Study on Pathology Common Data Model Using Natural Language Processing Toward Implementation of Integrated Platform for Clinical and Omics Data

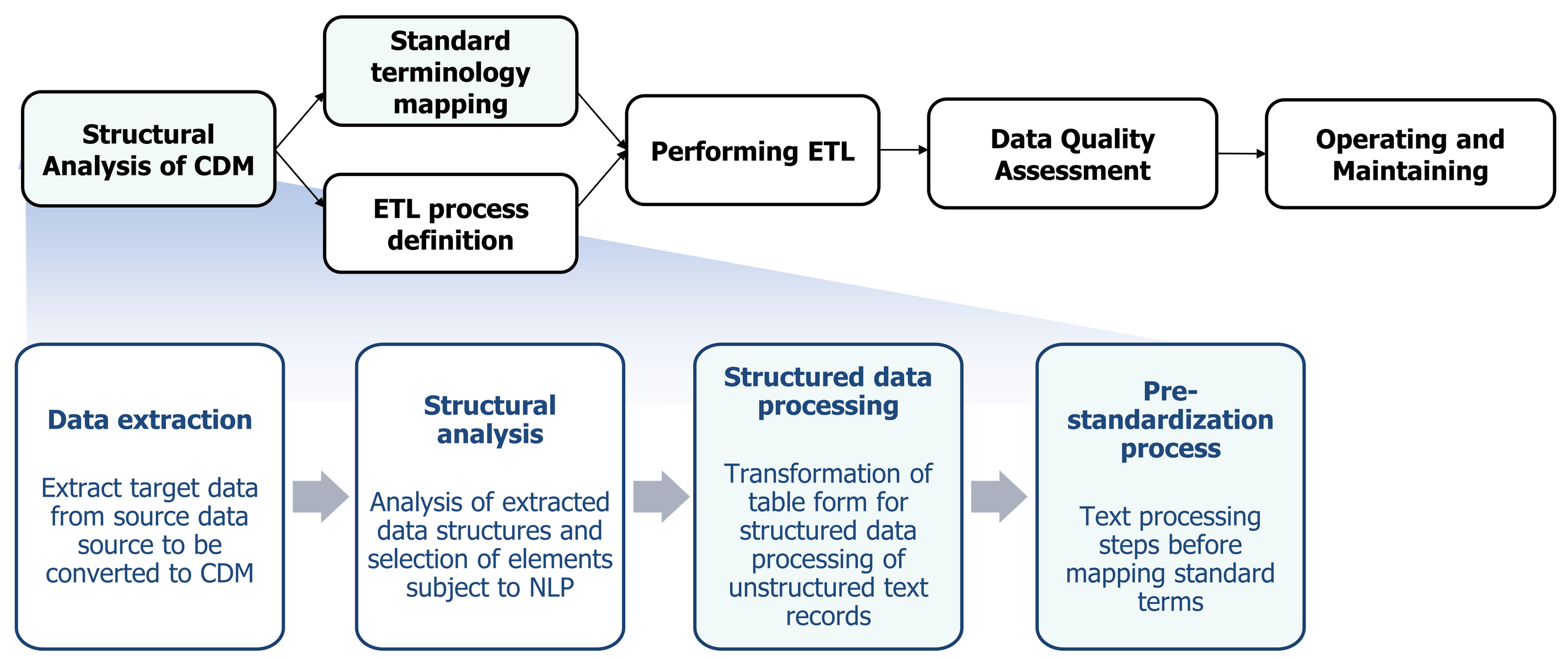
Borim Ryu, MS[1], Eunsil Yoon, MS[1], Seok Kim, MS[1], Hyunyoung Baek, MS[1], Sejoon Lee, PhD[1], Rong-Min Baek, MD,PhD[1], Sooyoung Yoo, PhD[1]

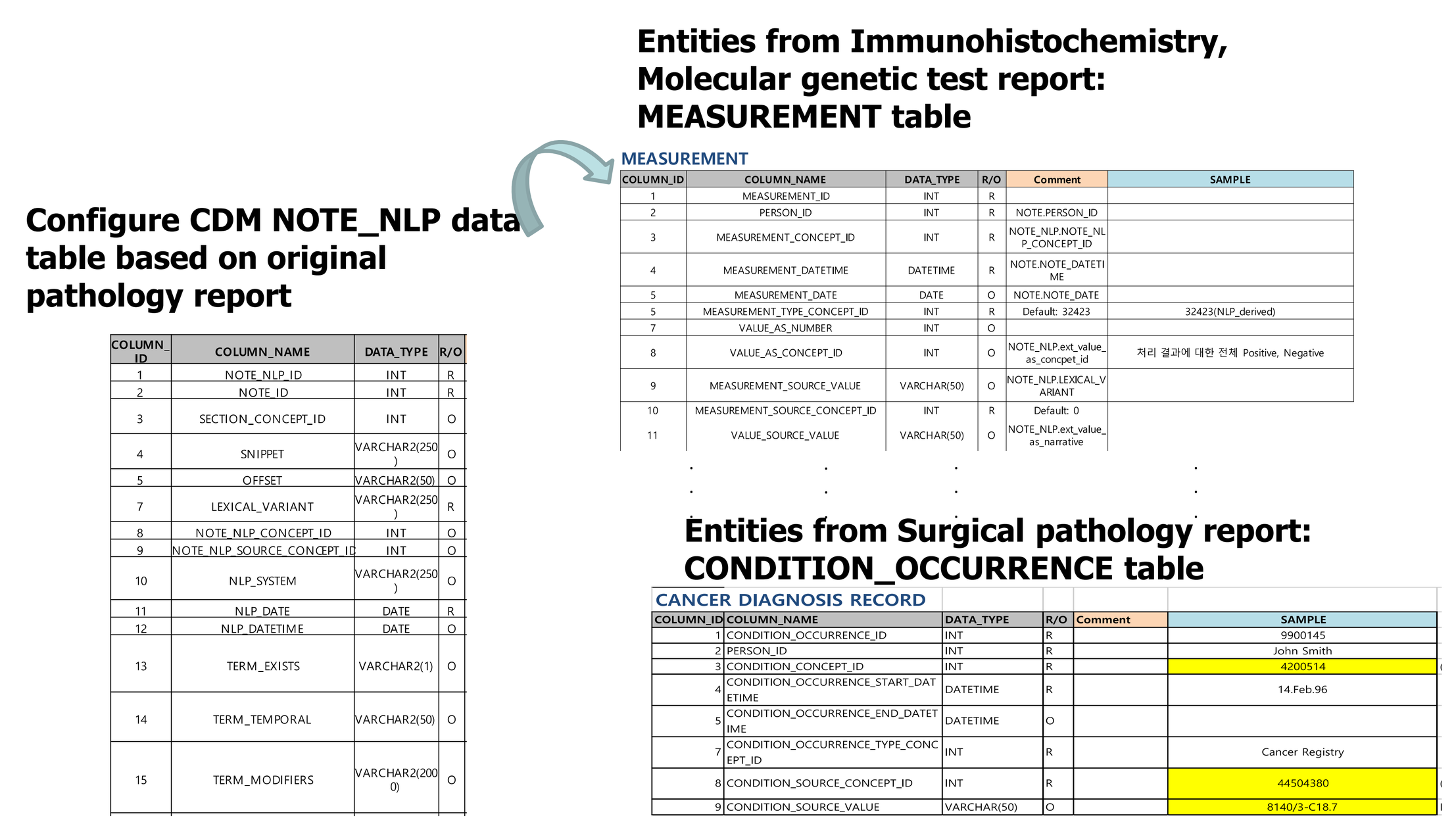
[1] Seoul National University Bundang Hospital, Seongnam-si, Gyeonggi-do, KOREA

Background

Three types of Pathology report written as free-text stored CDM NOTE database in SNUBH.

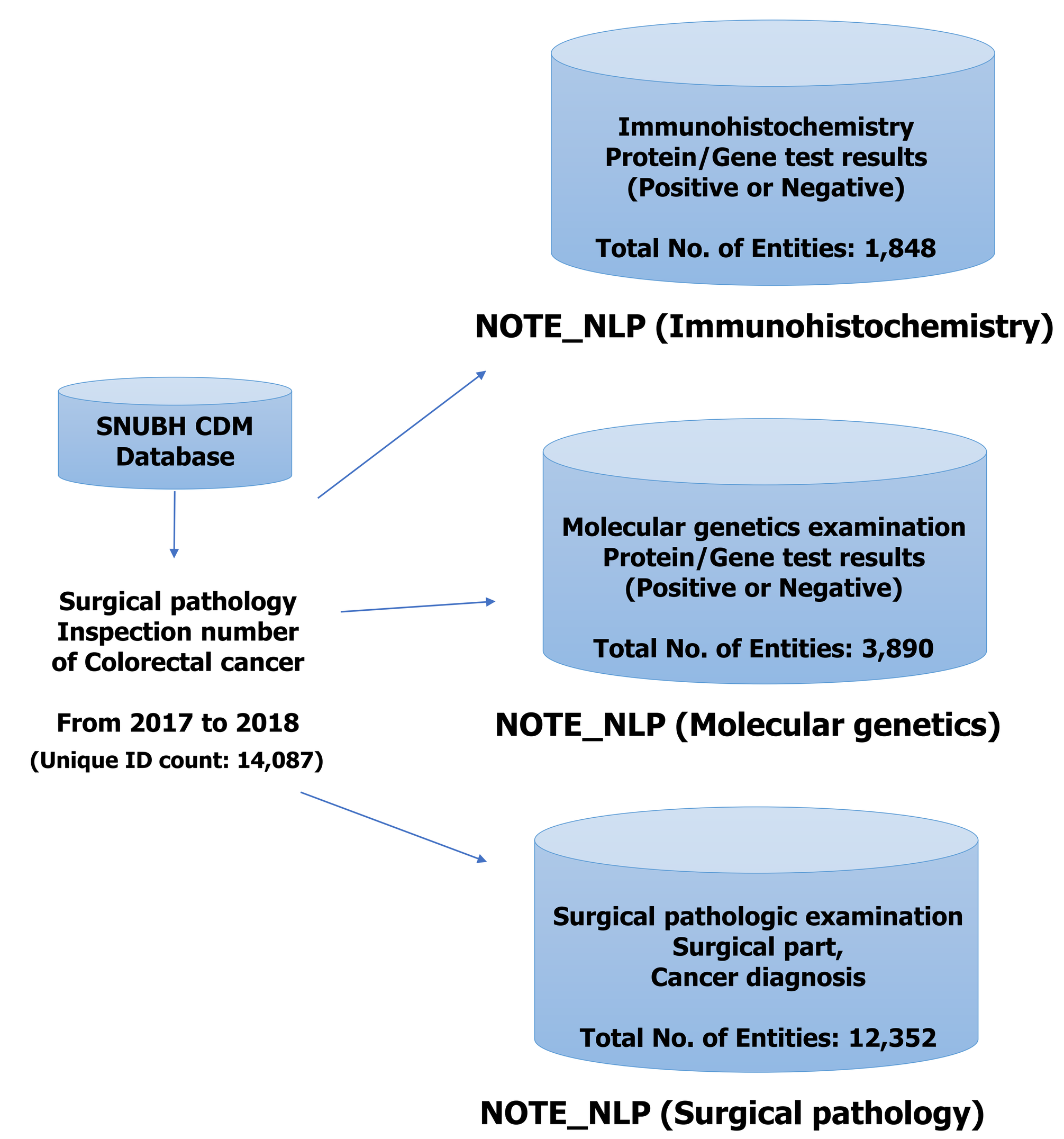
- Natural Language Processing (NLP) is commonly used to analyze these text narratives, to recognize or extract clinically important entities.
- To convert EHR data source into OMOP CDM data, re-formulation and standardization of data should be performed in advance as well.
- In this study, we transformed the pathology report data, which is written as English and Korean, from Seoul National University Bundang Hospital (SNUBH) into CDM NOTE_NLP data.
- The main objective of this research is to extract clinically meaningful text entities from the reports, convert them into CDM structure, and store in both NOTE_NLP table, MEASUREMENT table, or CONDITION_OCCURRENCE table.



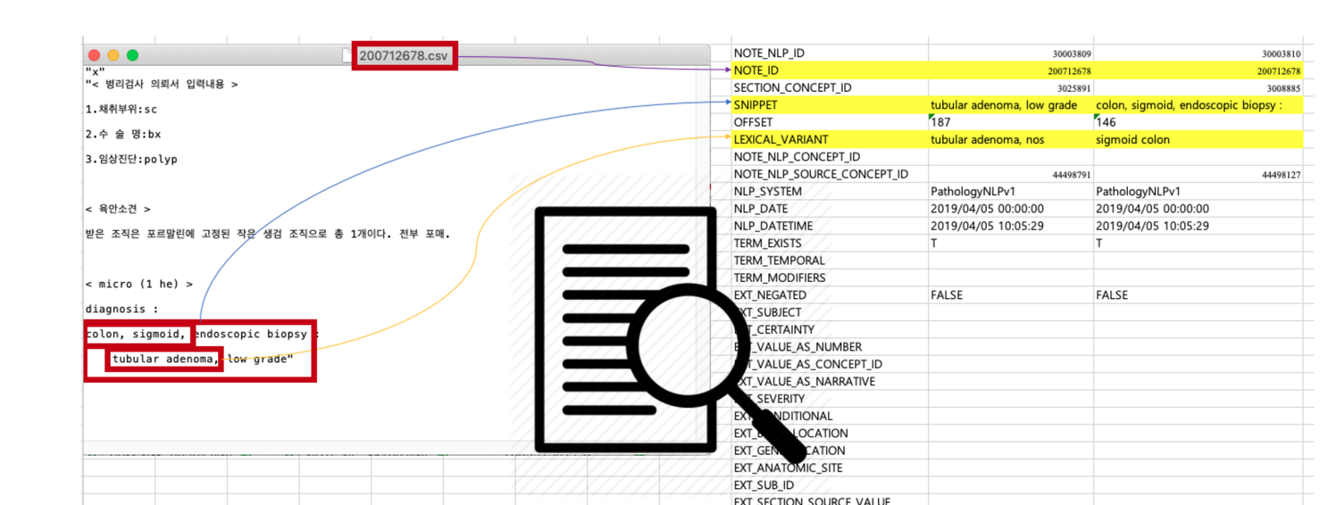


Results

Rules-based text processing based on regular expression was developed to extract major clinical entities of testing subject and its results information depending on the type of each report.



	Immunohistochemistry	Molecular genetics	Surgical pathology
	The number of Entities (No. of Reports)	The number of Entities (No. of Reports)	The number of Entities (No. of Reports)
NOTE_NLP	6,092 (1,848)	13,953 (3,890)	25,902 (12,352)



- Each 100 documents by type of the report were **reviewed manually** for verification of the study results.

Protein name	count
EGFR	74
BRAF	18
PTEN	18
hMLH1	17
C-erbB2	13
p53	13
Ki-67	8
CD3	5
CD8	5
Desmin	5
hMSH2	5
hMSH6	4
PMS2	4
Glucagon	3
Synaptophysin	3
C-Kit	2
CD31	2
CD34	2
Dog1	2
S100	2
SMA	2
CD20	1
CD4	1
CD56	1
CEA	1
Chromogranin	1
Cytokeratin	1
granzyme B	1
Total	**214**

Molecular genetic test results	count
Microsatellite instability test	**225**
BAT25	45
BAT26	45
D17S250	45
D2S123	45
D5S346	45
BRAF mutation test	**2**
K601E	1
V600E	1
KRAS gene mutation test	**96**
Codon 12 (exon 2)	32
Codon 13 (exon 2)	32
Codon 61 (exon 3)	32
NRAS gene mutation test	**66**
Codon 12 (exon 2)	22
Codon 13 (exon 2)	22
Codon 61 (exon 3)	22
Total	**389**

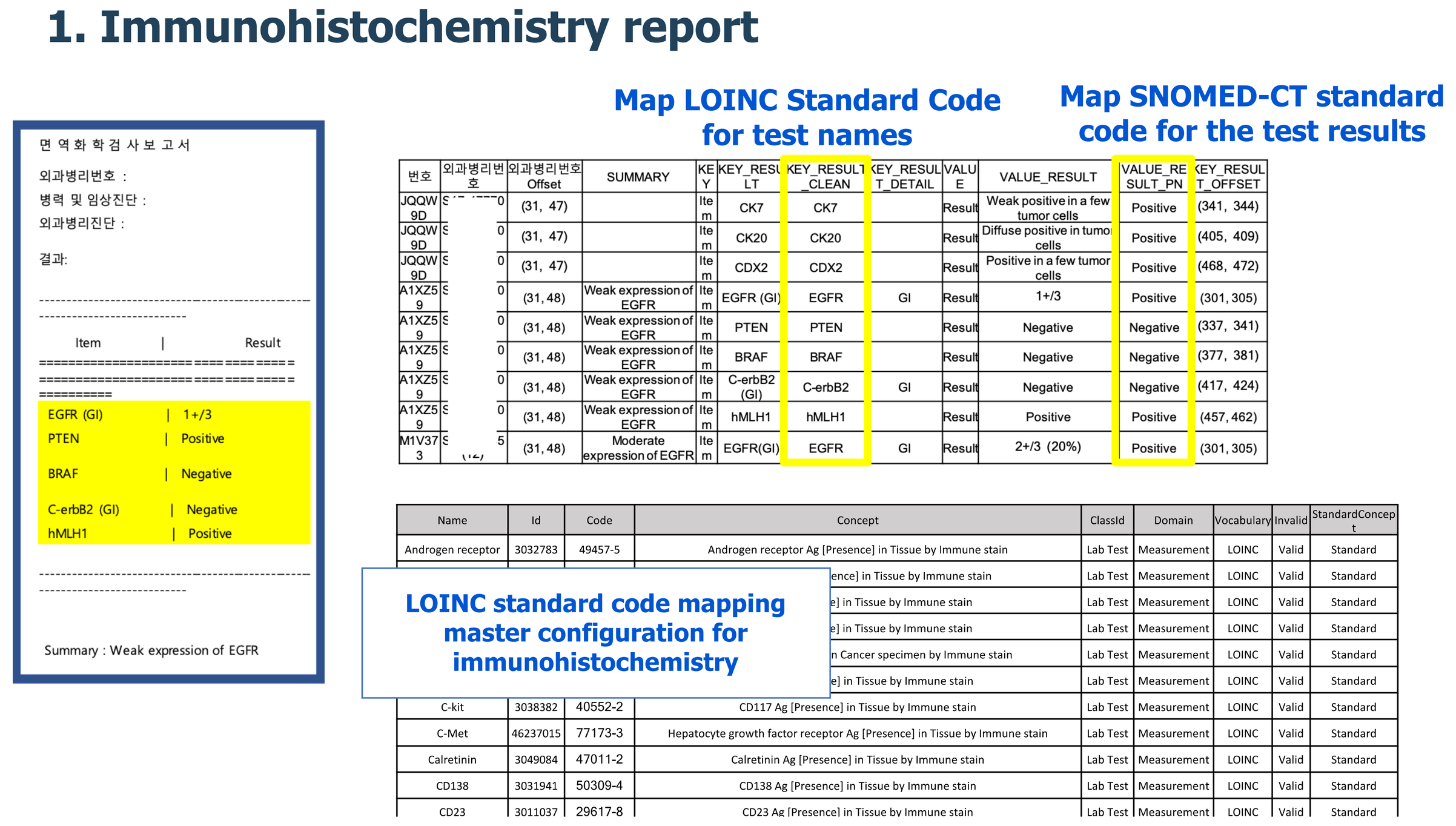
Cancer diagnosis	count
adenocarcinoma, nos	25
carcinoma, nos	1
neoplasm, nos	1
serrated adenoma	1
tubular adenoma, nos	66
tubulovillous adenoma, nos	2
tumor, nos	4
Total	**100**

Surgical part	count
appendix	6
ascending colon	23
cecum	10
colon	17
descending colon	3
hepatic flexure colon	5
rectum	26
sigmoid colon	14
transverse colon	10
Total	**114**

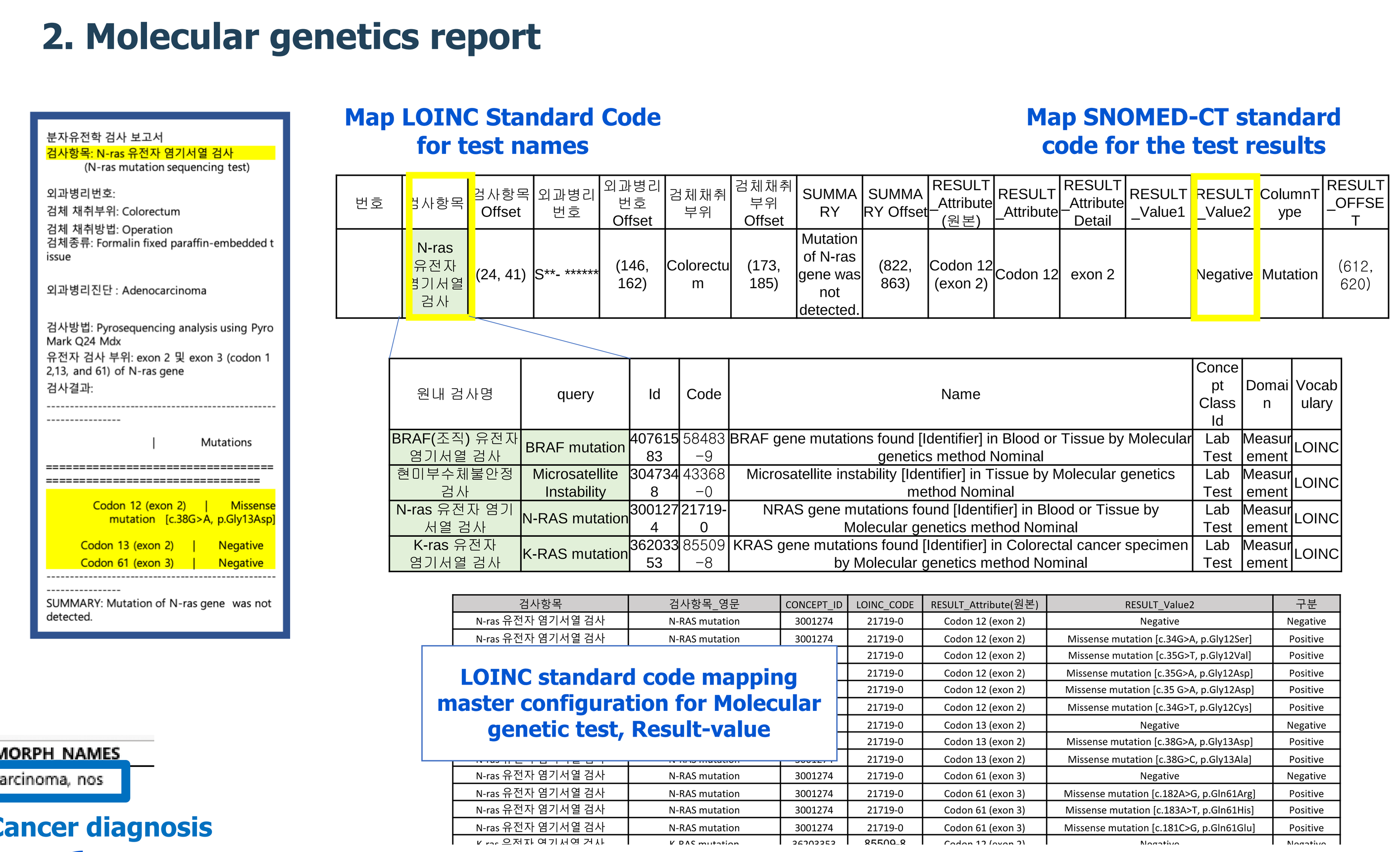
Methods

Pathology relevant major text entities were extracted and converted into OMOP CDM data structure.

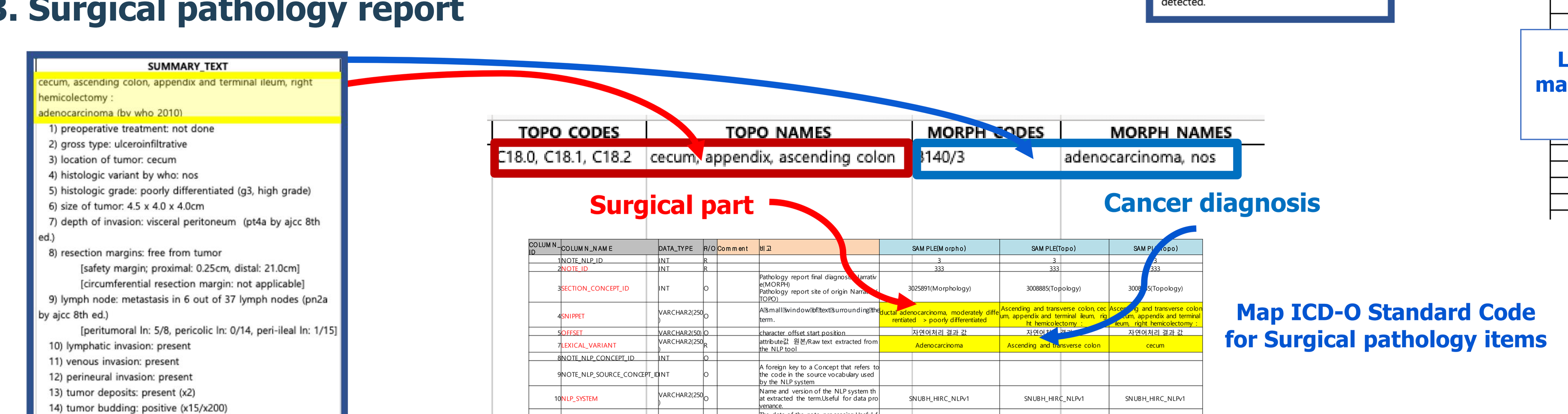
1. Immunohistochemistry report



2. Molecular genetics report



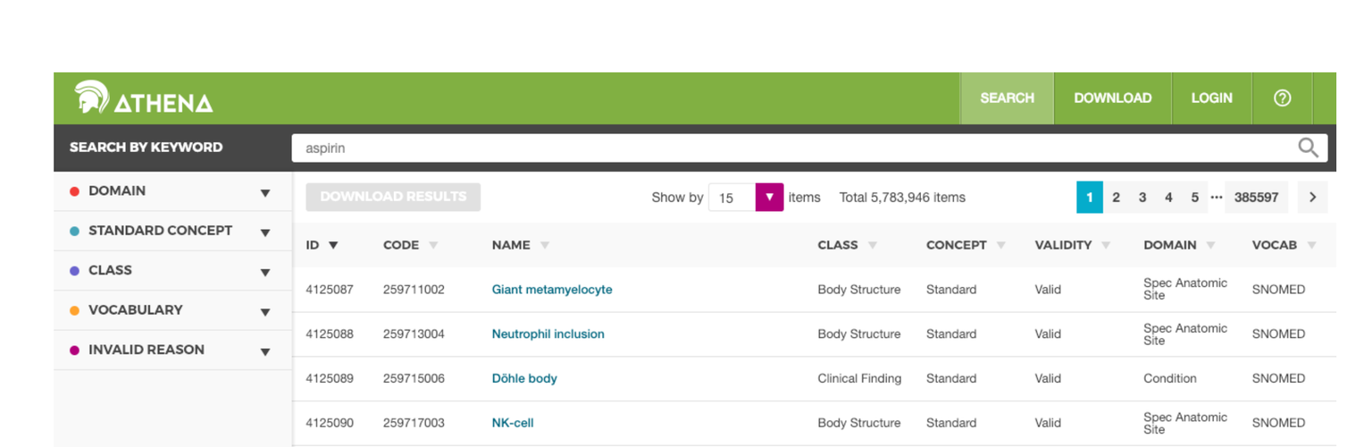
3. Surgical pathology report



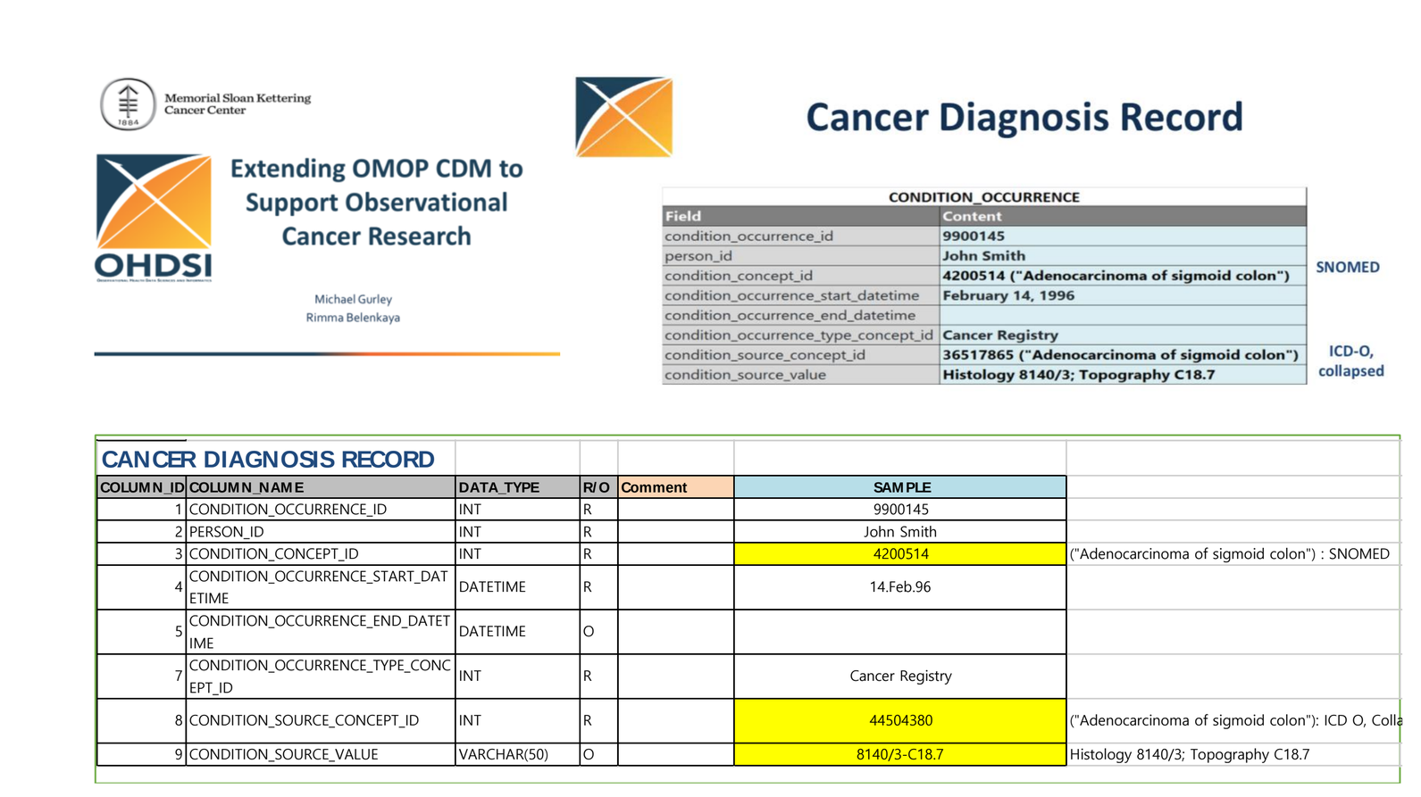
Conclusions

We aimed to prepare CDM data for research platform that can take advantage of all the omics clinical to patient data at SNUBH for colon cancer pathology.

- In this study, colorectal cancer related clinical entities were extracted from text data in the pathological examination report by Seoul National University Bundang Hospital for a research platform that can utilize both clinical and omics data, and CDM data was deployed through the term standardization.
- As a result, natural language processing and term standardization were conducted for 1,848 immunohistochemistry test reports, 3,890 molecular genetics test reports and 12,352 surgical pathology reports.
- The preparation of pathological data for cancer, genomic research, and the various elements of text data currently being recorded and managed have been able to derive problems for the CDM database, and further research on CDMs will ensure the utilization of a lot of data.



- Each data was mapped and reviewed according to ATHENA standard vocabularies



- For cancer diagnosis and surgical site information, our surgical pathologic CDM data was constructed by reflecting the Cancer Diagnosis Table proposed by the OHDSI oncology WG.
- Standard terminology was also mapped into ICD-O codes.

Acknowledgements

- This work was supported by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, Ministry of Science and ICT (MSIT) (No.NRF2017M3A9B6062319).