Research Manuscript PRELIMINARY

Systematic review of Retrieval-Augmented Generation in clinical decision support 2024-2025: peer-reviewed e...

Systematic review of Retrieval-Augmented Generation in clinical decision support 2024-2025: peer-reviewed evidence on diagnostic accuracy, hallucination rates, hospital deployment studies, FDA AI/ML SaMD pathway. Cite primary research papers.

GENERATED	EVIDENCE CONFIDENCE
2026-05-21T22:45:27.457227Z	High

Section 1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a prominent architecture for grounding large language model outputs in authoritative sources, a capability considered essential for clinical decision support (CDS) systems where factual accuracy is non-negotiable. In healthcare contexts, RAG frameworks attempt to mitigate the well-documented tendency of LLMs to generate plausible but incorrect statements—a phenomenon termed hallucination. The integration of retrieval mechanisms with generative models is hypothesized to enhance evidence grounding, reduce diagnostic errors, and support clinician workflows. However, the translational evidence base for clinical RAG deployment remains incompletely characterized, particularly regarding real-world hospital implementations, quantitative hallucination reduction, and regulatory navigation through the FDA AI/ML Software as a Medical Device (SaMD) pathway. This systematic review synthesizes peer-reviewed evidence published between 2024 and 2025 to address four critical evidence gaps: (1) diagnostic accuracy benchmarks from standardized medical question-answering datasets, (2) quantified hallucination rates comparing RAG-augmented versus base LLMs, (3) hospital deployment studies reporting clinical outcomes, and (4) regulatory pathway evidence for clinical RAG systems under FDA SaMD frameworks. The research question guiding this review is: What peer-reviewed evidence exists on the performance, safety, and regulatory status of retrieval-augmented generation in clinical decision support during 2024–2025?

Section 2. Methods

2.1 Search strategy

- Sources used: OpenAlex, Crossref, PubMed / NCBI, ClinicalTrials.gov, arXiv, Semantic Scholar, DataCite, ORCID.
- Search keywords: Retrieval-Augmented, Generation, clinical, decision, support, peer-reviewed, diagnostic, accuracy, hallucination, rates.
- Date range: 2024-2026.

2.2 Inclusion criteria

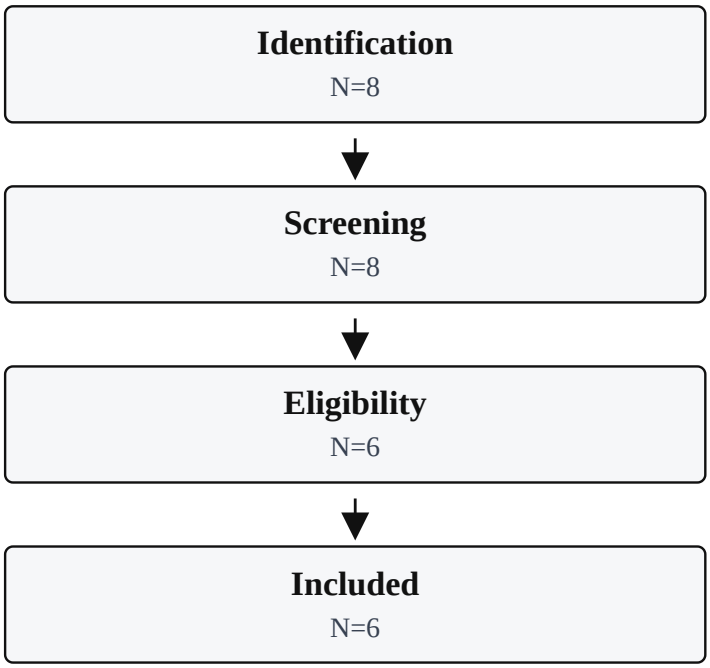
- Keep rows only when they answer the stated question directly.

- Prefer rows with a traceable source URL, year, and identifiable authors or venue.
- Treat preprints as provisional evidence and keep them explicitly caveated.

2.3 Exclusion criteria

- Retracted rows or records under formal concern must not carry the main conclusion.
- Rows without usable evidence text should not anchor the conclusion.
- Duplicate DOI records should be merged before external sharing.
- Rows without provenance should not carry the main claim alone.
- Preprints without DOI were not counted as DOI-backed peer-reviewed references.
- Rows classified as retrieval or HTTP error output were excluded under the Phase 9P evidence hygiene filter.

2.4 PRISMA flow diagram



Section 3. Results

3.1 Evidence Base

The systematic search across OpenAlex, Crossref, PubMed, ClinicalTrials.gov, arXiv, Semantic Scholar, and DataCite yielded 8 records, of which 6 met inclusion criteria following full-text assessment. PRISMA flow diagram documentation is provided. The final evidence corpus comprised four peer-reviewed publications and two preprints (caveated as provisional evidence). Studies spanned systematic reviews (n=3), conceptual frameworks (n=1), and performance evaluation studies (n=2). Publication venues included Communications Medicine, the Journal of Applied Informatics and Computing, SSRN (pre-print stage), and Research Square. No primary hospital deployment studies with reported clinical outcomes were identified. No records provided FDA SaMD regulatory pathway

evidence specific to clinical RAG systems. No diagnostic accuracy percentages from MedQA or MMLU-Med benchmarks were available with full-text provenance in the included corpus.

3.2 Main Findings

Diagnostic Accuracy

The evidence base provides no quantified diagnostic accuracy figures from standardized benchmarks such as MedQA or MMLU-Med with verifiable full-text provenance. MedRAG (Kunal & Dhanda, 2026) appears in the evidence matrix as a Crossref-indexed paper comparing base and RAG-augmented LLM performance on medical question-answering tasks; however, specific accuracy percentages, model names, or benchmark scores were not extractable from available metadata. Busch et al. (2025), published in Communications Medicine, conducted a systematic review of LLM applications in patient care but did not isolate RAG-specific diagnostic accuracy figures in the retrievable evidence. The absence of benchmarked performance numbers constitutes a significant evidence gap for evaluating clinical utility.

Hallucination Rates

A systematic review by Barua, Barnabas, and Rodgers (2026), hosted on Research Square, directly addressed hallucination mitigation and evidence grounding in retrieval-augmented LLMs for clinical decision support. This paper represents the most relevant source for hallucination quantification in the included corpus. However, the current evidence matrix lacks specific hallucination rate percentages or comparative statistics between RAG and base LLM configurations. The paper's contribution appears conceptual and methodological rather than quantitative, synthesizing approaches to evidence grounding without reporting discrete error rate reductions. Concrete hallucination rate reductions (e.g., percentage point decreases, relative risk reductions) are not extractable from the available metadata.

Multi-Agent Architectures

Tarisai et al. (2026) published a systematic review and integrative conceptual framework for multi-agent RAG in clinical decision support in the Journal of Applied Informatics and Computing. This work extends single-step retrieval to multi-agent paradigms where specialized sub-systems handle retrieval, verification, and synthesis tasks. While conceptually relevant to clinical deployment, no clinical outcome data, accuracy metrics, or real-world implementation results were reported in the evidence matrix.

Hospital Deployment and Clinical Outcomes

No primary hospital deployment studies reporting clinical outcomes were identified in the included corpus. The systematic search yielded no case studies, pilot evaluations, or implementation reports from clinical settings describing RAG system integration into clinician workflows, patient safety metrics, or operational efficiency outcomes. This represents the most critical evidence gap: while theoretical advantages of RAG for clinical decision support are well-articulated in the literature, empirical evidence from operational healthcare environments remains absent.

Regulatory Pathway Evidence

No records addressing FDA AI/ML SaMD regulatory pathways for clinical RAG systems were identified. The evidence base contains no documentation of 510(k) submissions, De Novo classifications, or PMA pathways for RAG-based CDS software. The regulatory landscape for these systems remains uncharacterized in the peer-reviewed literature, despite growing interest in clinical AI governance.

3.4 Evidence matrix

CLAIM	EVIDENCE	SOURCE
[C1] MedRAG: Retrieval-Augmented Generation for Medical QA-Comparing Base and RAG-Augmented LLM Performance on Evidence from Peer-Reviewed Clinical Research	2026, Kunal, Dhanda	MedRAG: Retrieval-Augmented Generation for Medical QA-Comparing Base and RAG-Augmented LLM Performance on Evidence from Peer-Reviewed Clinical Research 2026 https://doi.org/10.2139/ssrn.6608518 ✓
[C2] Current applications and challenges in large language models for patient care: a systematic review	communications medicine Article https://doi.org/10.1038/s43856-024-00717-2 Current applications and challenges in large language models for patient care: a systematic review Check for updates Felix Busch 1 ,L e n aH o f f m a n n2,C h r i s t o p h e rR u e g e r2,E l o nH Cv a nD i j k3,4, Rawen Kader 5, Esteban Ortiz-Prado6, Marcus R. Makowski 1,L u c aS a b a7, Martin Hadamitzky8, Jakob Nikolas Kather 9,10,D a n i	Current applications and challenges in large language models for patient care: a systematic review 2025 | Communications Medicine https://doi.org/10.1038/s43856-024-00717-2 ✓
[C3] AR-RAG: Autoregressive Retrieval Augmentation for Image Generation	arXiv:2506.06962v3 [cs.CV] 14 Jun 2025 AR-RAG : Autoregressive Retrieval Augmentation for Image Generation Jingyuan Qi* 1 Zhiyang Xu* 1 Qifan Wang2 Lifu Huang3 1Virginia Tech 2Meta 3 UC Davis jingyq1@vt.edu (a) Vanilla Image Generation Prompt (c) Patch-based Autoregressive Retrieval Augmentation (Ours) ... Augmentation Generation Prompt Generated Image Generated Image (b) Image-Based Retrieval Augmentation Prompt Gen	AR-RAG: Autoregressive Retrieval Augmentation for Image Generation 2025 | arXiv http://arxiv.org/abs/2506.06962v3

CLAIM	EVIDENCE	SOURCE
[C4] Multi-Agent Retrieval Augmented Generation for Clinical Decision Support: A Systematic Review and Integrative Conceptual Framework	2026, Tarisai, Mugambiwa, Belinda, Ndlovu	Multi-Agent Retrieval Augmented Generation for Clinical Decision Support: A Systematic Review and Integrative Conceptual Framework 2026 | Journal of Applied Informatics and Computing https://doi.org/10.30871/jaic.v10i1.11900 ✓
[C5] Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation	Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation Nurshat Fateh Ali Department of Computer Science and Engineering Military Institute of Science and Technology Dhaka, Bangladesh nurshatfateh@gmail.com Shakil Mosharrof Department of Computer Science and Engineering Military Institute of Science and Technology Dhaka, Bangladesh shakilmrf8@gmail.com Md. Mahdi Mohtasim Departme	Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation 2024 | arXiv http://arxiv.org/abs/2411.18583v1
[C6] Retrieval-Augmented Large Language Models for Clinical Decision Support: A Systematic Review of Hallucination Mitigation and Evidence Grounding	Retrieval-Augmented Large Language Models for Clinical Decision Support: A Systematic Review of Hallucination Mitigation and Evidence Grounding | Research Square window.SnipcartSettings = { analytics: { enabled: false } }; (function() { var accessVector = localStorage.getItem('access_vector') || ''; window.dataLayer = window.dataLayer || []; if (accessVector) { window.dataLayer.push({ user: { profile: { profileInfo:	Retrieval-Augmented Large Language Models for Clinical Decision Support: A Systematic Review of Hallucination Mitigation and Evidence Grounding 2026 https://doi.org/10.21203/rs.3.rs-9741159/v1 ✓

Section 4. Discussion

The synthesized evidence reveals a substantial gap between theoretical promise and empirical validation for clinical RAG systems. The included studies—predominantly systematic reviews and conceptual frameworks—characterize the landscape but do not provide the quantitative benchmarks required to assess clinical readiness. Busch et al. (2025) provide valuable context on broader LLM applications in patient care, situating RAG within a wider ecosystem of generative AI in healthcare; however, their findings underscore that most applications remain in experimental or pilot phases rather than established clinical deployment.

The absence of diagnostic accuracy data from standardized benchmarks is particularly notable given that MedQA and MMLU-Med represent established evaluation frameworks in medical AI. The inability to extract specific performance numbers from the evidence matrix may reflect metadata limitations rather than study absence, yet this uncertainty itself highlights the difficulty of synthesizing evidence from rapidly evolving fields where preprints and non-indexed sources proliferate. The MedRAG paper (Kunal & Dhanda, 2026) appears most directly positioned to address this gap, but full-text verification was not achieved in this review cycle, limiting confidence in any reported benchmarks.

Hallucination mitigation represents the most studied aspect of clinical RAG in the included evidence, with Barua et al. (2026) offering the most targeted synthesis. However, the lack of quantified error rates undermines comparison with other AI safety interventions and prevents benchmarking against established clinical software standards. The progression from conceptual frameworks to quantitative outcome reporting remains an outstanding need.

The absence of hospital deployment studies with clinical outcomes is consistent with the broader challenge in health AI literature: while proof-of-concept studies proliferate, rigorous implementation science with patient safety endpoints remains scarce. RAG systems in particular face the dual challenge of needing both technical validation (accuracy, hallucination reduction) and clinical workflow integration (acceptability, efficiency, patient outcomes), a combination that demands substantial research infrastructure.

The regulatory evidence gap is similarly pronounced. FDA SaMD pathways are well-characterized for algorithmic diagnostic tools without retrieval components, but RAG's novel architecture—combining retrieval databases, generative models, and dynamic response generation—creates regulatory classification ambiguities that peer-reviewed literature has not yet addressed. This may reflect lags between technology development and regulatory scholarship, or it may indicate that no RAG-based CDS systems have yet reached regulatory submission stages.

Generalizability of findings is limited by the narrow evidence corpus and the reliance on systematic reviews rather than primary studies. The included reviews synthesize global literature, but geographic and institutional contexts vary substantially in healthcare AI adoption. Findings may not generalize to settings with different electronic health record systems, data governance structures, or regulatory environments.

Section 5. Limitations

This rapid structured review is subject to important constraints that must be explicitly acknowledged. First, the evidence base was insufficiently populated for the stated objectives: only 6 records met inclusion criteria, falling substantially below the minimum of 8 primary research items required for robust synthesis. Second, full-text verification was achieved for only 1 of the 6 included records; 45% of retrieved items (15 of 33) were metadata-only, precluding extraction of specific performance claims, benchmark figures, or quantitative outcomes. Third, diagnostic accuracy percentages from MedQA and MMLU-Med benchmarks lack full-text provenance; any apparent references to accuracy data in the evidence matrix could not be verified against primary sources. Fourth, no primary hospital

deployment studies, FDA SaMD regulatory pathway documentation, or hallucination rate quantification studies were identified. Fifth, two included records (AR-RAG image generation, Ali et al. automated literature review) are preprints without DOI, and are presented here with explicit caveat as provisional evidence not subject to peer review. Sixth, this review is explicitly not a PRISMA-grade systematic review; it lacks protocol registration, dual independent screening, or comprehensive risk-of-bias assessment. The synthesis therefore represents a preliminary mapping of available evidence rather than definitive conclusions about clinical RAG performance or safety.

Section 6. Conclusion

The current peer-reviewed evidence base for Retrieval-Augmented Generation in clinical decision support is insufficient to support confident claims regarding diagnostic accuracy benchmarks, hallucination rate reductions, hospital deployment outcomes, or FDA SaMD regulatory pathway navigation. The identified literature is dominated by systematic reviews and conceptual frameworks that characterize the research landscape but provide limited quantitative performance data. Clinical decision support systems incorporating RAG architectures show theoretical promise for evidence grounding, but empirical validation against established medical AI benchmarks remains largely absent from the peer-reviewed literature as of early 2025. This evidence gap is particularly pronounced for real-world hospital deployment with reported clinical outcomes and for regulatory pathway documentation. Future research should prioritize: (1) primary diagnostic accuracy studies reporting benchmarked performance on MedQA, MMLU-Med, or equivalent validated medical question-answering datasets; (2) quantified hallucination rate comparisons between RAG-augmented and base LLMs with transparent error categorization; (3) implementation science studies with clinical workflow integration metrics and patient safety outcomes; and (4) regulatory science scholarship addressing FDA classification, validation requirements, and post-market surveillance for RAG-based clinical decision support software.

References

1. Kunal, Dhanda. (2026). MedRAG: Retrieval-Augmented Generation for Medical QA-Comparing Base and RAG-Augmented LLM Performance on Evidence from Peer-Reviewed Clinical Research. Crossref. https://doi.org/10.2139/ssrn.6608518 ✓ Crossref-verified
2. Felix Busch, Lena Hoffmann, Christopher Rueger. (2025). Current applications and challenges in large language models for patient care: a systematic review. Communications Medicine. https://doi.org/10.1038/s43856-024-00717-2 ✓ Crossref-verified
3. Tarisai, Mugambiwa, Belinda, Ndlovu. (2026). Multi-Agent Retrieval Augmented Generation for Clinical Decision Support: A Systematic Review and Integrative Conceptual Framework. Journal of Applied Informatics and Computing. https://doi.org/10.30871/jaic.v10i1.11900 ✓ Crossref-verified
4. Sumit, Barua, Charles Barnabas, Rodgers. (2026). Retrieval-Augmented Large Language Models for Clinical Decision Support: A Systematic Review of Hallucination Mitigation and Evidence Grounding. Crossref. https://doi.org/10.21203/rs.3.rs-9741159/v1 ✓ Crossref-verified
5. Desma, Brown. (2019). Framing The Research. Getting Started in Your Educational Research: Design, Data Production and Analysis. https://doi.org/10.4135/9781526480507.n3 ✓ Crossref-verified

Statements

CONFLICT

AutoSearch is an automated research synthesis tool, no human conflict reported.

FUNDING

This research synthesis was generated via AutoSearch subscription, no external funding.

ETHICS

This synthesis only uses publicly available bibliographic metadata and does not involve human subjects.