SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1
Leading health innovation through data

201914 Implementation of SNOMED CT for knowledge representation of biomedical literature: A case study for cancer behavioral risk factors knowledge base

Jiang Bian, University of Florida (US)

Co-authors

1. Hansi Zang
2. Yi Guo

Summary

We describe the curation of a Cancer Behavioral Risk Factors(CBRFs) Knowledge Base and provide a formal ontological representation for CBRFs with evidence-based information extracted from scientific literature. We will focus on our experience of using SNOMED CT to standardize the extracted knowledge

Audience

Research/academic

Learning Objectives

1. Learn the importance of accessing high-quality health information online.
2. Learn the benefits of having a cancer behavioral risk factors knowledge base.
3. Learn the benefits of implementing a controlled terminology provided by SNOMED CT to standardize the concepts extracted from biomedical literature

Abstract

SNOMED CT is a systematically organized computer processable collection of medical terms covering codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED CT is also accepted as a common global language for health terms in more than 50 countries. However, a term on its own does nothing unless it is implemented as part of an application and used. There are mainly three types of implementations of SNOMED CT: clinical records, knowledge representations and aggregations and analysis. In this abstract, we focus on using SNOMED CT expressions to represent or tag knowledge resources that maybe clinically relevant within the domain of cancer behavioral risk factors (CBRFs).

Varying level of evidence has linked the development of cancer to a wide range of risk factors. Many of these factors cannot be altered, such as age, sex and family history, while risky health behaviors including smoking, alcohol drinking, inadequate physical activity, and overweight can be avoided and managed. Recognized by well-established health behavior theories, an individual's health behavior (e.g., smoking) is determined by her intention, while intention is directly influenced by her knowledge, attitudes, among many other factors. Nevertheless, research has shown that the public's awareness of these CBRFs is poor and the public lacks the necessary knowledge towards a healthy lifestyle. According to a recent Pew Research report, 72% adults internet users in the United States searched online for health information. However, existing online health information about CBRFs is poorly organized, not

evidence-based, and confusing to lay health information consumers. In this study, we present a semantic web CBRF-knowledge base (CBRF-KB) that use a formal and standardized knowledge representation (i.e., ontology) to better organize and manage evidence-based CBRF-related information extracted from relevant, high-quality scientific literature (i.e., PubMed abstract).

The initial CBRF-KB was built based on CBRF-related factual statements manually extracted from relevant PubMed abstracts. We focused on the top 4 CBRFs including smoking, alcohol drinking, physical activity, and overweight and identified the relevant articles using the following keyword search strategy: a combination of cancer related keywords and their variations (e.g., "cancer", "neoplasm") and CBRF specific keywords and their variations (e.g., "tobacco", "smoking", "cigarette"). To ensure the quality of KB, we only extracted articles published in journals with an impact factor ≥ 8, which yielded 59 abstracts meeting the criteria. We then manually annotated these abstracts to extract concept classes (i.e., concepts related to CBRFs and cancer) and the relations among them and expressed them as factual statements in the form of triples (i.e., subject-predicate-object). Out of the 59 abstracts, we extracted 126 concept classes, 53 relations, and 374 triple statements.

The concepts (e.g., "alcohol drinking", "alcohol intake") and relations (e.g., "significantly increased risk for", "associated with a significantly increased risk of",) are not standardized, commonly used terms. Therefore, we built a CBRFs Ontology (CBRFO) to provide a controlled vocabulary that standardizes and organizes these concepts and relations. Following best practice in ontology development, we first considered reuse classes and relations from existing well-known ontologies if available and created new classes and relations only when it was necessary. To do so, we first identified high-quality candidate ontologies related to the 4 CBRFs (i.e., smoking, alcohol drinking, physical activity, and overweight) and cancer through National Center for Biomedical Ontology BioPortal (the largest repository for biomedical ontologies). An ontology is considered as a candidate if it contains the terms relevant to the 4 CBRFs or cancer. We excluded ontologies that were not yearly updated, and the last updates were earlier than 2018. Out of the 126 concept classes and 53 relations, we obtained 119 unique classes and 44 unique relations. The same concept (or relation) may exist in multiple ontologies; thus, we used an automated ontology alignment tool (i.e., LogMap) that links the same concept across different ontologies. We selected 3 main ontologies: National Cancer Institute Thesaurus (NCIt), Relation Ontology (RO), and Time Event Ontology (TEO) as the foundation for creating CBRFO, which covered 88.23% of the 119 concept classes and 27.27% of the 44 relations.

Knowledge bases are essential and have been widely used in clinical decision support (CDS) systems. One of the important implementations of our CBRF-KB is to support healthcare providers in making well-informed clinical decisions. For example, the CBRF-KB stores evidence-based factual statements describing how different behavioral risk factors impact certain cancers. Thus, by incorporating CBRF-KB to an electronic health record (EHR) system, it can assist clinicians in identifying patients at high risk of getting cancer based on patient's existing health behaviors and provide real-time assistant on delivering tailored educational information to patients for behavior changes. To do so, SNOMED CT is a perfect substrate for providing semantic interoperability to a wide range of EHR systems that already use SNOMED CT. A lot of works has been done to build knowledge bases using SNOMED CT to organize information or support clinical decision making. As the most comprehensive terminology in the world, SNOMED CT can help us organize, manage and map the concepts and relation extracted from biomedical literature to clinical terms commonly used in EHRs, thus, facilitating integrating CBRF-KB into EHR systems in the future.

In this study, we mapped concepts and relations in CBRF-KB (extracted from biomedical literature) with the concepts and relations in SNOMED CT. We found 70% (i.e., 83 out of 119) of the concepts and 14% (i.e., 6 out of 44) of the relations can be mapped to concepts and relations in SNOMED CT, respectively. Based on the concept hierarchy of the SNOMED CT, among the 83 mapped concept classes, 48 are clinical findings (e.g., "cancer", "obesity"), 12 are observable entities (e.g., "birth weight", "body mass index"), 8 are body structures (e.g., "polyp", "meningioma"), 6 are quantifier values (e.g., "early stage", "increased"), 4 are procedures (e.g., "chemotherapy", "radiotherapy"), 3 are social context (e.g., "adult", "woman") and 2 are environment or geographical locations (e.g., "United States of America", "India"). As we can see, clinical findings (i.e., related to disease) are the most commonly extracted concepts from the PubMed abstracts.

For those 36 concepts that cannot be mapped to SNOMED CT, we summarized the reasons as follows: 1) issues related to the granularity of a concept (e.g., we mapped the concept “biochemically recurrent prostate carcinoma” a subclass of “recurrent prostate carcinoma” in NCIt, while SNOMED CT only contains “recurrent prostate carcinoma”); and 2) the corresponding concept does not exist in SNOMED CT likely because that the concepts do not fit the scope of SNOMED CT (e.g., “cancer incidence”, “cancer related death”). We found that 25 of all unmapped concepts are not clinical terms. Since the implementation of the CBRF-KB is focusing on incorporating with the EHR system, our primary mission is to organize and manage clinical terms commonly used in EHRs to support clinical decision making. Therefore, these unmapped concepts, especially those that are not clinical terms, will not be an issue in capturing patients’ clinical information in the EHR system.

For the relation terms in CBRF-KB, the majority of them cannot be mapped to the relations in SNOMED CT. The most important reason is the granularity issue of the relation classes in SNOMED CT. For example, we extracted relations such as “associated with”, “significantly associated with”, “not significantly associated with”, “positively associated with”, and “inversely associated with” from scientific literature, while SNOMED CT only contains a high-level relation “associated with”. Nevertheless, to help clinicians make well-informed clinical decisions, it is important for the clinicians (as well as the patients) to understand the different levels of association between a risk factor and a cancer. For example, a clinician may consider addressing high-priority factors that are significantly associated with cancers than those that are only “positively associated with”. Similarly, when giving advices to patients, those positively associated risk factors should be differentiated from those negatively associated factors. Thus, in CBRF-KB, we primarily used relations with more granularity from RO and TEO and created relations in CBRFO for those that were not found in RO and TEO to represent extracted relations. Further, we grouped relations that are essentially the same but with different granularity and mapped these relation groups to the high-level SNOMED CT relations (e.g., “associated with”). Doing so will facilitate integration of the triple statements in CBRF-KB that use these relations with the clinical concepts expressed in SNOMED CT. Moreover, based on our experience in creating CBRF-KB, SNOMED CT may benefit from enriching its representation on relation class granularity to better support clinical decision making.

We also manually checked 83 mapped concepts in SNOMED CT and found that all these concepts in SNOMED CT have been mapped to other international standards and well-known ontologies (e.g., NCIt, ICD-9/10, and LOINC). Therefore, when we integrate our CBRF-KB with other KBs and EHR systems, SNOMED CT could ensure the semantic interoperability and enable relevant information to be recorded using consistent, common representations.