Development of an Oncology Subset of SNOMED CT Based on Patient Notes

Sina Madani, MD, PhD[1]; Jerry Henderson, MD[1]; Kin Wah Fung, MD, MS, MA[2]
[1]The University of Texas MD Anderson Cancer Center, Houston, TX
[2]National Library of Medicine, Bethesda, MD

Abstract

MD Anderson Cancer Center (MDA) is one of the world's largest institutions involved exclusively in cancer care, research, and prevention. More than 120,000 clinical transcribed documents are added to the MDA EMR system on a monthly basis. We used natural language processing methods to generate a subset of SNOMED CT concepts that are frequently documented as cancer diagnoses in patient notes.

Introduction

SNOMED CT is gaining momentum as an international clinical terminology as the membership of the International Health Terminology Standards Development Organization (IHTSDO) tripled from its inception in 2007 to 28 countries. In the U.S., SNOMED CT is the designated terminology for the problem list and procedures according to the Meaningful Use of the electronic health record incentive program, as well as for transmission of data to cancer registries[1]. Similar to the Clinical Observations Recording and Encoding (CORE) Problem List Subset of SNOMED CT[2, 3], a list of commonly used SNOMED CT concepts in oncology will save effort in mapping local terms to SNOMED CT, reduce variability in data capture and enhance data interoperability. We describe our experience in the creation of an oncology subset based on natural language processing (NLP) of patient data in a large cancer treatment center.

Methods

We retrospectively analyzed the content of more than 15 million clinical narratives entered in MD Anderson Cancer Center legacy EMR system using MetaMap NLP framework to extract active clinical problem and cancer-related concepts at the patient level. After careful examinations of various note types and consulting with subject matter experts, we decided to target the last ten instances of the targeted clinical notes for all patients inside our EMR repositories. The targeted note types included Discharge Summary, Emergency, History & Physical, Consultation, Primary Medical Evaluation, Progress, and Clinical notes. We defined a "Problem" or "Cancer Disease" as something that required a plan for "diagnosis" and/or "management". Therefore, from the selected note types mentioned above, all section headers related to the *Assessment & Plan* and *Diagnosis* sections, including subsections such as "Cancer Diagnosis", together with their contents were extracted and analyzed by MetaMap v2014. We empirically set the threshold of MetaMap output to a score of 580. We restricted the UMLS targets to the "Disorders" semantic group, which included the semantic type "Neoplastic Process" for cancer-related concepts. We have also developed a post-processing module that prevented particular trigger strings within the narratives being mapped incorrectly to UMLS concepts, specifically, when the target concept was represented in an abbreviated format (like "gist" for Gastrointestinal Stromal Tumor). Such functionality has since been added to the latest version of MetaMap (v2016). One Extensible Markup Language (XML) file was generated for each patient note and serialized into a relational database. Phrases that MetaMap failed to map were stored separately for review. We applied the CORE Problem List subset (v2015) as an initial filter for removing unwanted and/or irrelevant concepts from the MetaMap output. Two trained physicians reviewed all concepts that were not represented in the SNOMED CORE subset and identified the ones that were relevant to cancer diagnosis. We calculated concept usage index by the occurrence of a cancer concept in a patient's problem list by the total number of recorded cancer across all patients. To evaluate our NLP pipeline, we manually reviewed the output in two random patient samples (one for

general problems and the other for cancer diagnoses) to calculate the standard NLP performance metrics (precision, recall, F-measure).

Results

Based on the selection criteria mentioned in the method section, 554,801 unique patient records associated with 2,998,322 notes and 3,404,575 section headers were processed in the MetaMap pipeline. We identified 563 synonyms for the two categories of the target section headers. The performance metrics of the section header identification algorithm calculated as 97%, 99%, and 98% for precision, recall, and F-measure respectively. More than 3.7 million instances of cancer concepts corresponding to 2,698 unique concepts were extracted (Table 1).

UMLS CUI	Concept	SNOMED Code	Usage Index		
			by instance	by note	by patient
C0006142	Malignant tumor of breast	254837009	6.7	6.4	4.6
C0025202	Malignant melanoma	372244006	4.9	4.3	2.4
C0024299	Malignant lymphoma	118600007	2.7	2.4	1.9
C0376358	Malignant tumor of prostate	399068003	2.5	2.5	2.3
C0007131	Non-small cell lung cancer	254637007	2.1	2.2	1.7

Table 1. Top 5 cancer concepts and their usage index by instance, note, and patient frequencies.

NLP performance evaluation for the general problem list concepts (Disorders semantic group) on twenty randomly selected patients showed 94%, 90%, and 92% for recall, precision, and F-Measure respectively. We also evaluated performance metrics for only Neoplastic Process semantic type on a separate group of twenty randomly selected patients and calculated 100%, 83%, and 90% (recall, precision, and F-Measure). Nine cancer diagnosis discovered by MetaMap in this group (like Carcinosarcoma vs. Ovarian Carcinosarcoma) were considered as "too general" by the evaluating subject matter experts.

Discussion

We showed that it is feasible to use MetaMap to extract cancer-related SNOMED CT concepts from narrative patient notes. The oncology subset will be made available for download through NLM's website by any user with a SNOMED CT license. While the subset is not meant to be exhaustive, it can be used as a starter set as it is expected to cover the majority of SNOMED CT concept needed in most cancer treatment institutions. During the creation of the subset, we encountered some concepts that had considerable usage but were not in SNOMED CT. We would review the unmapped *concepts* for validity and submit them to IHTSDO as suggested additions to SNOMED CT. Furthermore, we plan to use NLP outputs for a gap analysis between a vendor's terminology and the internally identified cancer *synonyms* from MDA corpora for the enrichment of the existing provider friendly terminology incorporated within Epic system.

References

1. Health Information Technology Certification Criteria, U.S. Health and Human Services Department on 10/16/2015. Available from: https://www.federalregister.gov/articles/2015/10/16/2015-25597/2015-edition-health-information-technology-health-it-certification-criteria-2015-edition-base.
2. Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 2010;17(6):675-80. Epub 2010/10/22.
3. Fung KW, Xu J. An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT. J Am Med Inform Assoc. 2015;22(3):649-58. Epub 2015/03/01.