Original Paper

Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer

Borim Ryu[1], MSc; Eunsil Yoon[1], MSc; Seok Kim[1], MPH; Sejoon Lee[2], PhD; Hyunyoung Baek[1], MSc; Soyoung Yi[1], BSc; Hee Young Na[2], BSc; Ji-Won Kim[3], MD, PhD; Rong-Min Baek[4], MD, PhD; Hee Hwang[1], MD; Sooyoung Yoo[1], PhD

[1]Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[2]Department of Pathology and Translational Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[3]Division of Hematology and Medical Oncology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[4]Department of Plastic Surgery, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

Corresponding Author:
Sooyoung Yoo, PhD
Office of eHealth Research and Business
Seoul National University Bundang Hospital
82 173rd Street, Gumi-ro, Bundang-gu
Seongnam, 13620
Republic of Korea
Phone: 82 317878980
Email: yoosoo0@gmail.com

Abstract

Background: Common data models (CDMs) help standardize electronic health record data and facilitate outcome analysis for observational and longitudinal research. An analysis of pathology reports is required to establish fundamental information infrastructure for data-driven colon cancer research. The Observational Medical Outcomes Partnership (OMOP) CDM is used in distributed research networks for clinical data; however, it requires conversion of free text–based pathology reports into the CDM's format. There are few use cases of representing cancer data in CDM.

Objective: In this study, we aimed to construct a CDM database of colon cancer–related pathology with natural language processing (NLP) for a research platform that can utilize both clinical and omics data. The essential text entities from the pathology reports are extracted, standardized, and converted to the OMOP CDM format in order to utilize the pathology data in cancer research.

Methods: We extracted clinical text entities, mapped them to the standard concepts in the Observational Health Data Sciences and Informatics vocabularies, and built databases and defined relations for the CDM tables. Major clinical entities were extracted through NLP on pathology reports of surgical specimens, immunohistochemical studies, and molecular studies of colon cancer patients at a tertiary general hospital in South Korea. Items were extracted from each report using regular expressions in Python. Unstructured data, such as text that does not have a pattern, were handled with expert advice by adding regular expression rules. Our own dictionary was used for normalization and standardization to deal with biomarker and gene names and other ungrammatical expressions. The extracted clinical and genetic information was mapped to the Logical Observation Identifiers Names and Codes databases and the Systematized Nomenclature of Medicine (SNOMED) standard terminologies recommended by the OMOP CDM. The database-table relationships were newly defined through SNOMED standard terminology concepts. The standardized data were inserted into the CDM tables. For evaluation, 100 reports were randomly selected and independently annotated by a medical informatics expert and a nurse.

Results: We examined and standardized 1848 immunohistochemical study reports, 3890 molecular study reports, and 12,352 pathology reports of surgical specimens (from 2017 to 2018). The constructed and updated database contained the following extracted colorectal entities: (1) NOTE_NLP, (2) MEASUREMENT, (3) CONDITION_OCCURRENCE, (4) SPECIMEN, and (5) FACT_RELATIONSHIP of specimen with condition and measurement.

Conclusions: This study aimed to prepare CDM data for a research platform to take advantage of all omics clinical and patient data at Seoul National University Bundang Hospital for colon cancer pathology. A more sophisticated preparation of the pathology data is needed for further research on cancer genomics, and various types of text narratives are the next target for additional research on the use of data in the CDM.

(J Med Internet Res 2020;22(12):e18526) doi: 10.2196/18526

KEYWORDS

common data model; natural language processing; oncology module; colon cancer; electronic health record; oncology; pathology; clinical data

Introduction

Colorectal cancer is the third most common cancer in the world after lung cancer and breast cancer, and the second most common cause of cancer deaths in the world after lung cancer [1]. In addition, the incidence of colorectal cancer in Korea is continuously increasing owing to the westernization of diet and the widespread use of colonoscopy [2]. To determine treatment and prognosis, clinical and pathologic staging are both crucial. Pathology reports vary in format worldwide. The heterogeneity of pathology reports is not unique to colorectal cancer; there has been a growing need to standardize pathology reports [3]. Pathologic diagnosis is based on gross examination, microscopic examination, and sometimes molecular testing. Although some hospitals report molecular testing results in a structured format, biomarker results in the pathology report are usually recorded as unstructured free text or template-based semistructured text in electronic health record (EHR) systems. This unstructured document must be converted into a structure that can be processed by a computer.

Unstructured clinical narratives may summarize patients' medical history, diagnoses, medications, immunizations, allergies, radiology images, and laboratory test results in the form of progress notes, discharge reports, etc, providing a valuable resource for computational phenotyping [4]. Previous studies have applied controlled vocabularies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) collection to recognize various expressions of the same medical concepts in pathology and used the Unified Medical Language System (UMLS) as a metathesaurus [5-7]. Other significant efforts have been devoted to the implementation of open-source, standards-based systems to improve the portability of EHR-based phenotype definitions (eg, eMERGE [Electronic Medical Records and Genomics] [8] and PhEMA [Phenotype Execution and Modeling Architecture] [9]).

Common data models (CDMs) are healthcare data models with a standard structure. An example is the Observational Medical Outcomes Partnership (OMOP) CDM, adopted and maintained by Observational Health Data Sciences and Informatics (OHDSI) [10-13]. OHDSI is an open-science community that aims to improve health by empowering the community to collaboratively generate evidence that promotes better health decisions and care [14]. OHDSI conducts methodological research to establish scientific best practices for the appropriate use of OMOP CDM data and develops open-source analytics software for research use. The OHDSI oncology working group incorporated fundamental structural and semantic support into the OMOP CDM to represent clinical cancer disease and treatment data, significantly improving the specificity of cancer cohort definitions. To represent cancer diagnoses using the combination of histology and topography in the OMOP CDM Condition domain (Condition_Occurrence), without changes to the existing structure, they have proposed a precoordinated collapse of the International Classification of Diseases for Oncology (ICD-O) axes—histology and topography—to a single OMOP-originated concept representing a unique cancer diagnosis, which preserves linkages between these single codes and the ICD-O axes in the OMOP standard vocabulary.

To transform EHR data into the OMOP CDM format, data formalization and vocabulary mapping should be performed in advance. Thus, most structured clinical EHR data, such as diagnosis, medications, lab tests, and vital signs, are usually converted to the CDM first. However, a large amount of data in EHRs—including pathology reports—is recorded and stored in an unstructured or semistructured form. In particular, it is essential to extract and refine major clinical entities from the pathology reports through natural language processing (NLP) to utilize both clinical and genomics data in CDM-based cancer research.

In this study, we transformed and incorporated the colorectal cancer pathology reports of a tertiary general university hospital into the OMOP CDM format by developing an NLP module. The primary objective of this research is to (1) extract major biomarker entities from the pathology report, (2) convert them into the OMOP CDM format by mapping the vocabulary to standardized terminology, and (3) demonstrate the ability of the OMOP CDM to represent biomarker data using the OMOP CDM oncology extension module.

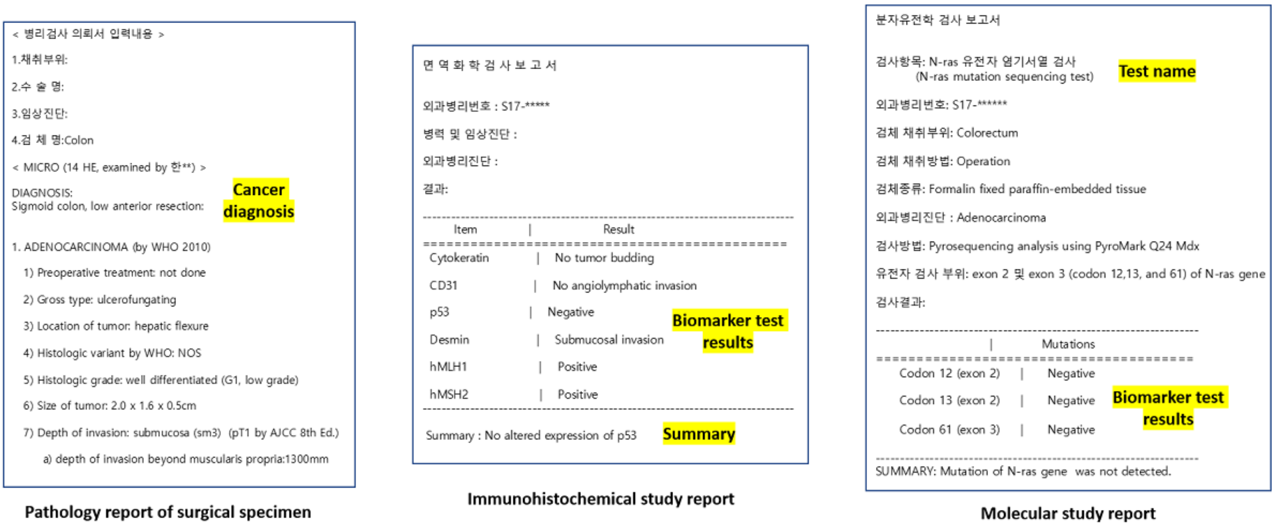
Methods

Study Data

The study site, Seoul National University Bundang Hospital (SNUBH), has converted EHR data from a 15-year period—from the opening of SNUBH until December 2018—into the OMOP CDM format. The CDM database contains data on more than 1.7 million patients. Although patient data such as basic demographic information, medical history, family history, diagnosis, drug exposures, test results, vital signs, surgeries, and procedures were converted into the OMOP CDM standard, free-text results, such as pathology reports, are stored in the form of free text in a table called NOTE. It is necessary to extract these data from the reports through text processing techniques for use in clinical research. In this study, the pathology reports of patients diagnosed with colorectal cancer from 2017 to 2018 were processed to investigate the ability of the OMOP CDM Oncology module to represent three types of pathology reports: pathology reports of surgical specimens, immunohistochemical study reports, and molecular study reports. This study was approved for exemption by the SNUBH Institutional Review Board.

The original data were derived from the NOTE table of the SNUBH OMOP CDM database and included the original free-text data of pathology reports of surgical specimens, immunohistochemical study reports, and molecular study reports (Figure 1). Surgical number, biomarker name, test result, and summary information were extracted from the immunohistochemical study tests. In pathology reports, a molecular study is a method of analyzing protein expression in cells; in this study, surgical number, test name, gene name, mutation examination result, and summary information were extracted from molecular study tests. The surgical pathology report is an essential part of patient care because it documents the pathologic findings in tissues removed from patients for diagnostic or therapeutic reasons [15]. Diagnosis names were extracted from pathology reports of surgical specimens.

Figure 1. Contents of information to extract by report type.



```
< 병리검사 의뢰서 입력내용 >
1.채취부위:
2.수 술 명:
3.임상진단:
4.검 체 명:Colon
< MICRO (14 HE, examined by 한**) >
DIAGNOSIS:
Sigmoid colon, low anterior resection:

1. ADENOCARCINOMA (by WHO 2010)
   1) Preoperative treatment: not done
   2) Gross type: ulcerofungating
   3) Location of tumor: hepatic flexure
   4) Histologic variant by WHO: NOS
   5) Histologic grade: well differentiated (G1, low grade)
   6) Size of tumor: 2.0 x 1.6 x 0.5cm
   7) Depth of invasion: submucosa (sm3)  (pT1 by AJCC 8th Ed.)
      a) depth of invasion beyond muscularis propria:1300mm
```

Cancer diagnosis

Pathology report of surgical specimen

```
면역화학검사보고서

외과병리번호 : S17-*****
병력 및 임상진단 :
외과병리진단 :
결과:
--------------------------------------------------------------
     Item          |              Result
==============================================================
 Cytokeratin       |  No tumor budding
 CD31              |  No angiolymphatic invasion
 p53               |  Negative
 Desmin            |  Submucosal invasion
 hMLH1             |  Positive
 hMSH2             |  Positive
--------------------------------------------------------------
 Summary : No altered expression of p53
```

Biomarker test results

Summary

Immunohistochemical study report

```
분자유전학 검사 보고서

검사항목: N-ras 유전자 염기서열 검사
      (N-ras mutation sequencing test)
외과병리번호: S17-******
검체 채취부위: Colorectum
검체 채취방법: Operation
검체종류: Formalin fixed paraffin-embedded tissue
외과병리진단 : Adenocarcinoma
검사방법: Pyrosequencing analysis using PyroMark Q24 Mdx
유전자 검사 부위: exon 2 및 exon 3 (codon 12,13, and 61) of N-ras gene
검사결과:
-------------------------------------------------------------
                   |          Mutations
=============================================================
   Codon 12 (exon 2)   |   Negative
   Codon 13 (exon 2)   |   Negative
   Codon 61 (exon 3)   |   Negative
-------------------------------------------------------------
SUMMARY: Mutation of N-ras gene  was not detected.
```

Test name

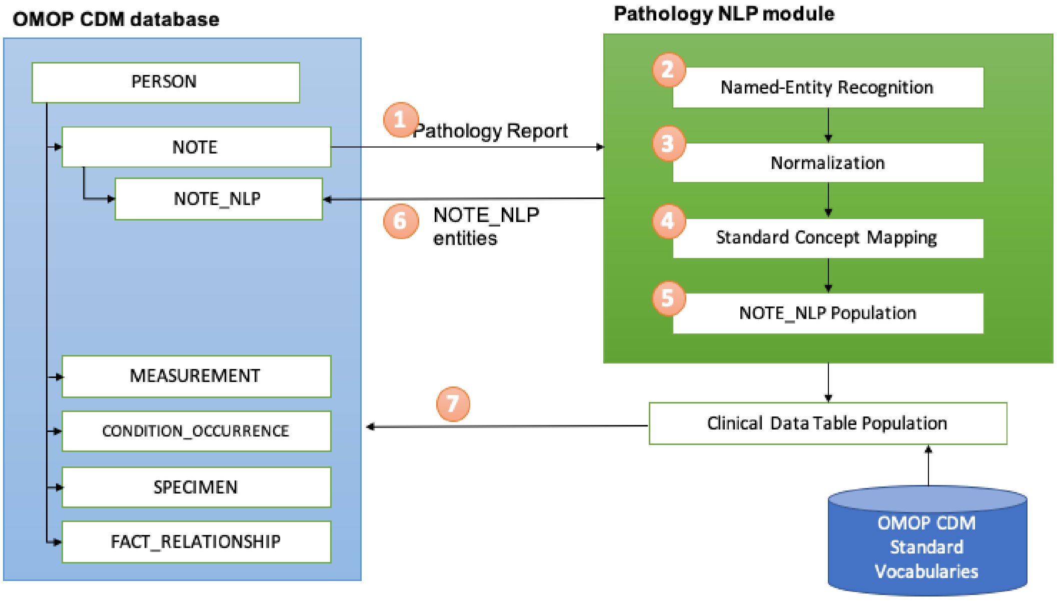
Biomarker test results

Molecular study report

Processing Model for Rule-Based Text Extraction

The types of key information recorded in pathology reports vary based on the pathological tests conducted. We first reviewed the format and text entities of the extracted original report and selected the target entities and values for information extraction. Rule-based text processing based on regular expressions was developed to extract the clinical entities of the test subject and result information depending on the type of report. In the immunohistochemical study reports and molecular study reports, target gene or protein information and the test result values were recognized and extracted, whereas pathology reports of surgical specimens were processed for information on the diagnosis and the region that was operated on. This rule-based approach is ideally suited for the retrieval of unstructured pathologic clinical entities and test result values in the SNUBH EHR notes. Figure 2 shows the NLP module developed in this study and depicts the overall study process, with the numbers in the circles indicating the sequence of steps.

Figure 2. Overall process of this study. CDM: common data model. NLP: natural language processing. OMOP: Observational Medical Outcomes Partnership.



The measurement value extraction system was packaged through a regular expression–based Python NLP pipeline using a method similar to that employed in previous research [16]. Our system recognizes clinical entities such as the biomarkers in a test and their values. Our system performs three broad steps: (1) it identifies the measurement name, such as gene or protein name, in the text; (2) it identifies the measurement value in the text; and (3) it transforms the appropriate value expressions according to their relationship. For immunohistochemical and molecular study tests, rules were created for extracting the target gene name, protein name information, and test result value. For example, surgical number, biomarker name, test result, and summary information were extracted from the immunohistochemical study tests (example in Figure 1). Text entities for a test result could include expressions such as “EGFR (GI)” (epidermal growth factor receptor; gastrointestinal) and “1+/3.” The term “EGFR (GI)” is regularized into “EGFR” and “epidermal growth factor receptor Ag [presence] in tissue by immune stain”. Through this clinical entity recognition, we could structure data into tables that can be transformed into the NOTE_NLP table after term standardization in the CDM (Table 1 and Figure 1).

Several modifiers were added to enter data extracted from the NOTE_NLP table. For the immunohistochemical study report and the molecular study test report, the raw text of the test results was entered in the value_as_narrative column, and the information on whether the normalized value was a negation expression and the concept_id were included in the value_as_concept_id column. The names of each report were added to the section_source_value column, and inspection item information was included for molecular study test reports. In addition, the surgical pathology number was added to the sub_id column to be linked to surgical specimen pathology reports.

We used Python version 3.6 and regular expressions to process three types of colon cancer pathology reports according to each report’s characteristics. For example, the specimen and diagnosis names were extracted from surgical specimen pathology reports, with the specimen name written in each report as “SAMPLE NAME: colon,” using a regular expression to extract the items after “SAMPLE NAME:.” Likewise, the diagnosis name is found after the entity expression “DIAGNOSIS:.” From the immunohistochemical study report, items following “SUMMARY:” and prior to “EXAMINER:” were extracted as summary information. In addition, biomarker results from the immunohistochemical study and molecular study pathology reports were extracted in a semistructured text table with continuous values such as hyphens (-). To deal with this text, data fields were separated based on the continuous value of hyphen (-), with biomarker name on the left of the vertical bar (|) and result value on the right. Refer to Multimedia Appendix 1 for processing rules (Table S1).

Table 1. NOTE_NLP table for pathology report data.

NOTE_NLP field	Surgical pathology report	Immunochemistry report	Molecular study report
NOTE_NLP_ID	12345	23456	34567
NOTE_ID	1	2	3
SNIPPET	ductal adenocarcinoma, moderately differentiated > poorly differentiated	EGFR (GI) | 1+/3	Codon 12 (exon 2) | Positive
OFFSET	560,632	170,188	200,231
SECTION_CONCEPT_ID	3025891 (Pathology report final diagnosis narrative)	40758358 (Immune stain study)	3001274 (NRAS gene mutations found [identifier] in blood or tissue by molecular genetics method nominal)
LEXICAL_VARIANT	Adenocarcinoma	EGFR	Codon 12 (exon 2)
NOTE_NLP_CONCEPT_ID	0	3016231 (Epidermal growth factor receptor Ag [presence] in tissue by immune stain)	0
NOTE_NLP_SOURCE_CONCEPT_ID	44498791 (Tubular adenoma, NOS[a])	0	0
TERM_MODIFIERS	Negated=FALSE; sub_id=S 110023456; section_source_value=Pathology reports of surgical specimen	Negated=FALSE; value_as_concept_id=9191; value_as_narrative=1+/3; sub_id=S 120034567; section_source_value=Immunochemistry test report	Negated=FALSE; value_as_concept_id=9191; value_as_narrative=Missense mutation [c.38G>A, p.Gly14Asp]; sub_id=S 130045678; section_source_value=N-ras Gene sequencing test

[a]NOS: not otherwise specified.

Vocabulary Standardization for Pathology Reports

To standardize the extracted pathologic data, protein or gene laboratory tests and pertinent test result values were mapped into the OMOP standard concept vocabulary and reviewed by a clinician, a nurse, and a bioinformatics expert. The test names were mapped to standard concepts of Logical Observation Identifiers Names and Codes (LOINC) in the MEASUREMENT domain, and the result values were normalized to those concepts (eg, categorized as positive or negative).

According to the OMOP CDM Oncology module, information on the cancer diagnosis and the region operated on, taken from the surgical specimen pathology reports, was mapped to the ICD-O, Third Edition, and then stored in the CONDITION_OCCURRENCE table with ICD-O-3 and its mapped SNOMED CT codes.

To map the extracted clinical and gene information to a standard concept_id, we constructed a dictionary table for the test names and result values that appear in the free-text pathology reports. For example, in immunohistochemical study reports, the test name for EGFR is concept_id 3016231 (Epidermal growth factor receptor Ag [presence] in tissue by immune stain), which is the LOINC code 32581-1 with a positive or negative result value. The result value of “1+/3” was normalized and mapped to the concept_id of 10828004 (Positive), which is SNOMED CT code 9191. The molecular study test report comprised results of four types of examinations: *NRAS* (OMIM 164790) mutation, *KRAS* (OMIM 190070) mutation, *BRAF* (OMIM 164757) mutation, and microsatellite instability. For each test result value, the normalization rules for test result values were separately defined and coded, as shown in Table 2.

Table 2. Defined mapping rules for OMOP CDM concept_id, LOINC code, and test results from the molecular study pathology reports.

Test name	CONCEPT_ID	LOINC code	RESULT_Attribute	Normalization rules for results value concept
NRAS mutation	3001274	21719-0	Codon 12 (exon 2), Codon 13 (exon 2), Codon 61 (exon 3)	If 1 or more positive values are present, treat as total positive
KRAS mutation	36203353	85509-8	Codon 12 (exon 2), Codon 13 (exon 2), Codon 61 (exon 3)	If 1 or more positive values are present, treat as total positive
MSI[a]	3047348	43368-0	BAT26, BAT25, D5S346, D17S250, D2S123	Categorized as MSI-H (high) if more than 1 value is positive, MSI-L (low) if 1 value is positive, and MSS[b] if no positive value is present
BRAF mutation	40761583	58483-9	V600E, K601E	If more than 1 positive value is present, treat as total positive

[a] MSI: microsatellite instability.

[b] MSS: microsatellite stable.

A total of 78 biomarkers were used in immunohistochemical reports, most of which are measurement domains, and the concept_id for each was retrieved and mapped from the OMOP CDM. However, there were 3 undetected biomarkers that were used with their own concept_id (3/78, 4%). Refer to Multimedia Appendix 2 for a list of mapped concepts (Table S2).

After term mapping, the NOTE_NLP table of the CDM was derived. To integrate the extracted data into the existing CDM data, the main results of immunohistochemical and molecular study tests were updated in the MEASUREMENT table. In addition, for the cancer diagnosis from surgical specimen pathology reports, relevant information was added to the CONDITION_OCCURRENCE table. To define the relationship between the MEASUREMENT, CONDITION_OCCURRENCE, and SPECIMEN tables, we populated the CDM FACT_RELATIONSHIP table. With the two existing concepts, “Specimen source identity (SNOMED)” and “Has specimen (SNOMED),” we could define the target specimen and its attribute.

System Validation

Two experts manually reviewed the transformed data in the NOTE_NLP table in 100 randomly selected pathology reports. To verify the accuracy of the extracted values for each table item, the nurse and medical information specialist manually checked the original pathology document and extracted text values. Furthermore, two reviewers manually reviewed the original document and checked the results of the processing rules.

Results

Major Text-Entity Extraction and Mapping Onto International Standard Terminology

Pathologic clinical information on colorectal cancer was extracted through NLP of immunohistochemical studies, molecular studies, and surgical specimen pathology reports to construct a CDM-formatted NOTE_NLP table. NLP extraction and terminology standardization were performed for 1848 immunohistochemical study reports, 3890 molecular study reports, and 12,352 surgical specimen pathology reports. The number of items, such as protein or gene name values, extracted from the immunohistochemical study reports was 6092, and the number extracted from the molecular study reports was 13,953; the summarized essential information was then updated in the existing MEASUREMENT table. Cancer diagnosis entities from the surgical specimen pathology reports (25,902) were delivered to the CONDITION_OCCURRENCE table. Figure 3 shows the number of each biomarker extracted from molecular study pathology reports by result, and Figure 4 shows the frequency of each diagnosis name and its surgical regions as a heatmap plot. *KRAS* mutations, *NRAS* mutations, *BRAF* mutations, and microsatellite instability are tested in the hospital. For each test result value, the normalization rules for test result values were separately defined and coded as described in Table 2. The majority of the test results are “negative” or “microsatellite instability high.” In Table 3, we included evaluation results for each attribute extracted and transformed into standard terminology concepts for each pathology report. The results of the attribute value extraction task are quite good. The attributes derived from surgical specimen pathology and molecular study reports were recognized 100% correctly among randomly selected test reports. “No endolymphatic tumor emboli” or “no lymphoepithelial lesion” values were recognized as positive in a few immunohistochemical study entities. This is because we have created rules that focus on processing the term “negative” when we process biomarker text.

According to Figure 4, there were 11 colon surgical regions at SNUBH from 2017 to 2018, and the most frequent diagnosis names were “tubular adenoma, nos” (not otherwise specified) and “adenocarcinoma, nos.” Positive and negative biomarker results extracted from immunohistochemical study reports are shown in Table 4.

Figure 3. Each biomarker and its results extracted from molecular study pathology reports.

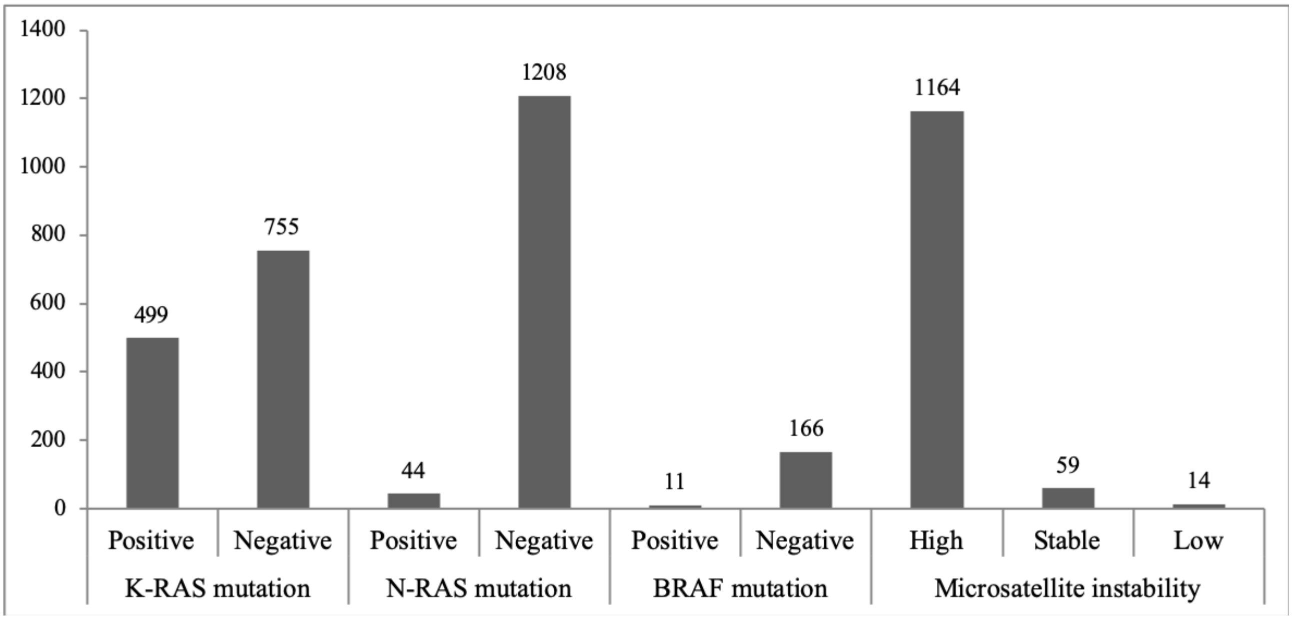


Figure 4. Frequency distribution in surgical pathology reports. NOS: not otherwise specified.

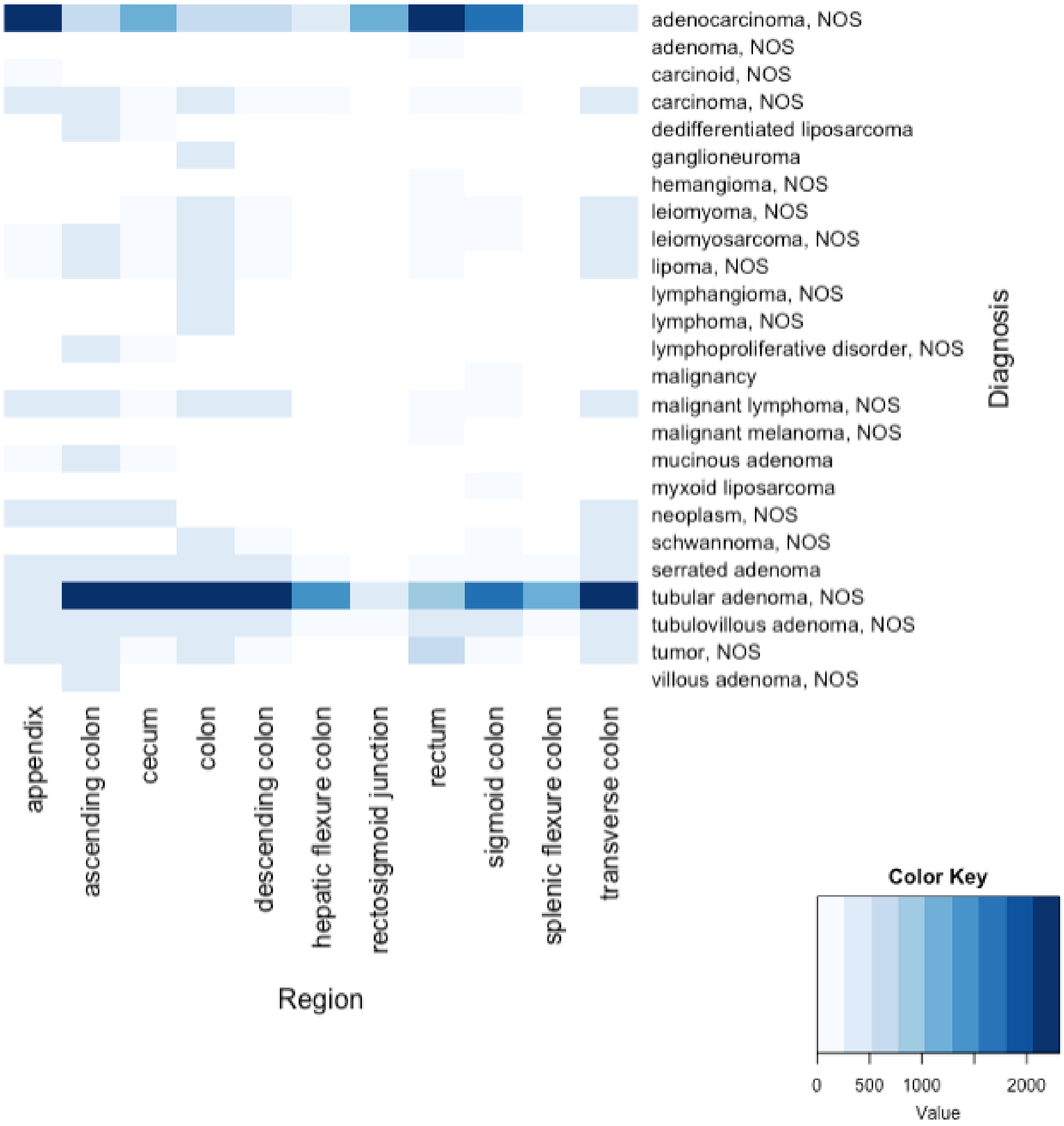


Table 3. Evaluation results for attributes found in 100 pathology test reports.

Type of report	Attribute example	Entities, N	Precision, n (%)	Recall, n (%)	F-measure
Surgical specimen pathology report	ductal adenocarcinoma, moderately differentiated > poorly differentiated	109	109 (100)	109 (100)	100
Immunohistochemical study	EGFR (GI) | 1+/3	367	362 (98.6)	367 (100)	99.31
Molecular study	Codon 12 (exon 2) | Positive	100	100 (100)	100 (100)	100

Table 4. Top 20 biomarker results from immunohistochemistry tests.

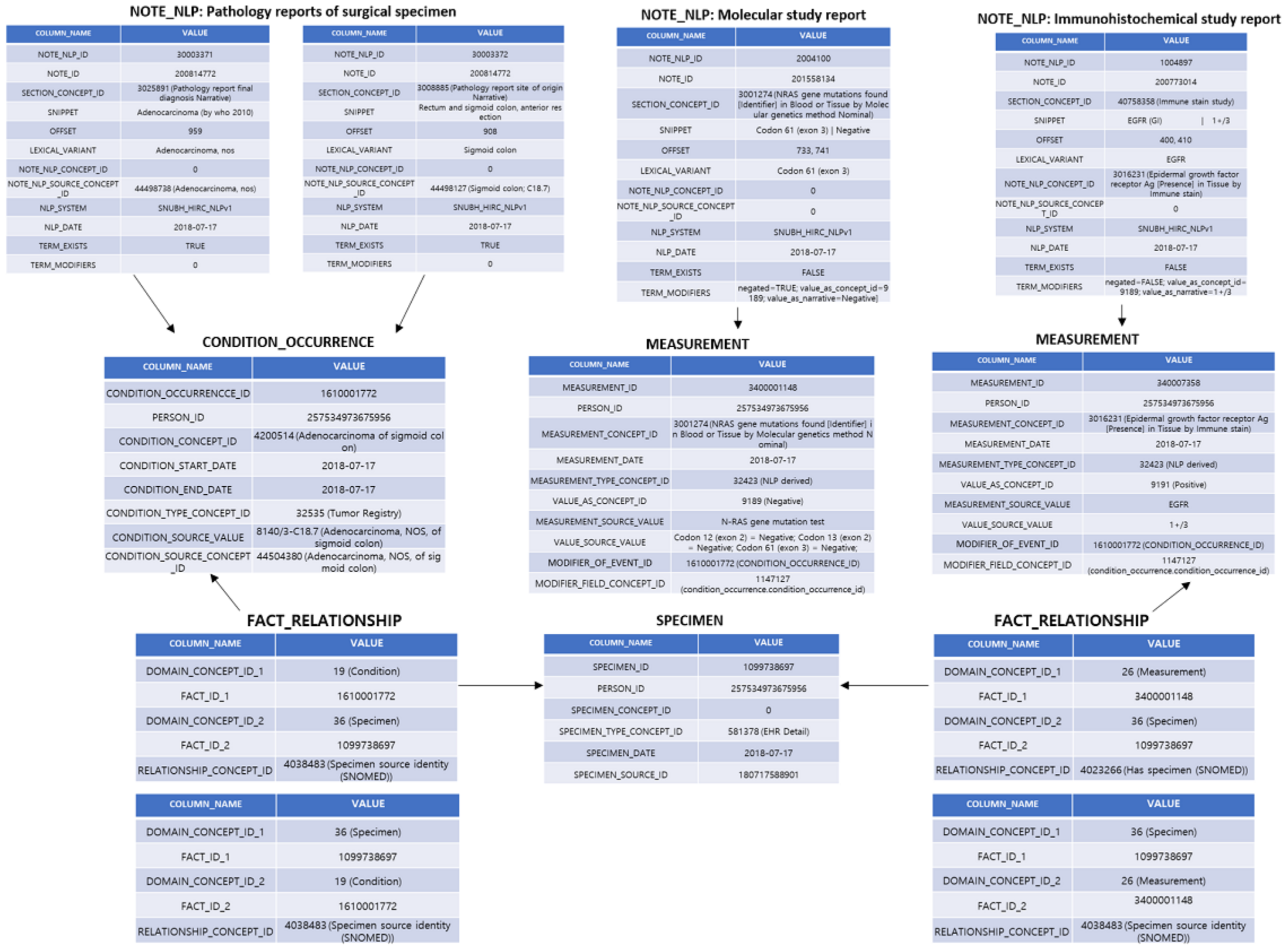
CONCEPT_ID	CONCEPT_NAME	Positive (N=4626), n (%)	Negative (N=2096), n (%)
3016231	EGFR	1360 (29.4)	33 (1.6)
21493968	hMLH1	361 (7.8)	25 (1.2)
3046605	Ki-67	359 (7.8)	0 (0.0)
21493983	PTEN	261 (5.6)	76 (3.6)
3017031	p53	111 (2.4)	212 (10.1)
21493982	BRAF	12 (0.3)	310 (14.8)
3027870	CD3	153 (3.3)	160 (7.6)
3019066	C-erbB2	75 (1.6)	194 (9.3)
3026213	CD20	171 (3.7)	37 (1.8)
21493969	hMSH2	185 (4.0)	18 (0.9)
3051327	bcl-6	100 (2.2)	56 (2.7)
21493970	hMSH6	137 (3.0)	18 (0.9)
21493971	PMS2	133 (2.9)	22 (1.1)
21492142	Cyclin D1	0 (0.0)	151 (7.2)
3002495	Desmin	132 (2.9)	18 (0.9)
3052827	CD8	140 (3.0)	8 (0.4)
3041284	CD10	86 (1.9)	60 (2.9)
3040360	Cytokeratin	79 (1.7)	67 (3.2)
3006921	Synaptophysin	110 (2.4)	32 (1.5)
3032734	MUM-1	48 (1.0)	55 (2.6)

Database and Relationships Following the OMOP CDM Format

The CDM’s NOTE_NLP table is composed of three types of pathology reports; the major entities extracted from the immunohistochemical study reports and molecular study reports contain result summaries in the CDM’s MEASUREMENT table. Surgical site and cancer diagnosis information extracted from the surgical specimen pathology reports was constructed by populating the cancer diagnosis information in the patient’s CONDITION_OCCURRENCE table.

The FACT_RELATIONSHIP table in the OMOP CDM contains records about the relationships between facts stored as records in any table of the CDM. Relationships can be defined between facts from the same domain or different domains. Examples of fact relationships include personal relationships (parent-child), care site relationships (hierarchical organizational structure of facilities within a health system), indication relationships (between drug exposures and associated conditions), usage relationships (of devices during the course of an associated procedure), or facts derived from one another (measurements derived from an associated specimen). To define the relationship between the MEASUREMENT, CONDITION_OCCURRENCE, and SPECIMEN tables, the FACT_RELATIONSHIP table of the CDM was populated (Figure 5).

Figure 5. Overall data table relationships developed by the study.



NOTE_NLP: Pathology reports of surgical specimen

COLUMN_NAME	VALUE
NOTE_NLP_ID	30003371
NOTE_ID	200814772
SECTION_CONCEPT_ID	3025891 (Pathology report final diagnosis Narrative)
SNIPPET	Adenocarcinoma (by who 2010)
OFFSET	959
LEXICAL_VARIANT	Adenocarcinoma, nos
NOTE_NLP_CONCEPT_ID	0
NOTE_NLP_SOURCE_CONCEPT_ID	44498738 (Adenocarcinoma, nos)
NLP_SYSTEM	SNUBH_HIRC_NLPv1
NLP_DATE	2018-07-17
TERM_EXISTS	TRUE
TERM_MODIFIERS	0

COLUMN_NAME	VALUE
NOTE_NLP_ID	30003372
NOTE_ID	200814772
SECTION_CONCEPT_ID	3008885 (Pathology report site of origin Narrative)
SNIPPET	Rectum and sigmoid colon, anterior resection
OFFSET	908
LEXICAL_VARIANT	Sigmoid colon
NOTE_NLP_CONCEPT_ID	0
NOTE_NLP_SOURCE_CONCEPT_ID	44498127 (Sigmoid colon; C18.7)
NLP_SYSTEM	SNUBH_HIRC_NLPv1
NLP_DATE	2018-07-17
TERM_EXISTS	TRUE
TERM_MODIFIERS	0

NOTE_NLP: Molecular study report

COLUMN_NAME	VALUE
NOTE_NLP_ID	2004100
NOTE_ID	201558134
SECTION_CONCEPT_ID	3001274 (NRAS gene mutations found [Identifier] in Blood or Tissue by Molecular genetics method Nominal)
SNIPPET	Codon 61 (exon 3) | Negative
OFFSET	733, 741
LEXICAL_VARIANT	Codon 61 (exon 3)
NOTE_NLP_CONCEPT_ID	0
NOTE_NLP_SOURCE_CONCEPT_ID	0
NLP_SYSTEM	SNUBH_HIRC_NLPv1
NLP_DATE	2018-07-17
TERM_EXISTS	FALSE
TERM_MODIFIERS	negated=TRUE; value_as_concept_id=9189; value_as_narrative=Negative)

NOTE_NLP: Immunohistochemical study report

COLUMN_NAME	VALUE
NOTE_NLP_ID	1004897
NOTE_ID	200773014
SECTION_CONCEPT_ID	40758358 (Immune stain study)
SNIPPET	EGFR (GI) | 1+/3
OFFSET	400, 410
LEXICAL_VARIANT	EGFR
NOTE_NLP_CONCEPT_ID	3016231 (Epidermal growth factor receptor Ag [Presence] in Tissue by Immune stain)
NOTE_NLP_SOURCE_CONCEPT_ID	0
NLP_SYSTEM	SNUBH_HIRC_NLPv1
NLP_DATE	2018-07-17
TERM_EXISTS	FALSE
TERM_MODIFIERS	negated=FALSE; value_as_concept_id=9189; value_as_narrative=1+/3

CONDITION_OCCURRENCE

COLUMN_NAME	VALUE
CONDITION_OCCURRENCCE_ID	1610001772
PERSON_ID	257534973675956
CONDITION_CONCEPT_ID	4200514 (Adenocarcinoma of sigmoid colon)
CONDITION_START_DATE	2018-07-17
CONDITION_END_DATE	2018-07-17
CONDITION_TYPE_CONCEPT_ID	32535 (Tumor Registry)
CONDITION_SOURCE_VALUE	8140/3-C18.7 (Adenocarcinoma, NOS, of sigmoid colon)
CONDITION_SOURCE_CONCEPT_ID	44504380 (Adenocarcinoma, NOS, of sigmoid colon)

MEASUREMENT

COLUMN_NAME	VALUE
MEASUREMENT_ID	3400001148
PERSON_ID	257534973675956
MEASUREMENT_CONCEPT_ID	3001274 (NRAS gene mutations found [Identifier] in Blood or Tissue by Molecular genetics method Nominal)
MEASUREMENT_DATE	2018-07-17
MEASUREMENT_TYPE_CONCEPT_ID	32423 (NLP derived)
VALUE_AS_CONCEPT_ID	9189 (Negative)
MEASUREMENT_SOURCE_VALUE	N-RAS gene mutation test
VALUE_SOURCE_VALUE	Codon 12 (exon 2) = Negative; Codon 13 (exon 2) = Negative; Codon 61 (exon 3) = Negative;
MODIFIER_OF_EVENT_ID	1610001772 (CONDITION_OCCURRENCE_ID)
MODIFIER_FIELD_CONCEPT_ID	1147127 (condition_occurrence.condition_occurrence_id)

MEASUREMENT

COLUMN_NAME	VALUE
MEASUREMENT_ID	340007358
PERSON_ID	257534973675956
MEASUREMENT_CONCEPT_ID	3016231 (Epidermal growth factor receptor Ag [Presence] in Tissue by Immune stain)
MEASUREMENT_DATE	2018-07-17
MEASUREMENT_TYPE_CONCEPT_ID	32423 (NLP derived)
VALUE_AS_CONCEPT_ID	9191 (Positive)
MEASUREMENT_SOURCE_VALUE	EGFR
VALUE_SOURCE_VALUE	1+/3
MODIFIER_OF_EVENT_ID	1610001772 (CONDITION_OCCURRENCE_ID)
MODIFIER_FIELD_CONCEPT_ID	1147127 (condition_occurrence.condition_occurrence_id)

FACT_RELATIONSHIP

COLUMN_NAME	VALUE
DOMAIN_CONCEPT_ID_1	19 (Condition)
FACT_ID_1	1610001772
DOMAIN_CONCEPT_ID_2	36 (Specimen)
FACT_ID_2	1099738697
RELATIONSHIP_CONCEPT_ID	4038483 (Specimen source identity (SNOMED))

COLUMN_NAME	VALUE
DOMAIN_CONCEPT_ID_1	36 (Specimen)
FACT_ID_1	1099738697
DOMAIN_CONCEPT_ID_2	19 (Condition)
FACT_ID_2	1610001772
RELATIONSHIP_CONCEPT_ID	4038483 (Specimen source identity (SNOMED))

SPECIMEN

COLUMN_NAME	VALUE
SPECIMEN_ID	1099738697
PERSON_ID	257534973675956
SPECIMEN_CONCEPT_ID	0
SPECIMEN_TYPE_CONCEPT_ID	581378 (EHR Detail)
SPECIMEN_DATE	2018-07-17
SPECIMEN_SOURCE_ID	180717588901

FACT_RELATIONSHIP

COLUMN_NAME	VALUE
DOMAIN_CONCEPT_ID_1	26 (Measurement)
FACT_ID_1	3400001148
DOMAIN_CONCEPT_ID_2	36 (Specimen)
FACT_ID_2	1099738697
RELATIONSHIP_CONCEPT_ID	4023266 (Has specimen (SNOMED))

COLUMN_NAME	VALUE
DOMAIN_CONCEPT_ID_1	36 (Specimen)
FACT_ID_1	1099738697
DOMAIN_CONCEPT_ID_2	26 (Measurement)
FACT_ID_2	3400001148
RELATIONSHIP_CONCEPT_ID	4038483 (Specimen source identity (SNOMED))

In Figure 5, in the patient's surgical specimen pathology reports, records regarding the surgical site (the sigmoid colon) and diagnosis ("adenocarcinoma, nos") are stored. These entities are delivered into the CONDITION_OCCURRENCE table, with cancer diagnosis information (adenocarcinoma of the sigmoid colon). Specimen information, including person_id and specimen_id during surgery, is stored in the SPECIMEN table. Here, the relationship definition is made using a FACT_RELATIONSHIP for the specimen with a diagnosis of CONDITION_OCCURRENCE. We used the "specimen source identity" concept to define and link the relationship between the two tables. The results are also stored in NOTE_NLP while performing the molecular study and immunohistochemical study tests with the specimen obtained from the operation. In our example, *NRAS* mutation test results in the molecular study report were negative; its details are "Codon 61 (exon 3)". In addition, the information "NLP_derived N-RAS gene mutations found [identifier] in blood or tissue by molecular genetics method nominal" was updated in the MEASUREMENT table. In the same way, the immunohistochemical study test result for the EGFR test ("Epidermal growth factor receptor Ag [Presence] in Tissue by immune stain"), including the contents of the result "1+/3", was saved in the NOTE_NLP table first; then, the NLP_derived EGFR positive summary is delivered to the MEASUREMENT table.

Discussion

Principal Findings

We demonstrate that we can store colon cancer–related textual entities processed through NLP, including those from concepts of standard vocabularies (eg, SNOMED CT, LOINC, and ICD-O), in CDM tables that can later be used for computational phenotyping. Our system can thus enable the development of new standard feature-based NLP systems and the reuse, portability, adaptation, and extension of other cancer-based reports. This is the first study to provide details on how to extract major entities and standardize each item in pathology examination reports. However, variations in text data used at different institutions complicate the application of our pattern-matching rule to other observational health care databases.

Concept recognition in biomedical text has been addressed by multiple systems such as MetaMap [17], the ConceptMapper Annotator of the Unstructured Information Management Architecture (UIMA) [18], and the Clinical Text Analysis and Knowledge Extraction System (cTAKES) [19]. MetaMap is a widely available program that provides access to concepts in the UMLS metathesaurus from biomedical text; it was introduced in an effort to improve biomedical text retrieval, specifically the retrieval of MEDLINE/PubMed citations. It also links the text of biomedical literature to knowledge, including synonymy relationships, embedded in the metathesaurus [17]. ConceptMapper Annotator was implemented as a UIMA component. It was designed to accurately map text onto controlled vocabularies, specified as dictionaries, including the association of any necessary properties from the controlled vocabulary as part of that mapping [18]. Although these tools allow the use of a custom dictionary, their approach to efficient task performance relies on the target text being grammatically well formed. Thus, these mappers rely on token-based matching, which is processed by sentence and phrase segmentation. Many of the abbreviations, typing errors, and implicit tables in clinical text blur sentence boundaries and make phrase segmentation challenging. Our proposed approach utilizes rule-based pattern matching and a set of terminology mappings among international standard vocabularies such as LOINC, SNOMED CT, and ICD-O that link both concepts and values.

Many previous studies have focused on developing new text processing technologies. For instance, Baghari et al [20] developed a UMLS-based biomedical semantic operator. Solt [21] developed a medication extraction system using combined conditional random fields and rule-based systems. NOBLE Coder [22] implements a general algorithm for matching terms to concepts from an arbitrary vocabulary set. Its developers benchmarked the system’s speed and accuracy against the Colorado Richly Annotated Full Text (CRAFT) [23,24] and Shared Annotated Resources (ShARe) [25] corpora as reference standards and compared it to other concept recognition systems for biomedical tasks. Moreover, the BioCreAtIvE (Critical Assessment of Information Extraction in Biology) competition for automated gene and protein name recognition consists of a community-wide effort to evaluate information extraction and text mining developments in the biological domain [26]. Mitsumori et al used the support vector machine algorithm as a learning method for gene and protein name recognition [27], investigating and evaluating the system’s performance when making partial dictionary pattern matches.

Regarding research on text engineering and CDMs, there has been a recent report of a cohort retrieval system that can execute textual cohort selection queries on both structured and unstructured EHR data using CDM data [28]. The system leveraged a combination of structured queries and information retrieval techniques in NLP results to improve cohort retrieval performance while adopting OMOP CDM to enhance model portability. The NLP component empowered by cTAKES was used to extract CDM concepts from textual queries, and a hierarchical index in Elasticsearch was generated to support CDM concept search using information retrieval techniques and frameworks [28].

Our study approach differs from those in previous studies based on the following characteristics. First, we analyzed three kinds of pathology reports and created pattern-matching rules for each. Second, we converted the colorectal cancer entities of colon cancer patients in our hospital into international standards to establish them in the OMOP CDM database. Third, a mapping dictionary was created to standardize terms for pathology reports; this is a reusable legacy asset and can be used in future studies. To our knowledge, this is the first time that relationships have been defined using FACT_RELATIONSHIP to link OMOP CDM to other tables; these are also recyclable assets that can be implemented in other studies.

In our study, text data extraction errors can occur during CDM conversion from the pathology reports, and errors may also be introduced in standard term mapping. To prevent pathology text data extraction errors, we had to check the correctness of all items to be extracted. Text extraction results of the NOTE_NLP table and the number of data stored in the CDM database table were compared to confirm errors in the CDM table. From randomly sampled documents, we measured the accuracy of each item extracted by comparing the original document with the NLP extraction result and confirmed that there were no errors. Standard term mapping errors were checked for by domain experts. Mapping errors were identified through expert reviews and double-checking by CDM researchers.

From this study, we derived the insight that domain experts such as pathologists and bioinformatics specialists are essential for accurate term mapping, as understanding genetic testing methods is required to examine pathology reports. Furthermore, it was difficult to automate mapping tasks because some terms could be mapped to many terms (1:N); another difficulty was confusion from incorrect standard term search results. Another important insight of this study is that the NOTE_NLP table does not have clear standard guidelines for modifiers, so further standardization studies on term modifiers are needed to express various kinds of information in the future.

There were several challenges faced in this study. More general application of the pattern-matching rules created in this study is difficult because report formats vary for every institution. In our hospital's surgical pathology report, the results of examination of the pathological tissues of various organs removed during surgery are written in one report. In this study, one surgical site and diagnostic information were extracted from one surgical pathology report, except for multi-organ results. This will have to be supplemented in further research. The text extraction process deals with the contents of one hospital's examination report and corresponding rules, which can make this system difficult to use in other studies; however, it is expected that the definition of the extracted entities and the standard term mapping data table will be available for other studies as well. Although text extraction is not universal, the mapping guidelines can be adjusted for reusability because it is possible to modify them to apply to other organizations. Since it is important to convert the data in the same standard concept_id in order to conduct multi-institution studies using CDMs, we expect that even small mapping sets that include the histologic type of invasive carcinoma, location information, and biomarker term information will be easier to map into standard terminology. In addition, this will be more useful because the data are from reports generated at a tertiary general hospital. It will be fully available as a reference for other institutions. Additionally, the textual components of cancer pathology in clinical documents are as diverse as the various noncolon cancer types, and they vary by location and over time with new forms and templates; therefore, further research is necessary for the normalization of different types of cancer records and for the establishment of an extended database. In addition, there are several ways to interpret the results of the immunohistochemical study report for colorectal cancer. In our study, biomarker results such as EGFR were mostly found in the results of the immunohistochemical study report of colorectal cancer. The EGFR expression and interpretation could be diverse based on pathologists' visual scoring or the fraction of carcinoma staining. In particular, the pathology department of our hospital chose the scoring method that maps EGFR results of 1+, 2+, and 3+ as "positive," so we converted accordingly. If the scoring method should be subdivided into different biomarker results such as C-erbB2, this is a limitation of our research and something we should do in a future study.

In this study, we extracted fundamental text entities from pathological examination reports of patients with colon cancer and built a CDM database through terminology standardization and database definition. The preparation of pathology data for cancer and genome research, as well as various textual data that are currently recorded and managed, could lead to various elemental problems; further research on the CDM is expected to utilize much data. Our system can extract and store key features from unstructured text as NLP annotations by using the format defined by the OMOP CDM. The essential text entities from pathology reports were extracted, standardized, and deployed. Furthermore, more sophisticated preparation of the pathology data is needed for further research on cancer genomics, and various types of text narratives are areas for additional research on the use of data in the CDM.

Limitations

This study could not be validated for an external organization for CDM. However, we attempted to build a consolidated CDM for dealing with narrative text, which enables researchers to more easily derive important information from unstructured data. Furthermore, with our concept-relation definition approach, other organizations can construct database structures. We hope to derive and apply more features of cancer-related text entities to utilize both clinical and omics data in further clinical studies based on this study.

Acknowledgments

This work was supported by the Bio & Medical Technology Development Program of the National Research Foundation of Korea funded by Korea's Ministry of Science and ICT (NRF2017M3A9B6062319). This work was also supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program) (20003883, Advancing and expanding CDM-based distributed biohealth data platform) funded by Korea's Ministry of Trade, Industry & Energy.

Authors' Contributions

BR analyzed the data and drafted the manuscript as the first author. EY contributed to the development of the NLP. SK, HB, and S Yi helped in preparing and evaluating the data. HN, SL, JWK, and HH contributed to pathology data discussions. RB initiated this study as the principal investigator of this project. S Yoo supervised the overall study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Pseudo code for processing colorectal cancer pathology report.
[DOCX File , 73 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Supplementary material for Immunochemistry concept mapping table.
[DOCX File , 16 KB-Multimedia Appendix 2]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424 [FREE Full text] [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Kim B, Kim JM, Kang GH, Chang HJ, Kang DW, Kim JH, et al. Standardized Pathology Report for Colorectal Cancer, 2nd Edition. J Pathol Transl Med 2020 Jan;54(1):1-19 [FREE Full text] [doi: 10.4132/jptm.2019.09.28] [Medline: 31722452]
3. Standards for Cancer Registries Volume V: Pathology Laboratory Electronic Reporting. URL: https://www.naaccr.org/wp-content/uploads/2016/11/Pathology-Laboratory-Electronic-Reporting-Version-4.0-April-2011.pdf [accessed 2019-06-12]
4. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):78 [FREE Full text] [doi: 10.1186/s12911-019-0786-z] [Medline: 30943974]
5. Crepaz R, Romeo C, Montanaro D, De Santis S. Long-term results of treatment with bosentan in adult Eisenmenger's syndrome patients with Down's syndrome related to congenital heart disease. BMC Cardiovasc Disord 2013 Sep 18;13:74 [FREE Full text] [doi: 10.1186/1471-2261-13-74] [Medline: 24047157]

6. Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA--U.S. Department of Veterans Affairs national-scale HIS. Int J Med Inform 2003 Mar;69(2-3):135-156. [doi: 10.1016/s1386-5056(02)00131-4] [Medline: 12810119]
7. Brown SH, Lincoln M, Hardenbrook S, Petukhova ON, Rosenbloom ST, Carpenter P, et al. Derivation and evaluation of a document-naming nomenclature. J Am Med Inform Assoc 2001;8(4):379-390 [FREE Full text] [doi: 10.1136/jamia.2001.0080379] [Medline: 11418545]
8. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 Oct;15(10):761-771 [FREE Full text] [doi: 10.1038/gim.2013.72] [Medline: 23743551]
9. Rasmussen LV, Kiefer RC, Mo H, Speltz P, Thompson WK, Jiang G, et al. A Modular Architecture for Electronic Health Record-Driven Phenotyping. AMIA Jt Summits Transl Sci Proc 2015;2015:147-151 [FREE Full text] [Medline: 26306258]
10. FitzHenry F, Resnic F, Robbins S, Denton J, Nookala L, Meeker D, et al. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform 2015;6(3):536-547 [FREE Full text] [doi: 10.4338/ACI-2014-12-CR-0121] [Medline: 26448797]
11. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
12. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
13. Park RW. Sharing Clinical Big Data While Protecting Confidentiality and Security: Observational Health Data Sciences and Informatics. Healthc Inform Res 2017 Jan;23(1):1-3 [FREE Full text] [doi: 10.4258/hir.2017.23.1.1] [Medline: 28261525]
14. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
15. Goldsmith J, Siegal G, Suster S, Wheeler T, Brown R. Reporting guidelines for clinical laboratory reports in surgical pathology. Arch Pathol Lab Med 2008 Oct;132(10):1608-1616 [FREE Full text] [doi: 10.1043/1543-2165(2008)132[1608:RGFCLR]2.0.CO;2] [Medline: 18834219]
16. Patterson OV, Freiberg MS, Skanderson M, J Fodeh S, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC Cardiovasc Disord 2017 Jun 12;17(1):151 [FREE Full text] [doi: 10.1186/s12872-017-0580-8] [Medline: 28606104]
17. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
18. Tanenblatt M, Coden A, Sominsky I. The ConceptMapper Approach to Named Entity Recognition Internet. 2010 Presented at: Seventh International Conference on Language Resources and Evaluation (LREC'10); May 19-21, 2010; Valletta, Malta URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/448_Paper.pdf
19. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
20. Cuzzola J, Jovanović J, Bagheri E. RysannMD: A biomedical semantic annotator balancing speed and accuracy. J Biomed Inform 2017 Jul;71:91-109 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.016] [Medline: 28552401]
21. Tikk D, Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. J Am Med Inform Assoc 2010;17(5):540-544 [FREE Full text] [doi: 10.1136/jamia.2010.004119] [Medline: 20819860]
22. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, Jacobson RS. NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. BMC Bioinformatics 2016 Jan 14;17:32 [FREE Full text] [doi: 10.1186/s12859-015-0871-y] [Medline: 26763894]
23. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics 2012 Jul 09;13:161 [FREE Full text] [doi: 10.1186/1471-2105-13-161] [Medline: 22776079]
24. CRAFT: The Colorado Richly Annotated Full Text Corpus. URL: http://bionlp-corpora.sourceforge.net/CRAFT/ [accessed 2019-06-13]
25. ShARe/CLEF eHealth Evaluation Lab. URL: https://sites.google.com/site/shareclefehealth/home [accessed 2019-06-13]
26. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005;6 Suppl 1:S1 [FREE Full text] [doi: 10.1186/1471-2105-6-S1-S1] [Medline: 15960821]
27. Mitsumori T, Fation S, Murata M, Doi K, Doi H. Gene/protein name recognition based on support vector machine using dictionary as features. BMC Bioinformatics 2005;6 Suppl 1:S8 [FREE Full text] [doi: 10.1186/1471-2105-6-S1-S8] [Medline: 15960842]
28. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Implementation of a Cohort Retrieval System for Clinical Data Repositories Using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-Concept System Validation. JMIR Med Inform 2020 Oct 06;8(10):e17376 [FREE Full text] [doi: 10.2196/17376] [Medline: 33021486]

Abbreviations

CDM: common data model
cTAKES: Clinical Text Analysis and Knowledge Extraction System
EGFR: epidermal growth factor receptor
EHR: electronic health record
ICD-O: International Classification of Diseases for Oncology
LOINC: Logical Observation Identifiers Names and Codes
NLP: natural language processing
NOS: not otherwise specified
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: SNOMED Clinical Terms
SNUBH: Seoul National University Bundang Hospital
UIMA: Unstructured Information Management Architecture
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 04.03.20; peer-reviewed by RW Park, Y Li; comments to author 25.03.20; revised version received 20.05.20; accepted 11.11.20; published 09.12.20

Please cite as:
Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, Na HY, Kim JW, Baek RM, Hwang H, Yoo S
Transformation of Pathology Reports Into the Common Data Model With Oncology Module: Use Case for Colon Cancer
J Med Internet Res 2020;22(12):e18526
URL: https://www.jmir.org/2020/12/e18526
doi: 10.2196/18526
PMID: 33295294