Article published online: 2022-06-15



# Transforming Thyroid Cancer Diagnosis and Staging Information from Unstructured Reports to the Observational Medical Outcome Partnership Common Data Model

Sooyoung Yoo[1] Eunsil Yoon[1] Dachung Boo[1] Borham Kim[1] Seok Kim[1] Jin Chul Paeng[2] Ie Ryung Yoo[3] In Young Choi[4,5] Kwangsoo Kim[6] Hyun Gee Ryoo[7,8] Sun Jung Lee[4,5] Eunhye Song[9] Young-Hwan Joo[10] Junmo Kim[11] Ho-Young Lee[1,2]

[1] Office of eHealth Research and Business, Healthcare Innovation Park, Seoul National University Bundang Hospital, Seongnam, South Korea
[2] Department of Nuclear Medicine, Seoul National University, College of Medicine, Seoul, South Korea
[3] Division of Nuclear Medicine, Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, South Korea
[4] Department of Medical Informatics, The Catholic University of Korea, College of Medicine, Seoul, South Korea
[5] Department of Biomedicine and Health Sciences, The Catholic University of Korea, College of Medicine, Seoul, South Korea
[6] Transdisciplinary Department of Medicine and Advanced Technology, Seoul National University Hospital, Seoul, South Korea
[7] Department of Nuclear Medicine, Seoul National University Hospital, Seoul, South Korea
[8] Department of Nuclear Medicine, Seoul National University Bundang Hospital, Seongnam, South Korea
[9] Department of Data Science Research, Innovative Medical Technology Research Institute, Seoul National University Hospital, Seoul, South Korea
[10] Biomedical Research Institute, Seoul National University Hospital, Seoul, South Korea
[11] Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, South Korea

Address for correspondence Ho-Young Lee, MD, PhD, Office of eHealth Research and Business, Healthcare Innovation Park, Seoul National University Bundang Hospital, 172 Dolma-ro, Bundang-gu, Seongnam-si, Gyeonggi-do 13605, South Korea (e-mail: debobkr@snubh.org).

Appl Clin Inform 2022;13:521–531.

## Abstract

**Background** Cancer staging information is an essential component of cancer research. However, the information is primarily stored as either a full or semistructured free-text clinical document which is limiting the data use. By transforming the cancer-specific data to the Observational Medical Outcome Partnership Common Data Model (OMOP CDM), the information can contribute to establish multicenter observational cancer studies. To the best of our knowledge, there have been no studies on OMOP CDM transformation and natural language processing (NLP) for thyroid cancer to date.

**Objective** We aimed to demonstrate the applicability of the OMOP CDM oncology extension module for thyroid cancer diagnosis and cancer stage information by processing free-text medical reports.

**Methods** Thyroid cancer diagnosis and stage-related modifiers were extracted with rule-based NLP from 63,795 thyroid cancer pathology reports and 56,239 Iodine whole-body scan reports from three medical institutions in the Observational Health Data Sciences and Informatics data network. The data were converted into the OMOP CDM

**Keywords**
- thyroid neoplasms
- natural language processing
- clinical documentation and communications
- electronic health records and systems
- standards

**received**
July 20, 2021
**accepted after revision**
December 31, 2021

DOI https://doi.org/10.1055/s-0042-1748144.
ISSN 1869-0327.



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

v6.0 according to the OMOP CDM oncology extension module. The cancer staging group was derived and populated using the transformed CDM data.

**Results** The extracted thyroid cancer data were completely converted into the OMOP CDM. The distributions of histopathological types of thyroid cancer were approximately 95.3 to 98.8% of papillary carcinoma, 0.9 to 3.7% of follicular carcinoma, 0.04 to 0.54% of adenocarcinoma, 0.17 to 0.81% of medullary carcinoma, and 0 to 0.3% of anaplastic carcinoma. Regarding cancer staging, stage-I thyroid cancer accounted for 55 to 64% of the cases, while stage III accounted for 24 to 26% of the cases. Stage-II and -IV thyroid cancers were detected at a low rate of 2 to 6%.

**Conclusion** As a first study on OMOP CDM transformation and NLP for thyroid cancer, this study will help other institutions to standardize thyroid cancer–specific data for retrospective observational research and participate in multicenter studies.

## Background and Significance

Electronic health records (EHRs) provide real-world data (RWD) to be used as clinical evidence for observational studies in cancer.[1] Particularly, pathology and test reports, included as free-text form in EHRs, contain information, such as cancer diagnosis and cancer stage which are vital to investigate survival, prognosis, and outcome in observational cancer research.

Medical organizations have been collecting structured data to expand cancer research data through manual chart review of EHR free-text data; however, it is a labor-intensive task and prone to human error. This problem was addressed with the development of natural language processing (NLP), a technique for processing clinical narratives by automatically extracting clinical data from free-text reports. Notes from radiology, pathology, and colonoscopy reports were most frequently used to extract clinical information for diagnostic reports.[2] In cancer research, NLP has been applied to obtain tumor characteristics, chemotherapy regimens, radiotherapy regimens, and operations, mainly in lung cancer,[3] breast cancer,[4–7] pulmonary cancer,[8] and colon cancer.[9] Free-text clinical notes on thyroid cancer have remained to be investigated with NLP.[10] Recently, there have been a few studies using NLP in the area of thyroid cancer. Idarraga and colleagues[11] extracted information from unstructured ultrasonography, fine-needle aspiration (FNA), and pathology reports using cTAKES tool to develop a predictive model of identifying false-negative FNA results. From pathology reports, they focused on the extraction of the nodule size, laterality, histologic type, and a final diagnosis, while we further tried to extract cancer subtype and cancer stage information in this study. Zhang and colleagues[12] developed deep NLP to diagnose thyroid cancer using sonographic text reports. Meanwhile, we focused on regular expression-based NLP of pathology reports and iodine whole-body scan reports for cancer staging.

The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) represents health care data from different data sources, including administrative claims and EHRs, in a consistent and standardized manner.[13–15] It enables a distributed research data network where researchers can run standardized analytical code and compile analytical result data without sharing patient-level information. The Observational Health Data Sciences and Informatics (OHDSI)[13] is an international, interdisciplinary collaboration which extended the OMOP CDM to include various types of data and created an open-source data analytic solution. As an OMOP CDM extension, the OMOP CDM oncology module was proposed to incorporate and standardize the key information on cancer to be utilized for observational cancer research. These data include cancer episode, tumor topography, histology, tumor attributes, and regimens[16] and was a part of the OMOP CDM v5.4 that was released on October 2021.

However, used cases of extracting information from a free-text report and converting it to the OMOP CDM oncology module for cancer diagnosis and cancer stage remains insufficient. To facilitate multicenter CDM-based cancer studies, it is important to keep conventions for vocabulary mapping and data transformation when converting CDM from unstructured data to ensure that cancer data are represented in a consistent way. In this study, we focused on extracting and presenting data on thyroid cancer diagnosis and cancer stage from unstructured reports using rule-based NLP and the OMOP CDM oncology module.

## Objectives

This study aimed to support CDM-based thyroid cancer research by sharing a conversion process that can derive cancer diagnosis and cancer stage information from thyroid cancer reports using NLP and the OMOP CDM oncology module. By using the same methods, the free-text reports were converted with the CDM module at three medical institutions to verify the possibility of multicenter expansion and whether this process can be used in other medical institutions.

## Methods

### Data Sources

Three metropolitan university hospitals from the OHDSI data network with EHR-based OMOP CDM data participated in this study. The study was led by Seoul National University Bundang Hospital (SNUBH), while Seoul National University Hospital (SNUH) and Catholic Medical Center (CMC) verified the conversion process of information extraction, vocabulary mapping, data loading to construct thyroid cancer diagnosis and cancer stage in the OMOP CDM. The study was approved and granted exemption of informed consent by the institutional review boards at each participating organization.

### Overall Process for Thyroid Cancer Diagnosis and Cancer Staging

The OMOP CDM is a patient-centric relational data model that consists of clinical data, health system data, health economics data, derived data, metadata, and vocabulary tables. For example, in the clinical data tables, the patients' demographics are present in the PERSON table, diagnosis data can be found in the CONDITION_OCCURRENCE table, medications are represented in the DRUG_EXPOSURE table, surgery and procedures are stored in the PROCEDURE_OCCURRENCE table, laboratory tests and quantitative findings are present in the MEASUREMENT table, clinical facts obtained in the context of examination or procedure can be found in the OBSERVATION table, and free-text clinical notes are present in the NOTE table. Vocabulary tables comprise the CONCEPT and CONCEPT_RELATIONSHIP tables that contain codes from various terminologies and relationship information between the concepts. The content of data tables should be represented through a standard concept in a *_concept_id field of its corresponding domain, while the source code should be maintained in a *_source_concept_id field.

Following the seventh edition of the American Joint Committee on Cancer (AJCC) tumor–node–metastasis (TNM) staging system,[17] the T-stage for thyroid cancer was determined by obtaining cancer diagnosis, the N-stage data from surgical pathology reports, while the M-stage information was derived from the whole-body scan report using rule-based NLP. Then, the extracted cancer diagnosis and cancer stage information was inserted into the OMOP CDM using the extended OMOP CDM oncology model, so that it could be used for cancer research, along with the existing clinical data from the OMOP CDM. The overall process of extracting and converting thyroid cancer diagnosis and cancer stage information from free-text medical reports is presented in ►Fig. 1 in which the data of the free-text reports stored in the NOTE table were extracted and converted into the corresponding OMOP CDM tables. The surgical pathology and iodine whole-body scan reports are presented in ►Fig. 2 (►Supplementary Table S1 [available in the online version] for the report samples from other institutions).

Two of the three institutions participating in this study used the same EHR and one used a different EHR. After sharing the regular expression and vocabulary mapping tables defined by one institution with other medical institutions, each institution applied the NLP according to their report formats, and converted the extracted data into OMOP CDM using the same conversion convention and vocabulary mapping. In this study, only limited regular expression-based approach was used for free-text parsing rather than a full-fledged natural language

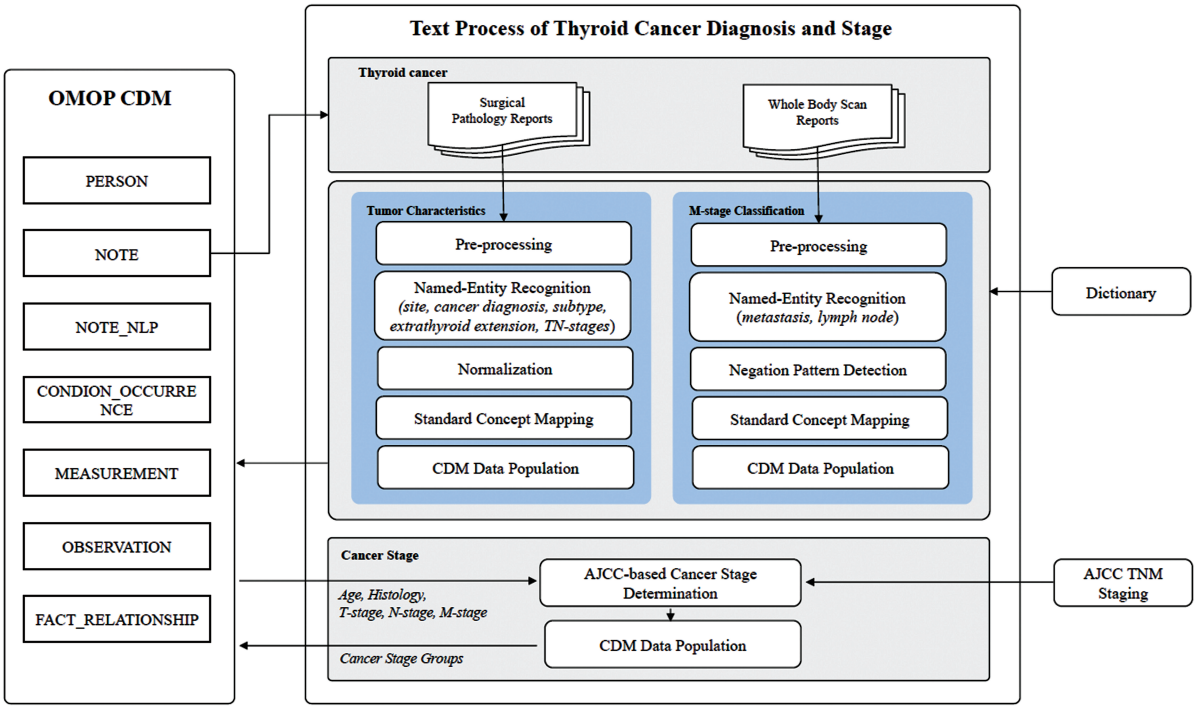



**Fig. 1** Overview of information extraction and CDM conversion of thyroid cancer diagnosis and cancer stage. OMPO CDM, Observational Medical Outcome Partnership's Common Data Model.

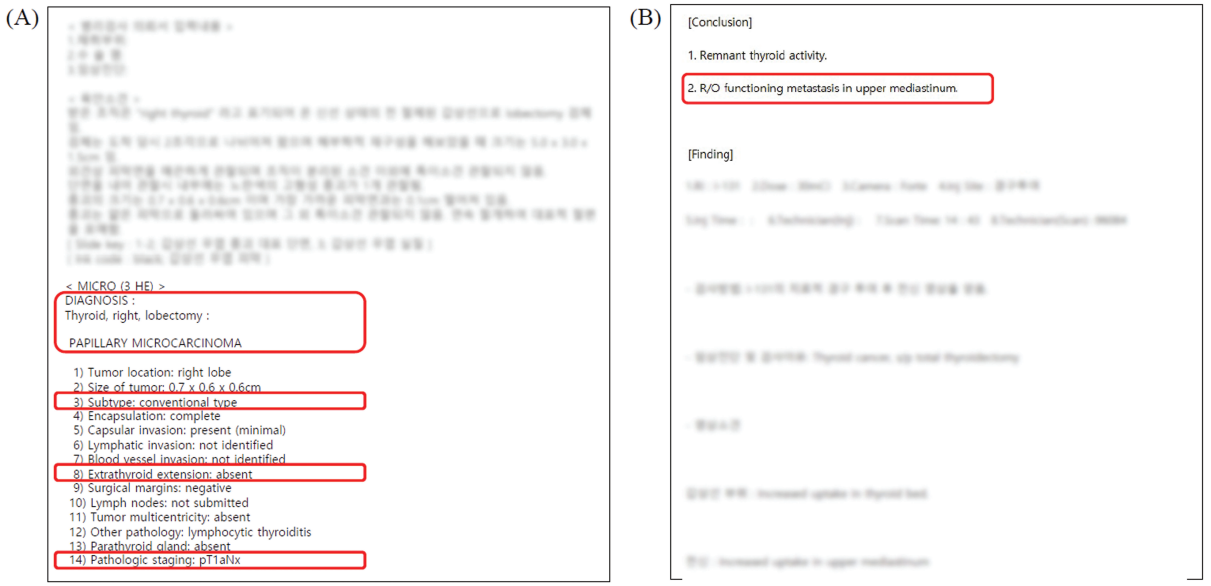



**Fig. 2** Example of free-text medical reports. (**A**) Surgical pathology reports and (**B**) Iodine whole-body scan reports. The text processing parts were marked by red-line box.

processing approach. The verification of the CDM conversion was performed by each institution.

### Processing Surgical Pathology Reports

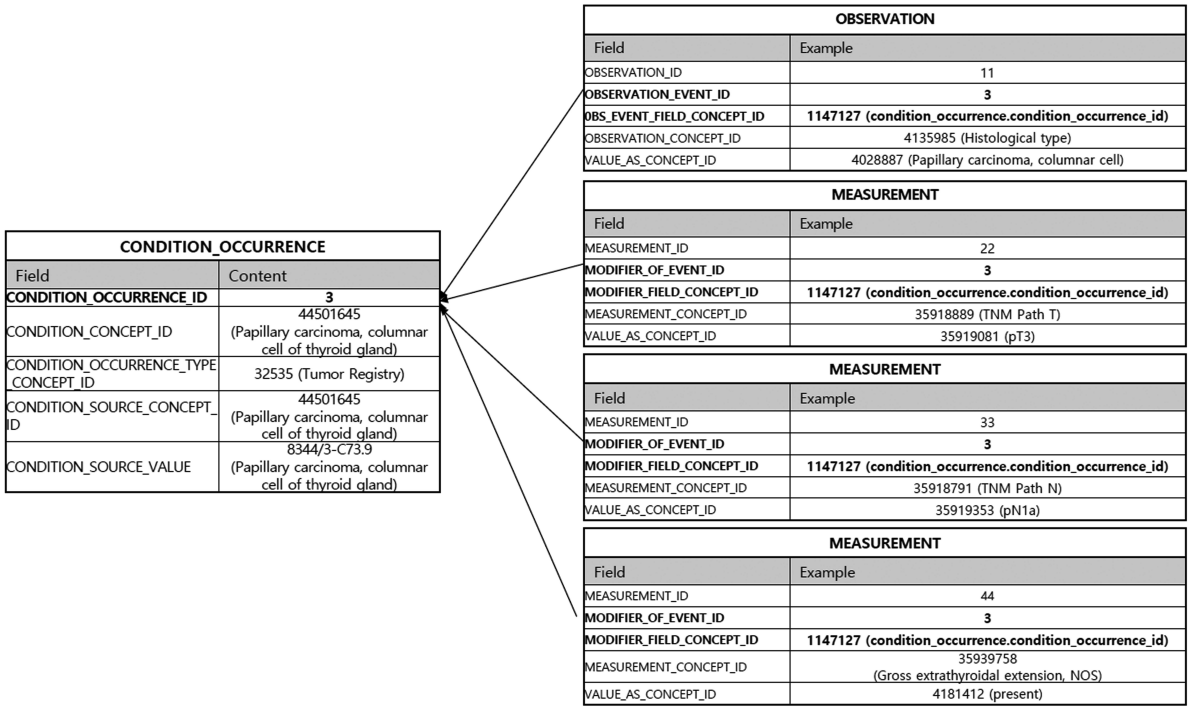
Tumor characteristics representing the topology (body site), morphology (histology), and pathological stages were extracted from the surgical pathology reports by using regular expression patterns. Specifically, information about specimen site, cancer diagnosis, cancer subtype, extrathyroidal extension, and pathological TN-stage information was extracted.

The specimen site was the area from which the tissue was collected, and cancer diagnosis was regarded as the disease diagnosed by a pathologist. Thyroid cancer can be classified into subtypes, including papillary, follicular, tall cell (columnar call), Hurthle's cell (oxyphilic and oncocytic), medullary, and anaplastic cancers. The pathological stage consists of a prefix, T-class, and N-class information. The prefix is whether the specimen sample was acquired in the pathological (p) process following treatment (yp) or after cancer recurrence (r). T-class provides the size and extent of the primary tumor, while the N-class gives information on lymph node metastasis. Extrathyroidal extensions indicate whether the primary tumor has invaded the structures surrounding the thyroid gland (e.g., the strap muscles, trachea, larynx, vasculature, esophagus, and recurrent laryngeal nerve).[18,19]

We extracted all pathology reports and selected only those containing the keyword "thyroid cancer" as the reports subject to NLP in this study. The pathology reports were a free-text that was automatically generated from semistructured templates with a free-text entry. They comprised a diagnosis section with a keyword "DIAGNOSIS" and numbered tumor characteristics items. The entities were extracted using the keyword and numbering features. When multiple samples were present in a single report, the sample number was also extracted and maintained. The regular expressions or extracting rules to recognize the entities were as follows:

- Specimen site: "Thyroid" term below the "DIAGNOSIS" section.
- Diagnosis: uppercase lines appearing on the lower line of the specimen site.
- Subtype: regular expression of "*|d|)|sSubtype|:|s(.*).*"
- Extrathyroid extension: regular expression of "*([Ee]xtrathyroid(al)? extension)|s*:|s*(.*)$.*"
- Pathological stage entity: regular expression of "*(pathologic stag(e|ing)).*|:|s*(.*)$.*"
- Pathological stage value: regular expression of extracting the TNM stage composed of prefix modifiers followed by alphanumeric codes staring with T, N, and M (see ►**Supplementary Fig. S1** [available in the online version] for an graphical example); "*^(yrp|ryp|yr|rp|yp|p|P|y|r)?T?((is)?[01234ll|x|X]?[abcd]?(mi)?[abcd]?[01234ll]?)|s?N?([01234ll|x|X|o|O]?(mi)?[abcd]?)M?([01ll|x|X]?[abc]?)|s?(.*).*"

When the pathological stage value of an exceptional case deviated from the regular expression, because of a typo, different notation, or missing input, it was corrected through clinical consultation. All the extracted values from the entities were normalized and mapped to form standard concepts using a mapping dictionary constructed by a domain expert (►**Supplementary Table S2** [available in the online version] for the normalization mapping list). Finally, all information extracted from the pathology report on the NOTE table was populated into the NOTE_NLP table. The NOTE_NLP data were then inserted into the corresponding clinical data tables according to the OMOP CDM oncology extension. Specifically, they were inserted based on the specimen site, diagnosis, and subtype. Cancer diagnosis was normalized to the International Classification of Diseases for



| CONDITION_OCCURRENCE | |
|---|---|
| Field | Content |
| **CONDITION_OCCURRENCE_ID** | **3** |
| CONDITION_CONCEPT_ID | 44501645 (Papillary carcinoma, columnar cell of thyroid gland) |
| CONDITION_OCCURRENCE_TYPE_CONCEPT_ID | 32535 (Tumor Registry) |
| CONDITION_SOURCE_CONCEPT_ID | 44501645 (Papillary carcinoma, columnar cell of thyroid gland) |
| CONDITION_SOURCE_VALUE | 8344/3-C73.9 (Papillary carcinoma, columnar cell of thyroid gland) |

| OBSERVATION | |
|---|---|
| Field | Example |
| OBSERVATION_ID | 11 |
| **OBSERVATION_EVENT_ID** | **3** |
| **OBS_EVENT_FIELD_CONCEPT_ID** | **1147127 (condition_occurrence.condition_occurrence_id)** |
| OBSERVATION_CONCEPT_ID | 4135985 (Histological type) |
| VALUE_AS_CONCEPT_ID | 4028887 (Papillary carcinoma, columnar cell) |

| MEASUREMENT | |
|---|---|
| Field | Example |
| MEASUREMENT_ID | 22 |
| **MODIFIER_OF_EVENT_ID** | **3** |
| **MODIFIER_FIELD_CONCEPT_ID** | **1147127 (condition_occurrence.condition_occurrence_id)** |
| MEASUREMENT_CONCEPT_ID | 35918889 (TNM Path T) |
| VALUE_AS_CONCEPT_ID | 35919081 (pT3) |

| MEASUREMENT | |
|---|---|
| Field | Example |
| MEASUREMENT_ID | 33 |
| **MODIFIER_OF_EVENT_ID** | **3** |
| **MODIFIER_FIELD_CONCEPT_ID** | **1147127 (condition_occurrence.condition_occurrence_id)** |
| MEASUREMENT_CONCEPT_ID | 35918791 (TNM Path N) |
| VALUE_AS_CONCEPT_ID | 35919353 (pN1a) |

| MEASUREMENT | |
|---|---|
| Field | Example |
| MEASUREMENT_ID | 44 |
| **MODIFIER_OF_EVENT_ID** | **3** |
| **MODIFIER_FIELD_CONCEPT_ID** | **1147127 (condition_occurrence.condition_occurrence_id)** |
| MEASUREMENT_CONCEPT_ID | 35939758 (Gross extrathyroidal extension, NOS) |
| VALUE_AS_CONCEPT_ID | 4181412 (present) |

**Fig. 3** An example of cancer diagnosis and modifiers that were converted into OMOP CDM Oncology Extension from surgical pathology reports. OMOP CDM, Observational Medical Outcome Partnership's Common Data Model.

Oncology, third edition (ICD-O-3) and was populated into the CONDITION_OCCURRENCE table in which the ICO-O-3 code was stored in the condition_source_concept_id field (►**Supplementary Table S3**, available in the online version). The standard concept mapped to it was stored in the condition_concept_id field, which can be Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) or an ICO-O-3 code using the mapping information of the CONCEPT_RELATIONSHIP table. The cancer subtype was inserted into the OBSERVATION table as SNOMED CT and linked to the cancer diagnosis, using the observation_event_id and obs_event_field_concept_id fields. The pathological T-stage, N-stage, and extrathyroidal extensions were stored into the MEASUREMENT table as the North American Association of Central Cancer Registries (NAACCR) code and were linked to the cancer diagnosis using the modifier_of_event_id and modifier_field_concept_id fields (►**Supplementary Table S4** [available in the online version] for the pathologic stage concepts). The example of populating cancer diagnosis data and modifiers extracted from surgical pathology reports is presented in ►**Fig. 3**.

## Processing Iodine Whole-Body Scan Reports

For thyroid cancer, an iodine whole-body scan is used to detect tumor metastasis or lesions using radioiodine. Iodine whole-body scan reports consisted of finding and conclusion sections with free-text entry, and information on distant metastasis was recorded in the conclusion section. We extracted metastasis information from the conclusion section using regular expression patterns. After extracting sentences containing the terms depicting the metastasis pattern, the sentences including negation patterns and lymph node patterns were excluded since they should be considered as non-metastasis. The following patterns were used to recognize metastasis.

- Metastasis pattern: "(.*)[*M*|*m*]*eta*.*."
- Negation patterns: "No evidence of," "Less, likely," "Disappeared," "no Functional meta.*$," "No remnant thyroid activity or functioning metastasis," and "Cannot."
- Lymph node patterns: "*LN*," "LNs," "Lymph Node," "Neck chain."

Based on the extracted metastasis information, the patient's TNM clinical M-stage was derived and classified as cM0 (no metastasis) or cM1 (metastasis) to be populated into the CDM. The clinical M-stage was normalized and inserted into the MEASUREMENT table as the NAACCR code of TNM Clin M (measurement_concept_id: 35918383) with values of cM1 (value_as_concept_id: 35919664) or cM0 (value_as_concept_id: 35919673).

## Deriving Thyroid Cancer Stage Grouping

Since we focused on differentiated carcinoma, stage groupings were created only for papillary or follicular cancers according to the AJCC seventh TNM staging system.[17] To categorize the stage grouping, we extracted age at diagnosis, histological type, TNM path T, TNM path N, and TNM clin M from the corresponding PERSON, CONDITION_OCCURRENCE, and MEASUREMENT tables. When the cancer stage grouping was categorized, T1, T1a, and T1b of the T-stage were defined as T1, and T3, T3a, and T3b were defined as T3. In the case of the M-stage, if metastasis was confirmed, at least once. during the whole-body scan within 2 years of thyroid cancer diagnosis, it was defined as cM1; otherwise, it was defined as cM0.

The derived cancer stage groupings, such as stages I, II, III, IVA, IVB, and IVC, were then inserted into the MEASUREMENT table with a measurement_concept_id of 35918286 (TNM path stage group) and normalized values of the NAACCR code (see ►Supplementary Table S5 [available in the online version] for the value concept_ids used in our study). To link the diagnosis name and M-stage information used in the stage categorization, a bidirectional relationship was added to the FACT_RELATIONSHIP table using "Has Measurement Component (SNOMED)" and "Measurement Component of (SNOMED)" relationship.

In this study, we derived information on the cancer stage groups for use in a population-level estimation analysis by creating and comparing subpopulations of a thyroid cancer cohort with the same stage

### Evaluation of Natural Language Processing

At the NLP development institution, the NLP rules were finalized by detecting and refining errors through two rounds of preevaluation with random samples of 100 documents for each report type before sharing the rules to two other institutions.

NLP was evaluated with random documents that were not used to define and preevaluate the rules. Specifically, for the evaluation of the results of NLP for surgical pathology reports, 100 reports were randomly selected, and a domain expert with clinical knowledge manually reviewed the results. For whole-body scan reports, each of the 50 reports with classified M0 and M1 stages were randomly selected and manually reviewed. Precision and recall metrics were measured as performance metrics. In this paper, the NLP evaluation was performed and reported by only one institution for the purpose of showing the evaluation process and metrics.

## Results

### Statistics for Processed Free-Text Reports

Overall, 52,133 surgical pathology reports and 56,239 iodine whole-body scan reports from three medical institutions were processed. Surgical pathology and whole-body scan reports had an average of approximately 1,607 and 471 words, respectively. The statistics of free-text reports of thyroid cancer patients from the three institutions that were used to convert into the OMOP CDM are summarized in ►Table 1.

### Evaluation Results of Natural Language Processing

NLP results of surgical pathology and whole-body scan reports of SNUBH were evaluated by comparing free-text reports through manual review. For surgical pathology reports, 100 notes were randomly selected. In the case of whole-body scan reports, 50 reports with metastasis and 50 reports without metastasis were randomly selected. Precision and recall were found to be 100% for all metastases and cancer diagnoses with subtype confirmation and its modifiers, including extrathyroidal extension, T-stage, and N-stage. The items to be extracted existed in a well-structured pattern (►Table 2).

### Distribution of Thyroid Cancer Diagnosis

The histopathological distributions of thyroid cancer in the three medical institutions were approximately 95.3 to 98.8, 0.9 to 3.7, 0.04 to 0.54, 0.17 to 0.81, and 0 to 0.3% of papillary carcinoma, follicular carcinoma, adenocarcinoma, medullary carcinoma, and anaplastic carcinoma, respectively. The histology distribution according to the detailed subtypes is presented in ►Table 3. Thyroid cancer accounted for 76 to 78% of female patients in the three institutions. The age-stratified distribution according to sex is presented in ►Fig. 4. Patients in their 30s and 50s accounted for approximately 70 to 77% of cases.

The distribution by stage followed the AJCC seventh TNM staging system, where stage-I thyroid cancer accounted for 55 to 64%, and stage III accounted for 24 to 26%. There were very low rates of stage-II and -IV thyroid cancer detection, with the low rate being 2 to 6%, according to real-world data. Cases with unknown stage grouping were found to appear at a rate of 3.6 to 12.3%. ►Table 4 shows the distribution of detailed pathological T- and N-stages, while the clinical M-stages were extracted from free-text reports. The stage grouping was derived from the TNM stage.

**Table 1** Statistics for free-text reports used in OMOP CDM conversion for thyroid cancer research

| Institution | Data period | Average word count (minimum–maximum) | No. of reports | No. of patients |
|---|---|---|---|---|
| Surgical pathology reports | | | | |
| SNUBH | May 12, 2003–June 29, 2019 | 1,738 (452–4,000) | 10,514 | 10,470 |
| SNUH | September 30, 2004–January 31, 2020 | 1,362 (125–5,574) | 32,800 | 16,838 |
| CMC | January 1, 2009–December 31, 2018 | 1,721 (209–2,000) | 8,819 | 7,708 |
| Iodine whole-body scan reports | | | | |
| SNUBH | September 3, 2003–June 28, 2019 | 459 (32–1,167) | 9,062 | 4,758 |
| SNUH | August 28, 2004–July 31, 2020 | 302 (63–1,255) | 18,500 | 8,656 |
| CMC | January 1, 2009–December 31, 2018 | 653 (69–2,000) | 28,677 | 5,578 |

Abbreviations: CMC, Catholic Medical Center; OMOP CDM, Observational Medical Outcome Partnership's Common Data Model; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.

**Table 2** Results of NLP for extracting cancer diagnosis and tumor characteristics

| Data element | Count | TP | FP | TN | FN | Precision[a] | Recall[b] |
|---|---|---|---|---|---|---|---|
| Surgical pathology reports ($n = 100$) | | | | | | | |
| Cancer diagnosis | 128 | 128 | 0 | – | 0 | 1.00 | 1.00 |
| T-stage | 101 | 101 | 0 | – | 0 | 1.00 | 1.00 |
| N-stage | 101 | 101 | 0 | – | 0 | 1.00 | 1.00 |
| Extrathyroid extension | 102 | 102 | 0 | – | 0 | 1.00 | 1.00 |
| Iodine whole-body scan reports ($n = 100$) | | | | | | | |
| M-stage (metastasis) | 50 | 50 | 0 | 0 | 0 | 1.00 | 1.00 |
| M-stage (nonmetastasis) | 50 | 0 | 0 | 50 | 0 | 1.00 | – |

Abbreviations: FN, false negative; FP, false positive; NLP, natural language processing; TN, true negative; TP, true positive.
[a]Precision = TP / (TP + FP) or TN / (TN + FN).
[b]Recall = TP / (TP + FN).

### Observational Medical Outcomes Partnership Common Data Model Conversion Results

Clinical data extracted from surgical pathology reports and whole-body scan reports for thyroid cancer were loaded into the CONDITION_OCCURRENCE, MEASUREMENT, and OBSERVATION tables of the OMOP CDM. The number of records added to the table from the three institutions is presented in ►Table 5. Tumor modifier information corresponding to the MEASUREMENT occupied >60% of the records. For each cancer extraction, approximately 1.3 records were populated for CONDITION_OCCURRENCE, 2.6 records for MEASUREMENT, and approximately 1.4 records for OBSERVATION.

## Discussion

In this study, OMOP CDM oncology extension module transformation was performed to support multicenter observational cancer research for thyroid cancer. Cancer diagnosis and tumor modifier information was extracted from the pathology report. Metastasis information was extracted from the iodine whole-body scan report to establish the thyroid cancer staging group. Through information extraction and transformation from three institutions, standard mapping for the OMOP CDM vocabulary and applicability of the OMOP CDM oncology extension module were successfully demonstrated. The three participating medical institutions were some of the largest hospitals located in the metropolitan area of Seoul in South Kore. All three institutions showed similar characteristics of thyroid cancer patients after OMOP CDM transformation and NLP. Although detailed reports on cancer subtypes were not available through the Korea Center Cancer Registry, compared with the previously reported distribution of thyroid cancer in South Korea,[20] it was found that the proportion of papillary carcinoma, follicular carcinoma, medullary carcinoma, and anaplastic carcinoma cases were similar in all three institutions.

The OMOP CDM oncology extension module is a comprehensive cancer data model comprised of cancer diagnosis, cancer treatments, and cancer episodes.[21] To the best of our knowledge, there are no studies on OMOP CDM transformation and NLP for thyroid cancer. In our previous work, we extracted pathological diagnosis and cancer-specific biomarker test results from colon cancer pathology reports and converted them into the OMOP CDM oncology extension module.[9] While the previous study focused on NLP and CDM conversion from pathology reports of immunohistochemical studies and molecular studies, this study is different in that it aimed at NLP and CDM conversion for cancer stage representation in addition to cancer diagnosis. For thyroid cancer in this study, the cancer stage was newly extracted and categorized from the surgical pathologic report and the iodine whole-body scan report. A new vocabulary mapping was also performed to convert the thyroid cancer diagnosis and cancer stage into OMOP CDM. With new information extraction items and different vocabulary mappings, we newly developed NLP for thyroid cancer. From this study, the possibility of using the OMOP CDM for cancer research was increased by deriving a cancer staging group that is important for phenotyping and prognosis prediction in thyroid cancer research.

In the NLP process for free-text report processing, NLP tools were not considered because the purpose of this study was to create an OMOP CDM oncology data for cancer research using precisely extracted data. Thus, using regular expression-based NLP, we focused on improving the accuracy of data extraction by consulting with medical professionals, since pathologists and clinicians used various terms and expressions. For instance, in the case of diagnosis subtype extraction, Hurthle's cell cancer was expressed as oxyphilic or oncocytic cancer. Additionally, in the case of TN stage extraction, there were expressions that did not conform to the rules. This was because of different notations (e.g., pT1aNx was recorded as pTIaNx), omission of stage information owing to a mistake, (e.g., pT N), and exceptions that required confirmation of the ADDENDUM document (e.g., pT2N0 vs. pT3N0 (see ADDENDUM)). In addition, rules for metastasis extraction were not able to be exactly defined as a specific regular expression pattern. Another limitation is that the NLP in our study was developed specifically for a particular institution. Although the regular expression developed in this study was shared with two other institutions, showing that it can be applied and customized in other institutions in Korea to populate OMOP CDM data in accordance with the oncology extension module, the regular

**Table 3** Histopathological distribution of thyroid cancer at three institutions

| Thyroid cancer diagnosis | SNUBH *n* (%) | | | SNUH *n* (%) | | | CMC *n* (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | Female | Male | Total | Female | Male | Total | Female | Male |
| Papillary carcinoma, columnar cell of thyroid gland | 686 (5.04) | 539 (5.21) | 147 (4.5) | 54 (0.37) | 45 (0.39) | 9 (0.28) | 492 (6.95) | 381 (6.99) | 111 (6.81) |
| Papillary carcinoma, encapsulated, of thyroid of thyroid gland | 136 (1) | 82 (0.79) | 54 (1.65) | 0 (0) | 0 (0) | 0 (0) | 180 (2.54) | 127 (2.33) | 53 (3.25) |
| Papillary carcinoma, follicular variant of thyroid gland | 1,317 (9.67) | 1,038 (10.03) | 279 (8.54) | 939 (6.38) | 734 (6.4) | 205 (6.31) | 655 (9.25) | 523 (9.6) | 132 (8.1) |
| Papillary carcinoma, numbers of thyroid gland | 10,799 (79.29) | 8,167 (78.89) | 2,632 (80.56) | 13,069 (88.78) | 10,209 (89) | 2,860 (88) | 5,656 (79.9) | 4,345 (79.74) | 1,311 (80.43) |
| Papillary carcinoma, oncocytic variant of thyroid gland | 38 (0.28) | 31 (0.3) | 7 (0.21) | 40 (0.27) | 32 (0.28) | 8 (0.25) | 9 (0.13) | 7 (0.13) | 2 (0.12) |
| Follicular adenoma, numbers of thyroid gland | 250 (1.84) | 203 (1.96) | 47 (1.44) | 450 (3.06) | 343 (2.99) | 107 (3.29) | 6 (0.08) | 4 (0.07) | 2 (0.12) |
| Follicular carcinoma, minimally invasive of thyroid gland | 2 (0.01) | 2 (0.02) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Follicular carcinoma, numbers of thyroid gland | 250 (1.84) | 190 (1.84) | 60 (1.84) | 0 (0) | 0 (0) | 0 (0) | 60 (0.85) | 47 (0.86) | 13 (0.8) |
| Oxyphilic adenocarcinoma of thyroid gland | 73 (0.54) | 58 (0.56) | 15 (0.46) | 6 (0.04) | 5 (0.04) | 1 (0.03) | 9 (0.13) | 8 (0.15) | 1 (0.06) |
| Medullary thyroid carcinoma of thyroid gland | 60 (0.44) | 36 (0.35) | 24 (0.73) | 119 (0.81) | 75 (0.65) | 44 (1.35) | 12 (0.17) | 7 (0.13) | 5 (0.31) |
| Carcinoma, anaplastic, numbers of thyroid gland | 9 (0.07) | 7 (0.07) | 2 (0.06) | 44 (0.3) | 28 (0.24) | 16 (0.49) | 0 (0) | 0 (0) | 0 (0) |
| Total | 13,620 (100) | 10,353 (100) | 3,267 (100) | 14,721 (100) | 11,471 (100) | 3,250 (100) | 7,079 (100) | 5,449 (100) | 1,630 (100) |

Abbreviations: CMC, Catholic Medical Center; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.

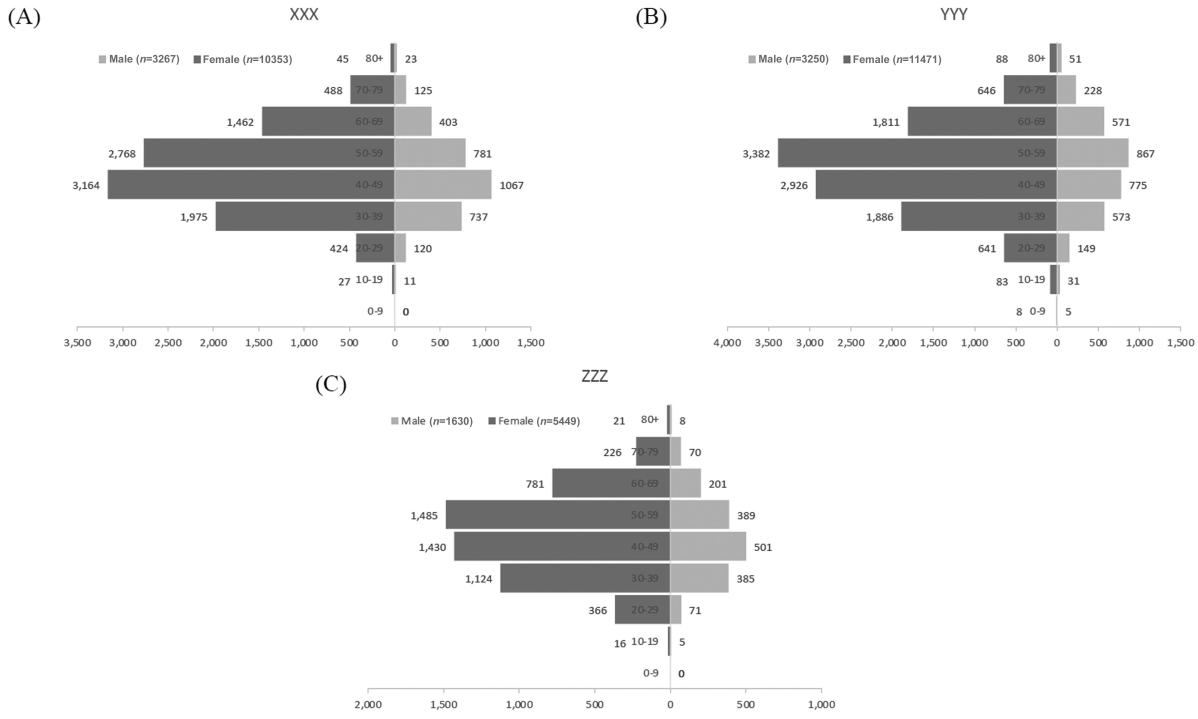



**Fig. 4** Age and sex distribution of thyroid cancer patients by three institutions: (**A**) SNUBH, (**B**) SNUH, and (**C**) CMC. The x-axis is the number of patients, and the y-axis is the age group. CMC, Catholic Medical Center; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.

**Table 4** Tumor characteristics according to the AJCC seventh TNM staging system

| Data element | SNUBH | | SNUH | | CMC | |
|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % |
| Pathologic stage T | | | | | | |
| pT1 | 4,877 | 46.4 | 5,503 | 45.4 | 2,142 | 58.3 |
| pT2 | 416 | 4.0 | 500 | 4.1 | 150 | 4.1 |
| pT3 | 5,169 | 49.2 | 6,067 | 50.0 | 1,351 | 36.8 |
| pT4 | 37 | 0.4 | 55 | 0.5 | 27 | 0.7 |
| pTx | 2 | 0.0 | 0 | 0.0 | 2 | 0.1 |
| Pathologic stage N | | | | | | |
| pN0 | 4,802 | 45.7 | 5,963 | 40.5 | 1,865 | 51.0 |
| pN1a | 3,630 | 34.6 | 3,271 | 22.2 | 1,460 | 39.9 |
| pN1b | 877 | 8.4 | 850 | 5.8 | 209 | 5.7 |
| pNx | 1,192 | 11.4 | 4,635 | 31.5 | 121 | 3.3 |
| Clinical stage M | | | | | | |
| cM0 | 8,482 | 93.6 | 17,007 | 90.2 | 27,511 | 95.9 |
| cM1 | 580 | 6.4 | 1,845 | 9.8 | 1,164 | 4.1 |
| Stage grouping | | | | | | |
| Stage I | 6,122 | 59.0 | 6,447 | 54.6 | 1,871 | 64.4 |
| Stage II | 247 | 2.4 | 425 | 3.6 | 61 | 2.1 |
| Stage III | 2,711 | 26.1 | 2,802 | 23.7 | 768 | 26.4 |
| Stage IV | 553 | 5.3 | 677 | 5.7 | 103 | 3.5 |
| Stage unknown | 748 | 7.2 | 1,455 | 12.3 | 104 | 3.6 |

Abbreviations: CMC, Catholic Medical Center; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital; TNM, tumor–node–metastasis.

**Table 5** The number of OMOP CDM records populated through NLP from free-text reports

| Table name | SNUBH | SNUH | CMC |
|---|---|---|---|
| CONDITION_OCCURRENCE | 14,047 | 14,721 | 7,079 |
| MEASUREMENT | 51,065 | 123,898 | 42,345 |
| OBSERVATION | 10,140 | 14,721 | 3,396 |
| FACT_RELATIONSHIP | 8,536 | 5,654 | 2,690 |
| NOTE_NLP | 22,176 | 29,442 | 7,366 |

Abbreviations: CMC, Catholic Medical Center; NLP, natural language processing; OMOP CDM, Observational Medical Outcome Partnership's Common Data Model; SNUBH, Seoul National University Bundang Hospital; SNUH, Seoul National University Hospital.

expression-based parsing of free-text reports generated from semistructured form, input is specific to the site and may not be replicable at other sites with fully free-text (e.g., dictation-based) reports, potentially reducing generalizability of the results. Nevertheless, we expect the results of OMOP CDM conversion of the extracted data are available to be extended to other institutions that are interested in observational thyroid cancer studies; thus, contributing to a CDM-based multicenter study. We think that it is better to perform NLP by customizing the process of extracting data from unstructured reports according to the report format specialized in each institution and standardize the extracted data using the same rules and vocabulary mapping in the OMOP data conversion process. By doing so, we believe that sharing the process of CDM conversion after NLP will be helpful for other institutions as well.

## Conclusion

In summary, the thyroid cancer staging group was derived from surgical pathology and iodine whole-body scan reports through rule-based NLP. By converting information of various thyroid cancer modifiers into the OMOP CDM oncology extension module, we tried to increase the opportunities for a multicenter observational study. As a follow-up study, we are conducting a multicenter study using the OMOP CDM oncology module on the incidence of secondary cancer after radiation exposure in thyroid cancer patients. In addition to such a population-level estimation research, we expect that our study can be extended to cohort identification, characterization, and patient-level prediction research using the OHDSI open-source analytics tool in the future. For the regular expression-based NLP in this study, further research will be needed to ensure a general applicability through the development of advanced NLP technology based on deep learning in the future.

## Clinical Relevance Statement

Our study aims to support observational thyroid cancer studies in the Observational Health Data Sciences and Informatics (OHDSI) community by converting thyroid cancer-specific data into the Observational Medical Outcome Partnership Common Data Model (OMOP CDM). The data extraction process and terminology mapping from free-text medical reports can help medical institutions to extend thyroid cancer patient data to the OMOP CDM. With CDM big data, reproducible multicenter research, including comparative effectiveness, patient prognostic prediction, and phenotyping study, can be performed for improving health and personalized treatment for thyroid cancer patients.

## Multiple Choice Questions

1. What is the common data model (CDM) used to represent thyroid cancer data in this study?
a. PCORnet
b. OMOP
c. i2b2
d. Sentinel

**Correct Answer:** The correct answer is option b. In this study, we used the Observational Medical Outcome Partnership CDM (OMOP CDM) oncology extension module to represent cancer diagnosis and modifiers.

2. From which medical report did the authors extract thyroid cancer metastasis information?
a. Surgical pathology report
b. Radiology report
c. Iodine whole-body scan report
d. Discharge summary

**Correct Answer:** The correct answer is option c. An iodine whole-body scan is performed to detect tumor metastasis or lesions for thyroid cancer patients; thus, the whole-body scan report included distant metastasis information.

### Protection of Human and Animal Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed and approved by each institutional review board of the three medical institutions participating in the study. The Common Data Model (CDM) database was retained at each medical institution, and only summary results were shared. No patient-level data were exported in this study.

### Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant

number: HI19C0378). This work was also partly supported by the Technology Innovation Program (grant number: 20003883, Advancing and expanding CDM based distributed biohealth data platform) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

## Conflict of Interest

None declared.

## References

1 Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst 2017;109(11): 1–5
2 Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77(77):34–49
3 Wang L, Luo L, Wang Y, Wampfler J, Yang P, Liu H. Natural language processing for populating lung cancer clinical research data. BMC Med Inform Decis Mak 2019;19(Suppl 5):239
4 Deshmukh PR, Phalnikar R. Anatomic stage extraction from medical reports of breast Cancer patients using natural language processing. Health Technol (Berl) 2020;10(06): 1555–1570
5 Johanna Johnsi Rani G, Gladis D, Manipadam MT, Ishitha G. Breast cancer staging using Natural Language Processing. 2015 Presented in International Conference on Advances in Computing, Communications and Informatics, ICACCI, August 10–13:2015. Kochi, India
6 Wieneke AE, Bowles EJ, Cronkite D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 2015;6(01):38
7 Yala A, Barzilay R, Salama L, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 2017;161(02): 203–211
8 Nobel JM, Puts S, Bakers FCH, Robben SGF, Dekker ALAJ. Natural language processing in dutch free text radiology reports: challenges in a small language area staging pulmonary oncology. J Digit Imaging 2020;33(04):1002–1008
9 Ryu B, Yoon E, Kim S, et al. 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
10 Spasić I, Livsey J, Keane JA, Nenadić G Text mining of cancer-related information: review of current status and future directions. Int J Med Inform 2014;83(09):605–623
11 Idarraga AJ, Luong G, Hsiao V, Schneider DF. False negative rates in benign thyroid nodule diagnosis: machine learning for detecting malignancy. J Surg Res 2021;268(268):562–569
12 Zhang Q, Zhang S, Li J, et al. Improved diagnosis of thyroid cancer aided with deep learning applied to sonographic text reports: a retrospective, multi-cohort, diagnostic study. Cancer Biol Med 2021;18j.issn.2095-3941.2020.0509
13 Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574–578
14 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(01):54–60
15 Seong Y, You SC, Ostropolets A, et al. Incorporation of korean electronic data interchange vocabulary into observational medical outcomes partnership vocabulary. Healthc Inform Res 2021; 27(01):29–38
16 Belenkaya R, Gurley M, Dymshyts D, et al. Standardized observational cancer research using the OMOP CDM oncology module. Stud Health Technol Inform 2019;264:1831–1832
17 Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FTrotti AAJCC Cancer Staging Manual. 7th ed New York, NY: Springer; 2010
18 Ortiz S, Rodríguez JM, Soria T, et al. Extrathyroid spread in papillary carcinoma of the thyroid: clinicopathological and prognostic study. Otolaryngol Head Neck Surg 2001;124(03):261–265
19 Andersen PE, Kinsella J, Loree TR, Shaha AR, Shah JP. Differentiated carcinoma of the thyroid with extrathyroidal extension. Am J Surg 1995;170(05):467–470
20 Ahn HY, Park YJ. Incidence and clinical characteristics of thyroid cancer in Korea. Korean J Med 2009;77(05):537–542
21 Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clin Cancer Inform 2021; 5:12–20