From Evidence to Excel: Generative AI for Automated SLR Data Extractions

MSR112

Barinder Singh[2], Shubhram Pandey[1], Rajdeep Kaur[1], Mrinal Mayank[1], Ritesh Dubey[1], Marjana Bharali[1]
[1]Pharmacoevidence, Mohali, India; [2]Pharmacoevidence, London, UK

Science-led HEOR & Tech-driven AI

INTRODUCTION & OBJECTIVES

- Data extraction is one of the most time and resource intensive steps in evidence generation process
- Leveraging Large Language Models (LLMs) can significantly streamline this process by reducing manual effort and improving efficiency
- This study aimed to evaluate a generative AI powered tool developed to extract structured information from unstructured data sources (Regulatory submissions, clinical study publications and guidelines) which are commonly used in health technology assessment (HTA) and Health economics outcomes research (HEOR)

METHODS

- The tool was developed using Python with AWS Amazon web services (AWS) Bedrock for language model processing retrieval-augmented generation (RAG) for unstructured data and PostgreSQL for structured data storage (**Figure 1**)
- Data from 20 publicly available publications of randomized controlled trials (RCTs) on diabetes, focusing on efficacy and safety outcomes were uploaded in RAG
- The uploaded files were standardized, where text content was converted into markdown format and tables and images were extracted and organized separately
- Custom extraction tables were defined by specifying field names (e.g., "Age", "Sample size"), data types (e.g., numerical, categorical), and extraction instructions (e.g., “extract mean age for all treatment groups”)
- For each defined field in the extraction tables, RAG identified the most relevant chunks across the selected documents to ensure accurate data capture
- Claude 3.7 Sonnet read the retrieved chunks, extracted the appropriate values, and formatted them according to the specifications provided for each document
- The extracted data was compiled into tables and exported as Excel workbooks, with structured data stored in PostgreSQL for long-term retention (**Figure 2**)
- Results were exported as Excel workbooks and validated by subject matter experts (SMEs) for completeness, clarity, and traceability of the extracted data (**Figure 3**)

Figure 1: Schematic diagram of data extraction process

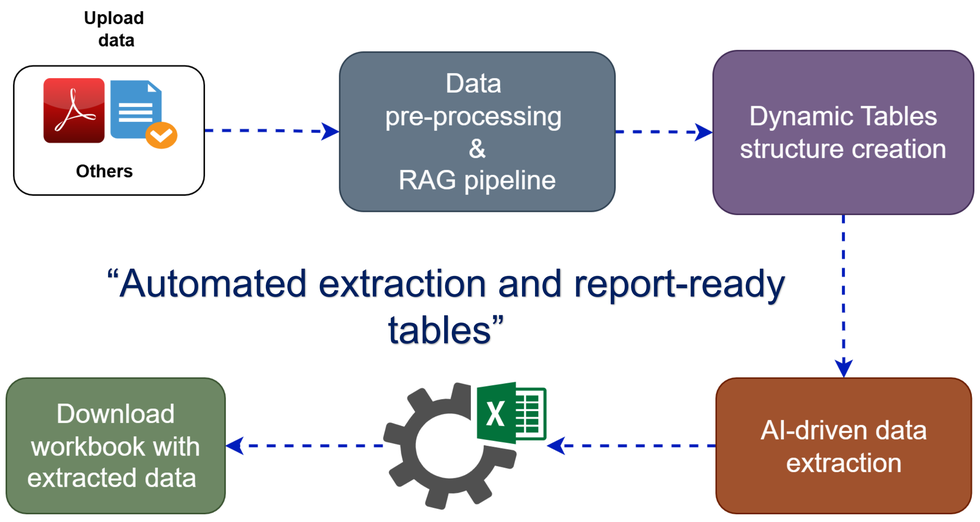
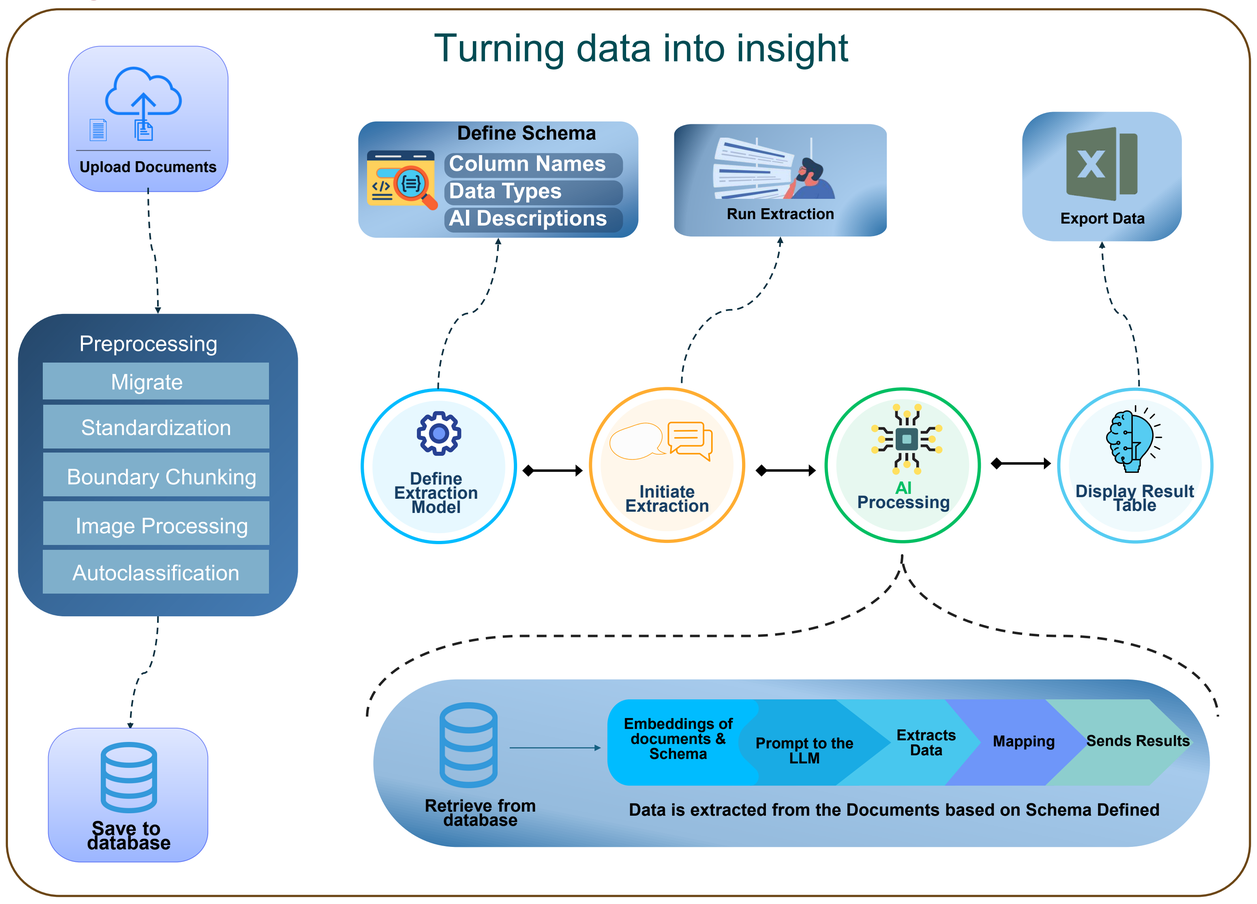
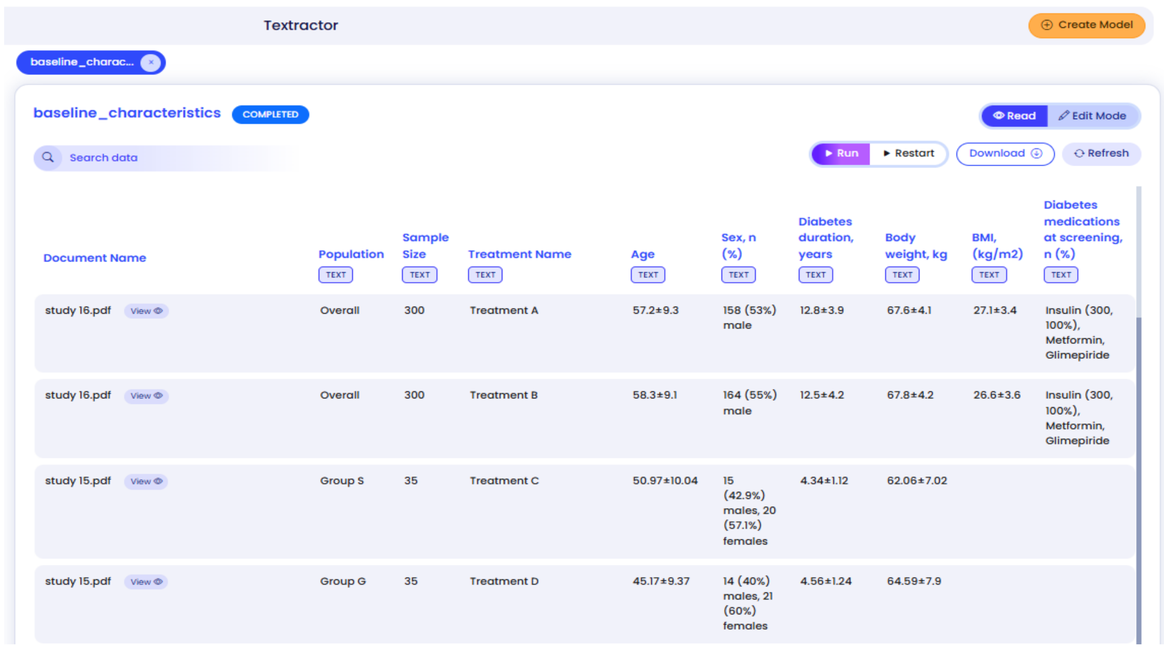


Figure 2: Technical workflow of the Textractor



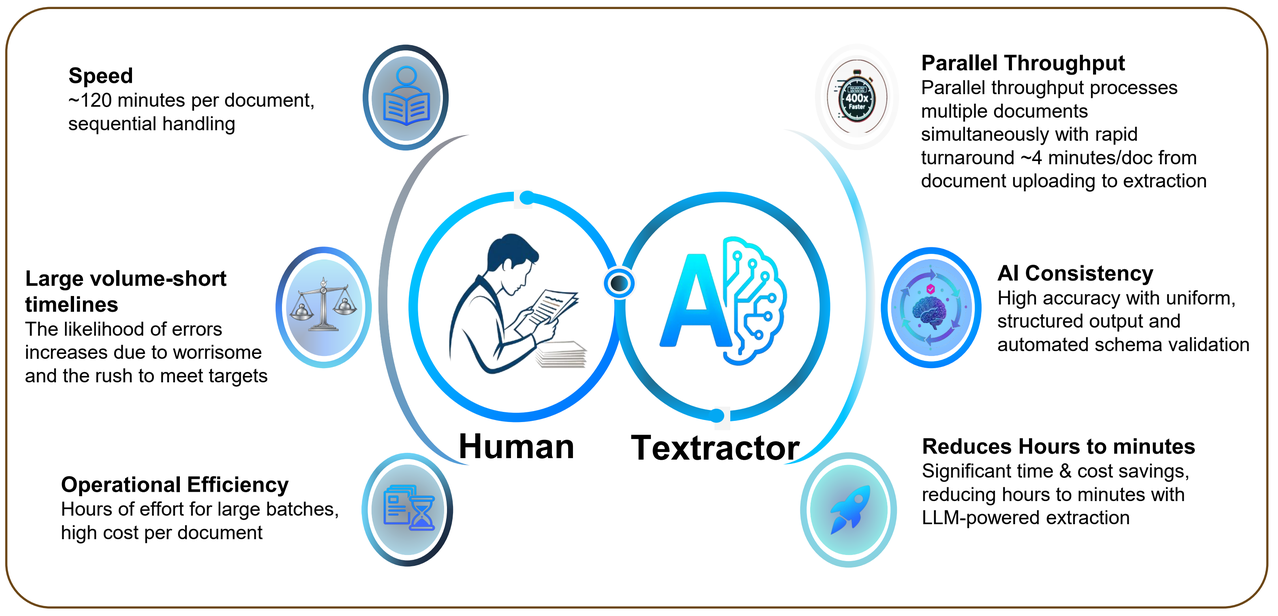
RESULTS

Figure 3: Visualization of data extraction table in tool



Document Name	Population	Sample Size	Treatment Name	Age	Sex, n (%)	Diabetes duration, years	Body weight, kg	BMI, (kg/m2)	Diabetes medications at screening, n (%)
	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT
study 16.pdf View	Overall	300	Treatment A	57.2±9.3	158 (53%) male	12.8±3.9	67.6±4.1	27.1±3.4	Insulin (300, 100%), Metformin, Glimepiride
study 16.pdf View	Overall	300	Treatment B	58.3±9.1	164 (55%) male	12.5±4.2	67.8±4.2	26.6±3.6	Insulin (300, 100%), Metformin, Glimepiride
study 15.pdf View	Group S	35	Treatment C	50.97±10.04	15 (42.9%) males, 20 (57.1%) females	4.34±1.12	62.06±7.02		
study 15.pdf View	Group G	35	Treatment D	45.17±9.37	14 (40%) males, 21 (60%) females	4.56±1.24	64.59±7.9		

Figure 4: Evaluation parameters of AI-generated responses



- Three separate data extraction tables were generated, capturing study characteristics, patient demographics, treatment details, clinical and economic outcomes
- SMEs verified that all data points related to study and patient characteristics were extracted with 100% accuracy, and complete traceability to the source documents
- A minor issue was noted in the clinical outcomes table, where the names of two secondary outcomes initially missing but were subsequently corrected manually
- Overall, SMEs confirmed that the tool effectively extracted structured data, enabling users to download analysis ready Excel workbooks and reduce manual effort by approximately 70% (**Figure 4)**

- ✓ RAG-assisted AI bridges automation and human expertise to create a new standard in evidence generation, i.e., transparent, auditable, and scientifically rigorous
- ✓ It accelerates data extraction, enables dynamic PICO simulations (for JCA and beyond), and streamlines data for ITC/NMA, reporting, dossiers, and modeling workflows, delivering unmatched speed and accuracy

CONCLUSIONS

The tool demonstrated strong potential to significantly reduce manual effort and save time by flexibly extracting data into user-defined tables. Its capability to download analysis-ready Excel outputs, further enhances the usability, supporting streamlined data processing across multiple workflows in HEOR including but not limited to reporting, model adaptation, and indirect treatment comparison

References
1. NICE position statement. Use of AI in evidence generation. Accessed 29 May 2025
2. CDA-AMC. New Position Statement. Accessed 29 May 2025

Correspondence: **Barinder Singh**; Barinder.singh@pharmacoevidence.com

Disclosure: **BS**, MM, RD, MB, RK and SP, the authors declare that they have no conflict of interest

ISPOR 2025; November 09-12; Glasgow, Scotland, UK