Multi-Agent Retrieval Augmented Generation for Clinical Decision Support: A Systematic Review and Integrative Conceptual Framework

Tarisai Mugambiwa [1]*, Belinda Ndlovu [2]**
*, **Informatics Department, National University of Science and Technology, Bulawayo, Zimbabwe
n02220886e@students.nust.ac.zw [1], belinda.ndlovu@nust.ac.zw [2]

Article Info

Article history:

Received 2025-11-30
Revised 2026-01-06
Accepted 2026-01-17

Keywords:

Retrieval Augmented Generation (RAG),
Multi agent systems,
Agentic AI Clinical decision support (CDSS),
Transparency,
Safety.

ABSTRACT

Multi agent retrieval augmented generation (RAG) systems are increasingly explored as advanced architectures for clinical decision support combining information retrieval, reasoning and verification through coordinated agent interactions. This study systematically reviews applications of agentic and multi agent RAG in clinical decision support systems (CDSS) and synthesizes an integrative conceptual framework linking technical design to technology adoption considerations. Following PRISMA guidelines, searches were conducted from PubMed, IEEE Xplore and ScienceDirect using structured Boolean strings combining terms for multi agent architectures, RAG and CDSS.The search yielded 12 studies published between 2020 and 2025 that met the inclusion criteria. The review synthesises evidence on multi agent role configurations retrieval and reasoning strategies, verification mechanisms and reported clinical contexts. Across studies, dominant challenges include data and corpus limitations retrieval quality dependency, limited clinical validation and computational overhead, alongside governance concerns such as privacy, bias and accountability. Building on the synthesis, we propose a four-agent CDSS framework retriever, reasoner, verifier, safety and map its deployment determinants to Technology Acceptance Model constructs perceived usefulness, perceived ease of use, trust and diffusion of Innovations attributes. The review concludes with design-oriented recommendations for safer, explainable, and adoption-ready multi-agent RAG CDSS, particularly for low-resource contexts.

I. Introduction

Artificial Intelligence (AI) is transforming the healthcare industry with its potential to enhance diagnostic precision, treatment planning and operational efficiency [1]. They help nurses and doctors with patient care, clinical guidelines and medical literature for diagnosis and treatment. Clinical decision support systems are limited by static rule based, and inflexible in responding to changing clinical evidence [2]. These limitations are especially evident in developing countries, where digital infrastructure, data quality, and AI literacy are often insufficient to support advanced decision-support tools [3][4].

LLMs can synthesise knowledge across vast medical corpora but are prone to hallucination, outdated pre-training and poor traceability of sources [5]. To address these problems, Retrieval-Augmented Generation (RAG) was introduced as a framework combining generative reasoning explicitly with non-parametric retrieval from verified external sources[6].In the field of medicine, retrieval augmented generation (RAG) has already been successful in the interpretation of guidelines, diagnosis of rare diseases and patient education, consistently decreasing hallucinations and improving knowledge relevance [7][8].

Despite retrieval, augmented generation (RAG) based models making progress towards accuracy and transparency in clinical reasoning, they are still limited in scope to single-agent architectures, unable to deal with more explanations that are complicated. Recently, we have proved that single-agent systems lack significant context understanding and do not include proper safety and ethical checks[9][10].To overcome these limitations, research has shifted toward multi-agent architectures, where multiple specialised agents collaborate through structured communication to enhance reasoning and reliability. These include planner, retriever and verifier agent patterns for diagnosis and treatment recommendations [11], multi-agent conversational frameworks for transparent reasoning and multi-agent retrieval augmented generation (RAG) systems for medical question answering [12].

In this review, a single agent RAG system denotes a single pipeline in which one primary language model performs retrieval and generate the final output. Agentic/multi agent RAG system uses two or more specialized agents that coordinate through explicit message passing, shared memory or tool call. Agents may operate sequentially to decompose the clinical task, retrieve evidence, verify claims and enforce safety constraints. This operational definition is applied consistently throughout the review to avoid conceptual ambiguity and to enable meaningful comparison across architectures. As a result, this systematic literature review will help fill this gap by synthesising and reviewing the current research in multi-agent decision-support systems with RAG. The study is guided by the following research questions:

RQ1.How have multi-agent RAG systems been applied in healthcare decision support, and how do they compare with traditional systems?

RQ2.In what ways do multi-agent RAG systems enhance explainability, transparency, and clinician trust?

RQ3. What measurable improvements (accuracy, recall, latency) do multi-agent RAG systems achieve over single-agent or classical approaches?

RQ4.What human infrastructural and ethical factors influence their adoption in low-resource contexts.

II. Methods

This systematic review followed PRISMA reporting guidance and applied a pre-specified review protocol defining the research questions, eligibility criteria, screening steps, and synthesis approach[14].

A. Search Strategy

A comprehensive search was conducted in three academic databases ScienceDirect, IEEE Xplore and PubMed. The search strategy combined key terms related to our research focus ("multi-agent system" OR "multi-agent architecture") AND ("retrieval augmented generation" OR RAG) AND ("decision support system" OR "clinical decision support"). The search covered January 2020 to September 2025, restricted to the English language. Studies were excluded if they were purely conceptual, non-healthcare focused, single agent systems or lacked decision support relevance. Title and abstracts were screened first, followed by full text eligibility assessment. Figure 1 summarises the PRISMA flow and selection decisions.

B. Inclusion and Exclusion Criteria

Table I
Inclusion And Exclusion Criteria

Criteria	Inclusion	Exclusion
Publication status	Published, peer-reviewed, journal articles ,	Unpublished work
Type of paper	Peer-reviewed journal articles, conference proceedings,	Editorials, opinion pieces, letters, books
Research Area	The primary focus is on healthcare, clinical decision	Studies focused on other domains, even if they use multi-agent RAG architectures.
Subject Areas	Computer science ,Nursing and health professionals	Finance, supply chain ,general education
Core Technology	RAG, multi agent architecture	discuss on of the core components multi agent system

C. Critical Appraisal

Each study was critically assessed using the Critical Appraisal Skills Programme (CASP) checklist to evaluate methodological validity, transparency, and potential sources of bias. Two reviewers independently appraised and scored each paper, achieving a Cohen's κ of 0.83, which indicates substantial inter-rate agreement. Studies that scored 8 out of 10 or higher on the CASP criteria were categorised as *high quality*, while those scoring 6–7 out of 10 were classified as *moderate quality*. Data extraction captured clinical domain, target users, knowledge sources, retrieval method, agent roles, evaluation design, metrics and reported deployment considerations. We performed a structured comparative analysis across architectures, clinical context and applied a thematic coding step to derive across cutting patterns. The proposed four agent framework was derived by mapping recurring functional roles observed in the included studies to a minimal, implementable CDSS architecture and then aligning external adoption factors with TAM and Diffusion of Innovations constructs.

D. Identification

The database search retrieved 191 research papers, 145 from ScienceDirect, 50 from IEE Xplore and 6 from PubMed. The research papers were exported in RIS format and imported into Mendeley Reference Manager to remove duplicates. There were zero duplicates.

E. Screening

191 research papers were imported into Mendeley Reference Manager for screening. Screening was then carried out at the title and abstract level based on the predefined inclusion and exclusion criteria derived from the research questions. Each study was evaluated for its relevance to multi-agent medical decision-support systems using Retrieval-Augmented Generation (RAG). Papers were tagged as Include_TA or Exclude_TA using Mendeley's tagging function. Studies tagged as *Include_TA* described RAG or retrieval-grounded large language models incorporating multi-agent architectures, applied within a healthcare or decision-support context, and published between 2020 and 2025. Papers were tagged as *Exclude_TA* if they lacked a retrieval component, did not involve multi-agent or agentic frameworks. The screening process excluded 156 papers, leaving 35 studies for full-text eligibility assessment.

F.Eligibility

Full text eligibility assessment was conducted for the 35 studies from the screening stage. Each paper was read carefully to see if it explained the method clearly and used reliable techniques. The paper was eligible if it used or explained a RAG system and applied it in a medical or clinical decision support setting. The system included multiple agents. The study evaluated or discussed the trust and safety of the system. Common reasons for excluding papers were no real RAG or multi-agent system, no multi-agent or decision support focus. After a full text, 12 studies met all eligibility criteria and were included, while 23 studies were excluded because they did not meet the eligibility criteria.

G.Included

After applying the screening and eligibility criteria, 12 studies were included in this systematic literature review. The included studies address the focus areas of the research questions on the role of multi-agent medical decision support systems using RAG.The limited number of included studies reflects the nascent state of empirical research on multi-agent RAG systems in healthcare rather than deficiencies in the search strategy, underscoring the early-stage maturity of the field.

III. Results And Discussion

The delimitation is shown in the following PRISMA flowchart by [15] in Figure 1:

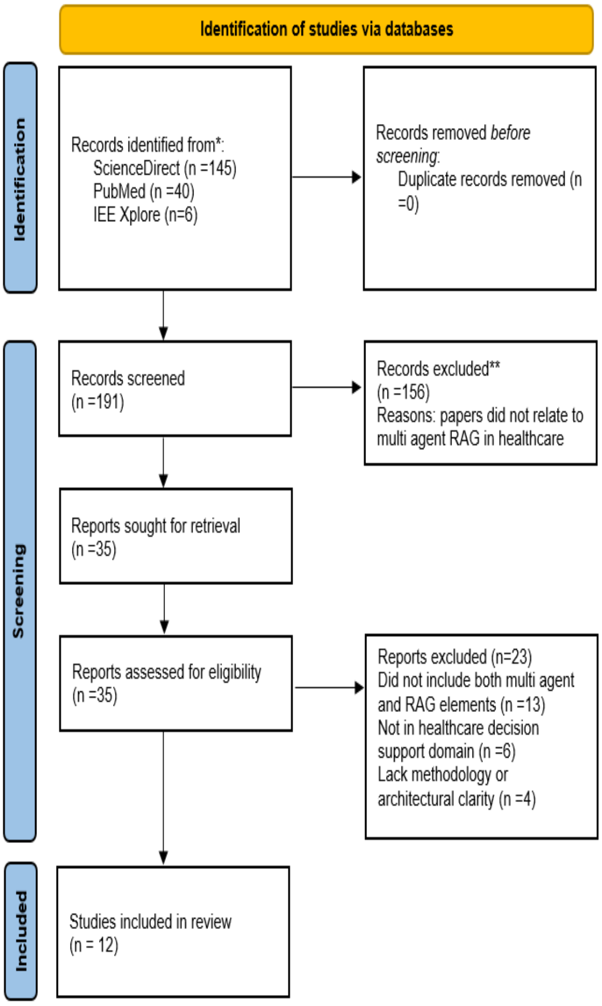
Identification of studies via databases
Identification
Screening
Included
Records identified from*:
ScienceDirect (n =145)
PubMed (n =40)
IEE Xplore (n=6)
Records removed before screening:
Duplicate records removed (n =0)
Records screened (n =191)
Records excluded** (n =156)
Reasons: papers did not relate to multi agent RAG in healthcare
Reports sought for retrieval (n =35)
Reports assessed for eligibility (n =35)
Reports excluded (n=23)
Did not include both multi agent and RAG elements (n =13)
Not in healthcare decision support domain (n =6)
Lack methodology or architectural clarity (n =4)
Studies included in review (n = 12)

Figure 1: PRISMA results

TABLE II
PAPERS THAT MET THE INCLUSION CRITERIA

AUTHOR/ YEAR	ORIGIN	DATASET	AIM OF STUDY	ALGORITHMS	OPPORTUNITIES	LIMITATIONS/ CHALLENGS
[16]	USA	Rare disease cases PubMed text	To Evaluate RAG enhanced ChatGPT (RareDxGPT) for rare disease diagnosis	• FAISS RAG retriever • GPT-3.5 • Prompt variants	• Potential to support rare-disease clinicians • enhance early diagnosis • reduce diagnostic odyssey in underserved regions	• Small dataset(30 disease) • Limited generalisability • GPT-3.5 constraints

[17]	Germany	Orthopaedic clinical guidelines	Develop RAG chatbot for orthopaedic guidance	• LGPT-4o • Qdrant • RAGAS evaluation	• Scalable patient education tool • can reduce clinician workload • adaptable to other specialties	• Corpus limited to German guidelines • Performance variations by topic
[18]	USA	Patient portal text emergency	Detect emergencies in patient messages using KG-RAG	• Multi-level KG-RAG with local and global search •	• Deployable triage automation • can reduce delays in emergency care • scalable to call center triage	• KG maintenance overhead • missing drug–drug interactions caused misclassification
[19]	Netherlands/Switzerland	HER data elements Controlled medical vocabularies	Map clinical data elements to controlled vocabularies	• Multi-agent RAG (query decomposition, ensemble retrievers, knowledge reservoir)	• Enabling interoperability; • foundational for EHR harmonisation & AI analytics pipelines	• Complexity of composite CDE linking • large vocabulary management
[20]	USA/Germany	Diagnostic dataset Biomedical ontology and literature	Broad diagnostic support for rare & long-tail diseases	• Retrieval based(PubMed and UMLS)RAG	• Powerful diagnosis support where labelled data is unavailable • scalable to global disease sets	• Retrieval quality critical • no real clinician evaluation
[21]	Morocco/ France	Multilingual clinical transcripts Speech to text	Improve transcription + compliance retrieval	• Whisper ASR • sentence RAG	• Supports multilingual clinical settings	• Multilingual ASR challenges • KG gaps
[22]	USA	Case vignettes Knowledge repositories	Evaluate multi-agent RAG diagnostic reasoning	• Planner,retriever,verifier agent pipeline	• Framework adaptable to specialty-specific AI assistants	• Limited clinical trial validation
[12]	Global	Multi task benchmark Clinical guidelines	Benchmark multi-agent RAG vs standard LLMs	• Multi-agent retriever, verifier, reasoning modules	• Basis for medical-grade reasoning agents • supports guideline aligned outputs	• Domain-specific corpora needed • possible latency trade-offs
[23]	USA	Radiology multimodal dataset	Combine imaging + text retrieval to enhance diagnosis	• Multimodal RAG with structured retrievers	• Potential for radiology copilots • interpretable image-text integration	• Large annotated imaging datasets are required
[7]	India	QA dataset Clinical guideline documents	Benchmark RAG vs standard LLM in medical QA	• Dense retrieval • RAG pipeline	• Can support clinical education, • guideline retrieval, rapid Q&A	• Limited scope (QA only)
[24]	China	Multimodal imaging dataset	Combine imaging + text reasoning	• 6-agent multimodal RAG	• Radiology co-pilots • interpretable multimodal reasoning	• Requires large annotated imaging datasets

A. Geographical Distribution of Included Studies

Figure 2 shows the geographical distribution of 12 studies on continents from 2020 to 2025

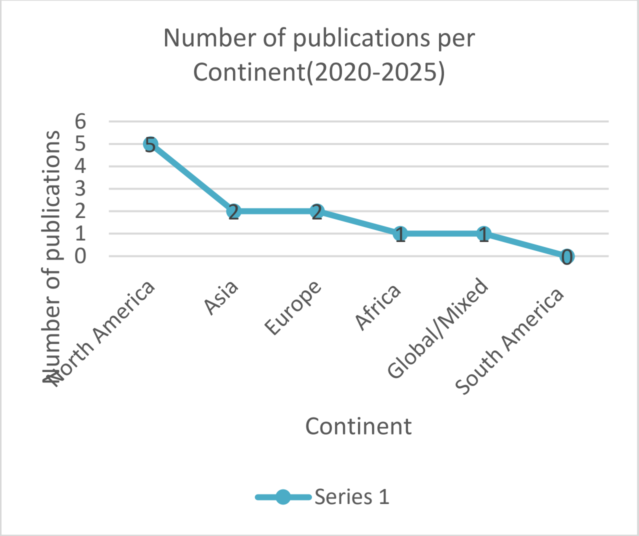

Figure 2: Distribution of Continents

Figure 2 shows the geographic distribution of 12 of the included studies. The number of publications was highest in North America, with five papers, indicating that the region has a high level of research on retrieval-augmented generation (RAG) and multi-agent clinical AI systems. Asia and Europe yielded two papers and their contributions were of a moderate nature, largely concentrated in clinical question answering and multimodal oncology models (Asia), orthopaedic patient education or data standardisation (Europe). Africa provided one publication through a Morocco/France partnership based on multilingual breast cancer RCPs.

B. Algorithm Category Used

Table III presents the algorithmic techniques applied across the included studies.

TABLE III
ALGORITHM CATEGORIES AND TYPES

Study	Vector RAG	KG-RAG	IR-based RAG	Multi Agent	Multimodal RAG
[16]	✓				
[17]	✓				
[18]	✓	✓			
[19]	✓			✓	
[20]			✓		
[21]	✓	✓			
[22]	✓			✓	
[12]	✓			✓	
[23]	✓				✓
[7]	✓		✓		
[24]	✓			✓	✓
[8]	✓				✓

The algorithms categorised in Table III are further visualised in Figure 3.

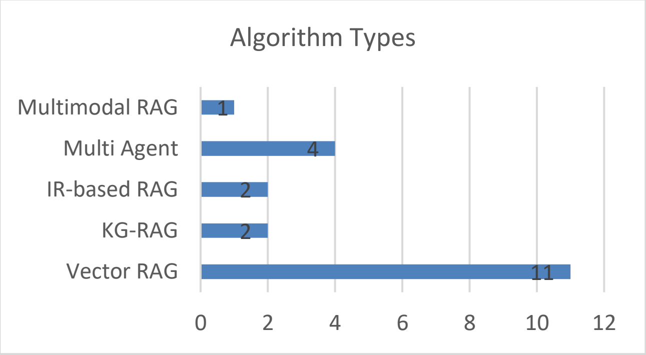

Figure 3: Algorithms

Figure 3 shows a distribution of the algorithm categories and types used in 12 studies. Vector-based RAG techniques are also the most prevalent and appear in 11 studies, emphasizing their relevance in the system design of clinical RAG at present. These are FAISS-based retrieval, Qdrant dense retrieval, PubMed/UMLS retrieval, Whisper + sentence-level retrieval, and other embedding-based retrieval pipelines. The prevalence represents the fact that most medical RAG systems heavily rely on dense vector search as the basis for model outputs. The next most common architecture was a multi-agent, implemented in four studies. Several well-coordinated agents, e.g., planners, retrievers, verifiers, or multimodal reasoning parts, supported these systems often. IR-based RAG and KG-RAG methods were not as frequent, existing only in two studies. IR-based RAG employed traditional sparse retrieval (e.g., PubMed, UMLS, keyword-based pipelines), and the KG-RAG framework leveraged domain-specific knowledge graphs to enhance semantic grounding, primarily in triage cases or multilingual clinical documentation contexts. A single study identified multimodal RAG incorporating imaging and text data, emphasizing that multimodal clinical RAG is in an early stage despite its massive potential for use in radiology and oncology

C. Challenges

Table 4 summarises the major categories of challenges identified across the included studies.

TABLE IV
CATEGORISING CHALLENGES

Challenges Category	Grouped Challenges
Data limitations and Generalisability	• Small dataset (30 diseases) • Limited generalisability • Corpus limited to German guidelines • Domain-specific corpora needed • Limited scope (QA only) • Domain narrower; limited real-world testing • Large annotated imaging datasets required
Knowledge source and KG constraints	• KG maintenance overhead • Missing drug–drug interactions caused misclassification • Dependent on knowledge graph completeness • KG must be continuously updated; domain gaps exist • Large vocabulary management • Complexity of composite CDE linking
Retrieval quality and Model dependency	• Retrieval quality critical • GPT-3.5 constraints • Domain-specific corpora needed • Performance variation by topic
Lack of clinical validation /real world evidence	• No real clinician evaluation • Limited clinical trial validation • No real-time clinical testing
Computational and scalability constraints	• Computationally demanding • Possible latency trade-offs
Safety, Risk and misclassification issues	• Missing drug–drug interactions caused misclassification • Dependent on incompleteness of corpora / KG gaps

Figure 4 illustrates the frequency of challenges categories reported

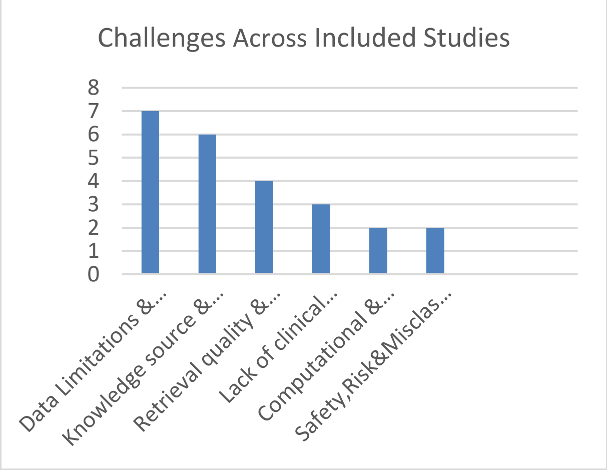

Figure 4: Challenges across Studies

The distribution of challenges identified for all 12 included studies is presented in Figure 4 and Table IV. Limitations in data were the most commonly reported issues, which were mentioned in seven studies, the analysis finds. Such constraints occurred through small datasets, language-restricted corpora, and the need for large annotated imaging or domain-specific resources, limiting generalisability. Knowledge-source constraints constituted the second largest category 6 studies, indicating chronic challenges in creating and maintaining knowledge graphs, incomplete vocabulary coverage, and the challenge of linking composite factors contained in various clinical data. However, retrieval and model-dependency problems were found in 4 studies in which system performance was significantly correlated with retriever quality, underlying LLM constraints, and even topic sensitivity. Another trio of studies found no clinical or real-world validation, suggesting that most systems are not tested in an environment other than research. Also, computational and scalability limitations were observed, especially in cases where multi-agent pipelines or large retrievers were required, which also was a factor for high computational overhead to use. Two studies reported risks of safety and misclassification, largely due to incomplete corpora, absent drug interactions and missing knowledge representations.

D.Opportunities

Table V outlines the main opportunity areas emerging from the reviewed studies.

TABLE IV
OPPORTUNITIES IN THEMES

Theme	Opportunities
Diagnostic Enhancement and Rare disease Clinician’s	• Potential to support rare-disease clinicians • Enhance early diagnosis • Reduce diagnostic odyssey in underserved regions • Powerful diagnosis support where labelled data is unavailable • Scalable to global disease sets
Patient Education and Communication Support	• Scalable patient education tool • Can reduce clinician workload • Adaptable to other specialities • Supports multilingual clinical settings
Triage Automation and Workflow Efficiency	• Deployable triage automation • Can reduce delays in emergency care • Scalable to call-centre triage
Interoperability and Health System Integration	• Enabling interoperability • Foundational for EHR harmonisation and AI analytics pipelines
Multimodal and Domain Specific Decision Support	• Potential for radiology copilots • Interpretable image-text integration • Interpretable multimodal reasoning • Can support clinical education • Supports guideline retrieval • Enables rapid Q&A

Table IV summarises the 23 opportunities that emerged from the included studies, which are grouped into six overall themes. Diagnostic enhancement emerged as a key opportunity, particularly when it comes to supporting rare-disease clinicians and enabling early detection, and improving diagnostic performance, especially in settings with low labels or in underserved areas of the population. There were some identified opportunities for patient education and communication support in multiple studies, in terms of scalable RAG-driven educational tools and multilingual support. Another major theme was triage automation, with systems showing promise in minimizing emergency care wait times and helping run call center. The interoperability oriented opportunities focused on the contribution of the RAG system to facilitate EHR data harmonization and enrich analytics pipelines. Agentic clinical reasoning and specialty-specific agents were seen as ushering in a new era of guided decision support.

E. Datasets

Table VI shows the datasets used across the included studies into four main groups.

TABLE V1
DATABASES USED IN STUDIES

Assigned Category	Dataset	Frequency
Clinical text dataset	-rare disease cases -Orthopaedic clinical guidelines -Patient portal text emergency -Diagnostic dataset -Multilingual clinical transcripts -Multi task benchmark -Clinical guidelines -QA dataset -Clinical guideline documents	11
Structured/HER & knowledge resources	-EHR data elements -Controlled medical vocabularies -Biomedical ontology and literature -Knowledge repositories	4
Multimodal imaging datasets	-Radiology multimodal dataset -Multimodal imaging dataset	2
Speech/ASR datasets	-Speech to text	1

Table VI summarise the distribution of dataset types used across the included studies. Most datasets (61.1%) 11 studies were clinical text resources, including rare-disease cases, PubMed articles, clinical guidelines, QA corpora, and case vignettes. A further (22.2%) 4 studies of datasets consisted of structured EHR elements and knowledge resources such as controlled medical vocabularies, ontologies, and knowledge repositories. Multimodal imaging datasets, mainly radiology image–report pairs, accounted for (11.1%) 2 studies of the resources, while speech data (oncology meeting recordings with speech-to-text output) represented (5.6%) 1 studies

Across the reviewed studies, multi-agent RAG systems demonstrated consistent advantages over single-agent and classical approaches in tasks requiring complex reasoning, evidence verification, and transparency. While single-agent RAG improved factual grounding, multi-agent architectures further reduced hallucinations, enhanced traceability, and enabled modular safety checks. However, these benefits were accompanied by increased computational overhead and infrastructure demands, indicating a trade-off between reasoning robustness and deployability. This comparative pattern highlights that multi-agent RAG systems are most suitable for high-risk, knowledge-intensive clinical decision support rather than lightweight or real-time applications.

B. Discussion

The following section discusses the findings from the sections A–E in relation to the four research questions guiding this review.

RQ1.How have multi-agent RAG systems been applied in healthcare decision support, and how do they compare with traditional systems?

The Netherlands/Switzerland applied a multi-agent RAG framework to represent clinical data elements in controlled vocabularies. Various agents executed query decomposition, ensemble retrieval, and knowledge consolidation, which allowed for the accurate mapping of atomic and composite CDEs [19]. This makes multi-agent RAG a decision support infrastructure tool, which promotes interoperable EHRs and analytics[25]. The rheumatology applied a planner, retriever, verifier pipeline in which one agent plans the diagnostic task, another retrieves targeted evidence, and a speciality specific approach promotes sophisticated stepwise clinical reasoning over straightforward question answering, and directly supports how clinicians assess competing hypotheses[26]. The global multi-agent model [12] was a benchmark of a multi-agent retriever , verifier , reasoning module compared to standard LLMs.The retrieval, reasoning, and verification integrated together within coordinated agents, the system would be able to manage heterogeneous tasks such as guideline interpretation, summarisation, and differential diagnosis, all within a single framework[27]. The Chinese HCC imaging [24]applied a six-agent multimodal RAG design, where agents specialised in handling imaging features, textual reports, and reasoning over combined evidence. Here, multi-agent RAG bridged radiology and clinical context, supporting decisions such as microvascular invasion assessment[23][24].

1) Diagnostic Enhancement and Rare disease support

Multi-agent and RAG-supported systems present substantial benefits in the context of diagnostic decision support for rare and long-tail diseases where classical algorithms are not applicable. In [16] retrieval augmentation enhanced the diagnostic accuracy from 37% to 40% by grounding model predictions in verified biomedical literature, which is not provided by classical rare-disease tools and standalone LLMs. [20] CliniqIR showed that retrieval-based decision support outpaced Clinical BERT for rare conditions, mitigating long-tail distribution obstacles that traditional machine learning models cannot accurately address. Multi-agent frameworks also advanced diagnostic reasoning [22] [24] by sharing tasks among planner, retriever, verifier, and multimodal agents that collectively produced more consistent, interpretable diagnostic explanations than traditional CNNs. In general, multi-agent RAG systems provide better factual support for diagnosis compared to their classical counterparts, the accuracy on rare cases is high, and the reasoning process is transparent[13][28].

2) Patient Educations and Communication Support

Compared to traditional patient information tools, RAG-enhanced patient education systems achieved significantly greater clarity, accuracy and trust. The orthopaedic RAG chatbot [7] obtained high accuracy (4.55/5), clarity (4.77/5), trust (4.23/5) scores, and significantly outperformed traditional template patient-information leaflets and rule-based FAQ systems. While standard educational tools provide generalised or static data, the RAG approach dynamically retrieves guideline-aligned content for the generation of customised, evidence-backed explanations. Unlike conventional chatbots that follow fixed scripts, RAG systems are transparent in their source citations. Multi-agent frameworks were not explicitly employed in [7], however the well-demonstrated strengths faithfulness, grounding and clarity lay the groundwork for future agentic extensions which support additional interpretability[14]. Overall, RAG-supported patient education tools have been shown to provide better communication quality and usability compared to traditional static or rule-based educational systems[29][30][31].

3) Triage Automation and Workflow Efficiency

In emergency triage and workflow automation, the RAG and KG-RAG systems exhibited enhanced performance relative to regular keyword based or machine-learning triage models[32].[18] Demonstrated excellent performance accuracy 0.99, sensitivity 0.98 and specificity 0.99 which significantly exceed clinician-coded rules and classifiers found in patient-portal triage[33][34]. Combining hierarchical KG retrieval, the system identified emergencies with better accuracy than conventional heuristics, which missed more subtle clinical cues. While not a complete multi-agent pipeline, modular KG-RAG design imitated multi-agent behaviour decomposing tasks and checking the signals through graph structures[35]. In summary, triage automation is very much part of a space in which RAG-enabled systems provide impressive improvement in performance and robust opportunities for multi-agent expansion for the future[13].

4) Interoperability and Health Integration

In both interoperability and controlled terminological mapping, multi-agent RAG systems significantly outperformed traditional rule [13][36]. [19] Adopted multi-agent architecture combined query decomposition, ensemble retrieval, and verification to achieve a 7.2% improvement to accuracy in their mapping over traditional terminological mappers. Classical platforms have been reported to fail on composite clinical information elements, multi-agent RAG combines distributed reasoning and cross-referencing for resolving complex mappings. In addition, in [21] multilingual and compliance-related retrieval showed that RAG is able to manage complex documentation environment that non classic approach of using the NLP system due to language barriers. In combination, these multi-agent and retrieval-augmented designs offer an alternative flexible, context-aware alternative to the brittle unidirectional classical analogous systems that fail to accommodate heterogeneous sources of clinical knowledge[9][10][37].

5) Agentic Clinical reasoning and Specialty Specific

Agentic RAG systems showed distinct advantages in clinical reasoning tasks requiring structured, explainable, guideline-aligned outputs. In [22], a planner, retriever and verifier pipeline improved reasoning stability by mirroring clinical thought processes. Defining the problem gathering evidence, and verifying against guidelines. Similarly, [12] (MAIN-RAG) used coordinated retriever, verifier, and reasoning agents to reduce hallucinations and increase guideline alignment across diverse clinical tasks. Traditional LLMs and classifier-based models cannot separate reasoning from verification, making them more prone to errors and unsupported claims[38][39].Multi-agent RAG architectures enforce hierarchical reasoning and evidence validation, producing outputs that are safer, traceable, and more clinically defensible[10][40]. These properties make multi-agent RAG systems a promising foundation for specialty specific AI assistants that require high levels of reliability and transparency[13].

6) Multimodal and Domain Specific Decision Support

The multimodal clinical decision support tasks, systems must be capable of combining imaging, text and structured data [13]. [24] introduced six-agent HCC imaging systems which achieved better prediction of microvascular invasion through the role of agents that was assigned for imaging features, *textual* evidence and cross modal fusion[41]. This allowed for image text reasoning, which cannot be achieved with standard CNNs or text-based algorithms. In Multimodal RAG accordingly is a significant improvement compared to conventional unimodal tools, which provide richer and more interpretable domain-specific decision support [23][42].

RQ2.In what ways do multi-agent RAG systems enhance explainability, transparency, and clinician trust?

The multi-agent RAG systems enhances explainability in the planner, retriever and verifier frameworks [22][12]. Decomposing the reasoning process into modular steps meant that for every diagnostic or interpretive decision, clinicians could follow how it resulted[43]. The planner agent initially formulated the clinical problem, the retriever retrieved domain-relevant evidence and the verifier agent verified the internal consistency of the output[13]. This stepwise reasoning is more intuitive for clinician thinking than end-to-end LLMs or classical predictive models, making complex decisions accessible[29]. The explainability in [24] was further enhanced with the use of multimodal agent specialisation where different agents handled imaging, textual data, and fusion reasoning. This made it clear how radiological patterns and textual evidence worked in relation to the overall prediction[23]. In contrast to traditional systems, which typically give a numeric prediction or a hard rule without rationale multi-agent RAG, systems generate interpretable paths of reasoning, ensuring that their decisions are more explainable[40].

Transparency was enhanced mainly thanks to output associated with evidence and auditing intermediate reasoning, components generally observed in multi-agent and RAG-based systems.[16] These explanations were directly connected to retrieved biomedical literature so clinicians could study sources of information behind diagnostic suggestions. [7] Evaluated transparency using RAGAS metrics and was found to be highly faithful with appropriate context. In [12][22]the intermediate agent outputs steps to plan, retrieved evidence, verification results were the subjects of analysis and would allow for auditability that was not possible with monolithic ML systems. Multimodal [23][24] also added transparency by demonstrating how the image features were connected to text retrieval in order to draw a conclusion.

Clinical trust was enhanced when systems were shown to be accurate, reliable and perform safely with multi-agent RAG architectures[13]. Trust enhanced in [7] due to the generation of correct responses, and a guideline-linked response, with clinicians rating clarity and trust greater than 4.2/5. Verification agents were vital in [22][12] clearing unsupported outputs and preventing hallucinations the single greatest barrier to building clinician confidence in AI. KG-RAG systems [18] enhanced trust even further, by centering triage decisions in biomedical ontologies with almost perfect sensitivity and specificity, something that the clinical community finds appealing during emergencies. Clinicians trust outputs more in multimodal domains [24] because multimodal agent reasoning lets clinicians see exactly how imaging observations are combined with text-based evidence informed predictions.

RQ3. What measurable improvements (accuracy, recall, latency) do multi-agent RAG systems achieve over single-agent or classical approaches?

Multi-agent and retrieval-augmented systems demonstrated measurable performance improvements over both single-agent and traditional machine learning approaches, particularly in accuracy, recall and reasoning reliability. Rare disease diagnosis improved from 37% to 40–43% with retrieval augmentation [16], while multi-agent frameworks for clinical data mapping achieved a 7.2% accuracy gain over rule-based methods [19], and retrieval-based long-tail diagnostic support outperformed Clinical BERT on rare classes [16]. Emergency triage using KG-RAG achieved near-perfect performance (Accuracy 0.99; Sensitivity 0.98; Specificity 0.99), far exceeding keyword-based triage systems [18]. Multi-agent systems such as planner, retriever and verifier pipelines reduced hallucinations and improved reasoning consistency compared with single-agent LLMs [22][12], while multimodal multi-agent models demonstrated superior diagnostic performance by integrating imaging and text more effectively than classical CNN-based or text-only systems [23][24]. Although multi-agent pipelines introduced modest latency increases in some settings [12], these trade-offs were outweighed by significant gains in accuracy, robustness, and factual grounding. Overall, the evidence shows that multi-agent RAG systems consistently outperform classical and single-agent models in clinical decision-support tasks, offering safer, more accurate and better-grounded outputs[44][45].

RQ4.What human infrastructural and ethical factors influence their adoption in low resource contexts.

High-level infrastructural constraints emerged as serious challenges of multi-agent RAG systems in low-resource settings for the implementation [46][47].It has been shown in several studies that multi-agent and multimodal models are resource consuming and difficult for under-resourced healthcare systems in which the computational capabilities are limited in scale. The multimodal HCC 6-agent system [24], for example, relied on high-performance GPUs and large annotated imaging datasets, restrictions that limit deployment beyond advanced radiology centres. Similarly, knowledge-graph-based RAG models [18] also necessitated constant graph maintenance in addition to ontology updates and reliable storage systems, such as the kind that are generally absent in rural and low-income healthcare settings. [12] also emphasised that multi-agent verification pipelines result in extra computational overheads that may strain limited server capacity. Infrastructure shortfalls go beyond just computation[21] demonstrated that multilingual oncology documentation systems rely on consistent internet connections, stable ASR pipelines, as well as robust data integration layers. Collectively, the result shows that low-resource healthcare environments often do not provide the digital infrastructure computing power, network reliability, data storage, EHR maturity, and technical tools for maintenance that modern multi-agent RAG systems require for safe and efficient deployment[48][49].

Ethical aspects contribute to the successful implementation of multi-agent RAG systems in low-resource environments that relate to data integrity, safety, and fairness. Several studies flagged biases in training data, including in [16] the rare-disease dataset contained only 30 conditions, leading to issues of representativeness and fairness, or in [18][21] the absence of knowledge-graph links or lack of multilingual coverage might create danger such as misclassification or unsafe recommendations. The addition of verification agents in [22][12]reduced hallucinations, but the lack of such safeguards in low-resource configurations may increase clinical risk. Ethical governance issues also relate to transparency for clinicians working in such settings, digital literacy may be scarce, meaning they may struggle to make sense of evidence-linked outputs or judge the trustworthiness of the system [50]. Privacy concerns are exacerbated in regions where cloud-based access is needed, or where external storage is required, and where data privacy regulations do not exist[51]. The importance of authentic clinician validation as an ethical principle may be lost on resource-constrained settings facing the fast digitisation challenges [20]. In general, ethical adoption requires data quality, no bias, and transparency[52].

To move from prototypes to deployable CDSS, ethical governance should be anchored to established healthcare requirements rather than generic principles. At a minimum, systems should implement medical device software risk management lifecycle controls, for example, SaMD-oriented risk management and software lifecycle processes [53]. Clinical safety management should include documented hazard analysis for misclassification, hallucination risks, and data protection compliance covering consent, access logging, retention, and cross-border data transfer [54].

G. Conceptual Framework for Multi Agent RAG in Clinical Decision Support

The proposed four-agent framework is not introduced as a speculative design but is systematically derived from the literature synthesis. Across the reviewed studies, recurring functional roles were identified, including evidence retrieval, clinical reasoning, factual verification, and risk or safety control. These roles appeared consistently across diverse implementations, albeit under different labels and configurations. By abstracting these recurring functions, the review consolidates them into a minimal yet implementable architecture comprising retriever, reasoner, verifier, and safety agents. The framework therefore represents a synthesis of empirical design patterns observed across studies rather than an independently proposed system.

The proposed framework is systematically derived from the literature synthesis and is summarised through the role to evidence mapping in Table VI

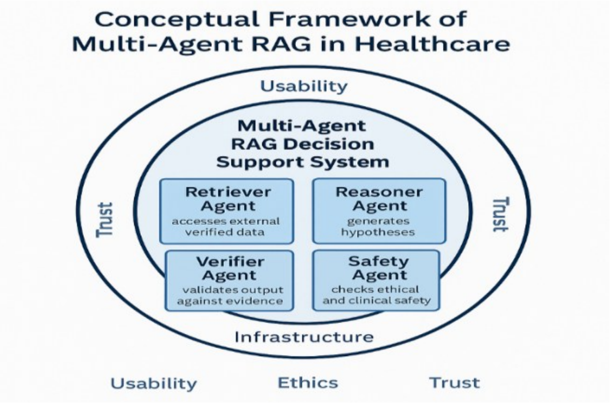

Figure 5: Conceptual Framework

Figure 5 displays an abstract of the Multi-Agent Retrieval-Augmented Generation (RAG) based approach to clinical decision support. In the core of the model, we propose a Multi-Agent RAG Decision Support System, with dedicated agents that act as facilitators of varied levels of clinical reasoning. In order to ground the model with credible evidence, the retriever agent can be grounded in authoritative external knowledge resources, such as clinical guidelines, biomedical literature. The reasoner agent aggregates these retrieved elements to develop hypotheses or preliminary clinical conclusions. The verifier agent examines these outputs for evidence, which reduces hallucinations and increases credibility. A safety agent who takes oversight is responsible for checking for ethical risks, clinical hazards and guideline violations, before any information is delivered to the user. The contextual factors that will affect the adoption and performance of the core system are usability, ethics, infrastructure and trust.

TABLE VI:
MAPPING OF REVIEW FINDING TO THE PROPOSED AGENTS

Framework	**Role derived**	**Representative studies**	**Design implications for CDSS**
Component	Retrieves guideline/literature/ontology evidence; may use vector, KG, or IR retrieval.	[16][17] [20][21] [8]	Use curated clinical sources; log retrieved passages; monitor retrieval quality and coverage.
Retriever Agent	Synthesizes retrieved evidence into clinical hypotheses, summaries	[12][22] [24]	Constrain reasoning to retrieved evidence; structure outputs to match clinician workflow.

Reasoner agent	Checks factual consistency, guideline alignment, and internal contradictions filters hallucinations.	[12][22] [19]	Implement claim checking, cross-retrieval, or multi-pass verification; surface uncertainty and provenance for audit.
Verifier Agent	Applies risk controls: contraindication checks, privacy controls, bias checks, escalation to clinician.	[18][24] [23]	Embed safety policies, red-flag triggers, and escalation pathways; align to local clinical governance data protection requirements.

1) Linking agentic/multi-agent components to TAM and Diffusion of Innovations

To address adoption determinants, we map technical components to TAM and Diffusion of Innovations constructs. Perceived usefulness is operationalized through measurable clinical value diagnostic accuracy, triage performance, guideline alignment enabled by the retriever and reasoner combination. Perceived ease of use is influenced by workflow fit, interaction design, latency, and explainability artefacts produced by stepwise agent outputs. Verifier and safety agents that provide provenance, contradiction checks, and escalation pathways strengthen trust. Within Diffusion of Innovations, relative advantage corresponds to improvements over rule-based CDSS compatibility reflects integration with existing EHR workflows complexity is driven by infrastructure demands and multi-agent orchestration.

H. Implications of the Study

1) Practical Implications

Practical implications for CDSS developers and healthcare organisations are as follows. Start with evidence curation restrict retrieval to authoritative guidelines, formularies, and local protocols, and continuously evaluate retrieval quality. Pilot in a low-risk where outputs are logged and reviewed by clinicians before influencing care. Incrementally introduce agent roles begin with retriever, reasoner, then add verifier and safety/governance. Implement clinician-centred interaction patterns: structured differential diagnosis, and transparent provenance links to retrieved sources. Budget for latency and compute multi-agent orchestration requires caching, asynchronous retrieval, and lightweight models to fit low-resource settings. Establish governance versioned policies, audit logs, bias monitoring, and clear accountability for human override. These steps translate the conceptual framework into implementable stages suitable for both high-resource and low-resource environments.

2) Theoretical Implications

This research contributes to the existing research on the relationship between multi-agent and retrieval augmented generation (RAG) for improving reasoning in artificial intelligence (AI). The study adds to theories of clinical decision support systems by defining multi agent based collaboration and retrieval as ways to facilitate more adaptive and explainable intelligence. The results confirm that the support the retrieval-augmented generation (RAG) provides to multi agent frameworks enables reasoning, knowledge verification and contextual learning as opposed to static rule based systems. This shifts artificial intelligence (AI) research in the healthcare sector from single model automation towards distributed cognitive architectures capable of dialogue, feedback and self-correction. These findings also set a conceptual framework for the exploration of trust, fairness and accountability issues in collaborative artificial intelligence (AI). These implications may serve as the foundation for future theoretical work to develop models that measure how multi agent RAG systems reasoning structures influence human, artificial intelligence collaboration and medical decision-making.

IV. Limitations Future Work And Conclusion

A) Limitations

This systematic review provides insights into multi-agent RAG clinical decision support systems in healthcare. The search was restricted to three databases Science Direct, IEE Xplore and PubMed. We considered only studies in English published from 2020 to 2025, which might have excluded other non-English studies. The overall number of studies reporting on healthcare based multi agent RAG systems was relatively small and limited the generalizability of finding across medical specialities. Most of the studies that were included were simulation based in comparison to those that were clinically implemented, which restricts the applicability in practice. A study bias may also be a concern as a number of included studies reported good in the majority while some have mentioned ethical considerations. The regional representation was uneven with few contributions from developing countries. These limitations indicate that although the review shows the potential of multi agent RAG systems, more empirical, multi database and generally representative studies should be conducted to confirm its successful action in real healthcare environment.

B) Future Works

Future research on multi-agent clinical decision-support systems using retrieval-augmented generation (RAG) should focus on translating promising experimental results into real-world clinical practice. Based on the studies reviewed, the majority of models were still being experimentally validated in controlled settings, emphasising the requirement for clinical pilot testing to assess usability, reliability and patient safety. To keep up with the emerging technologies systems need to take a standardised approach to the evaluation of diagnostic accuracy, reasoning transparency and computational efficiency to ensure that fair comparisons can be achieved on the scale of other domains. There is also a need for lightweight multi agent architectures to decrease system latency and to make them ready for low resource healthcare settings. The integration of knowledge graphs and feedback based learning increases contextual understanding. Expanding research participation from developing countries is essential to develop multi agent RAG systems that resonate with the local clinical and technological context. There is a need for a deeper understanding of human oversight models, bias detection and audit frameworks to guarantee fairness and accountability.

C) Conclusion

This systematic literature review examined the emerging use of multi agent retrieval augmented generation systems in clinical decision support. The synthesis shows that architectures employing explicit agent roles such as retrieval, reasoning, verification and safety offer conceptual advantages over single agent RAG systems by improving transparency, modularity and safeguards against unverified outputs. These properties align well with the requirements of clinical environments where traceability and trust are essential. Despite this potential current evidence remains limited in both scale and maturity. Most reviewed studies rely on small datasets, simulation based evaluations or narrow clinical contexts. Ethical considerations safety governance and regulatory alignment are frequently acknowledged but rarely operationalised within system designs. In addition, while the in integration of technology adoption perspectives provides valuable insight, empirical validation of clinical trust and usability remains largely unexplored.Multi agent RAG systems may offer more transparent and evidence grounded decision support compared to generative models. The findings emphasise the importance of modular architectures, verification mechanisms and explicit safety agents. For researchers and policymakers the review highlights the need for standardised evaluation protocols, regulatory aware system design and robust empirical studies conducted in real world clinical settings. Overall multi agent RAG systems should be regarded as a promising yet early stage approach requiring significant methodological ethical and practical refinement before widespread clinical adoption.

REFERENCES

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med. 2019 251*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[2] 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," Dec. 01, 2020, *Nature Research*. doi: 10.1038/s41746-020-0221-y.

[3] M. Mangundu, L. Roets, and E. J. van Rensberg, "Accessibility of healthcare in rural Zimbabwe: The perspective of nurses and healthcare users," *African J. Prim. Heal. Care Fam. Med.*, vol. 12, no. 1, pp. 1–7, 2020, doi: 10.4102/PHCFM.V12I1.2245;CTYPE:STRING:JOURNAL.

[4] Y. Kinfu, M. R. Dal Poz, H. Mercer, and D. B. Evans, "The health worker shortage in Africa: Are enough physicians and nurses being trained?," *Bull. World Health Organ.*, vol. 87, no. 3, pp. 225–230, 2009, doi: 10.2471/BLT.08.051599.

[5] A. B. Mbakwe, I. Lourentzou, L. Anthony, C. Id, J. Mechanic, and A. Dagan, "PLOS DIGITAL HEALTH ChatGPT passing USMLE shines a spotlight on the flaws of medical education," pp. 9–11, 2023, doi: 10.1371/journal.pdig.0000205.

[6] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[7] B. B. Ozmen *et al.*, "Development of a novel artificial intelligence clinical decision support tool for hand surgery: HandRAG.," *J. Hand Microsurg.*, vol. 17, no. 4, p. 100293, Jul. 2025, doi: 10.1016/j.jham.2025.100293.

[8] X. Zhao, S. Liu, S. Y. Yang, and C. Miao, *MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot*, vol. 1, no. 1. arXiv, 2025. doi: 10.1145/3696410.3714782.

[9] "How Multi-Agent RAG Systems Transform Healthcare Decision Support - Techkraft Inc." Accessed: Sep. 03, 2025. [Online]. Available: https://techkraftinc.com/how-multi-agent-rag-systems-transform-healthcare/

[10] G. de A. e Aquino *et al.*, "From RAG to Multi-Agent Systems: A Survey of Modern Approaches in LLM Development," Feb. 2025, doi: 10.20944/PREPRINTS202502.0406.V1.

[11] L. Deng, H. Hu, K. Lu, and P. He, "LLM-augmented multi-agent cooperative framework for medical case retrieval in cardiology," vol. 123, 2025.

[12] C. Chang *et al.*, "MAIN-RAG : Multi-Agent Filtering Retrieval-Augmented Generation".

[13] T. Nguyen, P. Chin, and Y.-W. Tai, "MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Reasoning," vol. 1, pp. 1–27, 2025, [Online]. Available: http://arxiv.org/abs/2505.20096

[14] Y. Miao, Y. Zhao, Y. Luo, H. Wang, and Y. Wu, "Improving Large Language Model Applications in the Medical and Nursing Domains With Retrieval-Augmented Generation: Scoping Review," *J. Med. Internet Res.*, vol. 27, 2025, doi: https://doi.org/10.2196/80557.

[15] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, 2009, doi: 10.1371/journal.pmed.1000097.

[16] C. Zelin, W. K. Chung, M. Jeanne, G. Zhang, and C. Weng, "Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT.," *J. Biomed. Inform.*, vol. 157, p. 104702, Sep. 2024, doi: 10.1016/j.jbi.2024.104702.

[17] D. Baur, J. Ansorg, C.-E. Heyde, and A. Voelker, "Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study.," *JMIR AI*, vol. 4, p. e75262, Oct. 2025, doi: 10.2196/75262.

[18] S. Liu, A. P. Wright, A. B. McCoy, S. S. Huang, B. Steitz, and A. Wright, "Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation.," *J. Am. Med. Inform. Assoc.*, vol. 32, no. 6, pp. 1032–1039, Jun. 2025, doi: 10.1093/jamia/ocaf059.

[19] K. Gilani, M. Verket, C. Peters, M. Dumontier, H.-P. B.-L. Rocca, and V. Urovi, "CDE-Mapper: Using retrieval-augmented language models for linking clinical data elements to controlled vocabularies.," *Comput. Biol. Med.*, vol. 196, no. Pt B, p. 110745, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110745.

[20] T. Abdullahi, L. Mercurio, R. Singh, and C. Eickhoff, "Retrieval-Based Diagnostic Decision Support: Mixed Methods Study," *JMIR Med. Informatics*, vol. 12, p. e50209, Jun. 2024, doi: 10.2196/50209.

[21] I. Emssaad, F.-E. Ben-Bouazza, I. Tafala, M. C. El Mezali, and B. Jioudi, "Leveraging multilingual RAG for breast cancer RCPs: AI-driven speech transcription and compliance in Darija-French

clinical discussions," *Comput. Methods Programs Biomed. Updat.*, vol. 8, p. 100221, 2025, doi: https://doi.org/10.1016/j.cmpbup.2025.100221.

[22] A. Madrid-García, D. Benavent, and B. Merino-Barbancho, "From chat to act: large language model agents and agentic AI as the next frontier of AI in rheumatology," *EULAR Rheumatol. Open*, vol. 1, no. 3, pp. 147–156, 2025, doi: https://doi.org/10.1016/j.ero.2025.06.012.

[23] E. Tzanis *et al.*, "Agentic systems in radiology: Principles, opportunities, privacy risks, regulation, and sustainability concerns," *Diagn. Interv. Imaging*, 2025, doi: https://doi.org/10.1016/j.diii.2025.10.002.

[24] L. Wang *et al.*, "A multimodal LLM-agent framework for personalized clinical decision-making in hepatocellular carcinoma," *Patterns*, p. 101364, 2025, doi: https://doi.org/10.1016/j.patter.2025.101364.

[25] B. Clay, H. I. Bergman, S. Salim, G. Pergola, J. Shalhoub, and A. H. Davies, "Natural language processing techniques applied to the electronic health record in clinical research and practice - an introduction to methodologies," *Comput. Biol. Med.*, vol. 188, p. 109808, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.109808.

[26] M. Wang, Y. Shen, B. Zhao, X. Zhou, L. Sun, and X. Liu, "Enhancing LLM-based clinical reasoning in anesthesiology via graph-augmented retrieval and explainable generation.," *Heal. Inf. Sci. Syst.*, vol. 13, no. 1, p. 62, Dec. 2025, doi: 10.1007/s13755-025-00379-x.

[27] S. A. Gebreab, K. Salah, R. Jayaraman, M. H. ur Rehman, and S. Ellaham, "LLM-Based Framework for Administrative Task Automation in Healthcare," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 2024, pp. 1–7. doi: 10.1109/ISDFS60797.2024.10527275.

[28] T.-C. Ureche, L. Boicescu, M. Raducanu, E.-C. Popovici, S. Halunga, and D. N. Vizireanu, "The Convergence of Emerging Technologies and AI-Driven Autonomous Cybersecurity in Smart Digital Ecosystems," in *2025 IEEE 2nd International Conference on Blockchain, Smart Healthcare and Emerging Technologies (SmartBlock4Health)*, 2025, pp. 1–6. doi: 10.1109/SmartBlock4Health64843.2025.11189602.

[29] J. C. L. Ong *et al.*, "Large language model as clinical decision support system augments medication safety in 16 clinical specialties.," *Cell reports. Med.*, vol. 6, no. 10, p. 102323, Oct. 2025, doi: 10.1016/j.xcrm.2025.102323.

[30] O. K. Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review.," *Digit. Heal.*, vol. 11, p. 20552076251337176, 2025, doi: 10.1177/20552076251337177.

[31] A. Mohammad, A. Bulbanat, and F. Aljassar, "Comparative Evaluation of Diagnostic and Management Capabilities of Infiniti AI and ChatGPT-4o in Corneal Diseases.," *Cureus*, vol. 17, no. 10, p. e95163, Oct. 2025, doi: 10.7759/cureus.95163.

[32] F. Gaber *et al.*, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis.," *NPJ Digit. Med.*, vol. 8, no. 1, p. 263, May 2025, doi: 10.1038/s41746-025-01684-1.

[33] G. Shan *et al.*, "Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis," *JMIR Med. Informatics*, vol. 13, 2025, doi: https://doi.org/10.2196/64963.

[34] K. Balakrishnan *et al.*, "Artificial Intelligence in Rural Healthcare Delivery: Bridging Gaps and Enhancing Equity through Innovation".

[35] S. Varshney, B. Jain, P. Singh, D. Rani, and S. Mehra, "Med-KGMA: A novel AI-driven medical support system leveraging knowledge graphs and medical advisors," *Comput. Biol. Med.*, vol. 197, p. 110929, 2025, doi: https://doi.org/10.1016/j.compbiomed.2025.110929.

[36] Y. Habchi *et al.*, "Advanced deep learning and large language models: Comprehensive insights for cancer detection," *Image Vis. Comput.*, vol. 157, p. 105495, 2025, doi: https://doi.org/10.1016/j.imavis.2025.105495.

[37] Z. Yao and H. Yu, "A Survey on LLM-based Multi-Agent AI Hospital," 2025, [Online]. Available: https://osf.io/bv5sg_v1

[38] Z. Wang, H. Li, D. Huang, H.-S. Kim, C.-W. Shin, and A. M. Rahmani, "HealthQ: Unveiling questioning capabilities of LLM chains in healthcare conversations," *Smart Heal.*, vol. 36, p. 100570, 2025, doi: https://doi.org/10.1016/j.smhl.2025.100570.

[39] L. A. Biesheuvel *et al.*, "Large language models in critical care," *J. Intensive Med.*, vol. 5, no. 2, pp. 113–118, 2025, doi: https://doi.org/10.1016/j.jointm.2024.12.001.

[40] G. Tippani, S. Durusoju, L. Mergoju, and P. V. Reddy, "MedicAI: AI-Powered System for Enhanced Medical Information Access," in *2025 9th International Conference on Inventive Systems and Control (ICISC)*, IEEE, Aug. 2025, pp. 804–809. doi: 10.1109/ICISC65841.2025.11188019.

[41] T. Çelikten and A. Onan, "Medcongtm: Interpretable multi-label clinical code prediction with dual-view graph contrastive topic modeling," *Knowledge-Based Syst.*, vol. 327, p. 114103, 2025, doi: https://doi.org/10.1016/j.knosys.2025.114103.

[42] Y. Hu, C. Xu, B. Lin, W. Yang, and Y. Y. Tang, "Medical multimodal large language models: A systematic review," *Intell. Oncol.*, vol. 1, no. 4, pp. 308–325, 2025, doi: https://doi.org/10.1016/j.intonc.2025.09.005.

[43] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," *Proc. Mach. Learn. Res.*, vol. 235, pp. 11733–11763, 2024.

[44] Chakraborty et. al, "Medical Application Using Multi Agent System - A Literature Survey Sougata Chakraborty *, Shibakali Gupta **," *Int. J. Eng. Res. Appl.*, vol. 4, no. 2, pp. 528–546, 2014.

[45] S. M. Palomares *et al.*, "The Impact of Artificial Intelligence Technologies on Nutritional Care in Patients With Chronic Kidney Disease: A Systematic Review," *J. Ren. Nutr.*, 2025, doi: https://doi.org/10.1053/j.jrn.2025.06.002.

[46] M. Mangundu, L. Roets, and E. J. van Rensberg, "Accessibility of healthcare in rural Zimbabwe: The perspective of nurses and healthcare users," *African J. Prim. Heal. Care Fam. Med.*, vol. 12, no. 1, p. 2245, 2020, doi: 10.4102/PHCFM.V12I1.2245.

[47] M. Alves, J. Seringa, T. Silvestre, and T. Magalhães, "Use of Artificial Intelligence tools in supporting decision-making in hospital management," *BMC Health Serv. Res.*, vol. 24, no. 1, 2024, doi: 10.1186/s12913-024-11602-y.

[48] S. Sun, Z. Zhong, N. Yu, X. Gong, and K. Yang, "HumanMoD: A multi-RAG collaborative LLM for inclusive urban public healthcare services," *Appl. Soft Comput.*, vol. 184, p. 113684, 2025, doi: https://doi.org/10.1016/j.asoc.2025.113684.

[49] V. Zuhair *et al.*, "Exploring the Impact of Artificial Intelligence on Global Health and Enhancing Healthcare in Developing Nations," *J. Prim. Care Community Heal.*, vol. 15, 2024, doi: 10.1177/21501319241245847.

[50] K. He *et al.*, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," *Inf. Fusion*, vol. 118, p. 102963, 2025, doi: https://doi.org/10.1016/j.inffus.2025.102963.

[51] M. Tarczyńska-łuniewska, "ScienceDirect ScienceDirect The procedure of building stable fundamental database with the use The procedure of building stable fundamental database with the use of Matlab software of Matlab software," *Procedia Comput. Sci.*, vol. 126, pp. 2163–2172, 2018, doi: 10.1016/j.procs.2018.07.232.

[52] C. Lin and C.-F. Kuo, "Roles and potential of Large language models in healthcare: A comprehensive review," *Biomed. J.*, vol. 48, no. 5, p. 100868, 2025, doi: https://doi.org/10.1016/j.bj.2025.100868.