

A Systematic Review of Transformer-Based Language Models for Clinical Decision Support in EMR-Based Systems: Implications for Epilepsy Care in Low-Resource Settings

***Kingangi Gitau¹, John M. Kihoro² & James Obuhuma Imende³**

¹Department of Computer Science and Information Technology in the School of Computing and Mathematics, The Co-operative University of Kenya

²Department of Mathematical Sciences, School of Computing and Mathematics, The Co-operative University of Kenya

³Department of Computer Science, School of Computing and Informatics, Maseno University

***Corresponding Email: jongitau@gmail.com**

Abstract

Clinical decision-making in low- and middle-income country (LMIC) healthcare settings is often constrained by the inability to efficiently synthesize longitudinal patient data stored within Electronic Medical Record (EMR) systems. While traditional clinical decision support systems (CDSS) rely on rule-based alerts, they lack the contextual reasoning required to interpret complex clinical notes for chronic conditions like epilepsy. Recent advances in transformer-based language models, including BERT-family models, and generative large language models (LLMs) offer promising approaches for clinician-initiated medical question-answering (QA). This study presents a systematic review evaluating the architectures, applications, and EMR integration of transformer- and LLM-based QA and related clinical decision-support systems, with implications for epilepsy management in resource-limited environments. A structured search of primary biomedical and engineering academic databases was conducted to identify peer reviewed studies published up to 23 June 2026 related to CDSS, EMR-based decision support, and LLM applications in healthcare. The synthesis indicates that transformer-based and generative language models demonstrate promising performance in controlled clinical QA evaluations, but evidence for their practical deployment as clinician-facing tools integrated with active EMR databases remains limited. Evidence is also limited concerning the use of these approaches to interpret complex, unstructured timelines required for longitudinal chronic disease management in LMIC infrastructure. LLM-based clinical QA represents a viable entry point for responsible machine learning in resource-constrained settings, but future research must shift from benchmark validation to assessing real-world workflow compatibility and EMR integration constraints.

Keywords: Transformer-Based Language Models; Large Language Models; Clinical Question-Answering; Clinical Decision Support Systems; Electronic Medical Records

Introduction

Data-driven decision-making is increasingly important in healthcare, where clinicians must interpret patient information collected across multiple encounters. Transformer-based language models have expanded the potential for natural language processing (NLP) to support the retrieval, interpretation, and synthesis of clinical information. These approaches include encoder-based models such as Bidirectional Encoder Representations from Transformers (BERT) and its biomedical and clinical variants, as well as generative large language models (LLMs) such as GPT-family models. Although these technologies have demonstrated capabilities in medical language understanding and clinical question answering, performance on general medical benchmarks does not necessarily establish their reliability when processing patient-specific electronic medical record (EMR) or electronic health record (EHR) data (Singhal et al., 2023; Sutton et al., 2020).

EMR and EHR systems contain structured information, including diagnoses, medications, and observations, alongside unstructured clinical notes and discharge summaries. Language-model-based clinical question answering may enable



clinicians to retrieve and synthesize information from these records through natural-language queries. For example, a clinician managing a patient with epilepsy may need to identify previously documented antiseizure medications, medication changes, adverse reactions, or treatment responses across multiple encounters. A patient-specific QA system could assist by locating and summarizing these documented events, provided that the information is available in the record and the output is evaluated for accuracy, completeness, and clinical relevance. Such systems are intended to support information access and clinical review rather than replace clinician judgment.

The potential value of these approaches is particularly relevant to chronic conditions such as epilepsy, where care may require consideration of seizure history, medication changes, treatment response, and information recorded over time. However, many language-model evaluations use general medical knowledge, examination questions, or curated datasets rather than patient-derived clinical records. These evaluations provide useful evidence of language-model capabilities but may not capture the challenges of retrieving accurate information from longitudinal EMR/EHR data, including incomplete documentation, conflicting entries, temporal relationships, and clinically important omissions. Research on clinical decision support also emphasizes the need to assess system benefits alongside safety, workflow, and implementation considerations (Sutton et al., 2020).

These issues are important in low- and middle-income countries (LMICs), including Kenya, where the applicability of clinical AI depends not only on model performance but also on the availability and quality of digital records, interoperability, infrastructure, governance, and integration into clinical workflows. Although language models may offer new ways to access patient-specific information, their feasibility and safety in routine clinical environments cannot be assumed from benchmark performance alone. Evidence from real-world EMR/EHR settings is therefore needed to understand how these systems are developed, evaluated, and implemented.

This systematic review identifies, appraises, and synthesizes evidence on transformer-based and generative language-model systems that use patient-derived EMR/EHR information for clinical question answering, retrieval, summarization, information synthesis, and related decision-support tasks. It examines model architectures, data sources, evaluation approaches, clinical involvement, safety, and implementation. Particular attention is given to the extent to which the available evidence informs the potential adaptation of clinical QA systems for epilepsy care in Kenya and other resource-constrained settings. The review does not assume epilepsy-specific effectiveness; rather, it identifies the evidence available and the research gaps that must be addressed before such systems can be considered for clinical use.

Methods

Study Design and Search Strategy

This study was conducted as a systematic literature review to identify, appraise, and synthesize evidence on transformer-based and generative language-model systems that use patient-derived electronic medical record (EMR) or electronic health record (EHR) information for clinical question answering (QA), information retrieval, summarization, information synthesis, clinical assistance, and clinical decision support (CDS). The review included encoder-based transformer models, such as BERT-family architectures, generative large language models (LLMs), including GPT- and LLaMA-family models, and retrieval-augmented or other hybrid architectures when evaluated using patient-derived clinical information.

The review was reported with reference to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guideline (Page et al., 2021). The review was guided by the PICOS framework, with particular emphasis on language-model interventions evaluated using patient-specific clinical information and their reported



clinical, technical, and implementation outcomes. The synthesis was designed to characterize the available evidence rather than to establish the superiority of a particular model family or clinical task.

PubMed/MEDLINE and IEEE Xplore were searched on 23 June 2026 for English-language, peer-reviewed studies published between January 2020 and 23 June 2026. The search strategy combined three main concept groups: transformer-based or generative language models; EMR/EHR systems or patient-record information; and clinical question answering, retrieval, summarization, clinical assistance, or decision support. Broader application terms were included to identify systems that process patient-specific clinical information even when their authors did not explicitly describe them as clinical QA systems.

In PubMed/MEDLINE, the search used title and abstract fields, represented by the [tiab] field tag, and applied publication-date, English-language, and abstract-availability filters. In IEEE Xplore, the search combined corresponding language-model, clinical-record, and clinical-application terms within document titles and abstracts. Boolean operators were used to combine related terms within each concept group and to connect the three concept groups. The complete database-specific search strings are provided below to support reproducibility.

The PubMed/MEDLINE search strategy was constructed using three concept groups combined with the Boolean operator AND. The first group identified transformer-based and generative language models, including BERT, BioBERT, ClinicalBERT, GPT, LLaMA, and related terms. The second group identified electronic medical records, electronic health records, clinical notes, and patient records. The third group identified clinical applications, including clinical decision support, question answering, information retrieval, retrieval-augmented generation (RAG), and clinical copilots.

Within each concept group, alternative search terms were connected using OR, allowing records containing any of the listed terms to be retrieved. The AND operator then required records to match all three concept groups, narrowing the results to studies addressing language models, patient-record information, and relevant clinical applications. The [tiab] field tag restricted these terms to article titles and abstracts. Additional filters limited retrieval to English-language records with abstracts and publication dates within the specified period. This combination was intended to identify studies evaluating language models using patient-derived clinical information rather than studies concerned exclusively with general medical knowledge or examination questions.

PubMed/MEDLINE search string

```
(
  (
    "large language model"[tiab] OR
    LLM[tiab] OR
    BERT[tiab] OR
    BioBERT[tiab] OR
    ClinicalBERT[tiab] OR
    GPT[tiab] OR
    "Med-PaLM"[tiab] OR
    LLaMA[tiab] OR
    transformer[tiab]
  )
  AND
  (
    "electronic medical record"[tiab] OR
    "electronic health record"[tiab] OR
    EMR[tiab] OR
```



```
EHR[tiab] OR
"clinical notes"[tiab] OR
"patient records"[tiab]
)
AND
(
"clinical decision support"[tiab] OR
"question answering"[tiab] OR
"clinical question answering"[tiab] OR
"clinical information retrieval"[tiab] OR
"retrieval augmented generation"[tiab] OR
RAG[tiab] OR
"clinical copilot"[tiab]
)
)
AND
("2020/01/01"[Date - Publication] : "2026/06/23"[Date - Publication])
AND English[Language]
AND hasabstract[text]
```

IEEE Xplore search string

```
(
(
"Document Title": "large language model" OR
"Abstract": "large language model" OR
"Document Title": LLM OR
"Abstract": LLM OR
"Document Title": BERT OR
"Abstract": BERT OR
"Document Title": BioBERT OR
"Abstract": BioBERT OR
"Document Title": ClinicalBERT OR
"Abstract": ClinicalBERT OR
"Document Title": GPT OR
"Abstract": GPT OR
"Document Title": transformer OR
"Abstract": transformer OR
"Document Title": "Med-PaLM" OR
"Abstract": "Med-PaLM" OR
"Document Title": "LLaMA" OR
"Abstract": "LLaMA"
)
)
AND
(
"Document Title": EMR OR
"Abstract": EMR OR
"Document Title": EHR OR
"Abstract": EHR OR
"Document Title": "electronic medical record" OR
"Abstract": "electronic medical record" OR
"Document Title": "electronic health record" OR
"Abstract": "electronic health record" OR
)
```



```
"Document Title": "clinical notes" OR
"Abstract": "clinical notes" OR
"Document Title": "patient records" OR
"Abstract": "patient records"
)
AND
(
"Document Title": "question answering" OR
"Abstract": "question answering" OR
"Document Title": "clinical information retrieval" OR
"Abstract": "clinical information retrieval" OR
"Document Title": "clinical assistant" OR
"Abstract": "clinical assistant" OR
"Document Title": "clinical copilot" OR
"Abstract": "clinical copilot" OR
"Document Title": "conversational AI" OR
"Abstract": "conversational AI" OR
"Document Title": "retrieval augmented generation" OR
"Abstract": "retrieval augmented generation" OR
"Document Title": "RAG" OR
"Abstract": "RAG" OR
"Document Title": "clinical decision support" OR
"Abstract": "clinical decision support" OR
"Document Title": "clinical question answering" OR
"Abstract": "clinical question answering"
)
)
```

The IEEE Xplore search strategy followed the same three-concept structure as the PubMed/MEDLINE search, but used the database's field-specific search syntax. The expressions "Document Title" and "Abstract" restricted individual terms to the document title and abstract fields. Within each concept group, OR connected alternative terms, while AND linked the three groups: language models, clinical records, and clinical applications. This approach aimed to identify engineering and computer-science studies evaluating transformer-based or generative language models for clinical question answering, retrieval, summarization, or decision support involving patient derived records. The two database strategies used equivalent concepts while accommodating differences in database search syntax.

Records retrieved from both databases were consolidated into a master screening dataset, and duplicate records were removed before title and abstract screening. Reference lists of eligible studies were also examined manually to identify potentially relevant primary studies that might not have been retrieved through the database searches. The number of records retrieved from each database and the subsequent study-selection process are reported in the Results section and illustrated in the PRISMA 2020 flow diagram.

Eligibility Criteria

Studies were eligible if they were peer-reviewed primary research; evaluated a transformer-based or generative language model; used patient-derived EMR/EHR information, clinical notes, discharge summaries, patient records, or structured or unstructured clinical data; and evaluated question answering, retrieval, summarization, information synthesis, clinical assistance, or clinical decision support. The eligibility criteria were defined a priori and reported in accordance with PRISMA 2020 guidance for specifying study inclusion and exclusion criteria (Page et al., 2021).

Studies using only general medical knowledge, examination questions, biomedical literature, or non-patient QA datasets were excluded, as were unrelated NLP tasks, reviews, editorials, protocols, commentaries, and preprints. Mixed-source studies were retained when the patient-derived component was independently evaluated; findings from external medical QA datasets were not treated as evidence of EMR-based performance (Page et al., 2021).

Study Selection and Eligibility Audit

A predefined protocol was developed before screening but was not prospectively registered. One reviewer conducted title/abstract screening, full-text assessment, extraction, and appraisal. Independent duplicate screening and inter-rater agreement were not performed (Page et al., 2021). Following an eligibility audit of the 22 initially included studies, 19 were retained, two were classified as Mixed/Retain, and one was excluded because its reported evaluation used synthetic rather than real patient-derived clinical records. The final synthesis therefore included 21 studies (Page et al., 2021).

Data Extraction and Quality Appraisal

Data were extracted on study characteristics, clinical setting, patient-derived data, model architecture, retrieval/RAG, clinical task, evaluation, clinician involvement, implementation, safety, and limitations. Appraisal was adapted to study design and considered methodological validity, data provenance, evaluation, clinical applicability, safety, and reporting transparency. No numerical quality score was assigned (Page et al., 2021; Viswanathan et al., 2018).

Data Synthesis

Because of substantial heterogeneity in clinical domains, datasets, architectures, tasks, and outcome measures, meta-analysis was not appropriate. A structured descriptive and thematic synthesis was performed, focusing on model capabilities, EMR-based information processing, retrieval and QA, clinical applicability, safety, and implementation (Page et al., 2021; Popay et al., 2006).

Results

Study Selection and Eligibility Audit

The systematic search identified 235 records: 171 from PubMed/MEDLINE and 64 from IEEE Xplore. After removal of seven duplicates, 228 records underwent title and abstract screening. Of these, 203 were excluded and 25 proceeded to full-text assessment. Three reports were excluded at the original full-text stage, leaving 22 studies for the eligibility audit.

The audit classified 19 studies as Retain, two as Mixed/Retain (Jeon and Kim, 2025; Guo et al., 2025), and one as Exclude (Jamil et al., 2026). The final qualitative synthesis therefore comprised 21 studies. The mixed-source studies were retained because their patient-derived EHR components were independently evaluated; Jamil et al. was excluded because its reported evaluation relied on synthetic patient-interaction scenarios rather than real patient-derived clinical records.

Characteristics of the Evidence Base

The 21 studies covered general clinical care, oncology, emergency medicine, diabetes, intensive care, radiology, and surgery. Patient-derived data included clinical notes, discharge summaries, structured EHR data, longitudinal records, and combinations of structured and unstructured information. Most studies were technical or experimental evaluations; fewer included clinician assessment, randomized evaluation, implementation assessment, or real-world clinical deployment.

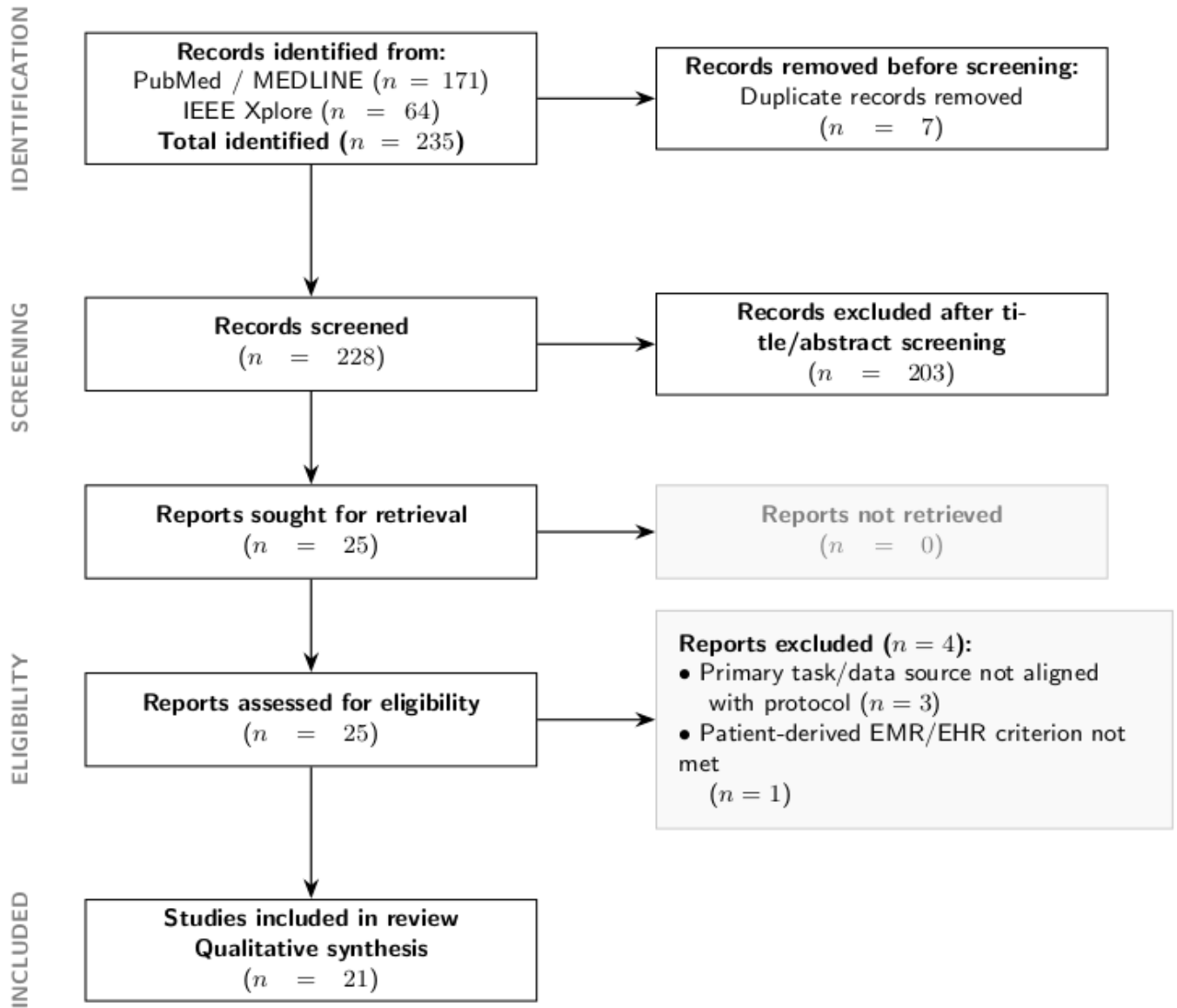


Figure 1: PRISMA 2020 flow diagram of the study selection process

The PRISMA flow diagram illustrates how the literature was systematically screened to identify studies relevant to transformer-based language models for clinical decision support using patient-derived EMR/EHR data. Of the 235 records identified from PubMed/MEDLINE and IEEE Xplore, 7 duplicates were removed, leaving 228 records for screening. Following title and abstract screening and full-text assessment, 22 studies underwent an eligibility audit, resulting in 21 studies being retained for the final synthesis. This selection process was important because the review focused on evidence from patient-derived clinical records rather than general medical question-answering benchmarks alone. The resulting evidence base was used to examine model capabilities, clinical reliability, integration into EMR workflows, and implementation considerations in low-resource settings. It also helped identify the limited evidence for epilepsy-specific clinical decision support, which provides the rationale for investigating a BERT-assisted EMR-based clinical question-answering system for epilepsy care in Kenya.

Table 1: Characteristics of Included Studies

Study	Clinical domain / patient-derived data	Model / approach	Clinical task	Evaluation	Clinical evaluation	Key limitation
Klang et al. (2024)	General care; EHR clinical notes	Multiple LLMs	QA and information extraction	>300,000 experiments; accuracy, failures, cost	Not reported	Performance and computational cost varied with task complexity
Ahn et al. (2025)	Emergency medicine; patient clinical records	Evidence-based generative LLM	Diagnostic CDS	Randomized controlled trial	Emergency clinicians	Detailed safety outcomes not fully reported
Fleming et al. (2024)	General care; 276 longitudinal EHRs	Six general-domain LLMs	EHR retrieval and instruction following	Response error rates; context sensitivity	15 clinicians	Errors increased with limited clinical context
Griot et al. (2025)	General care; EHR and clinical documents	LLM-EHR integration framework	Clinical information access and workflow support	Implementation assessment	Workflow-focused	Quantitative clinical evaluation limited
Marin et al. (2025)	General care; clinical notes and structured EHR data	Generative LLM + RAG	Clinical QA	3,500 QA items; F1 = 0.82	Not reported	Real-world safety and effectiveness not established
Wen et al. (2020)	General care; clinical notes	BERT/ClinBERT	Patient-specific clinical QA	13,545 QA items; exact/partial match	Physician-supported error analysis	Limited deeper reasoning
Lu et al. (2025)	Oncology; >6 million notes from 209,135 patients	Sentence-transformer + RAG	Clinical QA/retrieval	50 questions; NDCG, precision, recall	Not reported	Downstream clinical effectiveness not established
Jeon & Kim (2025)	General medicine; EHRNoteQA/MIMIC-IV plus external QA	GPT-4o-mini, GPT-3.5, o1-mini, Gemini	Medical and EHR QA	Comparative benchmark evaluation	Not reported	Mixed-source evaluation
Agweyu et al. (2026)	African primary care; 1,469 encounters	EMR-integrated LLM CDS	CDS and safety	Hallucination, guideline concordance and harmful-	Physician panel	Hallucinations and potentially harmful recommendations reported



				output assessment		
Li et al. (2026)	Spine surgery; 200 real-world cases	Multiple LLMs + RAG	Treatment recommendat ion/CDS	11-domain clinical/techni cal rubric	5 spine surgeons	Prospective clinical validation not reported
Wang et al. (2025) – diabetes	Diabetes; EHR, medication data and guidelines	BioGPT + RAG	Drug recommendat ion and dosage	PR-AUC, F1, faithfulness	Physician validation	Prospective safety evaluation limited
Jani et al. (2026)	Thoracic oncology; 2,433 patients	GPT-family models + RAG	Oncologic history summarizatio n	50-patient evaluation; accuracy, completeness, time and cost	Radiologist/re sident benchmark	Retrospective evaluation
Thio et al. (2026)	General care; MIMIC-IV, 10 patients	Graph-RAG	Patient-level QA	Factual recall and expert quality	Expert assessment	Very small evaluation set
Wang et al. (2025) – X- RAG	General care; EHR/EMR	GPT-4o + X- RAG	Medical information summarizatio n	Extraction and summarizatio n evaluation	Not reported	Heterogeneous data integration challenge
He et al. (2026)	Diabetes; de- identified outpatient EHRs	Llama, Qwen, GLM + RAG	Personalized treatment support	BLEU and ROUGE	Clinical plausibility assessment	Prospective clinical validation required
Mondal et al. (2025)	Diabetes; 50 real patient records	GPT-4	Management- plan generation	Completeness , necessity, dosage accuracy and errors	Physician comparison	Small sample
Moon et al. (2022)	General care; clinical notes	BERT-based transformer	Clinical QA	QA accuracy and transfer performance	Not reported	Performance affected by masking strategy
Patil et al. (2024)	General care; clinical notes/discharge summaries	Mistral/Llam a + QLoRA	QA, summarizatio n and clinical reasoning	Multiple NLP/reasonin g tasks	Not reported	External generalisability limited

Muralidharan et al. (2025)	General care; FHIR-standardized EHR	Mixtral + RAG + knowledge graph	History summarization and QA	Prototype evaluation	Not reported	Integration and interoperability challenges
Pattnayak et al. (2025)	General care; emrQA clinical notes	ClinicalBERT	QA and answer categorization	F1 and accuracy	Not reported	External generalizability not established
Guo et al. (2025)	ICU; real-world EHR plus external QA datasets	Knowledge graph + RAG + LLM	ICU CDS and reasoning	Comparative evaluation	Expert comparison	Mixed-source evaluation

Table 1 above summarizes the characteristics of the 21 included studies, including clinical setting, patient-derived data, model approach, clinical task, evaluation, clinical assessment, and key limitations. NR indicates information not reported in the primary study.

Model and Task Findings

Both BERT-family encoder models and generative LLMs were represented. Applications included patient-specific QA, retrieval, summarization, information synthesis, recommendation, and CDS. RAG, knowledge graphs, and other retrieval mechanisms were increasingly used with generative models. Direct comparisons under equivalent patient-record conditions were insufficient to establish superiority of one model family or architecture.

Clinical Applicability and Safety

Technical performance generally received greater attention than clinical outcomes. Common limitations included limited external validation, dataset or retrieval constraints, incomplete reporting, and reliance on accuracy or retrieval metrics without corresponding workflow or patient-level evaluation. Safety assessments varied and included clinician review, factuality, guideline concordance, and harmful-output analysis.

Translation to Real-World EMR Use

Evidence for routine EMR deployment was limited. Fleming et al. reported response error rates of 35–68% when evaluating longitudinal EHR information, illustrating the sensitivity of clinical QA to available context. Agweyu et al. (2026). provided direct evidence from 16 Kenyan primary-care clinics, including assessment of hallucinations and potentially harmful recommendations. However, evidence remained insufficient to establish minimum infrastructure requirements, long-term clinical effectiveness, or safety for epilepsy care.

Implications for Low-Resource Settings and Epilepsy Care

The findings suggest that epilepsy-oriented EMR QA should support retrieval and synthesis of longitudinal information such as seizure history, medication changes, treatment response, and clinical notes. This is a translational implication rather than evidence of demonstrated epilepsy-specific effectiveness. Further evaluation should address retrieval accuracy, temporal consistency, omissions, safety, clinician oversight, data governance, and workflow usability.



Results/Discussion

Clinical Reliability and Safety

Safety was less consistently evaluated than technical performance. Reported concerns included hallucinations, retrieval failures, factual errors, and potentially harmful recommendations. Agweyu et al. (2026) reported high guideline concordance alongside hallucinations and potentially harmful outputs, demonstrating that guideline concordance alone does not establish clinical safety. Although 99% of the evaluated recommendations aligned with local guidelines, 115 of 1,469 responses (7.8%) contained potentially harmful recommendations, and 67 of these appeared in the final clinical documentation.

Other studies used expert assessment and faithfulness measures, but safety outcomes were heterogeneous. Fleming et al. (2024) reported response error rates of 35–68% when evaluating longitudinal EHR information, demonstrating the challenges of maintaining factual reliability when processing patient records. Li et al. (2026) incorporated evidence citation and expert assessment, while Wang et al. (2025) reported a faithfulness measure alongside physician validation. However, these evaluation approaches assessed different aspects of reliability and did not provide a common measure of clinical safety. Important gaps included omission errors, temporal inconsistencies, uncertainty estimation, bias, privacy, and patient-level outcomes. Evaluation should therefore extend beyond accuracy to include factuality, completeness, retrieval reliability, inappropriate recommendations, traceability, and clinical safety (Fleming et al., 2024; Li et al., 2026; Wang et al., 2025; Agweyu et al., 2026).

Low-Resource Settings and Epilepsy Care

Evidence from low-resource settings remains limited. The Kenyan study by Agweyu et al. (2026) demonstrates that EMR-integrated language-model decision support can operate in routine primary care. However, computational requirements, connectivity, interoperability, data governance, and workflow integration remain important implementation considerations. The reviewed studies do not establish minimum technical requirements or demonstrate whether on-premise or cloud-based deployment is more suitable, as the available evidence does not provide a sufficient basis for comparing these deployment approaches in low-resource settings..

The implications for epilepsy care are primarily translational. Although longitudinal clinical information processing may be relevant to epilepsy management, no included study established the effectiveness or safety of language-model-based decision support specifically for epilepsy. A locally deployable, BERT-assisted EMR-based clinical QA system may be investigated as a research approach, but its clinical utility, safety, usability, and implementation requirements require independent evaluation.

Recommendations for Future Research

Future research should:

1. Evaluate language-model systems prospectively using real patient-derived EMR/EHR data and routine clinical workflows.
2. Assess longitudinal performance, including temporal ordering, medication-history accuracy, clinically important omissions, and consistency across encounters.
3. Evaluate safety through factuality, retrieval accuracy, inappropriate recommendations, uncertainty, traceability, human oversight, and patient-level outcomes.
4. Conduct external and multi-site validation, while assessing interoperability, computational requirements, data governance, privacy, and workflow integration in low-resource settings.

5. Evaluate epilepsy-specific clinical utility and safety independently rather than extrapolating from general clinical evidence.

Conclusion

This review found that transformer-based and generative language models show potential for patient-specific clinical question answering, information retrieval, summarization, and decision support using EMR/EHR data. However, limited prospective validation, heterogeneous evaluation methods, and unresolved safety concerns constrain conclusions about their clinical effectiveness and readiness for routine use.

Evidence for epilepsy-specific applications in low-resource settings remains insufficient. A BERT-assisted EMR-based clinical QA system for epilepsy care in Kenya represents a potential research direction, requiring independent evaluation of clinical utility, safety, and implementation feasibility.

References

- Agweyu, A., Mwaniki, P., Musau, W., Korom, R., Isaaka, L., Wanyama, C., et al. (2026). Safety of a large language model-based clinical decision support system in African primary healthcare. *Nature Health*. <https://doi.org/10.1038/s44360-026-00082-5>
- Ahn S, Park J, Hur S, Jung KY, Lee JH, Choi SW, Son MH, Kang MJ, Kim YJ, Park H, Cha WC, Yoo J. (2025). *Impact of an Evidence-Based Large Language Model (LLM) Diagnostic Decision Support System: A Randomised Controlled Trial*. *Studies in Health Technology and Informatics*, 329, 273–277. <https://doi.org/10.3233/SHTI250844>.
- Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E., Thapa, R., et al. (2024). MedAlign: A clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20), 22021–22030. <https://doi.org/10.1609/aaai.v38i20.30205>.
- Griot, M., Vanderdonckt, J., & Yuksel, D. (2025). Implementation of large language models in electronic health records. *PLOS Digital Health*, 4(12), e0001141. <https://doi.org/10.1371/journal.pdig.0001141>.
- Guo, J., Gao, Y., Liu, Z., Rao, Y., Hu, J., & Wang, H. (2025). EagleCare: Augmenting ICU monitoring with LLMs and knowledge graph reasoning. *2025 IEEE International Conference on Knowledge Graph (ICKG)*, 130–137. <https://doi.org/10.1109/ICKG66886.2025.00024>.
- He, S., Zhang, Y., & Li, J. (2026). Personalized diabetes treatment support using large language models fine-tuned on electronic health records: Development and evaluation study. *JMIR Formative Research*, 10, e71541. <https://doi.org/10.2196/71541>.
- Jani, K., Mattay, G., Narra, V., Shetty, A., Stancil, G., Bierhals, A., & Tadavarthi, Y. (2026). Zero-shot thoracic oncologic history generation for radiologists using retrieval-augmented large language model pipeline. *Radiology*, 319(3), e251581. <https://doi.org/10.1148/radiol.251581>.
- Jeon, S., & Kim, H.-G. (2025). A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Computers in Biology and Medicine*, 196, 110614. <https://doi.org/10.1016/j.combiomed.2025.110614>.
- Klang, E., Apakama, D., Abbott, E. E., Vaid, A., Lampert, J., Sakhuja, A., Freeman, R., Charney, A. W., Reich, D., et al. (2024). A strategy for cost-effective large language model use at health system-scale. *npj Digital Medicine*, 7, 320. <https://doi.org/10.1038/s41746-024-01315-1>.
- Li, Y., Lv, M., Cheng, L., Xie, Y., Gu, Q., He, H., Chen, Z., Fang, D., & Zhou, X. (2026). Explainable and evidence-linked recommendations for spine surgery via a retrieval-augmented LLM agent. *European Journal of Radiology*, 197, 112734. <https://doi.org/10.1016/j.ejrad.2026.112734>.
- Lu, K.-H., Mehdinia, S., Man, K., Wong, C. W., Mao, A., & Eftekhari, Z. (2025). Enhancing oncology-specific question answering with large language models through fine-tuned embeddings with synthetic data. *JCO Clinical Cancer Informatics*, 9, e2500011. <https://doi.org/10.1200/CCI-25-00011>.
- Marin F, Castaño J, Segalini A, Botta C, Benitez S, Otero C, Luna D. (2025). *Designing a Healthcare Co-Pilot with Generative AI*. *Studies in Health Technology and Informatics*, 329, 618–622. <https://doi.org/10.3233/SHTI250914>
- Mondal, A., Naskar, A., Roy Choudhury, B., Chakraborty, S., Biswas, T., Sinha, S., & Roy, S. (2025). Evaluating the performance and safety of large language models in generating type 2 diabetes mellitus management plans: A comparative study with physicians using real patient records. *Cureus*, 17(3), e80737. <https://doi.org/10.7759/cureus.80737>.
- Moon, S., He, H., & Fan, J. W. (2022). Effects of information masking in the task-specific finetuning of a transformers-based clinical question-answering framework. *20Question IEEE 10th International Conference on Healthcare Informatics (ICHI)*, 36–41.



<https://doi.org/10.1109/ICHIS4592.2022.00017>.

Muralidharan, A., A. S., A. A. R., Raj, A. B., & Manmadhan, S. (2025). MedSyn: AI-driven medical history summarization using EHR data. *2025 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, 1–7. <https://doi.org/10.1109/ACCTHPA65749.2025.11168697>.

Patil, S. S., Rathore, A. S., & Ramteke, M. (2024). QLoRA-based fine-tuning of LLMs on multiple medical reasoning tasks to enhance their comprehension of clinical notes. *2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. <https://doi.org/10.1109/ISCMI63661.2024.10851618>.

Pattnayak, P., Patel, H., Agarwal, A., Panda, S., Kumar, B., & Kumar, T. (2025). Clinical QA 2.0: Multi-task learning for answer extraction and categorization. *2025 IEEE International Conference on Electro Information Technology (eIT)*, 443–449. <https://doi.org/10.1109/eIT64391.2025.11103631>.

Thio, S., Lewis, M., Denaxas, S., & Dobson, R. J. B. (2026). Unlocking electronic health records: A hybrid graph RAG approach to safe clinical AI for patient QA. *Frontiers in Digital Health*, 8, 1780700. <https://doi.org/10.3389/fdgth.2026.1780700>.

Wang, J.-F., Chang, C.-C., Chiang, T.-M., Yeh, T.-C., Cheng, E., Lee, Y.-T., & Chen, H.-I. (2025). An innovative X-RAG technique combined with GPT-4o for summarizing medical information from EHR and EMR to assist doctors in clinical decision-making effectively and efficiently. *Health Informatics Journal*, 31(3), 14604582251381233. <https://doi.org/10.1177/14604582251381233>.

Wang, J.-F., Wei, M.-J., Chiang, T.-M., Yeh, T.-C., Cheng, E., Lee, Y.-T., & Chen, H.-I. (2025). Leveraging electronic health record data and up-to-date clinical guidelines for high-accuracy clinical diabetes drug and dosage recommendation. *Methods of Information in Medicine*, 64(1–02), 1–11. <https://doi.org/10.1055/a-2707-2862>. (PubMed)

Wen, A., Elwazir, M. Y., Moon, S., & Fan, J. W. (2020). Adapting and evaluating a deep learning language model for clinical why-question answering. *JAMIA Open*, 3(1), 16–20. <https://doi.org/10.1093/jamiaopen/ooz072>.

Page, M. J., McKenzie, J. E., 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., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., Agüera y Arcas, B., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>.

Singhal, K., Tu, T., Montgomery, J., et al. (2023). Towards expert-level clinical language understanding with Med-PaLM 2. *arXiv preprint arXiv:2305.09617*.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Medicine*, 3, 17. <https://doi.org/10.1038/s41746-020-0221-y>.