Title: **ZebraMap: A Multimodal Rare Disease Knowledge Map with Automated Data Aggregation & LLM-Enriched Information Extraction Pipeline**

Author Names: Md. Sanzidul Islam , Amani Jamal and Ali Alkhathlan

Abstract: Rare diseases often lead to delayed diagnosis because clinical knowledge is fragmented across unstructured research, individual case reports, and heterogeneous data formats. This study presents ZebraMap, a multimodal knowledge map created to consolidate rare disease information and transform narrative case evidence into structured, machine-readable data. **Methods:** Using Orphanet as the disease registry, we identified 1727 rare diseases and linked them to PubMed case reports. We retrieved 36,131 full-text case report articles that met predefined inclusion criteria and extracted publication metadata, patient demographics, clinical narratives (cases), and associated images. A central methodological contribution is an automated large language model (LLM) structuring pipeline, in which free-text case reports are parsed into standardized fields, such as symptoms, diagnostic methods, differential diagnoses, treatments, and outcome that produce structured case representations and image metadata matching the schema demonstrated in our extended dataset. In parallel, a retrieval-augmented generation (RAG) component generates concise summaries of epidemiology, etiology, clinical symptoms, and diagnostic techniques by retrieving peer-reviewed research to enhance missing disease-level descriptions. **Results:** The final dataset contains 69,146 structured patient-level case texts and 98,038 clinical images, each linked to a particular patient ID, disease entry, and publication. Overall cosine similarity between curated and generated text is 94.5% and performance in information extraction and structured data generation is satisfactory. **Conclusions:** ZebraMap provides the largest openly accessible multimodal resource for rare diseases and enables data-driven research by converting narrative evidence into computable knowledge.

Keywords: multimodal dataset; rare disease diagnosis; knowledge map; case reports; LLM applications; retrieval-augmented generation (RAG); clinical images; structured clinical records

https://doi.org/10.3390/diagnostics16010107

https://www.mdpi.com/2075-4418/16/1/107