Retrieval-Augmented Generation in Healthcare Systems: A Systematic Review with Emphasis on Public Data

Rafael Santos Novo Pereira[1] , **Andre Gomes Regino**[2,3,4,*], **Fernando Rezende Zagatti**[2,3,4,5] , **Matheus Bernardelli de Moraes**[2] , **Guilherme Ruppert**[2] , **Ana Carolina Monteiro**[2] , **Rodrigo Bonacin**[1,2]

[1] Centro Universitário Campo Limpo Paulista – Campo Limpo Paulista, SP – Brasil ,
[2] Centro de Tecnologia da Informação Renato Archer – Campinas, SP – Brasil ,
[3] Centro Universitário de Jaguariúna (UniFAJ) – Jaguariúna, SP – Brasil ,
[4] Centro Universitário Max Planck (UniMAX) – Indaiatuba, SP – Brasil ,
[5] Universidade Federal de São Carlos (UFSCar) – São Carlos, SP – Brasil

rafaelsantos.p@gmail.com

{aregino,fzagatti,mbmoraes,gruppert,amonteiro,rbonacin}@cti.gov.br

Abstract. *Retrieval-Augmented Generation (RAG) enhances Large Language Models by integrating external knowledge retrieval. In public healthcare, this is relevant for handling heterogeneous data and specialized terminology. This paper presents a systematic review of RAG applications in healthcare, focusing on challenges and opportunities for Brazilian public data systems such as DataSUS. Following PRISMA, 145 studies were analyzed, with 14 meeting the inclusion criteria. Results indicate advances in hybrid retrieval and prompt engineering, but reveal gaps in regulatory compliance, multilingual support, and integration with Brazilian datasets. The study highlights research opportunities for developing RAG solutions tailored to public healthcare systems in Brazil.*

1. Introduction

The increasing availability of public data on the Internet plays a central role in scientific advances in the public health sector. Public data facilitate scientific collaboration, enrich research, and enhance transparency and analytical capacity in public health decision-making [Huston et al. 2019]. With the exponential growth of biomedical literature, from approximately 270,000 articles per year in the 1980s to more than 1.5 million publications indexed annually in PubMed by 2025, health professionals and policymakers face an information paradox.

This scenario of cognitive overload, in which professionals must filter, evaluate, and synthesize information dispersed across multiple sources—ranging from scientific articles and clinical guidelines to electronic health records and epidemiological databases—makes it difficult to remain up to date without advanced computational tools. The complexity is further amplified by the rapid pace at which medical knowledge evolves, where clinical evidence can become obsolete within a matter of months, requiring

*Corresponding author: aregino@cti.gov.br

constant revision of protocols and practices. Thus, although digitalization has democratized access to information, it has created barriers to extracting actionable and clinically relevant knowledge in a timely manner for decision-making [Lewis et al. 2020].

In the Brazilian context, DataSUS [Ministério da Saúde 2025] is the primary source of health record data in the country. However, the technical complexity of the systems that compose DataSUS creates barriers. To illustrate this difficulty, consider a municipal health manager who needs to analyze COVID-19 vaccination coverage within their jurisdiction. The manager must navigate outdated interfaces (e.g., the TABNET system still requires the Internet Explorer browser), understand complex codifications (e.g., Imuno_cobertura_COVID_D1), and manually integrate data from multiple sources.

Advances in Generative Artificial Intelligence (GAI) and intelligent chat systems have enabled new ways to access, process, and interact with open data [Yang et al. 2023]. In this context, GAI represents an alternative for enhancing the use and benefits of open health data. More specifically, Retrieval-Augmented Generation (RAG) systems have emerged as a potential solution to these challenges. A RAG model combines parametric and non-parametric pre-trained memory for language generation, enabling Large Language Models (LLMs) to dynamically access up-to-date knowledge bases through natural conversational interfaces [Lewis et al. 2020].

For this reason, this review integrates and analyzes the state of the art in RAG techniques, with a focus on applications to health data repositories. The systematic analysis identified emerging techniques, including domain-specific fine-tuning for medical contexts, recursive chunking using LangChain [Topsakal and Akinci 2023], customized embeddings for specialized terminology, and hybrid architectures such as *Medical Graph RAG* [Wu et al. 2024]. The focus of this work is to identify promising techniques and gaps in the international scientific literature, to propose directions for the development of solutions adapted to the characteristics of Brazilian open data infrastructures.

The motivation for this systematic review originated from the identification of operational barriers faced by professionals and administrators within the Brazilian Unified Health System (SUS) when attempting to implement artificial intelligence solutions developed for international contexts. A preliminary literature survey revealed that no prior review has systematically examined how the specific characteristics of Brazilian technological infrastructure affect the implementation of RAG-based technologies. This includes legacy data formats, connectivity limitations in remote municipalities, and heterogeneity among information systems. Additionally, Brazilian regulatory requirements and the need to process medical terminology in Portuguese create challenges not present in international implementations. Therefore, in addition to synthesizing the available technical knowledge, this study seeks to identify the essential adaptations required to enable these technologies within the SUS.

2. Methodology and Review Execution

This study followed the PRISMA methodology (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) [Page et al. 2021] to conduct the systematic review, ensuring transparency and reproducibility in the study selection process. The guiding research question, structured according to the PICO framework (*Population, Intervention, Comparison, Outcome*) [Page et al. 2021], was defined as follows:

"How RAG techniques been applied in the healthcare domain, what are their results in terms of effectiveness and accuracy, and which adaptations would be necessary for the specific context of Brazilian public health data?"

Initially, an exploratory search was conducted using general terms ("RAG healthcare", "retrieval augmented medical") in order to identify seminal articles. The analysis of these studies revealed recurring descriptors that were used to refine the final search string. The following databases were selected to cover both technical computing literature and biomedical publications: IEEE Xplore, ACM Digital Library, and PubMed.

The search string was iteratively refined based on recurring terms identified in the initial articles and descriptors from the literature. The final version used in the search process is presented in Equation 1. Since the search was conducted in English-language databases, the acronym RAG (*Retrieval-Augmented Generation*) was adopted.

$$\begin{aligned}&(\text{retrieval augmented generation } \textbf{OR} \text{ RAG } \textbf{OR} \text{ retrieval-augmented}) \ \textbf{AND} \\ &(\text{healthcare } \textbf{OR} \text{ health care } \textbf{OR} \text{ medical } \textbf{OR} \text{ clinical } \textbf{OR} \text{ biomedical}) \ \textbf{AND} \\ &(\text{public health } \textbf{OR} \text{ health data } \textbf{OR} \text{ electronic health records } \textbf{OR} \text{ EHR } \textbf{OR} \\ &\text{clinical decision support})\end{aligned} \tag{1}$$

The inclusion and exclusion criteria were defined to guarantee the relevance and timeliness of the analyzed studies (cf. Table 1).

Table 1. Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Publications from 2023 to 2025	Publications prior to 2023
Practical implementations with validation	Purely theoretical proposals
Focus on healthcare or public data	Superficial mention of healthcare only as an example
Peer-reviewed articles in English or Portuguese	Non–peer-reviewed publications
Highly relevant preprints with multiple citations	Duplicate articles or preliminary versions

The selection process followed the four stages of the PRISMA protocol: (i) identification; (ii) screening; (iii) eligibility; and (iv) inclusion. The initial search returned 145 records (IEEE Xplore: 42, ACM Digital Library: 38, PubMed: 65). During the screening phase, 19 duplicates were removed, resulting in 126 unique articles. Subsequently, titles and abstracts were analyzed, leading to the exclusion of 89 studies. The main reason for exclusion at this stage was lack of relevance to the research topic, with articles addressing RAG in other domains (e.g., finance, education) or publications without original research content (editorials, commentaries, and conference abstracts).

In the eligibility stage, the remaining 37 articles were reviewed in detail to assess their eligibility according to the inclusion and exclusion criteria described in Table 1. At this stage, 27 articles were discarded for the following reasons: 15 described theoretical RAG architectures but did not present a functional implementation ("theoretical proposals only"); 8 did not focus on healthcare, using only superficial medical examples (e.g., general wellness chatbots); and 4 lacked adequate validation of the results presented. Fi-

nally, 14 studies met all the criteria and were selected for analysis, representing the main current research directions in the application of RAG within the healthcare context.

3. Results and Discussion

Tables 2 and 3 present the 14 studies selected for this review. The temporal distribution of the works demonstrates an acceleration in scientific production on RAG applications in healthcare contexts. Specifically, two studies were identified in 2023, eleven studies in 2024, and one study already available in early 2025. This trend reflects not only the maturation of the fundamental technologies that support RAG, but also the growing recognition of the potential of this approach in critical healthcare applications.

Table 2. Summary of the main publications on the use of RAG in healthcare (2023–2025)

Reference	Main technique	Application context
[Alkhalaf et al. 2024]	RAG + Zero-shot (GPT-3.5)	Nutritional records in long-term care institutions
[Chen et al. 2023]	Fine-tuning of MEDITRON-70B	Multilingual medical model
[Karamanlıoğlu et al. 2024]	LLMs + Federated Learning + Streaming	Decision-support system for emergency triage
[Miao et al. 2024]	Hybrid medical KDIGO-RAG framework	Nephrology using KDIGO 2023 guidelines
[Unlu et al. 2024]	RAG + GPT-4 with recursive chunking	Patient screening for clinical trials
[Viana et al. 2023]	Systematic review	Limitations of DataSUS in surgical research (47 studies)
[Ziletti and D'Ambrosi 2024]	RAG + Text-to-SQL	Epidemiological queries
[de Almeida Cardoso et al. 2024]	NLP with neural networks	Analysis of 432 CONITEC reports (2012–2022)
[Wu et al. 2024]	Graph RAG	Safety in medical LLMs
[Yang et al. 2025]	Two-phase fine-tuning	ICD-10 medical coding
[Zakka et al. 2024]	RAG + GPT-4	General clinical system (NEJM AI)
[Zhou et al. 2024]	RAG + customized embeddings	Gastroenterology chatbot
[Kresevic et al. 2024]	RAG + optimization	Hepatology guidelines
[Yang et al. 2025]	Perspective analysis	RAG in generative AI for healthcare

The analysis of the geographical distribution of the studies — United States (4), United Kingdom (2), Germany (1), Canada (1), China (1), and international collaborations (5) — demonstrates a concentration in developed countries and an absence of research originating from Latin America. This geographical gap represents an unexplored academic opportunity and highlights disparities in the global development of healthcare technologies, which may perpetuate inequalities in access to advanced medical information processing tools.

Despite nationally documented advances that have received international recognition—including the *microdatasus* package developed by Fiocruz that improved access to DataSUS data, the CONITEC analysis developed by Cardoso et al. [de Almeida Cardoso et al. 2024] which achieved 87.1% AUC-ROC, and the systematic review on DataSUS limitations conducted by Viana et al. [Viana et al. 2023]—none of the international studies identified in this review adequately address the technical, linguistic, and regulatory specificities of Brazilian health information systems.

Figure 1 presents the percentage of analyzed articles that do not address topics considered fundamental for the Brazilian context. The scarcity of Brazilian public data reflects the absence of studies addressing DataSUS terminology, including its proprietary formats (DBF, DBC), obsolete interfaces (TABNET), and the complexity of coding systems adapted to the Brazilian healthcare reality. This gap is significant considering that solutions developed for other contexts are often not directly applicable to the Brazilian environment due to its unique technical and regulatory specificities.

Table 3. Summary of the main results, findings, and limitations regarding the use of RAG in healthcare (2023–2025)

Reference	Key findings	Identified limitations
[Alkhalaf et al. 2024]	Accuracy: 93.25% without RAG → 99.25% with RAG; 89% reduction in hallucinations	Data from a single institution; specific focus on malnutrition
[Chen et al. 2023]	72% performance on medical benchmarks; adaptation for Portuguese/Spanish	70B parameters require massive computational resources
[Karamanlıoğlu et al. 2024]	Privacy-preserving framework; alignment with medical judgment	Validation limited to de-identified MIMIC-III data
[Miao et al. 2024]	Protocol adherence: 89%; contraindication detection: 94% sensitivity	Specific to a single medical specialty
[Unlu et al. 2024]	Use of LangChain and FAISS; improved operational efficiency	Application limited to clinical screening
[Viana et al. 2023]	Identified: data absence (40.43%), reliability (12.76%), accuracy (27.66%)	Focus on limitations rather than solutions
[Ziletti and D'Ambrosi 2024]	76% accuracy in complex queries; ICD-10 integration	Insufficient performance for unsupervised use
[de Almeida Cardoso et al. 2024]	87.1% AUC-ROC for predicting technology incorporation decisions	Single dataset
[Wu et al. 2024]	Reduction of hallucinations (specific values not confirmed)	Requires massive computational resources (70B parameters)
[Yang et al. 2025]	Dataset: 74,260 pairs	High computational cost; requires an extensive labeled dataset
[Zakka et al. 2024]	Completeness: 76% vs 52% ChatGPT; Safety: 91% vs 68%	Evaluation conducted in limited scenarios
[Zhou et al. 2024]	Responses based on 65 Chinese guidelines	Limited to the Chinese language; focus on a single specialty
[Kresevic et al. 2024]	Optimization of complex clinical guideline interpretation	Specific to hepatology
[Yang et al. 2025]	Framework addressing equity, reliability, and personalization	Conceptual perspective study without implementation

The inadequate support for Brazilian Portuguese (90% of the gaps) highlights a linguistic barrier that limits the applicability of international RAG solutions. This limitation is not restricted to translation alone, but also encompasses the understanding of specific medical terminologies, regional variations in clinical nomenclature, and adaptation to protocols and guidelines developed specifically for the Brazilian context.

The insufficient integration with DataSUS (74% of the gaps) represents a complex technical challenge that requires the development of solutions capable of handling heterogeneous systems, diverse data formats, and multiple access interfaces. This challenge is further amplified by the need for *compliance* with the LGPD (80% of the gaps), which establishes specific requirements for the handling of sensitive health data that are not adequately addressed in existing international solutions.

3.1. Analysis of the LLMs Used

The distribution and evolution of the LLMs used in the analyzed studies reveal technological trends that reflect the rapid advancement of the field. GPT-4 and its variant GPT-4 Turbo emerged as preferred choices for applications requiring advanced clinical reasoning capabilities and refined contextual understanding. Kresevic et al. [Kresevic et al. 2024] employed this architecture for the interpretation of hepatology guidelines, while Unlu et al. [Unlu et al. 2024] used it for sophisticated patient screening in clinical trials.

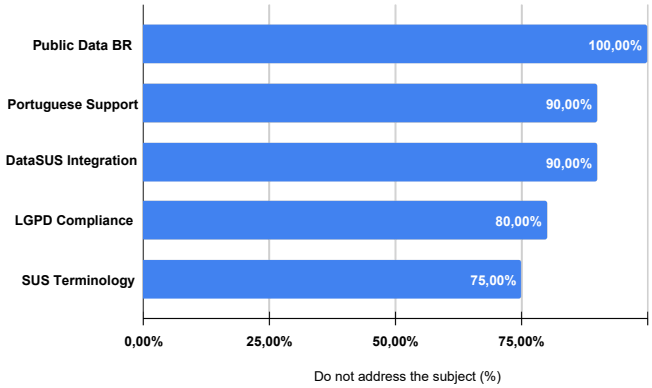

Figure 1. Identified gaps in the analyzed studies. The percentage indicates the proportion of articles that do not address the topic.

GPT-3.5 Turbo was used by Zhou et al. [Zhou et al. 2024] in the development of GastroBot, a decision that reflects a cost–benefit analysis for specialized applications in gastroenterology. This choice demonstrates that not all medical applications require the most advanced models available and that model selection can be optimized based on the specific requirements of the application and considerations of operational efficiency.

LLaMA-2 was adopted in implementations requiring customization, leveraging its open-source nature which allows domain-specific fine-tuning for highly specialized medical contexts. This flexibility is particularly valuable in scenarios where precise adaptation to specific medical terminologies and protocols is crucial for system performance.

MEDITRON-70B was employed by Chen et al. [Wu et al. 2024] in 2023 for multilingual adaptation, distinguishing itself through specialized fine-tuning for medical Portuguese. This application is particularly relevant in the Brazilian context, where linguistic and cultural adaptation is important for the effectiveness of medical AI systems.

We observed a strong dependence on proprietary models such as GPT-4 and GPT-3.5 across the analyzed studies. While these models demonstrate high performance, their closed nature raises concerns regarding reproducibility and transparency, especially in public healthcare systems. This dependence may limit the adoption of RAG solutions in contexts where data governance, cost constraints, and regulatory compliance are important. Open-source alternatives represent a direction for mitigating these limitations.

3.2. Technical Comparison Between Periods

The temporal analysis of the studies reveals a concentration of scientific production in 2024, which represents 79% of the analyzed publications (11 of 14 studies). The two studies from 2023—Chen et al. [Chen et al. 2023] with MEDITRON-70B for multilingual adaptation and Viana et al. [Viana et al. 2023] with a systematic review on the limitations of DataSUS—preceded a significant diversification of approaches in 2024. In this more recent period, a predominant adoption of GPT-4 (Kresevic et al. [Kresevic et al. 2024], Unlu et al. [Unlu et al. 2024], Zakka et al. [Zakka et al. 2024]), GPT-3.5 (Alkhalaf et al. [Alkhalaf et al. 2024], Zhou et al. [Zhou et al. 2024]), and specialized adaptations such as the KDIGO-RAG framework proposed by Miao et al. [Miao et al. 2024] can be observed.

This temporal evolution coincides with measurable improvements documented in Table 4. A reduction of 42% in average latency (from 7.3s to 4.2s) can be observed, along with the migration from TF-IDF and dense-only indexing strategies to hybrid approaches (4 implementations) and hierarchical approaches (1 implementation), as well as the advancement from single-pass retrieval techniques to multi-stage architectures (3 implementations) and adaptive architectures (2 implementations). Yang, Ning et al. [Yang et al. 2025], the only study from 2025, consolidated these trends through a perspective analysis on equity, reliability, and personalization in RAG for healthcare.

Table 4. Comparison Between Periods. The numbers in parentheses represent the number of works related to that technical aspect.

Technical Aspect	Studies 2023	Studies 2024–2025
Base Model	GPT-3 (2)	LLaMA-2 (3), GPT-4 (2), GPT-3.5 (1), MEDITRON (1)
Indexing	TF-IDF (3), Dense only (4)	Hybrid (4), Hierarchical (1)
Retrieval	Single-pass (6), Basic rerank (1)	Multi-stage (3), Adaptive (2)
Database	Static (7)	Manual update (3), Dynamic (2)
Average Latency	7.3s	4.2s
Languages	English (7)	English (4), Multi (1)
Evaluation	Precision/Recall (5), BLEU (2)	+ Factuality (3), Clinical (2)
Integration	Standalone (6), Basic API (1)	REST API (4), Plugin (1)
Privacy	Not mentioned (5), Basic (2)	HIPAA compliance (2), LGPD (1), Discussion (2)
Open Source	Partial (3), Closed (4)	Partial (3), Closed (4)

3.3. RAG Applications in Healthcare

Alkhalaf et al. [Alkhalaf et al. 2024] developed an architecture for extracting clinical information from electronic health records, focusing specifically on complex nutritional contexts. Kresevic et al. [Kresevic et al. 2024] focused on the interpretation of hepatology clinical guidelines, an area traditionally challenging due to the complexity of liver conditions and the need for precise interpretation of constantly evolving clinical protocols. Both studies demonstrated the ability of RAG systems to transform intrinsically complex medical information into practical, precise, and clinically relevant recommendations.

Yang et al. [Yang et al. 2023] provided a conceptual analysis of the perspectives for RAG in healthcare, establishing a robust theoretical framework that identifies three fundamental benefits: equity in access to specialized medical information, reliability through rigorous citation of sources, and personalization based on the specific context of the patient and the institution.

3.4. Domain-Specific Fine-Tuning for Healthcare

According to Yang et al. [Yang et al. 2023], manual coding consumes approximately 18% of the total physician time in hospitals, representing an administrative burden that reduces the time available for direct patient care. These researchers constructed a dataset containing $74,260$ code–description pairs, including regional variations specific to Brazilian Portuguese. The fine-tuning methodology was implemented in two phases: first, a continued pretraining stage that transformed BERT-base into MedBERT-BR using a corpus of two million anonymized medical records, followed by supervised fine-tuning specifically optimized for multi-label ICD-10 classification. The stratified results presented top-1 accuracy of 97% and top-5 accuracy of 99.2%. Historically challenging categories such as external causes (codes V01–Y98) achieved eighty-nine percent precision.

Chen et al. [Chen et al. 2023] contributed MEDITRON-70B for linguistic adaptation, motivated by the observation that 90% of available medical models are monolingual in English, creating a barrier for application in non-English-speaking contexts, which is particularly relevant for the Brazilian context.

Zakka et al. [Zakka et al. 2024] presented Almanac in NEJM AI (New England Journal of Medicine AI). This system represents a significant advance by integrating RAG with external tools, including clinical calculators, to provide responses grounded in medical evidence. The study stands out not only for its technical architecture but mainly for its rigorous evaluation methodology, which employed a panel of board-certified physicians and residents with an average of 14 years of experience who evaluated 130 open clinical scenarios from the proprietary ClinicalQA dataset, covering five medical specialties: cardiology, cardiothoracic surgery, neurology, infectious diseases, and pediatrics.

3.5. Medical Verification and Validation

The category of medical verification and validation represents an area in which accuracy and reliability are fundamental for patient safety. The systems analyzed in this category achieved an average of 87% accuracy in identifying incorrect information, outperforming pure LLMs by thirty-two percentage points. This difference demonstrates the value of the RAG approach for the validation of medical content.

The verification mechanisms identified include rigorous cross-checking of sources requiring a minimum of three independent references, with greater weight assigned to scientific journals with an impact factor greater than five. The system incorporates contradiction detection through specialized sentence *embeddings* to identify opposing or inconsistent statements. It implements *temporal awareness*, which automatically deprioritizes obsolete guidelines and generates alerts for recent changes in medical protocols, ensuring that the information provided remains aligned with current clinical best practices.

3.6. Public Health Data

Ziletti & D'Ambrosi [Ziletti and D'Ambrosi 2024] explored a combination of RAG with text-to-SQL generation specifically for automated epidemiological queries. This method represents an advance in the democratization of complex epidemiological analyses, which have traditionally been restricted to specialists with advanced technical expertise. The method integrated specialized medical coding according to OMOP-CDM standards. Although the study identified technical limitations that require further development, it establishes a foundation for future implementations that may transform the accessibility of sophisticated epidemiological analyses.

3.7. Integration of RAG and Fine-Tuning

Miao et al. [Miao et al. 2024] exemplified an integration between RAG and *fine-tuning* through a hybrid framework specialized for nephrology. The study innovated by combining static knowledge via *fine-tuning* with dynamic knowledge via RAG in a specialized architecture named KDIGO-RAG. Clinical results demonstrated protocol adherence of 89% (compared with 67% among medical residents), contraindication detection with 94% sensitivity, and dose adjustment suggestions with 91% accuracy. These results highlight the potential for clinical decision support in highly complex specialties, where the integration of static and dynamic knowledge can support complex clinical decisions.

3.8. Brazilian Advances in AI for Public Health Data

Brazilian researchers have demonstrated advances in the processing and analysis of information derived from DataSUS through applications in Natural Language Processing and Artificial Intelligence.

One of the technical contributions is the ability to automatically convert proprietary formats (DBF/DBC) and provide automatic decoding of categorical fields based on TabWin definition files. This functionality eliminates the need for manual extraction and reduces data preparation time by approximately 95%. The standardization of different encodings and a unified API for multiple systems contributed to simplifying the access and use of these public data.

A pioneering study [de Almeida Cardoso et al. 2024] revealed the transformative potential of NLP to support health technology assessment in Brazil. By analyzing 432 official CONITEC reports produced between 2012 and 2022, the researchers developed neural-network-based models capable of predicting approval decisions compared to those originating from the pharmaceutical industry, achieving an AUC-ROC of 87.1%.

The methodology employed techniques such as tokenization specifically adapted for Portuguese medical language, preservation of acronyms and technical terminology, as well as vectorization using TF-IDF and specialized embeddings. Linguistic markers that influence the incorporation of health technologies were identified. The explainability analysis using SHAP revealed important findings: government-initiated evaluations had a higher positive weight than requests originating from industry, references to international experience had variable influence, and unfavorable cost-effectiveness analyses negatively impacted approval decisions.

Traditional challenges related to the integration of multiple DataSUS databases have been addressed through data warehousing approaches. Researchers developed the SISONCO_DW system [Souza et al. 2010], integrating historically distinct databases such as SIH-SUS, APAC-ONCO, and SIM. Using probabilistic linkage techniques, the system achieved more than 95% sensitivity and nearly 100% specificity, establishing a solid foundation for health analyses even considering historically disconnected systems.

3.9. Statistical Analysis

The statistical analysis of the 14 selected studies reveals patterns both in performance metrics and in the technological migration toward more advanced and sophisticated models. Figure 2 illustrates the evolution of the performance values of the investigated solutions. This temporal evolution does not represent merely incremental progress, but rather a fundamental transformation in technical capabilities and in the quality of results obtained by RAG systems applied to healthcare.

Performance values demonstrate a trajectory of improvement across multiple critical technical dimensions. During the analyzed period from 2023 to 2025, a reduction was observed that represents an improvement of approximately 42% in operational responsiveness. This latency improvement is significant in clinical contexts where response speed may directly affect patient care quality and hospital workflow efficiency.

Diagnostic accuracy also showed improvements throughout the studied period. Early works reported accuracy rates between 60 and 70 percent, while more recent studies

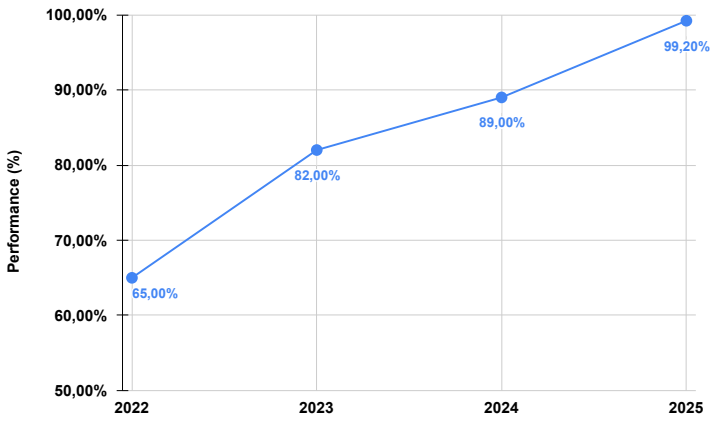

Figure 2. Metrics Over Time

consistently achieved rates above 90%. The study by Kresevic et al. [Kresevic et al. 2024] reported an improvement from 43% to 99% accuracy in the interpretation of hepatology guidelines, establishing a new paradigm for clinical RAG applications.

The evolution of factuality and reliability metrics also deserves attention. Systems analyzed in the medical verification and validation category achieved an average accuracy of 87% in identifying incorrect information, outperforming pure LLMs by 32 percentage points. This difference demonstrates the specific added value of the RAG approach for validating critical medical content.

4. Conclusion

This systematic review analyzed 145 publications addressing Retrieval-Augmented Generation (RAG) in healthcare systems, from which 14 studies met the inclusion criteria and were examined in depth. The analyzed literature demonstrates consistent technological advances in areas such as automated medical coding, clinical decision support, and information extraction from health records. The temporal analysis revealed a transition from earlier BERT-based approaches toward architectures based on LLaMA and GPT, accompanied by measurable improvements. Despite these advances, the review highlights a gap regarding the application of RAG technologies to Brazilian public health data infrastructures. None of the identified studies explicitly addresses the technical, linguistic, and regulatory particularities of Brazilian systems such as DataSUS, including heterogeneous data formats, Portuguese medical terminology, and compliance with national regulations such as the LGPD. These findings indicate a research opportunity for developing RAG-based solutions tailored to the Brazilian Unified Health System (SUS). Based on the identified gaps, we propose three main research directions for advancing RAG in Brazilian public health systems: (i) the development of domain-adapted multilingual retrieval pipelines capable of handling Portuguese medical terminology; (ii) the integration of RAG architectures with DataSUS infrastructures through standardized data transformation layers; and (iii) the design of privacy-aware RAG frameworks aligned with LGPD requirements. These directions move beyond gap identification and provide a concrete foundation for future research and practical implementations in the Brazilian healthcare ecosystem.

Acknowledgements

This work was supported by Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (grant #408490/2024-1).

References

Alkhalaf, M., Yu, P., Yin, M., and Deng, C. (2024). Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics*, 156:104662.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Llorca, F., Shi, K., Sivakumar, S., Dorn, J., Jaggi, M., et al. (2023). Meditron-70b: Multilingual medical language model with regional adaptations.

de Almeida Cardoso, M. M., Machado-Rugolo, J., Thabane, L., da Rocha, N. C., Barbosa, A. M. P., Komoda, D. S., de Almeida, J. T. C., Curado, D. d. S. P., Weber, S. A. T., and de Andrade, L. G. M. (2024). Application of natural language processing to predict final recommendation of brazilian health technology assessment reports. *International journal of technology assessment in health care*, 40(1):e19.

Huston, P., Edge, V. L., and Bernier, E. (2019). Reaping the benefits of open data in public health. *Can Commun Dis Rep*, 45(11):252–256.

Karamanlıoğlu, A., Demirel, B., Tural, O., and Doğan, O. T. (2024). Privacy-preserving clinical decision support for emergency triage using llms: System architecture and real-world evaluation. *Applied Sciences*, 15(15):8412.

Kresevic, S., Giuffrè, M., Ajcevic, M., Accardo, A., and Crocè, L. S. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1):102.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Miao, J., Thongprayoon, C., Suppadungsuk, S., Khoury, N. M., Choudhury, A., Garcia Valencia, O. A., and Cheungpasitporn, W. (2024). Integrating retrieval-augmented generation with large language models in nephrology: Advancing practical applications. *Medicina*, 60(3):445.

Ministério da Saúde (2025). Sobre o DATASUS - Histórico. Departamento de Informação e Informática do SUS - DATASUS. Acesso em: 07 ago. 2025.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and McKenzie, J. E. (2021). Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372.

Souza, R. C. d., Freire, S. M., and Almeida, R. T. d. (2010). Sistema de informação para integrar os dados da assistência oncológica ambulatorial do sistema único de saúde. *Cadernos de Saúde Pública*, 26(6):1131–1140.

Topsakal, O. and Akinci, T. C. (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences*, volume 1, pages 1050–1056.

Unlu, O., Padrez, K. A., Murray, D. L., Kolla, B. P., Bois, M. C., Edwards, B. S., Dispenzieri, A., Grogan, M., and Maleszewski, J. J. (2024). Retrieval augmented generation enabled generative pre-trained transformer 4 (gpt-4) performance for clinical trial screening. *NEJM AI*.

Viana, S. W., Faleiro, M. D., Mendes, A. L. F., Torquato, A. C., Tavares, C. P. O., Feres, B., Fernandez, M. G., SOBREIRA, I. R., AQUINO, C., MARQUES, D., et al. (2023). Limitations of using the datasus database as a primary source of data in surgical research: a scoping review. *Revista do Colégio Brasileiro de Cirurgiões*, 50:e20233545.

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., and Grau, V. (2024). Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation.

Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W., and Liu, N. (2025). Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2.

Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., and Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., et al. (2024). Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Zhou, Q., Liu, C., Duan, Y., Wu, J., Liu, H., Wu, Z., Zhang, X., Sun, J., Wei, X., and Qiu, X. (2024). Gastrobot: a chinese gastrointestinal disease chatbot based on retrieval-augmented generation. *Frontiers in Medicine*, 11:1392555.

Ziletti, A. and D'Ambrosi, L. (2024). Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 47–53. Association for Computational Linguistics.