Leveraging Large Language Model for Populating OMOP Oncology CDM from the EHR : Feasibility Study

PRESENTER: Seng Chan You

INTRODUCTION

- The Oncology CDM Working Group developed the OMOP Oncology Extension to support the integration of cancer-specific information into the OMOP CDM.
- Despite these advancements, much of the cancer-data in EHR remains in unstructured formats, making it challenging to utilize and standardize.
- Generative large language models (LLMs) present a promising solution to these challenges, by leveraging the in-context learning capabilities of LLMs.
- In this study, we developed strategy to extract the cancer information from unstructured pathology and radiology reports of patients with colorectal cancer using state-of-the-art LLM.
- Among several candidate applications to validate feasibility, we focused on whether LLM-derived cancer data can be used to define cancer stage at diagnosis in accordance with updates to the AJCC staging system.

METHODS

Data sources

- We obtained unstructured pathology and radiology reports for patients diagnosed with colorectal cancer at Severance Hospital between 2010 and 2023.
- A random sample of 1,000 individuals was selected for inclusion in the study. We used 1,579 radiology and 2,632 pathology reports documented within 30 days before or 120 days after initial cancer diagnosis.

Prompt design

- We interacted with GPT-4o via zero-shot prompting through the OpenAI API. A total of 20 reports were sampled to develop prompts to extract cancer data (**Table 1**). All output was compiled into a JSON format.

Evaluation

- We classified the cancer stage at diagnosis using based on the 8th edition of the AJCC TNM staging system. We compared the LLM-derived cancer stage at diagnosis with the TNM values retrieved from the EHRs database.
- Additionally, we defined the cancer stage using both the 7th and 8th editions of the AJCC staging system and illustrated the changes in cancer stage, demonstrating the usefulness and flexibility of the LLM-derived cancer information.

Generative LLM can be used to populate Oncology CDM from the unstructured EHRs

Table 1. Oncologic data extracted from pathology and radiology reports

Pathology reports			
Category	**Descriptor**	**Category**	**Descriptor**
Feature	Size	Lymph node	Metastasis site
	Histologic grade		Metastasis count
	Histologic type	Biomarker	BRAF mutation
	Location		KRAS mutation
	Procedure		Ki-67 index
	Tumor status		MLH1
Invasion	Depth of invasion		MSH2
	Lymphovascular invasion		MSH6
	Perineural invasion		Microsatellite instability
	Tumor budding		Mitotic count
	Tumor deposits		NRAS mutation
Margin	Basal margin		PMS2
	Circumferential margin	Other	Post treatment/Procedure status
	Distal margin		
	Lateral margin		
	Proximal margin		
	Resection margin		
Radiology reports			
Feature	Tumor location		
	Tumor status		
	Size		

Figure 1. Overall performance of GPT-4o on classification of cancer stage

AJCC staging from LLM \ AJCC staging from EHR	0	I	IIA	IIB	IIC	IIIA	IIIB	IIIC	IVA	IVB	IVC
0	69	1	1	0	0	0	0	1	3	0	0
I	2	234	7	0	0	0	1	0	0	1	0
IIA	1	1	116	1	0	0	3	0	1	0	0
IIB	0	0	0	11	0	0	0	0	1	1	0
IIC	1	0	0	0	3	0	0	0	0	0	0
IIIA	0	1	0	0	0	18	0	0	0	0	0
IIIB	0	0	1	0	0	0	99	4	0	0	1
IIIC	0	0	0	0	0	0	1	14	0	0	0
IVA	0	15	13	0	0	1	5	2	9	2	1
IVB	0	1	2	1	0	0	0	0	0	2	1
IVC	0	0	5	1	1	0	2	2	4	2	2

Figure 2. Comparison of TNM staging according to the AJCC editions

TNM staging from AJCC 8th edition \ TNM staging from AJCC 7th edition	0	I	IIA	IIB	IIC	IIIA	IIIB	IIIC	IVA	IVB
0	75	0	0	0	0	0	0	0	0	0
I	0	245	0	0	0	0	0	0	0	0
IIA	0	0	123	0	0	0	0	0	0	0
IIB	0	0	0	13	0	0	0	0	0	0
IIC	0	0	0	0	4	0	0	0	0	0
IIIA	0	0	0	0	0	19	0	0	0	0
IIIB	0	0	0	0	0	0	105	0	0	0
IIIC	0	0	0	0	0	0	0	15	0	0
IVA	0	0	0	0	0	0	0	0	48	0
IVB	0	0	0	0	0	0	0	0	0	7
IVC	0	0	0	0	0	0	0	0	10	9

RESULTS

- A total of 4,211 pathology and radiology reports from 1,000 patients were analyzed.
- The agreement between LLM-derived AJCC stage and AJCC stage from structured EHRs is presented using confusion matrix in Figure 1. The overall accuracy of LLM-derived staging was 0.86. Cohen's Kappa was 0.82 (95% confidence interval [CI], 0.78-0.85).
- Figure 2 shows the comparison of TNM staging groups according to the AJCC 7th and 8th edition.
- A major difference between 7th and 8th edition is that the inclusion of new stage involving peritoneal metastasis (stage IVC).
- As a result, 19 patients, originally classified as stage IVA or IVB under the 7th edition, were reclassified as stage IVC.

CONCLUSION

- This is ongoing study. Generative LLMs demonstrate feasibility in automating the extraction of structured cancer information from unstructured EHRs.
- This approach has the potential to construct well-fined resources for future research, reducing the workload of human experts.
- By leveraging generative LLM, we will standardize the cancer-specific data from the EHR based on the OMOP Oncology Extension.

Subin Kim[1,2], Jeong Eun Choi[1,2], Chang Jun Ko[3], Seng Chan You[1,2]

[1]Dept. of Biomedical Systems Informatics, Yonsei University College of Medicine

[2]Institute for Innovation in Digital Health Care, Yonsei University

[3]Dept. of Health Informatics and Biostatistics, Graduate School of Public Health, Yonsei University