An AI-powered RAG based Framework for Data Extractions in Systematic Literature Reviews

MSR149

Rajdeep Kaur[1], Barinder Singh[2], Pankaj Rai[1], Vedant Soni[1], Sunil Kumar[1], Mrinal Mayank[1]
[1]Pharmacoevidence, Mohali, India; [2]Pharmacoevidence, London, UK

Pharmaco® Evidence

CONCLUSION

- The development of the AI-powered RAG-based framework represented a significant advancement in automating the extraction phase of the SLRs
- The AI-driven extraction pipeline led to a >70% reduction in turnaround time, streamlining what is traditionally a time- and labor-intensive process. This efficiency was achieved without compromising quality, with accuracy and reliability at acceptable levels relative to human reviewers

Introduction

- Systematic literature reviews (SLRs) play a vital role in the health economics and outcomes research domain, providing critical insights for healthcare decision-making
- Data extraction in SLRs is an important step to collect insightful information from the relevant studies
- Conventional literature search and review processes are largely undertaken manually and typically require substantial time and resources[1,2]
- AI methods have the potential to automate various steps in these processes[1,2]
- Large language models with Retrieval-Augmented Generation (RAG) can be used to automate data extraction from studies

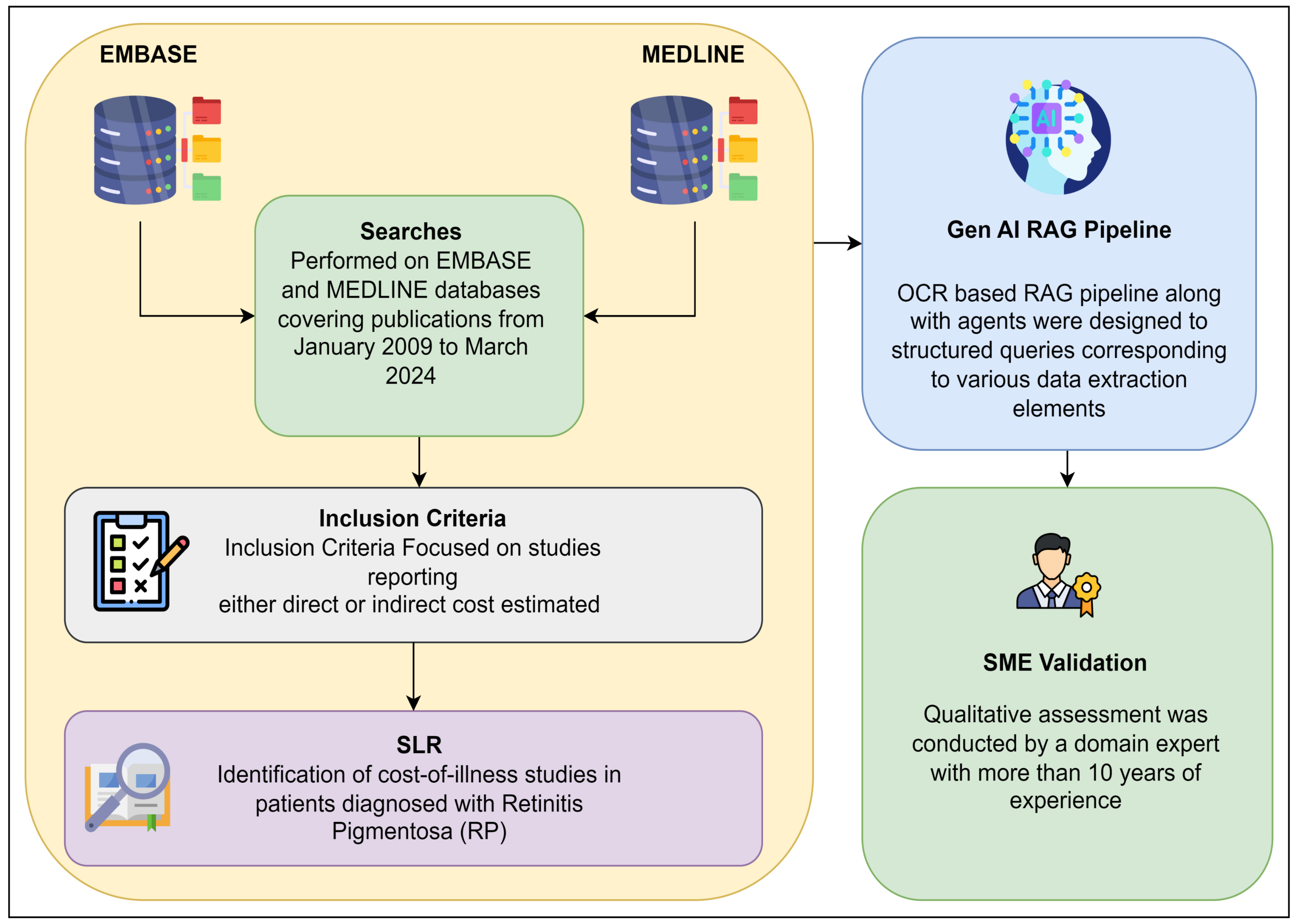
Objective

- The aim of the study was to develop an AI automated RAG driven platform to streamline the extraction of relevant information from included studies in SLRs, reducing the time required for data extractions

Methodology

- A systematic literature review was conducted to identify cost-of-illness studies in patients with Retinitis Pigmentosa (RP)
- Searches were performed in Embase and MEDLINE databases, covering publications from January 2009 to March 2024
- Inclusion criteria focused on studies reporting either the direct or indirect cost estimates associated with RP in any geographic setting
- A retrieval-augmented generation (RAG) pipeline was developed to enable automated data extraction based on pre-defined parameters
- The process began with a standardization engine incorporating high-accuracy Optical Character Recognition (OCR) to convert PDF documents into clean, machine-readable text
- This standardized content was then segmented into smaller chunks, and corresponding text embeddings were generated and stored in a vector database for efficient retrieval and analysis
- A multi-agent AI system was designed to respond to structured queries corresponding to various data extraction variables
- The prompts for AI agents were carefully constructed to extract relevant information such as study setting, design, cost components, and population characteristics
- Domain experts with at least 10 years of knowledge evaluated the data extraction results and conducted cross verification against the source documents to ensure the accuracy and consistency

Figure 1. Methodological flow



Results

- The SLR included a total of six studies conducted across the United States (US) (n=2), Japan (n=2), Spain (n=1), and globally (US and Canada, n=1)
- The AI platform was used to extract the study characteristics, population characteristics, direct and indirect cost outcomes, and key findings from the included studies
- Subject matter experts (SMEs) developed structured prompts to extract data from the RAG system and evaluated the accuracy of AI-generated outputs
- Quality was assessed using three categories: 'Strongly Agree' (complete alignment with human-extracted data), 'Minor Deviation' (minor contextual differences or low-impact hallucinations), and 'Disagree' (incorrect or misleading extraction)
- Of the 75 extracted data points, 90% were rated as 'Strongly Agree', 7% as 'Minor Deviation', and 3% (2 parameters) as 'Disagree'
- Overall, the AI-extracted content showed strong consistency in terminology and structure, closely mirroring human-extracted results
- The process achieved over 70% reduction in turnaround time, with the RAG-powered language model enhancing context retention and minimizing hallucinations

"AI-driven SLR data extractions achieved accuracy and consistency at an acceptable level compared to conventional human-led processes, while reducing turnaround time by over 70%"

Figure 2. Study characteristics of studies

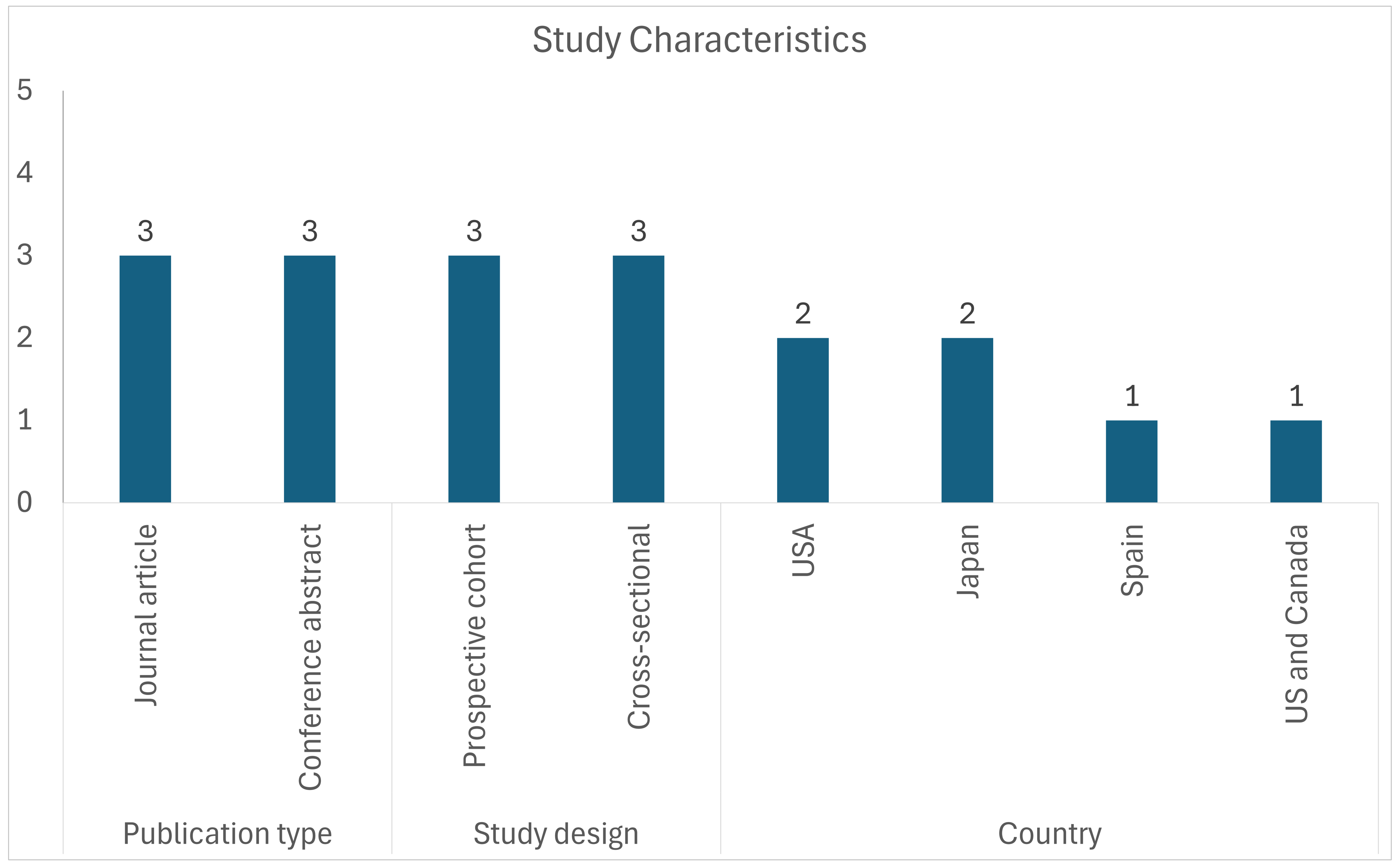


Table 1: Overview of automated data extraction variables and performance Metrics

Category	Variables/observation

Types of Information Extracted	• Study characteristics: country, publication year, design, sample size, age/gender
	• Population characteristics: RP subtype, severity, eligibility
	• Cost outcomes:
	  • Direct medical (e.g., hospitalization, treatment)
	  • Direct non-medical (e.g., transport, assistive devices)
	  • Indirect (e.g., productivity loss, caregiver time)
	• Key findings on economic and clinical burden

Accuracy (Expert Review)	90% "Strongly Agree" (rated accurate and relevant)
Errors / Limitations	- 7% showed hallucinations or added noise - 3% instances of incorrect/missing data
Additional Insights	- Maintained consistency in terminology - Handled diverse cost categories well - Slightly impacted by OCR noise and mixed-format tables

References

1. CDA-AMC. New Position Statement Aims to Guide the Use of AI Methods in Health Technology Assessment. 2025
2. NICE. Use of AI in evidence generation: NICE position statement. 202