

Building an Integrated Oncology Common Data Model Using Electronic Health Records for Real-World Cancer Research

Subin Kim^{1,2}, Jeong Eun Choi^{1,2}, Chang Jun Ko^{1,2}, Seng Chan You^{1,2}

¹Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul,
Republic of Korea

²Yonsei Institute for Digital Health, Yonsei University, Seoul, Republic of Korea

Background

The volume of healthcare data generated in oncology has increased substantially with the widespread adoption of electronic health records (EHRs).¹ These data contain rich clinical information, including diagnoses, treatments, pathology findings, and longitudinal patient histories, providing valuable opportunities for cancer research and real-world evidence generation. The development of the OMOP Common Data Model (CDM) has enabled the standardization of oncology-related data across institutions.² However, the effective use of EHR data remains challenging, as a substantial proportion of oncology-related information is stored in unstructured formats and dispersed across multiple data sources.³ To fully leverage EHR data for oncology research, unstructured information must be systematically extracted, standardized, and integrated with structured data to create a comprehensive dataset for downstream analyses.

Furthermore, institution-specific EHR systems often provide an incomplete view of the patient journey. Patients who discontinue follow-up or receive subsequent care at other institutions may no longer be captured within the local EHR, resulting in incomplete ascertainment of disease progression, recurrence, and survival outcomes. The Korean National Cancer Registry offers complementary nationwide data by systematically collecting cancer-related information across healthcare institutions. This registry captures key variables including cancer incidence, disease extent, treatment, progression, and survival outcomes, enabling more complete longitudinal follow-up and identification of cancer-related events that may not be recorded within a single institution's EHR.

Therefore, this study aimed to build a comprehensive oncology database within the OMOP CDM by extracting and standardizing oncology-specific information from unstructured EHR data and linking it with national cancer registry data to support comprehensive cancer research.

Methods

We constructed a comprehensive oncology database by integrating pathology-derived information, treatment regimen data, and national cancer registry data into an existing institutional OMOP CDM conforming to OMOP CDM version 5.4 (**Figure 1**). The study population consisted of adult patients (≥19 years) with breast, colorectal and lung cancer who visited the Severance Hospital in Korea between 2010 and 2023. Patients with benign tumors and records with invalid national cancer registry identifiers were excluded.

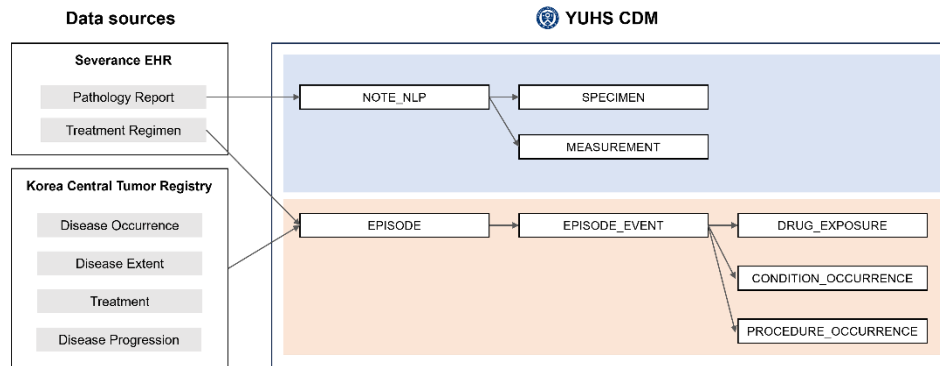


Figure 1. Construction of a Comprehensive Oncology database Derived from EHR and National Cancer Registry Data

To incorporate pathology information with the OMOP CDM, pathology reports were processed using a large language model (LLM)-based natural language processing pipeline, which was built on a previous study.⁴ The pipeline parsed pathology reports and extracted cancer-specific variables, including pathological diagnosis, tumor size, invasion status, and biomarker information, based on the College of American Pathologists cancer protocols (**Figure 2**). To evaluate the performance of the NLP pipeline, 100 pathology reports for each cancer type were randomly selected, and we calculated the variable-level accuracy of the NLP pipeline. The extracted variables were transformed into tabular data and mapped to OMOP standard concepts and subsequently loaded into the SPECIMEN and MEASUREMENT domains.

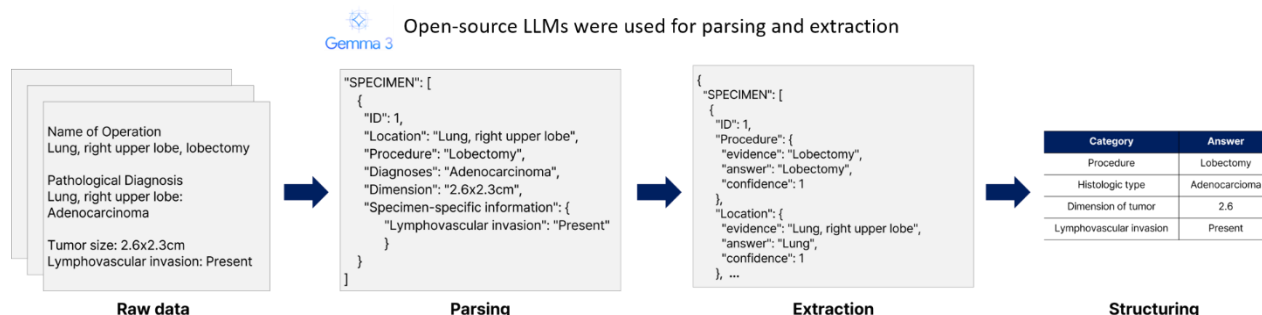


Figure 2. Natural Language Processing Pipeline for Extraction and Standardization of Pathology Information

Treatment regimen information was derived from EHRs and directly mapped to HemOnc. Longitudinal cancer registry data from the Korean National Cancer Registry, which capture disease occurrence, extent, treatment, and progression across healthcare institutions, were standardized and integrated into the database. These data were stored in the EPISODE and EPISODE_EVENT tables and linked to corresponding CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, and DRUG_EXPOSURE records. The quality of the integrated database was evaluated using the OHDSI Data Quality Dashboard (DQD), which performs standardized assessments of data conformance, completeness, and plausibility.⁵

To demonstrate the practical applications of our database, we conducted both cohort-level and patient-level analyses. First, chemotherapy treatment pathways were analyzed across the three types of

cancer using treatment regimen and line-of-therapy information. First-, second-, and third-line treatment pathways were visualized using sunburst plots. Second, we randomly selected a colorectal cancer patient and visualized the patient's longitudinal trajectory by integrating pathology findings, treatment episodes, disease progression events, laboratory measurements, and survival outcomes.

Results

A total of 54,679 patients with breast, colorectal, or lung cancer were identified from the Severance Hospital. The study population included 17,765 patients with breast cancer, 18,413 patients with colorectal cancer, and 18,501 patients with lung cancer (**Table 1**). At the time of diagnosis, 20,319 patients had stage I or earlier disease, while 9,526, 9,252, and 12,410 patients were diagnosed with stage II, III, and IV, respectively.

	No. of Patients
Total	54,679
Cancer Type	
Breast	17,765
Colorectum	18,413
Lung	18,501
Cancer stage at diagnosis	
≤I	20,319
II	9,526
III	9,252
IV	12,410
Unknown	3,172
Chemotherapy	22,557
Clinical Outcomes	
Recurrence	2,796
Death	9,205

Table 1. Clinical Characteristics and Outcomes

Pathology reports were processed using the LLM-based pipeline, generating 190,929 NOTE_NLP records and 156,536 MEASUREMENT records. The accuracy of NLP pipeline was 98.4% for colorectal cancer, 96.5% for breast cancer, and 93.8% for lung cancer. Integration of chemotherapy regimen and national cancer registry data resulted in 763,973 EPISODE records and 404,604 EPISODE_EVENT records. **Figure 3** summarizes the DQD assessment results. An overall pass rate of 97% (2,265/2,328) was achieved, with domain-specific pass rates of 99% for plausibility, 96% for conformance, and 97% for completeness.

DATA QUALITY ASSESSMENT

SEVERANCE_OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP COMMON DTATA MODEL

DataQualityDashboard Version: 2.8.6

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	509	9	518	98%	291	0	291	100%	800	9	809	99%
Conformance	851	41	892	95%	137	0	137	100%	988	41	1029	96%
Completeness	462	11	473	98%	15	2	17	88%	477	13	490	97%
Total	1822	61	1883	97%	443	2	445	100%	2265	63	2328	97%

Figure 3. The Results of Data Quality Assessment

Figure 4 illustrates the treatment pathways for each cancer type. In breast cancer, doxorubicin plus cyclophosphamide (AC) was the predominant first-line regimen, accounting for 72.2% of patients. In colorectal cancer, modified folinic acid, fluorouracil, and oxaliplatin (mFOLFOX6) was the most prescribed first-line regimen (33.7%), followed by capecitabine with radiation therapy (10.5%) and mFOLFOX6 plus bevacizumab (10.1%). Among patients who initiated treatment with mFOLFOX6, the most frequently observed second-line regimens were folinic acid, fluorouracil, and irinotecan (FOLFIRI)-based therapies. For lung cancer patients, the most frequently prescribed regimens were carboplatin and paclitaxel with radiation therapy (10.8%) and cisplatin and vinorelbine (10.6%).

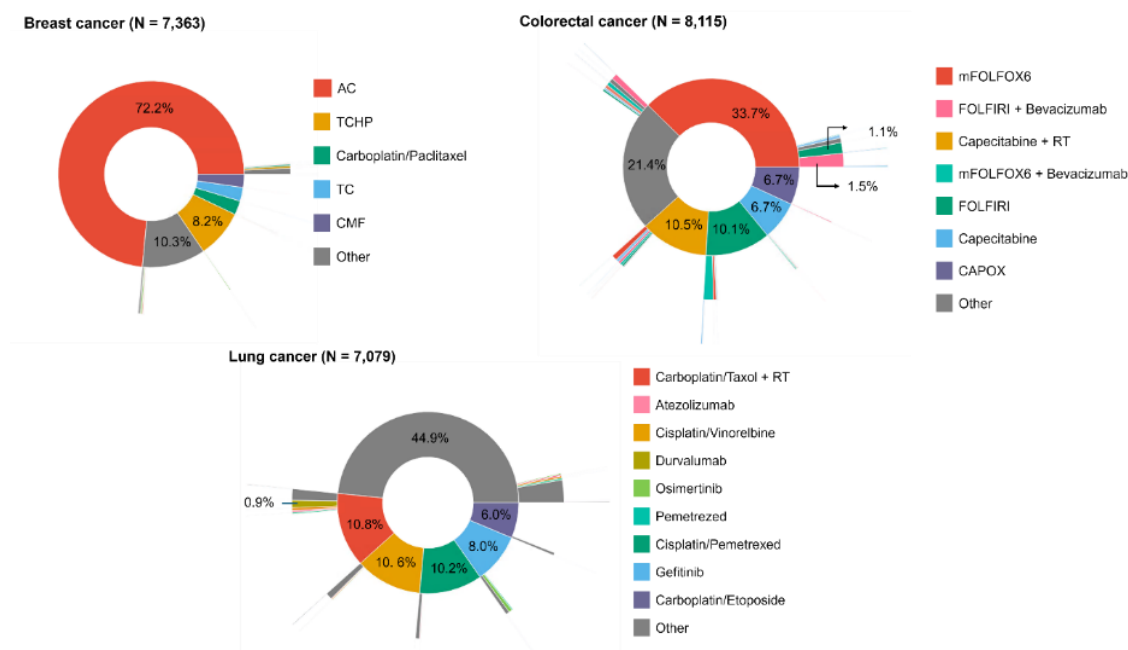


Figure 4. Treatment Regimen Patterns in Patients with Breast, Colorectal and Lung Cancer

We further visualized the longitudinal trajectory of a randomly selected colorectal cancer patient (**Figure 5**). This patient was diagnosed with stage IVB adenocarcinoma in sigmoid colon, developed metastatic progression involving the lung and liver, received FOLFIRI chemotherapy, and died 125 days after diagnosis.

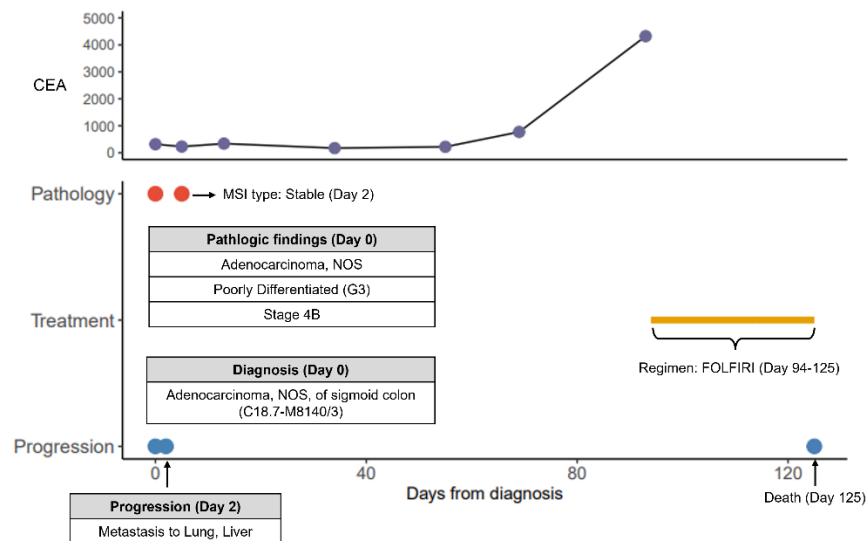


Figure 5. Example of Longitudinal Patient Trajectory

Conclusion

By integrating EHR-derived pathology information, chemotherapy regimen data, and national cancer registry records, we built an oncology database following the OMOP CDM. The resulting database supports longitudinal characterization of cancer patients from diagnosis through treatment and outcome and may facilitate large-scale observational studies and real-world evidence generation in cancer research.

References

1. Abernethy, A. P., Etheredge, L. M., Ganz, P. A., Wallace, P., German, R. R., Neti, C., ... & Murphy, S. B. (2010). Rapid-learning system for cancer care. *Journal of Clinical Oncology*, 28(27), 4268-4274.
2. Belenkaya, R., Gurley, M. J., Golozar, A., Dymshyts, D., Miller, R. T., Williams, A. E., ... & Reich, C. (2021). Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clinical Cancer Informatics*, 5, 12-20.
3. Penberthy, L. T., Rivera, D. R., Lund, J. L., Bruno, M. A., & Meyer, A. M. (2022). An overview of real-world data sources for oncology and considerations for research. *CA: A Cancer Journal for Clinicians*, 72(3), 287-300.
4. Park, J., You, S. C., Jeong, E., Weng, C., Park, D., Roh, J., ... & Park, R. W. (2021). A framework (SOCRAteX) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. *JMIR medical informatics*, 9(3), e23983.
5. Blacketer, C., Defalco, F. J., Ryan, P. B., & Rijnbeek, P. R. (2021). Increasing trust in real-world evidence

through evaluation of observational data quality. *Journal of the American Medical Informatics Association*, 28(10), 2251-2257.