Advancing Computational Cancer Phenotyping: A Comparative Study of Structured-Only Unsupervised Methods vs. Multimodal RAG Hosted and Local Pipelines

Victor M. Castro, MS[1], Martin J. Rees[1], Nich Wattanasin, MS[1], Vivian S. Gainer, MS[1], Shawn N. Murphy, MD, PhD[1]
[1]*Research Information Science and Computing, Mass General Brigham, Somerville, MA, USA*

Mass General Brigham

Introduction

- Computational phenotyping (CP) underpins large-scale observational studies in EHR data.
- Most CP algorithms operate on structured EHR data with limited information
- RAG-based LLM pipelines can incorporate clinical text to augment CP algorithms.

Methods

- All clinical notes for each patient were embedded using either a local open-source model (MedEmbed-large-v0.1) or a proprietary model (OpenAI text-embedding-3-large).
- We deployed GPT-4o as a hosted proprietary model through Azure OpenAI services, while Qwen3-8B was run as an open-source model locally.
- As a structured-only baseline, we implemented the KOMAP algorithm (Xiong, 2023).
- A licensed clinical nurse performed manual chart review for a 50-patient subset to determine the presence or absence of melanoma.
- These adjudications served as the gold standard for assessing sensitivity, positive predictive value (PPV), negative predictive value (NPV), and overall agreement.

Results

Method	Data Modality	Model Setup	*Comparison to Clinician Chart Review (N=50)*				
			Agreement	**Cohen's κ**	**Sensitivity**	**PPV**	**NPV**
KOMAP	Structured EHR only	Unsupervised modeling	0.792	0.271	0.946	0.814	0.600
Local embedding / OpenAI LLM	Notes (embeddings)	MedEmbed-large-v0.1 + GPT-4o (hosted)	**0.960**	**0.875**	**1.000**	0.951	**1.000**
OpenAI embedding / OpenAI LLM	Notes (embeddings)	text-embedding-3-large + GPT-4o (hosted)	0.900	0.718	0.923	0.947	0.750
Local embeddings / Local LLM	Notes (embeddings)	MedEmbed-large-v0.1 + Qwen3-8B (local)	0.920	0.794	0.897	**1.000**	0.733

Conclusions

- RAG LLM pipelines consistently outperform structured-only unsupervised methods for melanoma case identification.
- Local MedEmbed+Qwen3-8B configuration achieved performance comparable to GPT-4o, supporting the feasibility of privacy-preserving, institution-controlled deployments.