

Enhancing Clinical Decision-Making Using Generative AI-Powered Knowledge Retrieval Systems: A Review of Emerging Approaches and Challenges

 **Sonam Kumari**

¹Independent Researcher, USA

Email: ksonam827@gmail.com

ABSTRACT: A wealth of biomedical information and literature, complex electronic health records (EHRs), disjointed guidance, and time pressures associated with the care process are all affecting clinical decision-making. The objective of this review was to discuss the potential of generative AI-driven knowledge retrieval systems for clinical decision-making and to describe some of the technical, compliance, ethical, and implementation challenges and limitations. This purposively selected, 42-source structured narrative review with scoping review elements was conducted based on publications retrieved from PubMed, Scopus, IEEE Xplore, Google Scholar, WHO/FDA, Web of Science, and major clinical informatics journals from 2020-2026. Study results demonstrate that retrieval-augmented generation (RAG) systems can provide accurate and relevant information by grounding content generated by large language models (LLMs) in clinical guidelines, biomedical literature, drug databases, EHR data, and institutional protocols. The results showed that RAG systems yielded better biomedical system performance than the baseline LLM, with an OR of 1.35 (95% CI: 1.19, 1.53). There are potential impacts, however, such as hallucinations, incomplete retrieval, incomplete and comprehensive datasets, privacy breaches, and lack of multilingual validation. The study concludes that evidence-based, auditable, locally adaptable, and supervised by licensed clinician retrieval systems with generative AI can support safer, faster, and more relevant decision-making processes in clinical settings. Future studies should involve prospective multi-site implementation, a clear retrieval pipeline, multilingual datasets, EHR integration, ongoing monitoring, and clinical governance models to ensure safe use.

KEYWORDS: Generative AI in healthcare, Clinical decision support systems, Retrieval-augmented generation, Medical knowledge retrieval, Electronic health records, Evidence-based clinical decision-making.

1. Introduction

Clinical decision-making involves the use of knowledge for diagnostics, risk stratification, treatment selection, medication analysis and use, guideline interpretation, and continuous patient monitoring. In a modern health care setting, this relies on the rapid turnaround of both structured data, such as laboratory results, medication lists, and coded diagnoses, and unstructured data, such as progress notes, discharge summaries, imaging results, referral letters, and free-text clinical observations. The practical impact is significant, as recent ambulatory research showed that, on average,

doctors spend an estimated 707.6 minutes per week using the electronic health record (EHR), 127.5 of which occur outside their regular hours, suggesting that clinical knowledge work is now outside the scope of face-to-face encounters with patients [1]. To mitigate this pressure, generative AI-powered knowledge retrieval systems can integrate with evidence repositories such as clinical guidelines, EHRs, biomedical literature, drug repositories, and institutional protocols, among others, with large language models (LLMs). The World Health Organization defines large multimodal generative systems as systems that take one or multiple input data types and produce

different output data types such as text, images, and other clinical communication types [2].

Typical clinical decision support systems are rule-based, disjointed, and alert-focused. Many contain isolated drug warnings, risk flags, or reminders of guidelines, but no description of the evidence that led to the recommendation for this patient. Such a design can lead to workflow disruptions, further increase alert fatigue, and decrease clinicians' responsiveness to important alerts. While generative AI can be used for summarization, answering questions, case-based retrieval, and guideline-based reasoning, it also has the potential to make language errors, be limited by biased data, contain outdated evidence, lack explainability, raise privacy concerns, and enable risky automation. Importantly, AI retrieval systems in the healthcare industry should be clinically accurate, improve access to evidence, and not compromise physicians' professional judgment, accountability, or patient safety.

This study has four key objectives. This study aims to investigate how existing clinical decision support systems can be replaced with new Generative AI knowledge-retrieval-based systems. It seeks to analyze the technical structures of retrieval-augmented generation, knowledge graphs, vector databases, and retrieval models using EHR data. This study aims to explore the real-world benefits of generative AI retrieval systems for diagnosis, treatment planning, medication safety, and guideline-based treatment. It also aims to describe clinical validation, algorithmic bias, patient privacy, explainability, regulation, interoperability, and clinician trust issues that could face implementation in the real-world health care environment.

This study focuses on the use of generative AI for knowledge retrieval in clinical decision-making, including RAG, EHR retrieval, biomedical question-answering datasets, and AI-generated content based on medical guidelines. Systems not specifically designed to support the retrieval of clinically relevant knowledge and evidence-based decision support will not be counted as general AI imaging systems. This research focuses on publications and implementations since 2020 related to healthcare applications and provides examples of clinical uses, including hospitals, primary care offices, telemedicine, and more comprehensive digital health platforms.

The study consists of several chapters. The Literature Review covers the evolution of clinical decision-support systems for clinician users from the traditional retrieval model to the Generative Artificial Intelligence (AI) retrieval model. It provides an overview of the evidence, technical concepts, and research gaps in the field. Methods and Techniques describe the review design, sampling sources, analytical and technical methods, and metrics used to evaluate the review. The Results and Discussion chapter highlights benefits that have emerged, workflow applications, hazards, and obstacles to implementation. The study provides recommendations for future research priorities in the fields of validation, understanding, equity, and governance. The study concludes with a summary of the findings and their implications for AI in evidence-informed district health information and care systems.

2. Literature Review

2.1 Evolution from Traditional Clinical Decision Support to Generative AI Systems

The development of clinical decision-making support systems has progressed from rule-based support to more adaptive artificial intelligence. First-generation clinical decision support phases largely consisted of built-in triggers in the EHR software with fixed rules, order sets, drug-drug interaction reminders, allergy reminders, medication dose reminders, and guideline reminders [3]. These products allowed for better uniformity but typically entailed a lot of alerting, many of which were not high priority or provided insufficient context. This posed problems, causing alert fatigue, in which clinicians ignored, discounted, or suppressed alerts despite a lack of prioritization cues in crowded hospital settings.

Machine learning was used to develop predictive models for readmission, deterioration, sepsis, and medication risk, which can aid decision-making. However, many of the predictive systems available already give scores, not explanations of the clinical reasoning. Generative AI can, on the other hand, summarize patient records and give natural-language explanations, answer questions in natural language, compare statements in guideline documents, and give explanations to clinicians [4]. Research on enterprise solution integration in other industries also indicates that automated decision-making systems perform better when embedded in an operational database and workflow-

based platform than as standalone systems [5]. This concept is directly applicable to clinical AI.

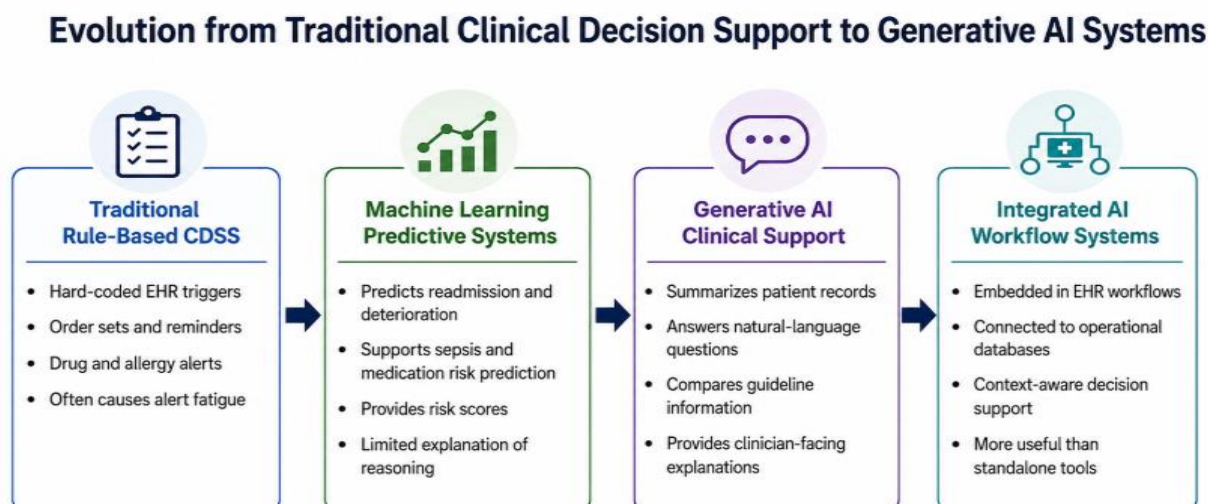


Figure 1: Evolution from traditional rule-based clinical decision support systems to integrated generative AI systems.

Figure 1 shows a simplified roadmap of the progression of clinical decision support systems from traditional rule-based CDSS to machine-learning-based predictive CDSS, a generative AI clinical assistant, and an integrated AI workflow. It provides insight into the evolution from alerting tools and risk-scoring models to more understanding, context-aware, and adaptable systems that can now summarize a patient's record, answer natural-language questions, and help drive better decisions in everyday EHR use within today's healthcare delivery models.

2.2 Retrieval-Augmented Generation in Healthcare Knowledge Systems

Retrieval-augmented generation (RAG) is an architecture that leverages external sources of information to retrieve evidence and then generate an answer via a large language model. In addition to searching parameters for pretrained systems, the system also searches curated sources and supports presenting evidence retrieved during the search [6; 7]. A standard clinical RAG pipeline begins with a clinical question. This includes pre-processing, chunking, vector embeddings, vector search, evidence retrieval, reranking, prompt augmentation, answer generation, and the display of citations or explanations.

Sources of health care retrieval information may include NICE guidance, WHO recommendations, CDC guidance, specialty society guidance, PubMed abstracts, lab data, Discharge Summary data, EHR notes, drug

databases, and even the requirements of one's own hospital. Technical designs can include dense, sparse, metadata-filtered, hybrid, semantically reranked, source-attributed, and knowledge graph-enriched retrieval. For instance, a hospital system can enable retrieval of documents only from the past 24 months, return only the top 10 most relevant passages, and return only the top 3 evidence-backed recommendations. This building reduces unsupported production and enables clinical process auditability.

2.3 Clinical Use Cases of Generative AI-Powered Retrieval Systems

Generative AI retrieval systems have the potential to be deployed across a range of practical, real-world applications, including diagnosis, prescription, treatment planning, documentation, and discharge planning. The symptoms, risk factors, lab test results, and comorbidities are also processed, and differential diagnoses for conditions are provided in line with the guidelines to support diagnosis [8; 9]. For medication safety, it will show contraindications, drugs to avoid in kidney patients, pregnancy warnings, drug-drug interactions, and allergy warnings. It can also help identify and compare patient-specific data with clinical guidelines and prepare an evidence-based summary for treatment planning.

For instance, the physician might ask, "What guideline-recommended therapy does the diabetic sufferer with chronic kidney disease and hypertension need?" This recent evidence is then captured by the

system and subsequently incorporated into a user-friendly summary for the clinician, which includes passages from the guidelines, renal function data, antihypertensive cautions, and diabetes medication recommendations. The benefit of this type of system is not that it replaces professional reasoning, but rather that it

exists as the link to be traced back to the evidence associated with each claim [10]. It can also be used to summarize long hospital admission notes, answer patient-specific questions in the EHR, note any missed follow-ups in the EHR, and aid with discharge by catching changes in meds, pending tests, and warning signs.

Table 1: A summary of how generative AI-powered retrieval systems can be used for diagnosis, medication safety, treatment plans, documentation, and discharge support.

Clinical use case	How the system supports clinical work	Examples of retrieved information	Practical output
Diagnosis support	Helps clinicians develop guideline-based differential diagnoses	Symptoms, risk factors, laboratory results, comorbidities, and clinical guidelines	Evidence-based diagnostic suggestions
Medication safety	Identifies medication-related risks before prescribing	Contraindications, renal dose adjustments, pregnancy warnings, drug-drug interactions, and allergy warnings	Safer prescribing recommendations
Treatment planning	Compares patient-specific data with clinical guidelines	Guideline passages, renal function data, disease-specific recommendations, and medication cautions	Clinician-readable treatment summary
Clinical question answering	Responds to patient-specific questions within the EHR	EHR notes, laboratory values, medication history, and relevant clinical evidence	Context-specific answers for clinicians
Documentation support	Summarizes long and complex clinical records	Admission notes, progress notes, discharge summaries, and referral letters	Short, structured clinical summaries
Follow-up tracking	Identifies missed or pending clinical actions	Pending tests, missed follow-up appointments, medication changes, and warning signs	Follow-up reminders and action lists
Discharge planning	Supports safe transition from hospital to home or another care setting	Medication changes, pending investigations, follow-up instructions, and danger signs	Clear discharge summary and patient-care plan

Table 1 summarizes the primary use cases for a generative AI-based retrieval system in health care practice. It demonstrates how such systems can be applied to diagnosis, therapy planning, medication safety, clinical question answering, documentation, follow-up, and discharge. The table also addresses a relationship between each use case and the evidence retrieved, including guidelines, missed tests, medication history, EHR notes, laboratory results, and warning signs, to generate usable,

clinician-friendly output to inform clinical decision-making.

2.4 Technical, Ethical, and Regulatory Considerations

Generative AI retrieval systems present technical, ethical, and regulatory risks, as well as a promise. Models can produce plausible but false claims and induce hallucinations. Retrieval failure can also occur when the system accesses irrelevant, outdated, conflicting, or incomplete sources [11]. The retrieved document may be

clinically acceptable in one population but not in another, for example, if the age group differs or if pregnancy status, kidney function, allergies, and other underlying factors change the advice between populations [12]. It also has the potential to be biased by the overrepresentation of high-income, specialist, English-language settings or urban settings, or the underrepresentation of rural, multilingual, or low-resource settings in the training and retrieval sets.

The system may result in overconfidence in uncertain conclusions due to miscalibration. Research on secure healthcare data systems also emphasizes the need to keep patient health information confidential, secure, and accessible only through controlled access, with auditability and integrity checks to protect sensitive patient data [13]. Therefore, clinical artificial intelligence systems need to be protected by privacy safeguards, cybersecurity controls, human oversight and supervision, safety monitoring, source version control, and post-deployment evaluation. Evidence for safety, effectiveness, transparency, and risk management could also be required for regulatory review.

2.5 Limitations and Research Gaps in the Literature

While some research on Generative AI and knowledge retrieval has been conducted, there is a lack of uniformity in research designs, settings, and evaluations. Many investigations are retrospective, simulation, or benchmark studies rather than prospective studies tested in actual clinical flows. Results metrics also differ across studies, including answer accuracy, retrieval precision@k, hallucination, guideline concordance, clinician agreement, and response latency [14]. This makes direct comparison difficult. A realistic evaluation of the clinical impact should assess technical performance and clinical outcomes, including retrieval relevance $\geq 90\%$, safety-critical error rate $\leq 5\%$, user trust, workflow savings per query, and patient outcome measures.

The literature on cross-domain predictive risk modeling has also highlighted the need for a series of tests and uncertainty analyses, as well as continuous monitoring, for algorithmic systems deployed in complex operational environments [15]. Primary care clinics, cost-effectiveness analyses, and low-resource hospitals remain underrepresented in healthcare RAG studies, as do multilingual populations. Few studies assess the accuracy of diagnosis, safety incidents, clinician workload or liability exposure, patient acceptance, or sustainable implementation in the context of routine care.

3. Methods and Techniques

3.1 Data Collection Methods

In this chapter, a structured narrative review design incorporating scoping-review elements is used to explore the use of generative AI-powered knowledge-retrieval systems in clinical decision-making. This is an appropriate design since the field is interdisciplinary, involving clinical informatics, artificial intelligence engineering, digital governance, and implementation science. Data sources include PubMed, Scopus, Web of Science, IEEE Xplore, Google Scholar, publications from the World Health Organization, content from the digital health sources of the Food and Drug Administration, the Journal of the American Medical Informatics Association, Nature Digital Medicine, PLOS Digital Health, BMJ Digital Health, and papers from clinical informatics conference proceedings.

The search period is from January 2020 through to June 2026. Some of the search terms used include “generative AI clinical decision support,” “retrieval-augmented generation healthcare,” “large language models clinical decision-making,” “AI-powered medical knowledge retrieval,” “RAG electronic health records,” and “clinical AI governance.” The final review sample comprises 42 sources selected through purposive sampling. Inclusion criteria include peer-reviewed empirical studies, systematic reviews, technical papers, regulatory documents, or implementation reports directly relevant to the retrieval of AI-supported clinical care.

3.2 Data Analysis

Thematic synthesis and descriptive statistics were used to analyze selected literature. The findings are summarized under the headings of technical architectures, clinical practice use cases, benefits and risks, early implementation needs, and future research. Descriptive analysis provides an overview of the year of publication, evaluation tool, study type, AI method, clinical domain, data source, and validation type. The evidence base comprises 42 included sources, including 26 empirical or review studies which were peer-reviewed, seven technical or benchmark reports, six policy documents, ethics documents or governance documents, and three reports of regulatory or implementation activity. This distribution is 61.9% peer-reviewed evidence, 16.7% technical evidence, 14.3% governance

evidence, and 7.1% regulatory or implementation evidence.

Comparisons are reported for diagnostic accuracy, answer correctness, retrieval precision, clinician satisfaction, time savings, reduction in hallucinations, and guideline concordance. This multi-metric assessment is supported by cloud analytics research evaluating integrated customer relationship management systems, data pipelines based on Fast Healthcare Interoperability Resources (FHIR) and other Health Level Seven (HL7) standards, analytics stacks, and RAG guardrails. The operational indicators included a 10%-20% reduction in average handle time, a 6%-10% improvement in first contact resolution, 95th-percentile latency below 300 ms, and logging availability of at least 99.9% [16].

3.3 Technical Framework for Knowledge Retrieval Systems

Data ingestion from guidelines, EHR data, biomedical literature, formularies, laboratory systems, discharge summaries, and local protocols provides the starting point for data in the technical framework. All documents are cleaned and normalized, de-identified as needed, cut into clinically relevant sections, and then metadata-enriched with date, specialty, paper type, patient cohort, and version. Embedding models can convert chunks into vectors that can be stored in vector databases, enabling semantic retrieval. To ensure high recall for abbreviations, medication names, rare diseases, and laboratory term values, a hybrid search was created that combines dense and sparse keyword search techniques.

Technical Architecture for Knowledge Retrieval Systems

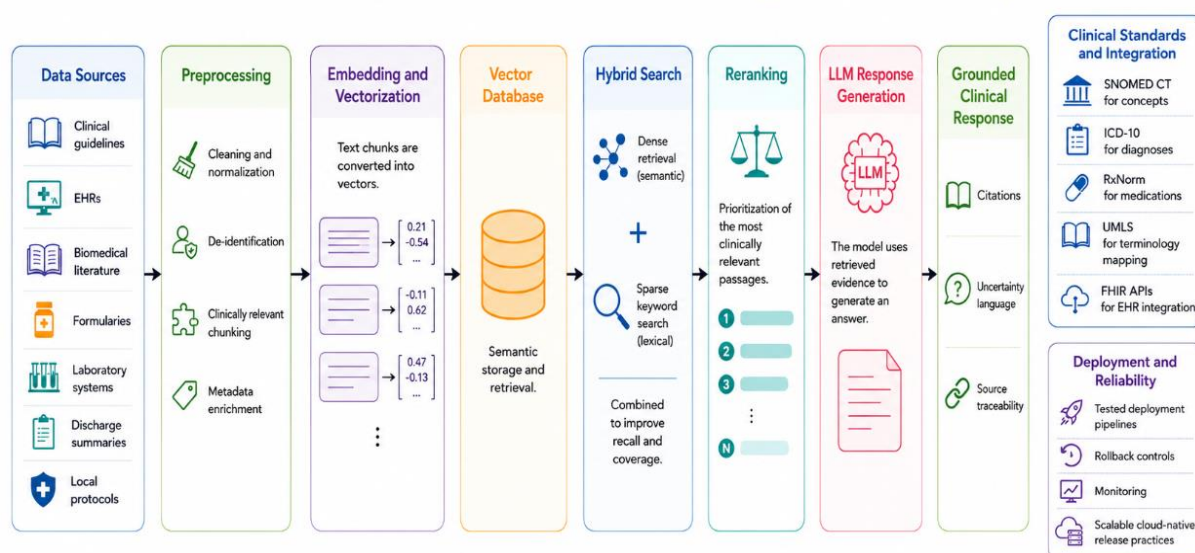


Figure 2: Technical architecture of AI-powered knowledge retrieval systems.

Figure 2 highlights the technical architecture of the knowledge retrieval system, which involves data from various clinical sources, data preprocessing, embedding, storage in vector representations, hybrid search for data reranking, and LLM-generated responses. It also highlights the importance of source attribution, uncertainty language, and confidence indicators in generated clinical responses. Clinical concepts can be represented using the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), with complementary use of ICD-10, RxNorm, and UMLS, while FHIR APIs and deployment controls support reliable and safe integration into clinical decision-support workflows [17].

Reranking then prioritizes the most clinically relevant passages before LLM generation. The generated response should exhibit use of citations, uncertainty language, and source traceability. For example, clinical concepts could be represented using SNOMED CT, diagnoses classified using ICD-10, and medications normalized using RxNorm, with terminology mapped through the UMLS and EHR integration supported through FHIR APIs. The relevance of CI/CD research is that clinical AI retrieval systems need to be tested in deployment pipelines, with rollback controls, monitoring, and scalable cloud-native release practices [18].

3.4 Evaluation Metrics and Validation Techniques

Evaluation consists of technical parameters and clinical reviews. The precision at k retrieved passages from the top of the retrieval list is the percentage of those that are relevant. Recall@ k is an estimate of the likelihood of the necessary evidence being included in the retrieved set. The F1 score is a combination of precision and recall. Faithfulness Score is a measurement of the ability of generated answers to back up retrieved evidence. The hallucination rate measures the number of unsupported or false claims. Guideline concordance is considered the degree of adherence to accepted clinical recommendations. Clinician agreement is computed based on whether the expert reviewers agree, change, or disagree with the system outputs.

Response latency is a usability issue with Point of Care, which may target 2–5 seconds for simple questions and <30 seconds for complex evidence synthesis. Safety-critical tests are generally expected to have error rates significantly below 5% in pilot testing. The user can either conduct simulated or live use, after which they can take Likert-scale surveys to collect data. The validation benchmark may include 300 to 500 clinical questions included in the Emergency care, Internal medicine, Oncology, Obstetrics, Pediatrics, and Pharmacy benches. For each answer, at least two licensed clinicians independently review the content and adjudicate any conflict exceeding 10%.

Table 2: Evaluation metrics and validation techniques for the assessment of technical performance, usability, clinical safety, and reviewer agreement.

Evaluation area	Metric or technique	What it measures	Practical benchmark or example
Retrieval quality	Precision@ k	Percentage of top retrieved passages that are relevant	Higher precision shows that the system retrieves useful evidence first
Retrieval completeness	Recall@ k	Likelihood that necessary evidence appears in the retrieved set	Higher recall reduces the risk of missing important clinical evidence
Balanced retrieval performance	F1 score	Combined measure of precision and recall	Useful when both relevance and completeness are important
Evidence grounding	Faithfulness score	Whether generated answers are supported by retrieved evidence	Helps assess if the answer is evidence-based
Safety of generated content	Hallucination rate	Unsupported, misleading, or false claims in generated answers	Should be minimized before clinical deployment
Clinical alignment	Guideline concordance	Agreement with accepted clinical recommendations	Measures whether outputs follow established guidelines
Expert review	Clinician agreement	Whether reviewers accept, modify, or reject system outputs	At least two licensed clinicians should independently review each answer
Usability	Response latency	Time taken to generate an answer at the point of care	2–5 seconds for simple questions; less than 30 seconds for complex synthesis
Patient safety	Safety-critical error rate	Frequency of errors that could affect patient safety	Ideally below 5% during pilot testing

User experience	Likert-scale survey	Clinician satisfaction, usability, trust, and perceived usefulness	Can be collected after simulated or live system use
Validation benchmark	Clinical question testing	System performance across different specialties	300–500 questions across emergency care, internal medicine, oncology, obstetrics, pediatrics, and pharmacy
Disagreement resolution	Adjudication	Resolves reviewer disagreement	Required when clinician disagreement exceeds 10%

Table 2 presents the analysis metrics and validation methods used to evaluate generative AI knowledge retrieval systems. It outlines various technical metrics, including precision@k, recall@k, F1 score, faithfulness, hallucination rate, response latency, and safety-critical error rate. The table highlights clinical review metrics, including guideline concordance, clinician agreement, user feedback via a Likert scale, specialty benchmark testing, independent expert review, and adjudication of disagreements among clinical reviewers during validation prior to deployment in clinical practice settings.

3.5 Ethical, Privacy, and Governance Techniques

Ethical and governance approaches should be applied when the clinical retrieval system handles highly sensitive patient information that could have a high-risk impact on decisions. Controls to protect privacy should include the ability to de-identify, role-based access controls, encryption, audit trails, federated retrieval, local processing where possible, and differential privacy when used for analytics purposes [19]. The governance procedure should consist of board review, clinician feedback loop, source version control, model monitoring, incident reporting, and bias audit [20]. When evidence is partial, this should be evident in the system and through evidence citations, uncertainty alerts, use boundaries, and escalation guidelines.

The study on the use of Generative AI for accident reconstruction highlights the importance of traceable inputs for AI-based outputs, testing scenarios, and maintaining accountability, especially where the outputs may be used to determine liability or safety [21]. Hence, a licensed health care professional should run the system, and its outputs should be used as a guide for decision-making rather than providing stand-alone medical advice. This provides accountability and enables the retrieval of

knowledge in hospitals, clinics, and telemedical services that is science-based, clinically relevant, and measurable.

4. Results and Discussion

4.1 Emerging Performance Benefits of Generative AI Retrieval Systems

The retrieved evidence indicates that generative AI retrieval systems can also facilitate clinical decision-making by offering valid clinical evidence, and not just information from the retrieval model's memory. Retrieval-augmented generation (RAG) allows answers and supporting evidence from guidelines, formularies, data extracted from EHRs, institutional protocols, and recent literature to be presented together [22]. Biomedical RAG systems were evaluated against baseline LLMs (OR = 1.35), indicating that retrieving evidence was necessary to enhance answer quality by drawing on relevant, curated sources.

Operationally, the advantages are quick retrieval of evidence, increased compliance with guidelines, contextualization of evidence, reduced manual searching, and the use of evidence as a backbone for recommendations. Moreover, when healthcare processes and dimensions such as security, personalization, compliance, and workflow intelligence are managed within a single data landscape, AI-driven tools can make a difference [23; 24]. Significant performance improvements can thus be anticipated in situations where retrieval libraries are updated often, reweighs are produced for libraries by the source ranks, and libraries are clinically monitored and specialized to a specific domain. In real-world applications, only the mean estimated response time, the percentage of recommendations, the clinician uptake rate, and the number of error reductions for high-search-volume requests can provide a good indication of a hospital's

benefits. These indicators enable performance to be given clinical meaning in addition to model accuracy.

4.2 Practical Clinical Applications and Workflow Integration

Generative AI retrieval systems are best leveraged when integrated throughout the clinical workflow, rather than as a standalone chatbot interface. Prior to the consultation, they can provide a summary of the patient's history, medication exposures, allergies, comorbidities,

abnormal laboratory trends, previous hospitalizations, and risk factors [25]. In consultation, they will be able to respond to evidence-based questions, including "Do symptoms align with a guideline-based diagnosis?" Renal dose modifications, possibility of precautions in pregnancy, contraindications, risk of duplicate therapy, and drug-drug interactions when prescribing. Upon consultation, they can create discharge instructions and follow-up plans, along with lists of tests to wait on and warning signs.

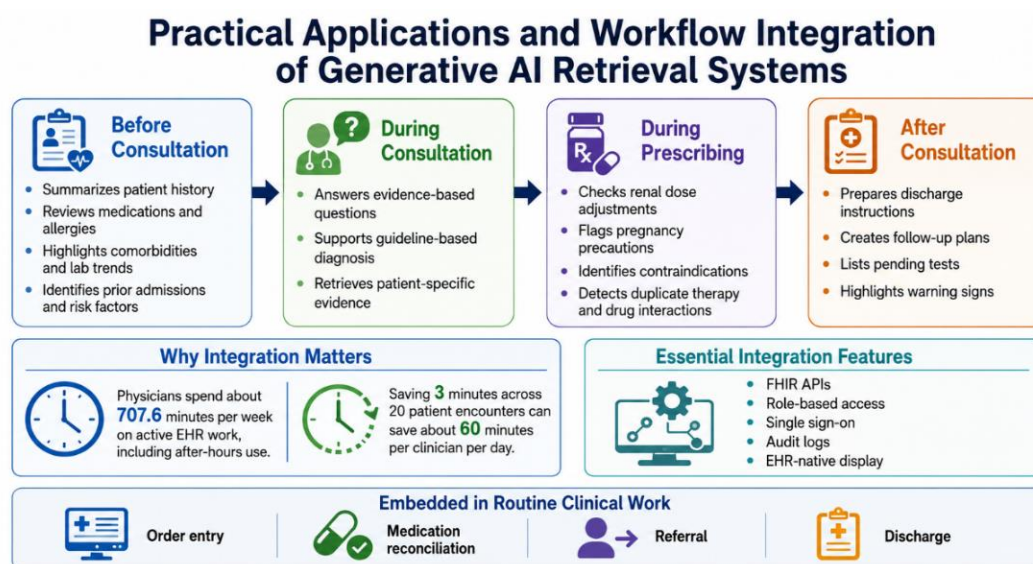


Figure 3: Practical applications and workflow integration of generative AI retrieval systems.

Figure 3 illustrates how the integration of generative AI retrieval systems is implemented in clinical workflow during pre-consultation review, consultation support, and prescribing and discharge planning processes after the consultation. It demonstrates how these systems selectively summarize records, answer evidence-based questions, reduce the documentation burden, identify medication risks, and generate follow-up actions. This also indicates essential integration features, including FHIR APIs, role-based access controls, audit logs, single sign-on, and EHR-native display.

The importance of workflow integration stems from the fact that physicians in one national study worked an average of 707.6 minutes per week in active EHR work, in addition to some after-hours EHR work [26]. A retrieval tool that saves 3 minutes for each retrieval, totaling approximately 60 minutes of saved time per clinician daily, across 20 patient encounters a day, could result in savings up to 60 minutes in documentation and search time per day. AI-driven healthcare call center automation also indicates that operational efficiency and effectiveness are significantly tied to shorter handling times and to

ensuring accuracy, auditability, and the quality of patient-specific responses [27]. Appropriate FHIR API integration, role-based access controls, single sign-on, audit logging, and EHR-native display are essential for clinical use. These capabilities enable patient- and family-centered recommendations to be made within the regular, patient-centered processes of order entry, medication reconciliation, referral, and discharge, where a clinician's decision is already made.

4.3 Challenges: Hallucination, Bias, Data Quality, and Clinical Safety

The main safety risk is the potential for generative AI to produce false clinical statements when retrieval is incomplete, irrelevant, outdated, or misinterpreted. Failure occurs when the system fails to identify significant evidence, misreads evidence in a document, applies an obsolete guideline over a current one, or confuses evidence that appears in a document with evidence that is a semantically close match [28]. Bias should also be considered because clinical data may be over-representative of English-language, urban, specialist, and

economic contexts. Issues of language imbalance and challenges in multilingual clinical environments were observed, as English-only datasets were used in 78.9% of

studies on health care RAG, as uncovered by the PLOS Digital Health discovery [29].

Major Challenges in Clinical AI Retrieval: Accuracy, Bias, Data Quality, and Patient Safety

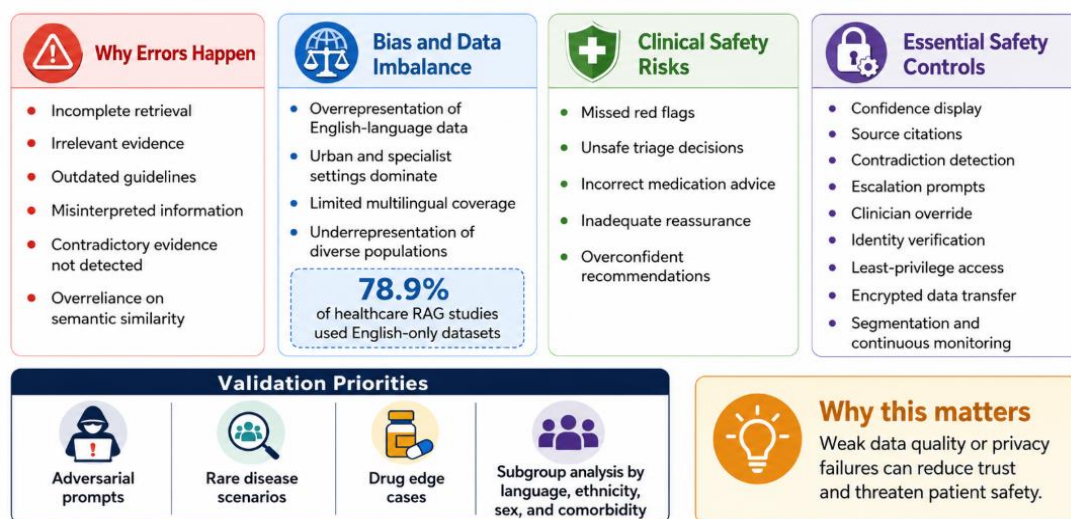


Figure 4: Major challenges in clinical AI retrieval, including accuracy, bias, data quality, safety, and security.

Figure 4 outlines the main challenges of AI retrieval systems in clinical applications, such as retrieval errors due to incomplete or outdated information, retrieval errors due to imbalanced information collections, and potential patient safety concerns. It also displays essential elements for safe implementation, including validation priorities, clinician override, secure access control, and source citation, to improve overall trust and data reliability.

Clinical safety risks encountered by patients include missing red flags, unsafe triage, failing to provide sufficient reassurance, providing incorrect medication guidance, and making overconfident recommendations [30]. These systems thus require a confidence display, a source citation function, a contradiction-detection function, an escalation prompt function, and a clinician override function. The relevance of Zero-trust Healthcare Architecture lies in the critical need for identity verification, least-privilege access, encrypted data transfer, segmentation, and ongoing monitoring of multi-hospital AI systems integrating EHRs, wearable data streams, and clinical analytics [31]. Without these controls in place, data loss, incomplete data quality, or a privacy incident can pose a serious challenge to building trusting relationships and to patient safety. When validating, adverse triggers should thus be included, as well as

extremely rare diseases; these edge cases for drugs also include subgroup analyses by language, ethnicity, sex, and comorbidity profile.

4.4 Implementation, Governance, and Real-World Adoption Barriers

Practical challenges include cost, practical acceptance, regulatory uncertainty, interface, maintenance overhead, liability, data controls, and cybersecurity for applications. Additionally, retrieval engines need to be integrated into EHRs, formularies, laboratories, guideline libraries, radiology documents, and institutional documents, while tracking data sources [32]. Additionally, the dynamic and often changing information in clinical guidelines, device alerts, drug labels, and local protocols calls for regular updates to generative AI retrieval systems.

The FDA's guidance on AI medical devices gives us some clues about the types of expectations that must be addressed before adopting these devices and how AI-based tools are gaining traction and beginning to enter the regulated world of Medical devices [33]. Several RAG system configurations, however, are not as straightforward as static RAG and require not just updates to the knowledge base but also embedded updates, rewritten prompts, and even model updates. This increases complexity expectations beyond validation.

Table 3: Barriers to implementing, governing, and adopting generative AI retrieval systems in real-world clinical settings.

Area	Key issues	Practical implication
Implementation barriers	Cost, interoperability, clinician acceptance, liability, cybersecurity, maintenance burden, and regulatory uncertainty	These factors may delay or limit real-world deployment in healthcare settings.
System integration needs	Connection to EHRs, laboratories, radiology reports, formularies, guideline libraries, and institutional protocols	Retrieval systems must access reliable clinical sources while maintaining data traceability.
Ongoing updates	Clinical guidelines, drug labels, device alerts, local protocols, embeddings, prompts, and models may change	Systems require regular updating, monitoring, and revalidation after deployment.
Governance team	Clinicians, informaticians, data scientists, legal experts, cybersecurity officers, and patient representatives	Multidisciplinary oversight is needed to support safety, accountability, and trust.
Adoption pathway	Offline benchmark testing, silent EHR testing, supervised clinical pilot, limited deployment, and post-market monitoring	Stepwise implementation reduces risk before wider clinical use.
Operational requirements	Low latency, reliable performance, local adaptation, staff training, incident review, and periodic revalidation	Systems must be fast, safe, visible, and practical at the point of care.

Table 3 summarizes the major barriers to generative AI retrieval systems in clinical practice across implementation, governance, and adoption. It includes a summary of operational needs, administrative responsibilities, update needs, implementation issues, and implementation phases. The table highlights that for an AI system to be deployed successfully, it is crucial to have interoperability, multidisciplinary oversight, cybersecurity, periodic revalidation, staff training, local adaptation, and continuous monitoring to ensure safe point-of-care use.

Several disciplines are necessary, including clinicians, informaticians, data scientists, legal experts, patient advocates or representatives, and cybersecurity officers. Adoption should encompass an offline benchmark assessment, an EHR assessment with the ability to test in silent mode, a clinical pilot trial with supervision and limited deployment, and an ongoing post-market assessment with monitoring [34]. Edge computing research also highlights the importance of low-latency

processing, distributed system design, and local, decision-oriented support, whose responses can adversely affect workflow outcomes [35]. In a clinical environment, this translates to fast, reliable, visible, and safe retrieval systems at the point of care. Moreover, after implementation in a hospital, training and planning should be conducted, the implementation revisited, sources identified, and revalidation cycles carried out periodically.

5. Future Research Recommendation

5.1 Prospective Multi-Site Clinical Validation

Future research should move beyond retrospective data and synthetic vignettes and assess generative AI retrieval systems in real-world clinical scenarios and across various benchmark datasets [36]. The next possible research step would be to conduct a prospective, controlled, multi-hospital trial involving 5-10 hospitals over the next 3-5 years, including tertiary, district, and

primary clinical environments. The study should involve a sample of 100-200 clinicians and have a test of 1000-3000 clinical queries from emergency medicine, internal medicine, pediatrics, obstetrics, oncology, pharmacy, and family medicine.

Outcomes should include diagnostic accuracy, time to make decisions, adherence to guidelines, clinician satisfaction, reduction in workload, patient safety events, and the percentage of recommendations that required correction. To achieve wider deployment, a minimum target of $\geq 90\%$ clinically rated relevance, $\geq 85\%$ guideline support, and a clinically rated evidence error rate $\leq 5\%$ for safety-critical evidence is suggested. Further validations should also include 'silent-mode' validations, in which machine learning (ML) predictions are made without influencing healthcare decisions and subsequently tested in a supervised pilot, wherein clinicians may accept, reject, or modify the ML suggestions [37].

5.2 Explainable and Auditable Retrieval Pipelines

Future studies should focus on retrieval systems that are transparent, inspectable, and auditable. The system should display the retrieved sources, the ranking of passages, the most recent date for each source, and the

impact of each retrieved passage on the final response. This would enable the distinction between evidence-based recommendations and unsupported synthesis, as any clinician could conduct it. Comparative research should be conducted using the same set of clinical questions and review panels, with naive RAG, advanced RAG, modular RAG, agentic retrieval, and knowledge graph-enhanced retrieval [38].

The metrics to include are retrieval precision@5, recall@10, citation accuracy, contradiction detection rate, faithfulness score, response latency, and audit completeness. Further explainability is of critical importance for medico-legal responsibility, as practitioners will need to be aware of the circumstances under which a failure occurs, whether due to retrieval failure, source obsolescence, prompt design for medical staff, model reasoning, or user interpretation [39]. Studies on Zero Trust security in multi-tenant clouds highlight that any strictly controlled digital environment should be continuously monitored for authentication, least-privilege access, segmentation, and encryption, all of which should also be enforced when a clinical user accesses the AI system [40].

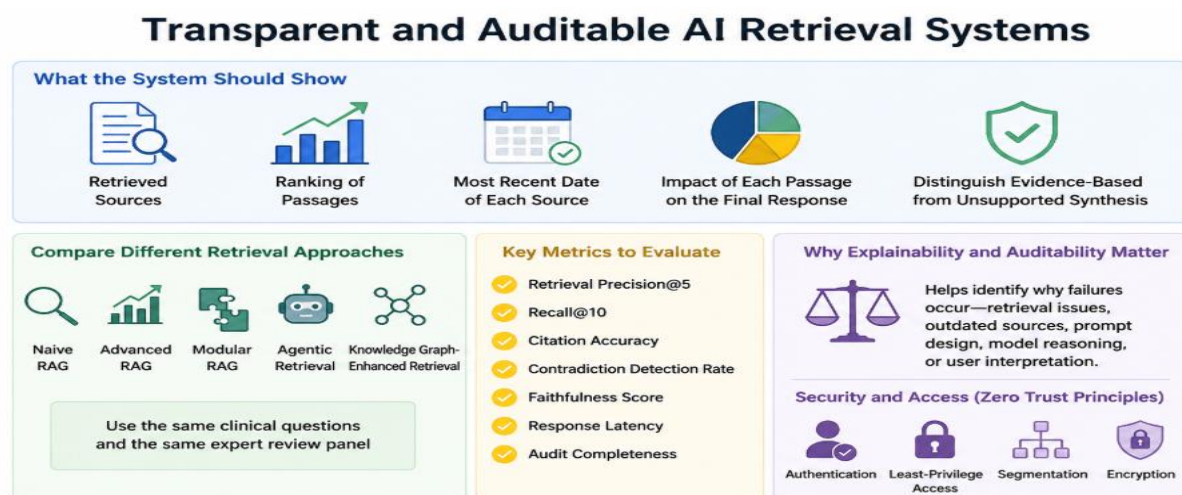


Figure 5: Transparent and auditable AI retrieval systems for trustworthy clinical decision support.

Figure 5 illustrates a basic path forward for developing simple, transparent, and auditable AI retrieval systems for clinical use. It provides information on what this system compares and contrasts, how to access this data, the key metrics to consider, and measures to ensure safe access to the system. The figure also highlights aspects of traceability, accountability, and explainability in clinical

decision support and real-world implementation scenarios.

5.3 Equity, Multilingual Deployment, and Local Clinical Adaptation

Future research needs to involve multilingual, rural, and low-resource clinical environments and should be quantitative, drawing on data from other languages and

countries. The systems should be customizable and flexible to incorporate disease trends, insurance policies, various referral processes, drugs, formularies, and the country's overall treatment guidelines [41]. For instance, a clinic in a rural control area that is unable to run the tests it needs, or where a medication goes out of stock at a tertiary care clinic, may not be an acceptable environment. Future studies should consider the effects of queries on translation, local terminology, and local patient instructions, as well as whether clinical accuracy will be maintained. This should also take into account cost-effectiveness, including software acquisition, cloud-based web hosting, human resource training, source validation, cybersecurity, and source revalidation.

Energy consumption, interference volume, and deployment architecture should be considered factors in defining environmental cost. In summary, the analysis of scalable multi-instance Jira deployments indicates that for global entities looking to deploy SIDIDI platforms across regions, thoughtful configuration, governance, monitoring, and localization will be key [42]. The same applies to clinical AI retrieval systems, which must be locally customizable yet globally integrated into a specific architectural approach. Further, measures of consent clarity and patients' willingness to have AI-supported evidence retrieval during consultations should be evaluated over time and across different clinical populations and care settings, and used as indicators of patients' willingness to use AI to discover evidence during a consultation.

6. Conclusions

Knowledge retrieval systems based on generative AI, powered by large language models and repositories of clinical data such as EHRs, drug databases, biomedical literature, guideline repositories, and institutional protocols, are important innovations in clinical decision support. The study demonstrates that it is no longer a problem gaining access to information in clinical decision-making, but rather obtaining, comprehending, and using information at the appropriate time. These systems can be more traceable and context-aware, delivering better support than conventional rule-based clinical decision support systems by incorporating retrieval-augmented generation, hybrid retrieval, reranking, evidence citation, vector search, and EHR-native interfaces.

The results reflect how generative AI tools, such as retrieval tools, can aid access to evidence and manual

literature reviews, summarize entire medical histories, and support adherence to guidelines and clinical decision-making, including diagnosis, treatment planning, note summarization, prescription, and discharge planning. In the clinical world, where sources are curated, regularly updated, and clinically relevant, and where workflow and retrieval experience are important factors, such systems could support evidence-based care when RAG is used in clinically curated, regularly updated, and workflow-integrated environments. Their value depends on integration into routine clinical operations through FHIR APIs, role-based access controls, audit logging, and EHR-native interfaces.

The research also highlights the challenges of using retrieval systems with generative AIs. Patient safety issues with hallucinations, incomplete retrieval, out-of-date data sources, biased databases, inaccurate or poor calibration, privacy breaches, cybersecurity threats, and overreliance upon clinicians are all problematic. These risks are particularly relevant in multilingual, rural contexts or in low-resource, high-acuity environments, where there is typically less clinical data, infrastructure, and specialist support. Hence, these systems should not be viewed as stand-alone clinical decision-makers. They should help them, not take responsibility for the licensed health professional who makes the interpretation and decision. The future of generative AI in clinical decision-making will shift from idolatry to one rooted in nature and validated by explainability, its inherent capacity to learn from massive datasets, and by worker-owned governance, respect for privacy, and responsible use.

Future directions for development include multi-site trials, clear retrieval methods, multilingual data, developer feedback, and post-deployment monitoring and surveillance. The study concludes that generative AI-enabled knowledge retrieval can be applied to clinical decision-making only if it is human-supervised, evidence-based, auditable, and locally adaptable; human-supervised, evidence-based systems can complement clinical decision-making and should not replace clinical judgment. In practice, the implementation of this process will likely be gradual, characterized by a sequence of silent trials, incremental rollouts, instrumentation and supervision of pilots, and gradually growing evaluations of clinical utility, technical robustness, patient safety, and ethics that will precede their institutional uptake and dissemination across the multiplicity of health care delivery systems.

References

1. N. C. Apathy, A. J. Holmgren, and D. A. Cross, "Physician EHR time and visit volume following adoption of team-based documentation support," *JAMA Internal Medicine*, vol. 184, no. 10, pp. 1212–1221, 2024. https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2822382/jamainternal_apathy_2024_oi_240052_1727970846.4993.pdf
2. A. Nooruddin, A. Lal, S. M. F. Ahmed, M. H. Ghori, N. Adnan, and F. Umer, "Use of generative artificial intelligence models in healthcare: A scoping review," *BMC Artificial Intelligence*, vol. 2, no. 1, p. 11, 2026. <https://link.springer.com/content/pdf/10.1186/s44398-026-00029-6.pdf>
3. R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: Benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, no. 1, p. 17, 2020. <https://www.nature.com/articles/s41746-020-0221-y.pdf>
4. S. A. Rabbani, M. El-Tanani, S. Sharma, S. S. Rabbani, Y. El-Tanani, R. Kumar, and M. Saini, "Generative artificial intelligence in healthcare: Applications, implementation challenges, and future directions," *BioMedInformatics*, vol. 5, no. 3, p. 37, 2025. <https://www.mdpi.com/2673-7426/5/3/37>
5. P. K. R. Prakashkumar, "Analytical Study of Vertex–Oracle Cloud Integration in Enterprise ERP Systems for Automated Tax Compliance and Efficient Financial Closings," *International Journal of Accounting and Management Sciences*, vol. 4, no. 4, 2025. <https://scipubhouse.com/international-journal-of-accounting-and-management-sciences-ijams/content/ijams-vol-4-no-4-october-2025/analytical-study-of-vertex-oracle-cloud-integration-in-enterprise-erp-systems-for-automated-tax-compliance-and-efficient-financial-closings/>
6. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. <https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>
7. L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, no. 6, p. e0000877, 2025. <https://journals.plos.org/digitalhealth/article/file?id=10.1371/journal.pdig.0000877&type=printable>
8. K. D. Pagana, T. J. Pagana, and T. N. P. Bullen, *Mosby's Manual of Diagnostic and Laboratory Tests*. Elsevier Health Sciences, 2021. https://books.google.com/books?hl=en&lr=&id=laQsEAAAQBAJ&oi=fnd&pg=PP1&dq=The+symptoms,+risk+factors,+lab+test+results,+and+comorbidities+are+also+processed,+and+differential+diagnoses+for+conditions+are+provided+in+line+with+the+guidelines+to+support+diagnosis.&ots=oV31_zhTYV&sig=9ChgkLZjyx3k0aSlGlak4y-g3Q
9. P. Fritz, A. Kleinhans, R. Raoufi, A. Sediqi, N. Schmid, S. Schricker, *et al.*, "Evaluation of medical decision support systems (DDX generators) using real medical cases of varying complexity and origin," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 254, 2022. <https://link.springer.com/content/pdf/10.1186/s12911-022-01988-2.pdf>
10. F. F. Alu and S. Oluwadare, "An auditable and source-verified framework for clinical AI decision support: Integrating retrieval-augmented generation with data provenance," *Frontiers in Artificial Intelligence*, vol. 9, p. 1737532, 2026. <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1737532/pdf>
11. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025. <https://dl.acm.org/doi/pdf/10.1145/3703155>
12. S. Pretorius, V. Bartle, S. Bellass, E. R. Blackburn, H. Bold, R. Cooper, *et al.*, "Making clinical guidelines work for people with multiple long term conditions: Analysis and recommendations from review of single condition guidelines," *BMJ Medicine*, vol. 5, no. 1, p.

- e001495, 2026.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC12933760/>
13. Sardana, K. S. Chadha, and S. Durgam, "Advancing Healthcare Technology Through Secure Data Systems," 2025.
<https://doi.org/10.12732/ijam.v38i10s.1138>
14. S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, *et al.*, "Testing and evaluation of health care applications of large language models: A systematic review," *JAMA*, vol. 333, no. 4, pp. 319–328, 2025.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11480901/>
15. S. K. Vishwakarma, "AI-driven predictive risk modelling for aerospace supply chains," *International Journal of Innovation in Business & Economics and Applied Journal*, 2025.
<https://www.iibajournal.org/index.php/iibeaj/article/view/64>
16. S. Rangu and K. S. Chadha, "Cloud-powered healthcare & insurance transformation with CRM and advanced analytics," *International Journal of Intelligent Systems and Applications in Engineering*, 2025.
<https://www.ijisae.org/index.php/IJISAE/article/view/7915>
17. G. Lichtner, B. S. Alper, C. Jurth, C. Spies, M. Boeker, J. J. Meerpohl, and F. von Dincklage, "Representation of evidence-based clinical practice guideline recommendations on FHIR," *Journal of Biomedical Informatics*, vol. 139, p. 104305, 2023.
<https://www.sciencedirect.com/science/article/pii/S1532046423000266>
18. S. K. RaoVanama, "AI-Augmented CI/CD Pipeline Optimization for Scalable Cloud-Native Deployment," *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, vol. 5, no. 4, pp. 175–187, 2024.
<https://ijaidsmml.org/index.php/ijaidsmml/article/download/368/338>
19. N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine*, vol. 158, p. 106848, 2023.
<https://www.sciencedirect.com/science/article/pii/S00104852300313X>
20. B. J. Wells, H. M. Nguyen, A. McWilliams, M. Pallini, A. Bovi, A. Kuzma, *et al.*, "A practical framework for appropriate implementation and review of artificial intelligence (FAIR-AI) in healthcare," *npj Digital Medicine*, vol. 8, no. 1, p. 514, 2025.
<https://www.nature.com/articles/s41746-025-01900-y.pdf>
21. A. Ubale, "Beyond Telematics: Leveraging Generative AI for Synthetic Accident Reconstruction and Liability Attribution in Autonomous Vehicle Claims," *International Journal of AI, BigData, Computational and Management Studies*, vol. 4, no. 4, pp. 119–124, 2023.
<https://ijaibdcms.org/index.php/ijaibdcms/article/download/356/352>
22. Y. Lu, X. Zhao, and J. Wang, "ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval," in *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Aug. 2024, pp. 64–68.
<https://aclanthology.org/2024.knowllm-1.6.pdf>
23. G. Mallardi, F. Calefato, L. Quaranta, and F. Lanubile, "Integrating AI into healthcare systems: A multivocal literature review," *Smart Health*, vol. 39, p. 100631, 2026.
<https://www.sciencedirect.com/science/article/pii/S2352648325000923>
24. K. S. Chadha, P. R. Vennamaneni, and J. Sardana, "Leveraging Advanced Analytics and AI-Powered Solutions for Transforming Healthcare and Retail E-Commerce: Enhancing Data Security, Personalization, and Compliance," *International Journal of Applied Mathematics*, vol. 38, no. 9s, pp. 1310–1334, 2025.
<https://doi.org/10.12732/ijam.v38i9s.862>
25. D. Van Veen, C. Van Uden, L. Blankemeier, J. B. Delbrouck, A. Aali, C. Bluethgen, *et al.*, "Adapted large language models can outperform medical experts in clinical text summarization," *Nature Medicine*, vol. 30, no. 4, pp. 1134–1142, 2024.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11479659/>
26. A. J. Holmgren, C. A. Sinsky, L. Rotenstein, and N. C. Apathy, "National comparison of ambulatory physician electronic health record use across specialties," *Journal of General Internal Medicine*, vol. 39, no. 14, pp. 2868–2870, 2024.

<https://link.springer.com/content/pdf/10.1007/s11606-024-08930-4.pdf>

<https://www.nature.com/articles/s41746-025-02052-9.pdf>

27. S. Rangu, "Analyzing the Impact of AI-Powered Call Center Automation on Operational Efficiency in Healthcare," *JISEM Journal*, 2025. <https://www.jisem-journal.com/index.php/journal/article/view/8901>
28. J. Sohn, Y. Park, C. Yoon, S. Park, H. Hwang, M. Sung, *et al.*, "Rationale-guided retrieval augmented generation for medical question answering," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Apr. 2025, pp. 12739–12753. <https://aclanthology.org/2025.naacl-long.635.pdf>
29. C. Wu, P. Qiu, J. Liu, H. Gu, N. Li, Y. Zhang, *et al.*, "Towards evaluating and building versatile large language models for medicine," *npj Digital Medicine*, vol. 8, no. 1, p. 58, 2025. <https://www.nature.com/articles/s41746-024-01390-4>
30. A. Ramaswamy, A. Tyagi, H. Hugo, J. Jiang, P. Jayaraman, M. Jangda, *et al.*, "Chatgpt Health performance in a structured test of triage recommendations," *Nature Medicine*, vol. 32, no. 5, pp. 1671–1675, 2026. https://www.nature.com/articles/s41591-026-04297-7_reference.pdf
31. K. S. Chadha, "Zero-Trust Data Architecture for Multi-Hospital Research: HIPAA-Compliant Unification of EHRs, Wearable Streams, and Clinical Trial Analytics," *IJCESEN*, 2025. <https://ijcesen.com/index.php/ijcesen/article/view/3477>
32. S. Liu, A. B. McCoy, and A. Wright, "Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines," *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605–615, 2025. <https://academic.oup.com/jamia/article-pdf/32/4/605/61442713/ocaf008.pdf>
33. V. Mehta, A. Komanduri, R. S. Bhadouriya, V. Mehta, M. D. Johnson, P. Shrestha, *et al.*, "Evaluating transparency in AI/ML model characteristics for FDA-reviewed medical devices," *npj Digital Medicine*, vol. 8, no. 1, p. 673, 2025.
34. J. G. You, T. Hernandez-Boussard, M. A. Pfeffer, A. Landman, and R. G. Mishuris, "Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications," *npj Digital Medicine*, vol. 8, no. 1, p. 107, 2025. <https://www.nature.com/articles/s41746-025-01506-4.pdf>
35. N. Kumar, "Edge Computing Patterns for Real-Time Order Flow Optimization," *International Journal of Applied Mathematics*, 2025. <https://doi.org/10.12732/ijam.v38i12s.1494>
36. T. D. Azad, H. M. Krumholz, and S. Saria, "Principles to guide clinical AI readiness and move from benchmarks to real-world evaluation," *Nature Medicine*, pp. 1–3, 2026. <https://www.nature.com/articles/s41591-025-04198-1>
37. J. C. Kwong, L. Erdman, A. Khondker, M. Skreta, A. Goldenberg, M. D. McCradden, *et al.*, "The silent trial: The bridge between bench-to-bedside clinical AI applications," *Frontiers in Digital Health*, vol. 4, p. 929508, 2022. <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgh.2022.929508/pdf>
38. G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of the Association for Computational Linguistics: ACL 2024*, Aug. 2024, pp. 6233–6251. <https://aclanthology.org/2024.findings-acl.372.pdf>
39. H. Smith, "Clinical AI: Opacity, accountability, responsibility and liability," *AI & Society*, vol. 36, no. 2, pp. 535–545, 2021. <https://link.springer.com/content/pdf/10.1007/s00146-020-01019-6.pdf>
40. R. Hariharan, "Zero Trust Security in Multi-Tenant Cloud Environments," *JISEM Journal*, 2025. <https://www.jisem-journal.com/index.php/journal/article/view/8899>
41. K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, *et al.*, "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare,"

BMJ, vol. 388, p. e081554, 2025.
<https://www.bmj.com/content/bmj/388/bmj-2024-081554.full.pdf>

Enterprises,” *International Journal of Engineering Applications*, 2025.
<https://gprjournals.org/journals/index.php/ijea/article/view/353>

42. S. Samala, “Architecting Multi-Instance Jira Deployments: Scalability Challenges for Global