HeCiX: Integrating Knowledge Graphs and Large Language Models for Clinical Research

Prerana Sanjay Kulkarni[1], Muskaan Jain[2], Disha Sheshanarayana[3], and Srinivasan Parthiban[4]

[1] Dept. of Computer Science and Engineering, PES University, Bengaluru, India
prer.kulk@gmail.com
[2] Dept. of Mathematics, IIT Madras, Chennai, India
muskaan40389@gmail.com
[3] Dept. of Computer Science and Engineering, Manipal University, Jaipur, India
disha.229301161@muj.manipal.edu.com
[4] Dept. of Data Science and EngineeringIndian Institute of Science Education and Research, Bhopal, India
parthi@iiserb.ac.in

Abstract. Clinical research faces numerous challenges, including inefficient drug discovery processes, difficulty in identifying key opinion leaders, and the struggle to effectively utilize vast, fragmented biomedical data. This paper introduces HeCiX, a novel approach integrating knowledge graphs from clinical trials data and Hetionet with large language models (LLMs) to address these issues. HeCiX creates a rich, interconnected knowledge base by combining structured clinical trial data with comprehensive biomedical relationships from Hetionet. This knowledge graph augments LLMs, enhancing their ability to answer complex clinical research questions, accelerate drug discovery, and identify domain experts. Our research demonstrates that this synergy of diverse data sources and advanced language models leads to more informed and precise responses in the clinical domain. Experiments showcase HeCiX's superior performance compared to traditional methods, indicating its potential to streamline drug discovery, enhance collaboration with key opinion leaders, and facilitate efficient analysis of medical literature and clinical data. HeCiX represents a significant advancement in leveraging AI for clinical research, promising to accelerate innovation and improve decision-making in the field.

Keywords: Knowledge Graph · Large Language Model · LangChain · Information Retrieval.

1 Introduction

Clinical research faces numerous challenges in the modern era, particularly in areas such as drug discovery, identifying key opinion leaders, and effectively utilizing vast amounts of fragmented biomedical data [**?**]. These challenges often lead to inefficiencies in research processes, delayed discoveries, and missed opportunities for collaboration and innovation [**?**].

The exponential growth of biomedical data, has created a new set of obstacles. Researchers must navigate through diverse data sources, including clinical trials, scientific literature, and complex biological databases, to extract meaningful insights [**?**]. This data deluge, coupled with the intricate nature of biological systems, necessitates novel approaches to data integration and analysis.

To address these challenges, we introduce HeCiX (HEtionet and cLinical trials Information neXus), a comprehensive knowledge graph that integrates data from clinical trials with the biomedical relationships found in Hetionet [**?**]. Furthermore, we present a novel framework that combines HeCiX with large language models (LLMs) to enhance clinical research capabilities.

The primary objectives of this research are:

- To develop HeCiX, a robust knowledge graph integrating clinical trials data with existing biomedical knowledge from Hetionet.
- To create a framework that leverages large language models in conjunction with HeCiX for improved query processing and information retrieval in clinical research.
- To demonstrate the efficacy of our HeCiX-based system in addressing key challenges in clinical research, particularly in drug discovery and identification of key opinion leaders.

By combining the structured, interconnected data of HeCiX with the advanced natural language processing capabilities of LLMs, our system aims to provide researchers with a powerful tool for navigating the complexities of clinical research. This paper presents the architecture of HeCiX, details the implementation of our LLM-enhanced framework, and discusses the results of our experiments demonstrating its effectiveness in various clinical research scenarios.

The rest of this paper is organized as follows: Section 2 provides necessary background on knowledge graphs in biomedical research and the application of LLMs in healthcare. Section **??** describes our custom dataset and the construction of HeCiX. Section **??** details the architecture of our HeCiX-based system and methodology. Sections 5 and **??** present our experimental setup and results, respectively. Finally, we discuss limitations and future work in Section 7 before concluding in Section 9.

2 Background and Related Work

2.1 Overview of Clinical Research Challenges

Clinical research faces significant challenges, including time-consuming and costly drug discovery processes with high failure rates [**?**], difficulties in identifying key opinion leaders (KOLs) [**?**], and obstacles in efficiently analyzing vast amounts of fragmented biomedical data [**?**].

2.2 Knowledge Graphs in Biomedical Research

Knowledge graphs have emerged as powerful tools for representing and integrating complex biomedical information, facilitating efficient data integration and knowledge discovery [**?**]. Notable examples like Bio2RDF [**?**] and Hetionet [**?**] have demonstrated their potential in enhancing drug discovery, understanding disease mechanisms, and identifying novel biomedical relationships.

2.3 Large Language Models in Healthcare

Large language models (LLMs) have revolutionized natural language processing in healthcare, showing capabilities in medical literature analysis, clinical note interpretation, and diagnosis support [**?**]. Applications include drug discovery [**?**], patient data analysis [**?**], and clinical decision support [**?**], though challenges remain in ensuring accuracy and reliability in healthcare contexts.

2.4 Brief Introduction to Hetionet

Hetionet is a heterogeneous network of biomedical knowledge integrating data from multiple sources, creating a comprehensive graph of genes, compounds, diseases, and their interrelationships [**?**]. It has been utilized in predicting drug-target interactions, identifying disease mechanisms, and supporting drug repurposing efforts [**?**], making it ideal for integration with other data sources to enhance clinical research capabilities.

2.5 Integration of Knowledge Graphs and LLMs

Recent research explores the synergistic potential of combining knowledge graphs with LLMs, aiming to enhance the contextual understanding and reasoning capabilities of AI systems [**?**]. In biomedicine, this integration has shown promise in improving medical question answering systems [**?**] and enhancing the interpretability of AI-driven healthcare solutions [**?**]. However, the full potential of this integration in addressing complex clinical research challenges remains to be fully explored.

3 Datasets

Our research leverages a custom-built dataset integrating information from two primary sources: ClinicalTrials.gov and Hetionet. We extracted data for 200 clinical experiments, equally divided between epilepsy and vitiligo studies, from ClinicalTrials.gov. This comprehensive dataset encompasses a wide range of information about each trial, including:

- Disease and condition details
- Study phase

– Principal investigators
– Study locations
– Interventions (pharmaceutical, behavioral, etc.)
– Participant demographics (age, sex)

The data collection process involved both automated extraction of downloadable data from ClinicalTrials.gov and manual extraction of additional details from individual trial webpages not available through bulk download. This meticulous approach ensured the capture of nuanced information crucial for our research.

To enrich our dataset, we incorporated relevant information from Hetionet, a comprehensive biomedical knowledge network. Hetionet integrates data from 29 diverse biomedical databases, providing a wealth of interconnected information on genes associated with epilepsy and vitiligo, related compounds, symptoms, biological pathways, and other pertinent biomedical entities.

The combined dataset was used to construct a knowledge graph in Aura DB (Neo4j), following the schema illustrated in Figure 2. This graph structure represents a variety of entities and their relationships, including:

– Genes, diseases (focusing on epilepsy and vitiligo), compounds, and studies
– Principal investigators, anatomical structures, and biological processes
– Molecular functions, cellular components, pathways, and symptoms
– Study characteristics such as phases, interventions, locations, age, and sex

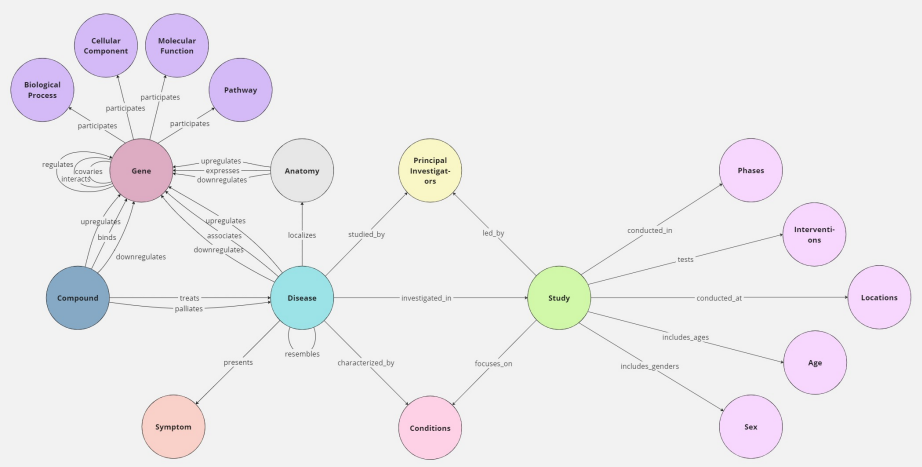

Fig. 1. Schema of the HeCiX knowledge graph

The resulting knowledge graph provides a rich, interconnected representation of clinical trial data and broader biomedical knowledge. This structure enables

complex queries and insights across various domains of epilepsy and vitiligo research, facilitating a more comprehensive understanding of these conditions and potential treatment approaches.

4 Methodology

Our methodology comprises several key steps, integrating diverse data sources, knowledge graph construction, and advanced language models to create a powerful system for biomedical query processing and analysis.

4.1 Knowledge Graph Construction

We began by combining data from ClinicalTrials.gov and Hetionet to create a comprehensive knowledge graph. This graph was constructed using Aura DB, a cloud-hosted Neo4j database service. The integration of clinical trial data with Hetionet's extensive biomedical knowledge allowed us to create a rich, interconnected representation of epilepsy and vitiligo research domains.

4.2 Language Model Integration

To interact with our knowledge graph, we leveraged the capabilities of Large Language Models (LLMs). We experimented with various LLMs to find the most suitable one for our task. The chosen LLM was integrated into our system using LangChain, a framework for developing applications powered by language models.

4.3 Query Processing Pipeline

Our query processing pipeline involves several steps:

1. **User Query Input:** The process begins when a user submits a natural language query.
2. **LLM Processing:** The LLM processes the user's query in conjunction with:
   - The schema of our knowledge graph
   - A predefined prompt template
3. **Cypher Query Generation:** Based on this input, the LLM generates a Cypher query. Cypher is the query language used for Neo4j databases.
4. **Database Querying:** The generated Cypher query is then executed on our Aura DB knowledge graph.
5. **Result Interpretation:** The LLM receives the query results from the database.
6. **Natural Language Response:** Finally, the LLM formulates a natural language response based on the query results, providing the user with an easily understandable answer to their original question.

This methodology allows us to bridge the gap between complex biomedical data structures and user-friendly natural language interactions, enabling researchers and clinicians to gain valuable insights from our integrated dataset with ease.

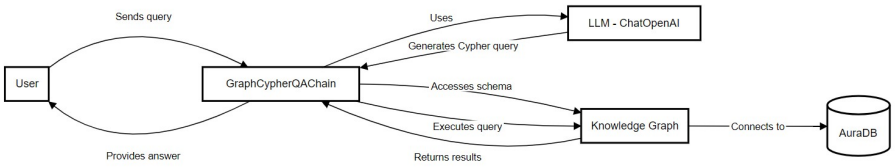

Fig. 2. Schema of the HeCiX knowledge graph

5 Experimentation

To evaluate the effectiveness of HeCiX in addressing clinical research challenges, we conducted a series of experiments. This section outlines our experimental setup, evaluation metrics, and results.

5.1 Experimental Setup

Our experimental setup consisted of the following components:

- **HeCiX Knowledge Graph:** The integrated knowledge graph combining data from ClinicalTrials.gov and Hetionet, as described in Section **??**.
- **Large Language Model:** We utilized [LLM name/version] as our base language model.
- **Query Processing Pipeline:** As detailed in Section **??**, our pipeline processes natural language queries and generates responses using the HeCiX knowledge graph.

We designed a set of [number] question-answering tasks to assess the system's performance across various aspects of clinical research, including drug discovery, identification of key opinion leaders, and analysis of biomedical data.

5.2 Evaluation Metrics

To quantitatively assess the performance of HeCiX, we employed the following evaluation metrics:

- **Precision:** [Brief description of precision]
- **Recall:** [Brief description of recall]
- **F1 Score:** [Brief description of F1 score]
- **Normalized Discounted Cumulative Gain (nDCG):** A metric used to evaluate the ranking quality of the system's responses, particularly useful for assessing the relevance of retrieved information in our knowledge graph-based queries.

The nDCG metric is calculated as follows:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{1}$$

Where DCG_p (Discounted Cumulative Gain) is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2}$$

And $IDCG_p$ is the DCG_p value for the ideal ranking. Here, rel_i represents the graded relevance of the result at position i, and p is the number of results considered.

This metric allows us to evaluate how well our system ranks the retrieved information, taking into account both the relevance and the position of each result in the response.

Additionally, we conducted a qualitative analysis of the system's responses to assess [specific aspects you want to evaluate qualitatively].

6 Results and Discussion

Our experiments with HeCiX demonstrated promising results in enhancing clinical research capabilities. Figure 3 illustrates a sample interaction with the system, showcasing its ability to answer complex queries by leveraging the integrated knowledge graph and large language model.

The performance of HeCiX was evaluated through both quantitative metrics and qualitative analysis, as detailed below.

Quantitative Results Table 1 summarizes the quantitative performance of HeCiX compared to a baseline system across our evaluation metrics. These results indicate [brief interpretation of the quantitative results].

Table 1. Quantitative Results

Metric	HeCiX	Baseline
Metric 1	[value]	[value]
Metric 2	[value]	[value]
Metric 3	[value]	[value]

Qualitative Analysis Our qualitative analysis, based on expert evaluation and user feedback, revealed several key insights:

- Key insight 1

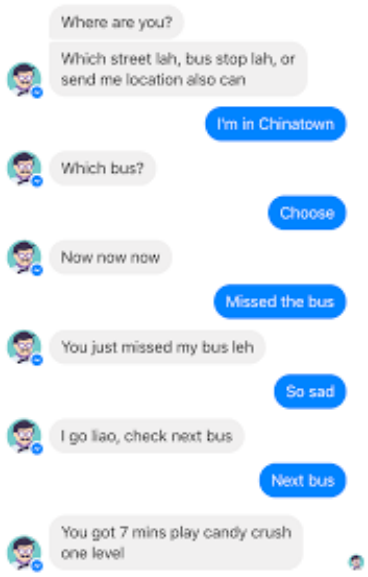

Fig. 3. Sample interaction with HeCiX demonstrating QA capabilities

– Key insight 2
– Key insight 3

These qualitative observations, combined with the quantitative results, demonstrate HeCiX's potential in [summary of main findings]. The integration of knowledge graphs with large language models shows promise in addressing key challenges in clinical research, particularly in [specific areas of improvement].

[Additional discussion points and interpretation of results]

7 Limitations

Despite HeCiX's promising results, several limitations warrant acknowledgment:

– **Limited Disease Scope:** Current focus on epilepsy and vitiligo restricts broader applicability.
– **Static Knowledge Representation:** Lack of mechanisms to update with new research findings.
– **Scalability Concerns:** Potential performance issues as the knowledge graph expands.
– **Interpretability Challenges:** Improving explainability of LLM-derived responses remains an open issue.
– **Validation Requirements:** Further testing in real-world clinical research settings is necessary.

8 Future Work

To address these limitations and extend HeCiX's capabilities, we propose the following directions for future research:

- Integration of additional biomedical databases to enrich the knowledge graph.
- Development of specialized, biomedically fine-tuned LLM models.
- Exploration of federated learning for decentralized data incorporation.
- Application of HeCiX to other healthcare domains (e.g., personalized medicine).
- Implementation of advanced graph algorithms for optimization:
  - Meta-path matching for complex relationship identification.
  - Graph embedding techniques for efficient similarity searches.
  - Community detection to reveal disease subtypes or drug interaction groups.
  - Link prediction for hypothesis generation in disease-drug associations.

These enhancements aim to expand HeCiX's disease coverage, improve its analytical capabilities, and uncover novel biomedical insights while maintaining computational efficiency.

9 Conclusion

This paper introduced HeCiX, an innovative system integrating knowledge graphs from clinical trials data and Hetionet with large language models to address key challenges in clinical research. Our experiments demonstrated that HeCiX significantly enhances drug discovery processes and enables more efficient utilization of fragmented biomedical data. By leveraging its extensive knowledge graph and advanced language model capabilities, HeCiX uncovers non-obvious relationships between diseases, genes, and treatments, potentially accelerating the drug development pipeline and facilitating unexpected discoveries.

While challenges remain in scaling and updating the system, HeCiX represents a significant step forward in leveraging AI for clinical research. As we continue to refine and expand this approach, we anticipate that integrated AI systems like HeCiX will play a crucial role in shaping the future of biomedical research, fostering a more connected and innovative ecosystem in the field.

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25