-in-differences design. Adoption was associated with 0.04 additional RVUs per encounter, 1.81 additional RVUs per physician-week, 0.80 additional encounters per physician-week, and no difference in denial rates. The authors translated the RVU estimate to approximately \$3,044 per physician-year under the 2025 Medicare Physician Fee Schedule. [18]

This estimate remains a billing and capacity proxy. The analysis did not subtract license, implementation, integration, review, security, monitoring, governance, or incident costs and did not measure incremental collected cash. It also could not distinguish additional services, more complete coding, and upcoding. The available evidence therefore supports a possible pathway to financial value, not a general conclusion of positive net ROI.

3.5. Patient Message Drafting

Patient-message studies showed a gap between draft quality and operational use. In a five-week prospective implementation study, clinicians used 20% of generated drafts. Task load and work exhaustion improved, but reply, writing, and reading time did not change. [19] A separate audit-log study found 19.4% utilization and a 6.76% improvement in turnaround after prompt refinement. Automatic generation for messages that ultimately received no response also created avoidable review work. [20] A systematic review of 23 studies reported often favorable quality and empathy ratings but inconsistent performance, limited real-world utilization, safety concerns, and substantial design heterogeneity. [21]

3.6. Summarization and Coding

In a prospective pilot of an agentic hospital-course summarizer, clinicians used AI-generated content in 57% of discharges. Among 100 summaries with feedback, 25% contained omissions, 20% contained inaccuracies, and 2% contained hallucinations; one summary was judged likely to cause moderate harm. [22] Selective feedback, the single-arm design, and the small sample limit inference, but the findings demonstrate why review time and correction work must be included in operational and financial evaluation.

A benchmark using 50 deidentified inpatient notes found that general-purpose large language models performed poorly for zero-shot ICD-10-CM extraction. The best overall agreement with a human coder was 15.2%, kappa values were near zero, and errors included unconfirmed diagnoses, nonspecific codes, symptom coding, and hallucinations. [23] This result does not establish that every specialized coding system fails. It does show that unsupervised deployment of a general-purpose model is not supported by the reviewed evidence.

3.7. Other Administrative Workflows

For claims, prior authorization, scheduling, referrals, revenue-cycle management, human resources, procurement, and compliance, the located evidence was mainly policy context, technical evaluation, proposed practice, or vendor assertion. Federal assessments similarly describe possible benefits alongside limited independent accuracy, cost, and spending evidence. [27, 30] Independent, controlled evaluations with complete cost, quality, and safety measurement remain a priority. Results from ambient documentation should not be extrapolated to workflows with different denominators, incentives, error costs, and decision rights. The evidence by use case is summarized in Table 1.

Table 1. Evidence summary by healthcare administrative use case

Use case	Operational evidence	Financial evidence	Interpretation
Ambient documentation	Two randomized trials plus controlled and pre-post studies; generally modest time or well-being gains; heterogeneous volume findings	RVU and encounter proxies in one controlled study; complete TCO and net cash not measured	Promising for burden reduction; requires local controlled evaluation and human review [9, 10, 11, 12, 13, 14, 15, 16, 17, 18]
Patient messages	Real-world utilization near 20%; selected burden or turnaround gains; no consistent time saving	No convincing independent net financial evidence	Targeting, draft necessity, and review burden determine value [19, 20, 21]
Hospital-course summaries	Early single-site pilot with frequent use and observed omissions and inaccuracies	No direct financial evidence	Human review and safety sampling are mandatory [22]
Coding and billing	General-purpose zero-shot benchmark performed poorly; documentation studies reported some coding or RVU changes	Coding intensity, denials, recoupment, and net collections remain unresolved	Use task-specific systems with coder oversight and compliance monitoring [18, 23]

Hospital adoption	National survey demonstrates rapid but unequal diffusion; outcomes were not measured	No estimate	financial-effect	Adoption is not evidence of effectiveness [4]
Claims and other administrative workflows	Sparse independent real-world GenAI evaluations located	No generalizable net ROI		High-priority evidence gap; do not extrapolate from ambient documentation [27, 30]

4. Evaluation Framework

The empirical findings support five sequential decision gates, shown in Figure 1. A use case should progress only when the evidence required at the current gate is adequate for its risk level. Material changes to the model, prompt, vendor, data source, integration, or workflow should trigger reassessment.

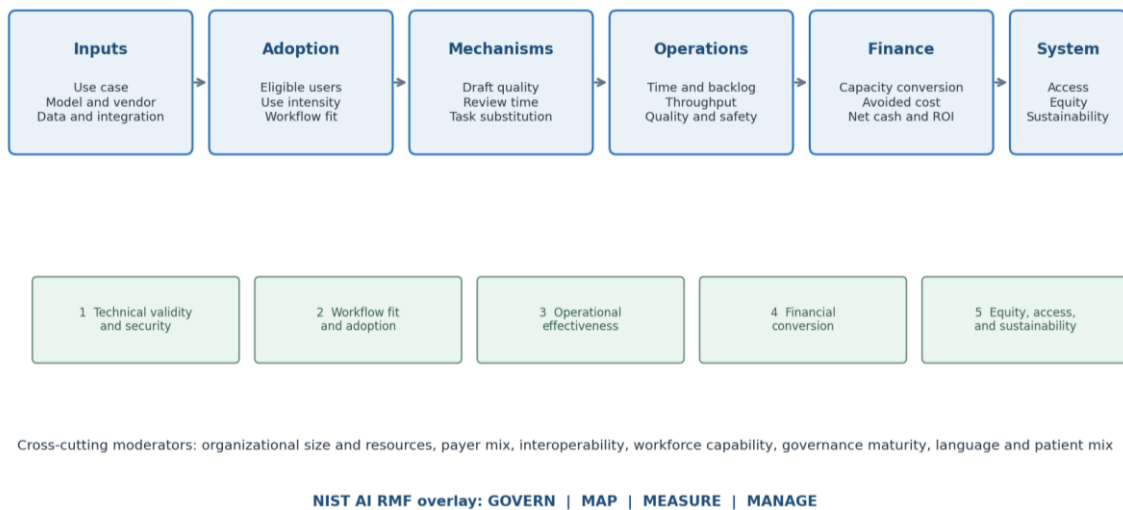


Figure 1. Evidence-informed pathway and sequential evaluation gates for GenAI-enabled healthcare workflows. The framework integrates sociotechnical, implementation, economic-reporting, and AI risk-management principles [5, 6, 7, 24, 25, 26].

4.1. Technical Validity and Security

The first gate evaluates task-specific accuracy, completeness, robustness, privacy, cybersecurity, and failure modes. Validation should use local cases that represent expected populations, languages, settings, and edge conditions. High average quality cannot compensate for rare but consequential omissions, fabricated content, privacy exposure, or insecure data handling.

4.2. Workflow Fit and Adoption

The second gate measures who is eligible, who activates the tool, how often it is used for eligible tasks, and how it changes review, editing, handoffs, overrides, workarounds, and patient experience. Adoption should be expressed with a valid denominator. Draft acceptance alone is insufficient because an accepted draft may still require substantial checking or shift work to another role.

4.3. Operational Effectiveness

The third gate requires a credible comparison for time, turnaround, backlog, throughput, quality, safety, and access. Randomized rollout, stepped-wedge implementation, interrupted time series, or difference-in-differences designs are preferable when feasible. Event-log measures should be validated against time-motion samples when idle time or background processes can distort recorded duration.

4.4. Financial Conversion

The fourth gate connects verified operational changes to realized economic value. TCO should include licenses, implementation, integration, training, support, human review, quality assurance, security, governance, monitoring, incident response, and exit costs. Net benefit equals incremental collected revenue plus realized avoided cost minus TCO. ROI equals net benefit divided by TCO, multiplied by 100. Economic reporting should state the perspective, time horizon, price year, discounting assumptions, uncertainty, and missing cost categories. [24]

A reduction in task time creates capacity, not automatically cash. Saved time has financial value only when it changes overtime, contractor use, hiring, throughput, collections, or another measurable resource decision. RVUs and gross charges require payer-specific conversion to allowed amounts, collection rates, denials, recoupments, and variable contribution margin. Scenario and break-even analyses should vary use intensity, usable-draft rate, review time, conversion of released capacity, price, implementation life, monitoring cost, payer mix, and error-related rework.

4.5. Equity and Sustainability

The final gate examines subgroup performance, access, language coverage, workload distribution, environmental and staffing effects, long-term benefit, monitoring burden, vendor change control, interoperability, and exit readiness. NIST governance functions provide a practical cross-cutting structure for ownership, risk mapping, measurement, monitoring, corrective action, suspension, and retirement. [25, 26]

5. Discussion

5.1. Principal Findings

The reviewed evidence supports selective operational benefit, especially for ambient documentation, but not a general claim of financial transformation. Randomized and controlled observational studies consistently point toward modest documentation-time reductions for some products and settings. Effects vary by product, clinician, specialty, use intensity, and outcome. Visit and productivity changes are smaller and less consistent than time measures, and direct net financial evidence is rare.

The distinction between operational and financial evidence is central. A health system can create capacity without reducing expenditure when released time is not converted into additional service, lower staffing cost, avoided hiring, or another resource change. It can generate more RVUs without receiving equivalent cash because payer contracts, denials, collection lags, coding review, and recoupment intervene. It can also increase collected revenue while producing negative net benefit when implementation and ongoing control costs are high.

5.2. Human Review and Measurement

Human review is part of the intervention and belongs in both the operational denominator and the cost model. Unnecessary message drafts and summaries containing omissions or inaccuracies show how automation can create new checking and correction work. [20, 22] Evaluations should measure net task time, edit burden, rework, exception handling, downstream correction, and error consequences rather than counting generated outputs.

5.3. Generalizability and Equity

Most outcome studies came from academic or integrated systems with mature EHR infrastructure and evaluation capacity. National data indicate that independent, rural, critical-access, and high-Medicaid hospitals may adopt later or face different resource constraints. [4] Implementation cost, interoperability, vendor leverage, staffing, language coverage, and local payer mix may therefore alter both benefit and risk. Subgroup and site-level reporting should accompany overall estimates so that aggregate improvement does not conceal unequal performance or workload transfer.

5.4. Implications for Health System Leaders

- Define one use case, one eligible population, and a falsifiable business case before procurement or scale-up.
- Fund telemetry, version tracking, independent quality assurance, data engineering, finance linkage, and protected analyst time as implementation costs.
- Negotiate access to task-level logs, model and configuration changes, audit rights, data-use and retention terms, security obligations, incident support, and exit assistance.
- Set quality, safety, privacy, and equity thresholds before throughput or adoption targets.
- Report operational outcomes and financial outcomes separately, then connect them through an explicit and auditable conversion model.

5.5. Research Priorities

Future research should extend beyond short-term ambient-scribe studies. Priority designs include pragmatic randomized or stepped-wedge evaluations, longer interrupted time series, and multicenter studies that link task-level exposure to validated event logs, staff time, patient outcomes, denials, collections, costs, and incidents. Claims, prior authorization, scheduling, referrals, revenue-cycle management, and other administrative workflows need use-case-specific studies rather than assumptions borrowed from documentation research. Protocols should prespecify adoption denominators, subgroup analyses, version changes, missing-data handling, cost perspective, time horizon, and sponsor roles.

5.6. Limitations

This review was structured but not systematic. Licensed database searches were not completed, screening and extraction were not performed independently by two reviewers, and exact search-flow counts were unavailable. Relevant studies may therefore have been missed. Selective publication and outcome reporting are plausible, and funding or product-provision relationships may influence a rapidly developing evidence base.

Most primary studies evaluated ambient documentation in short-term United States academic or integrated settings. Products, prompts, workflow configurations, and model versions can change faster than publication. Outcome definitions and denominators were heterogeneous, and well-being measures were often self-reported. No organization-level dataset was analyzed for this article. The proposed framework is evidence informed but has not undergone consensus validation, psychometric testing, or prospective multicenter evaluation.

6. Conclusions

Generative artificial intelligence can improve selected healthcare administrative and hybrid workflows, particularly ambient documentation, but benefits depend on the product, use intensity, setting, and surrounding work system. The evidence for operational relief is stronger than the evidence for direct financial return.

Health systems should evaluate GenAI through sequential gates for technical validity, workflow fit, operational effectiveness, financial conversion, and equity and sustainability. Time saved, capacity, RVUs, gross charges, avoided hiring, realized cash savings, and net benefit must remain separate constructs. Until complete local cost and cash data are linked to a credible counterfactual, ROI should be reported as a scenario rather than a fact.

Abbreviations

AI, artificial intelligence; EHR, electronic health record; GenAI, generative artificial intelligence; NIST, National Institute of Standards and Technology; ROI, return on investment; RVU, relative value unit; TCO, total cost of ownership; wRVU, work relative value unit.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] Tseng, P.; Kaplan, R.S.; Richman, B.D.; Shah, M.A.; Schulman, K.A. Administrative costs associated with physician billing and insurance-related activities at an academic health care system. *JAMA* 2018, 319, 691-697. <https://doi.org/10.1001/jama.2017.19148>
- [2] Arndt, B.G.; Beasley, J.W.; Watkinson, M.D.; et al. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. *Ann. Fam. Med.* 2017, 15, 419-426. <https://doi.org/10.1370/afm.2121>
- [3] Gaffney, A.; Woolhandler, S.; Cai, C.; et al. Medical documentation burden among US office-based physicians in 2019: A national study. *JAMA Intern. Med.* 2022, 182, 564-566. <https://doi.org/10.1001/jamainternmed.2022.0372>
- [4] Everson, J.; Nong, P.; Richwine, C. Uptake of generative AI integrated with electronic health records in US hospitals. *JAMA Netw. Open* 2025, 8, e2549463. <https://doi.org/10.1001/jamanetworkopen.2025.49463>
- [5] Sittig, D.F.; Singh, H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual. Saf. Health Care* 2010, 19, i68-i74. <https://doi.org/10.1136/qshc.2010.042085>
- [6] Feldstein, A.C.; Glasgow, R.E. A practical, robust implementation and sustainability model for integrating research findings into practice. *Jt. Comm. J. Qual. Patient Saf.* 2008, 34, 228-243. [https://doi.org/10.1016/S1553-7250\(08\)34030-6](https://doi.org/10.1016/S1553-7250(08)34030-6)
- [7] Glasgow, R.E.; Vogt, T.M.; Boles, S.M. Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *Am. J. Public Health* 1999, 89, 1322-1327. <https://doi.org/10.2105/AJPH.89.9.1322>
- [8] Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA: A scale for the quality assessment of narrative review articles. *Res. Integr. Peer Rev.* 2019, 4, 5. <https://doi.org/10.1186/s41073-019-0064-8>
- [9] Rotenstein, L.S.; Holmgren, A.J.; Thombley, R.; et al. Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: A multisite study. *JAMA* 2026, 335, 1408-1417. <https://doi.org/10.1001/jama.2026.2253>
- [10] Lukac, P.J.; Turner, W.; Vangala, S.; et al. Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI* 2025, 2. <https://doi.org/10.1056/Aloa2501000>

- [11] Afshar, M.; Baumann, M.R.; Resnik, F.; et al. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI* 2025, 2. <https://doi.org/10.1056/Aloa2500945>
- [12] Pearlman, K.; Wan, W.; Shah, S.; Laiteerapong, N. Use of an AI scribe and electronic health record efficiency. *JAMA Netw. Open* 2025, 8, e2537000. <https://doi.org/10.1001/jamanetworkopen.2025.37000>
- [13] Husa, R.A.; Haggerty, J.; Nute, A.W.; et al. Ambient artificial intelligence use and clinician documentation burden, productivity, and efficiency. *JAMA Netw. Open* 2026, 9, e2615762. <https://doi.org/10.1001/jamanetworkopen.2026.15762>
- [14] Dutta, S.; You, J.G.T.; Dunham, L.; et al. Medical scribe and ambient artificial intelligence impact on emergency physician documentation burden and clinical productivity. *Ann. Emerg. Med.* 2026, in press. <https://doi.org/10.1016/j.annemergmed.2026.04.022>
- [15] Ma, S.P.; Liang, A.S.; Shah, S.J.; et al. Ambient artificial intelligence scribes: Utilization and impact on documentation time. *J. Am. Med. Inform. Assoc.* 2025, 32, 381-385. <https://doi.org/10.1093/jamia/ocae304>
- [16] Duggan, M.J.; Gervase, J.; Schoenbaum, A.; et al. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. *JAMA Netw. Open* 2025, 8, e2460637. <https://doi.org/10.1001/jamanetworkopen.2024.60637>
- [17] Olson, K.D.; Meeker, D.; Troup, M.; et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Netw. Open* 2025, 8, e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976>
- [18] Holmgren, A.J.; Fenton, C.L.; Thombly, R.; et al. Ambient artificial intelligence scribes and physician financial productivity. *JAMA Netw. Open* 2026, 9, e2553233. <https://doi.org/10.1001/jamanetworkopen.2025.53233>
- [19] Garcia, P.; Ma, S.P.; Shah, S.; et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw. Open* 2024, 7, e243201. <https://doi.org/10.1001/jamanetworkopen.2024.3201>
- [20] Mandal, S.; Wiesenfeld, B.M.; Szerencsy, A.C.; et al. Utilization of generative AI-drafted responses for managing patient-provider communication. *NPJ Digit. Med.* 2025, 8, 591. <https://doi.org/10.1038/s41746-025-01972-w>
- [21] 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, 27. <https://doi.org/10.1038/s44401-025-00032-5>
- [22] Grolleau, F.; Liang, A.S.; Keyes, T.; et al. Physician-reported safety outcomes of AI-generated hospital course summaries. *JAMA Netw. Open* 2026, 9, e2616556. <https://doi.org/10.1001/jamanetworkopen.2026.16556>
- [23] Simmons, A.; Takkavatakarn, K.; McDougal, M.; et al. Extracting International Classification of Diseases codes from clinical documentation using large language models. *Appl. Clin. Inform.* 2025, 16, 337-344. <https://doi.org/10.1055/a-2491-3872>
- [24] Husereau, D.; Drummond, M.; Augustovski, F.; et al. Consolidated Health Economic Evaluation Reporting Standards 2022 statement: Updated reporting guidance for health economic evaluations. *BMJ* 2022, 376, e067975. <https://doi.org/10.1136/bmj-2021-067975>
- [25] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0); NIST AI 100-1; NIST: Gaithersburg, MD, USA, 2023. <https://doi.org/10.6028/NIST.AI.100-1>
- [26] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile; NIST AI 600-1; NIST: Gaithersburg, MD, USA, 2024. <https://doi.org/10.6028/NIST.AI.600-1>
- [27] U.S. Government Accountability Office. Science and Technology Spotlight: AI for Medical Notes and Coding; GAO-26-109116; GAO: Washington, DC, USA, 2026. <https://www.gao.gov/products/gao-26-109116>
- [28] U.S. Department of Health and Human Services, Office for Civil Rights. Guidance on HIPAA and Cloud Computing. 2022. <https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html>
- [29] Office of the National Coordinator for Health Information Technology. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Final Rule. 2024. <https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program>
- [30] Centers for Medicare & Medicaid Services. Interoperability and Prior Authorization Final Rule CMS-0057-F. 2024. <https://www.cms.gov/priorities/key-initiatives/burden-reduction/interoperability/final-rule-cms-0057-f>