Leveraging Generative Large Language Model to Populate OMOP Oncology CDM from the EHR : Feasibility Study

PRESENTER: Subin Kim, Seng Chan You

INTRODUCTION

- Converting unstructured cancer data into a standardized format is essential to enhance the utility of EHRs in cancer research.
- In this study, we developed an **NLP pipeline** to extract cancer-specific information from unstructured pathology reports using an open-source generative LLM. Further, we **integrated the extracted information** into the **current OMOP-CDM** database.

METHODS

Data sources

- Pathology reports were retrieved from patients with colorectal, breast, or lung cancer at Severance Hospital (2010-2023). From a total of 57,433 eligible patients, 10,000 patients were randomly selected as a study population.

Development of NLP Pipeline

- We used 120 pathology reports per cancer type as the training dataset and validated performance using 100 randomly sampled pathology reports for each type of cancer.
- The **NLP pipeline** was designed with three key stages: **parsing, extraction, and structuring** (Fig. 1).
- Through this workflow, cancer-specific attributes such as tumor location, histology, tumor size, invasion status, and biomarker were extracted.

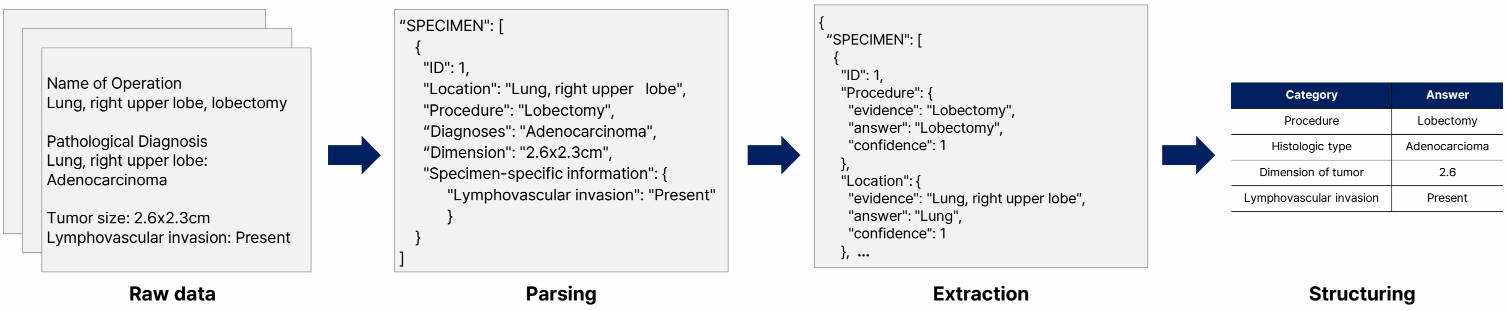
Data integration

- The extracted variables were processed through an ETL pipeline to integrate into the OMOP CDM.
- Each variable was mapped to an OMOP standard vocabulary with a corresponding concept ID.
- Information extracted from the pathology reports was inserted into the NOTE_NLP table. Then, the data were loaded into the MEASUREMENT table according to the OMOP CDM Oncology Extension.

Proof-of-concept study

- We compared overall survival between Stage II–III colorectal cancer patients with deficient mismatch repair (dMMR) and those with proficient mismatch repair (pMMR), **using integrated dataset**.
- A Kaplan–Meier survival curve was used to visualize survival differences between these two groups.

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



Raw data

Name of Operation
Lung, right upper lobe, lobectomy

Pathological Diagnosis
Lung, right upper lobe:
Adenocarcinoma

Tumor size: 2.6x2.3cm
Lymphovascular invasion: Present

Parsing

```
“SPECIMEN": [
  {
    "ID": 1,
    "Location": "Lung, right upper   lobe",
    "Procedure": "Lobectomy",
    “Diagnoses": "Adenocarcinoma",
    “Dimension": "2.6x2.3cm",
    "Specimen-specific information": {
        "Lymphovascular invasion": "Present"
        }
  }
]
```

Extraction

```
{
 “SPECIMEN": [
  {
   "ID": 1,
   "Procedure": {
    "evidence": "Lobectomy",
    "answer": "Lobectomy",
    "confidence": 1
   },
   "Location": {
    "evidence": "Lung, right upper lobe",
    "answer": "Lung",
    "confidence": 1
   },  …
```

Structuring

Category	Answer
Procedure	Lobectomy
Histologic type	Adenocarcioma
Dimension of tumor	2.6
Lymphovascular invasion	Present

Figure 1. NLP Pipeline for Extracting Cancer-specific Information from Pathology Reports

RESULTS

NLP Pipeline

- The accuracy of our NLP pipeline was 98.4% for colorectal cancer, 96.5% for breast cancer, and 93.8% for lung cancer (Table 1).
- Using the integrated dataset, we analyzed the distribution of histological subtypes in 3,334 colorectal cancer patients.
- “Adenocarcinoma, not otherwise specified” was the most prevalent subtype (88.2%). Other subtypes included mucinous adenocarcinoma (5.5%), neuroendocrine tumors (3.3%), and signet-ring cell carcinoma (1.8%).

Table 1. Performance of NLP Pipeline

Type	No. of reports	No. of variables	Accuracy	Precision	Recall
Colorectum	100	1,637	98.4	98.8	99.6
Breast	100	2,614	96.5	96.7	99.9
Lung	100	1,528	93.8	94.1	99.7

Proof-of-concept study

- Figure 2 shows the Kaplan–Meier survival curves comparing overall survival between colorectal cancer patients with dMMR and those with pMMR.
- There was no significant difference in overall survival between the dMMR and pMMR groups in patients with Stage II–III colorectal cancer.

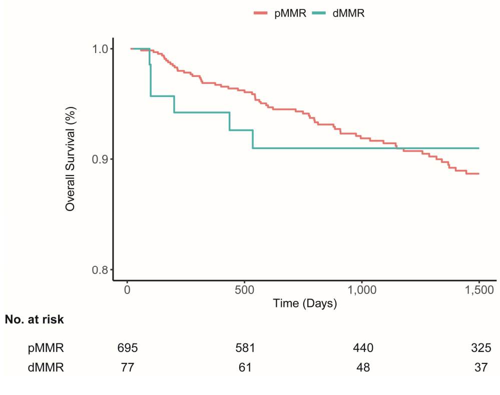


Figure 2. Overall Survival between dMMR and pMMR

CONCLUSION

- Generative LLM demonstrates feasibility in automating the extraction of structured cancer information from unstructured EHRs.
- This approach has the potential to construct robust resources for future research, significantly reducing the workload of human.
- Continued refinement and validation of this approach will be essential to ensure accuracy, generalizability, and clinical applicability in real-world settings.

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]Yonsei Institute for Digital Health, Yonsei University

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