SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

RadDB-Agent: An Automatic Agentic System for Efficiently Building Large-scale Trustworthy Cancer Patient Databases

Zheren Zhu, PhD Student, University of California, San Francisco, University of California, Berkeley; Kang Wang, MD, PhD; Yang Yang, PhD

Introduction/Background

Large-scale and trustworthy cancer patient databases are essential for advancing data-driven cancer research. However, many institutional data warehouses, despite containing hundreds of millions of clinical notes (Fig.1), often lack patient-level classification for specific cancer types. Identifying patients using ICD codes is unreliable due to varied sources and low detail—e.g., distinguishing PDAC from ampullary tumors. The volume of notes also makes manual review infeasible and overwhelms simple LLMs.

Methods/Intervention

We designed RadDB-Agent, an agentic LLM system to automatically build large-scale, patient-level databases of any cancer type with high precision and efficiency. The system (Fig.2) begins with a DB query agent: starting from patients with given ICD-9/10 codes for a specific cancer type, all pathology/cytology and progress notes are grouped by patient and arranged chronologically. A randomly sampled internal validation set was labeled as positive/negative diagnosis by a radiologist and LLM to benchmark LLM's accuracy and guide keyword filtering. Pathology/cytology notes are first processed by LLM to identify positive patients. For unconfirmed cases, an automatic keyword filtering agent is applied to progress notes using LLM-generated keywords, iteratively refined (Fig.2) on the validation set. This removes the need for manual term curation and generalizes across cancer types. Only keyword-matched notes are processed for additional positives, reducing computational load.

Results/Outcome

In this study, we deployed RadDB-Agent on UCSF's clinical data warehouse, targeting patients with pancreatic ductal adenocarcinoma (PDAC). ICD-based querying returned 11,080 patients, with 325,873 progress notes (covering 9,503 patients) and 52,460 pathology/cytology notes (covering 6,464 patients). To evaluate, a 10-patient sub-cohort was randomly sampled (221 progress notes, 68 pathology/cytology notes). An internal validation set of 48 notes was labeled by both a radiologist and LLM, yielding LLM's perfect accuracy (1.0) in identifying positive/negative notes (Fig.3). The keyword agent met stopping criteria in one iteration (0/33 positive notes in the "without keyword" segment in validation). Five patients were identified as positive, with a radiologist-verified sensitivity of 1.0. The system skipped 240/289 notes, greatly reducing LLM load by 83%.

Conclusion

RadDB-Agent can automatically identify cancer patients with high precision and efficiency.

Statement of Impact

This system enables scalable, efficient, and potentially fully automated construction of high-quality cancer patient databases.

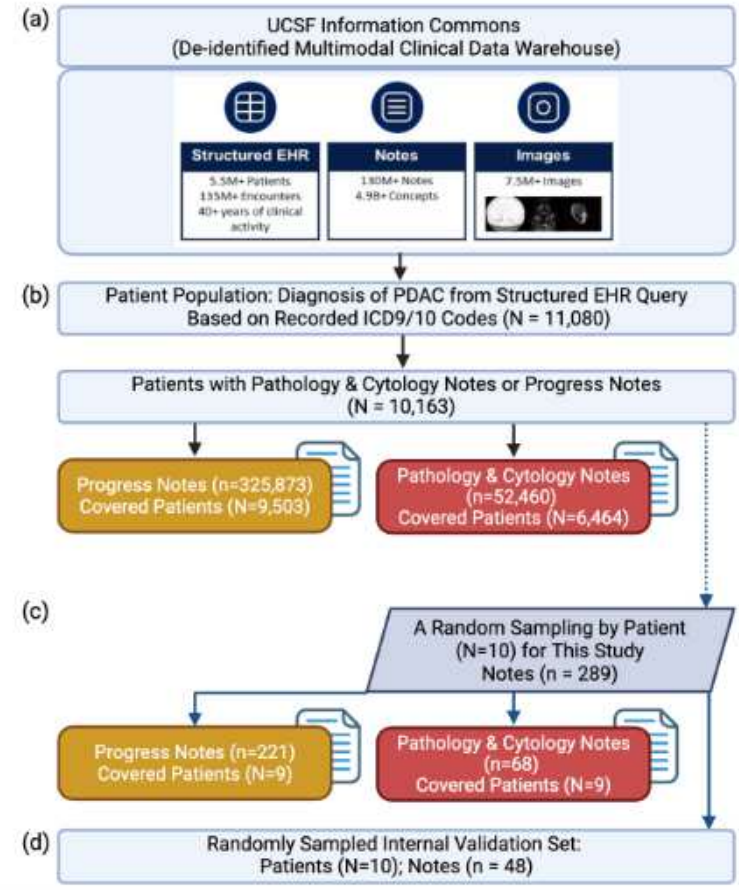

Fig.1 (a) Illustration of a clinical data warehouse at UCSF (b) Entire queried cohort for PDAC patients. (c) Cohort of a randomly sampled subset of patients used in this study (d) Internal validation set subsampled from the cohort in this study to benchmark LLM's accuracy in identifying diagnosis in notes and guide agentic keyword filtering.

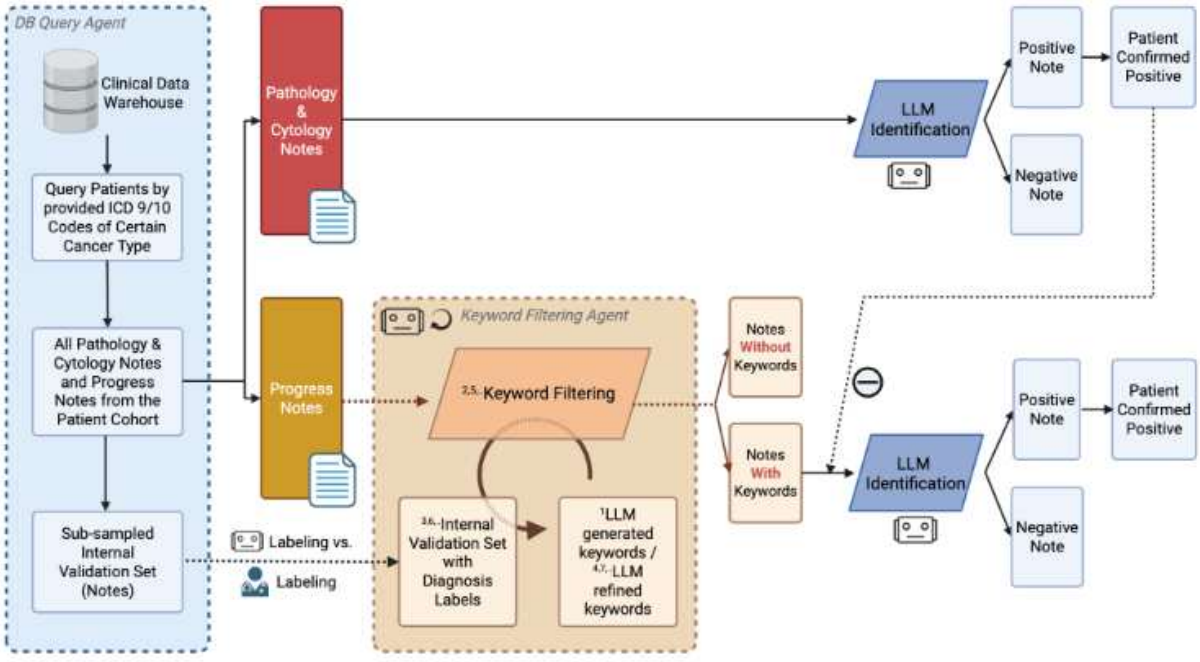

Fig.2 Pipeline of RadDB-Agent system. LLM used in this system is a secured institutional deployment of GPT-4.1 (version gpt-4.1-2025-04-14) accessed through Azure OpenAI service. For the Keyword Filtering Agent, the numeric corner mark at the top left of the steps represents the iteration sequence: 1. LLM generates a list of keywords for the specific cancer type. 2. Use the keyword list to filter the internal validation set. 3. Test the proportion of notes with labeled positive diagnoses filtered to the

'without keywords' segment. 4. Refine the keyword list based on the proportion and the keywords from the previous iteration. Refinement stop criteria: fewer than 2% of positive diagnoses remain in notes without keywords. After the stopping criterion is met, the agent will filter the entire set of progress notes using the final keyword list.

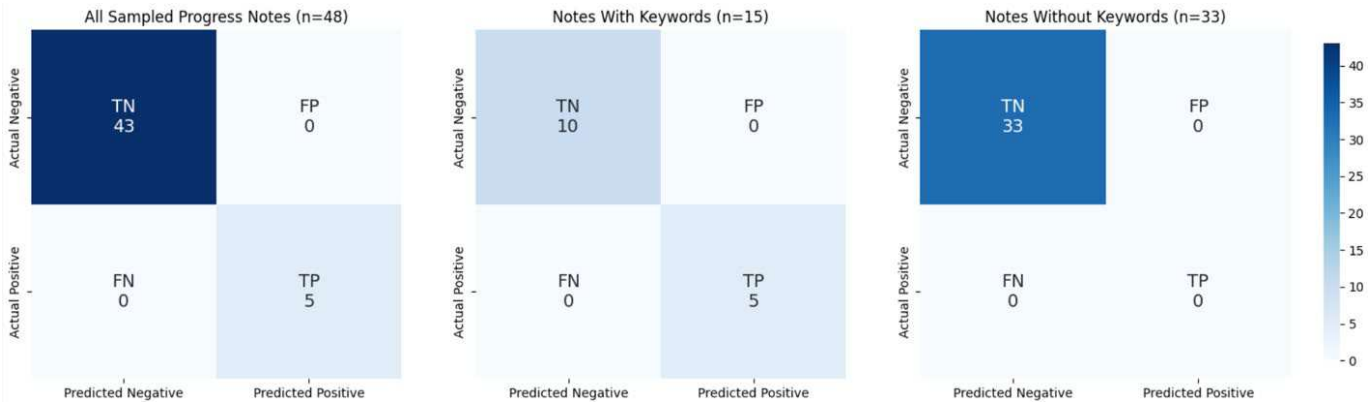

Fig.3 Confusion metrics for a radiologist's identifications (ground truths) and LLM's identifications on progress notes with positive/negative diagnosis of PDAC over the entire internal validation set, notes with final keywords in the validation set, and notes without final keywords in the validation set.

Keywords

Large Language Model; AI Agent; Trustworthy and Large-Scale Patient Database; Radiology Notes; Computer-aided Diagnosis