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

Jiang Bian[1], Hansi Zhang[1], Yi Guo[1] | Department of Health Outcomes & Biomedical Informatics, University of Florida

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS SPONSORED BY:
tpp

INTRODUCTION

END SHOW

TITLE SLIDE

Click on a bubble to navigate the e-poster

introduction

methods

results

discussion

Abstract

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.

Background

Why we build the cancer CBRF-KB?

- An immense amount of evidence from research studies has linked the development of cancer to a wide range of risk factors[1,2]
- The general public's awareness of cancer behavioral risk factors (CBRFs) is poor; and even when they are aware, they lack the necessary knowledge towards a healthy lifestyle.
- Given that 72% adult internet users in the United States searched online for health information, the Internet is a great venue to disseminate CBRF information[3]. However, existing CBRF information online is poorly organized, not evidenced-based, and confusing to health information consumers.

4 IN 10 CANCER CASES CAN BE PREVENTED...
...MAKE A CHANGE TO REDUCE THE RISK OF CANCER
Be smoke free
Keep a healthy weight
Be safe in the sun
Avoid certain substances at work
such as asbestos
Protect against certain infections
such as HPV and H.Pylori
Drink less alcohol
Eat a high fibre diet
Avoid unnecessary radiation
including radon gas and x-rays
Cut down on processed meat
Avoid air pollution
Breastfeed if possible
Be more active
Minimise HRT use
Larger circles indicate more UK cancer cases
Circle size here is not relative to other infographics based on Brown et al 2018.
Source: Brown et al, British Journal of Cancer, 2018.
LET'S BEAT CANCER SOONER
cruk.org
CANCER RESEARCH UK

Why we use SNOMED CT?

- One of the important implementations of our CBRF-KB is to support healthcare providers in making well-informed clinical decisions
- **Our ultimate goal:** 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
- SNOMED CT is a perfect substrate for providing semantic interoperability to a wide range of EHR systems that already use SNOMED CT.

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

Jiang Bian[1], Hansi Zhang[1], Yi Guo[1] | Department of Health Outcomes & Biomedical Informatics, University of Florida

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS SPONSORED BY:
tpp

METHOD

END SHOW

TITLE SLIDE

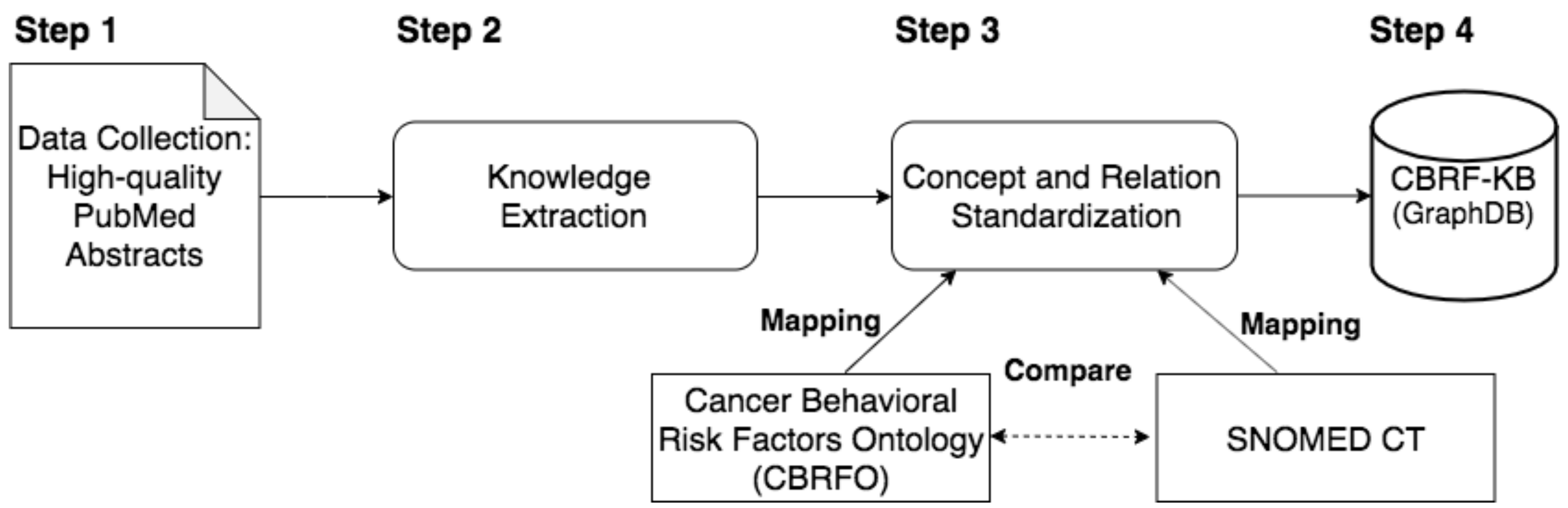
Click on a bubble to navigate the e-poster

introduction

methods

results

discussion


Step 1
Step 2
Step 3
Step 4
Data Collection: High-quality PubMed Abstracts
Knowledge Extraction
Concept and Relation Standardization
CBRF-KB (GraphDB)
Mapping
Mapping
Compare
Cancer Behavioral Risk Factors Ontology (CBRFO)
SNOMED CT

Step 1: Data collection.

- **Risk factors**: smoking, alcohol drinking, physical activity, and overweight
- **Search source and query**: For each risk factor, we searched **PubMed** using **risk factor keywords** (e.g., "smoking", "cigarette") in combination with **cancer keywords** (e.g., "cancer", "neoplasm") considering the synonyms for each keyword.
- **Quality assurance**: review articles with factors > 8

Step 2: Knowledge extraction.

- **Extraction process:** Two annotators reviewed each abstract and extracted information relevant to either cancer or CBRFs, and expressed them as factual statements in the form of triples (i.e., subject-predicate-object)

Step 3: Concept and relation standardization.

- **CBRFO mapping**: We built a CBRF Ontology (CBRFO) to provide a controlled vocabulary to standardize the extracted terms (e.g., "alcohol drinking", "alcohol intake") and relations (e.g., "significantly increased risk for", "associated with a significantly increased risk of").
- **SNOMED mapping:** We also mapped concepts and relations in CBRF-KB (extracted from biomedical literature) with the concepts and relations in SNOMED CT.

Step 4: Triple and associated provenance data management.

- **Triple format:** Nanopublication **-** A nanopublication has three basic elements

(1) an assertion (e.g., smoking – significant associated with – lung cancer risk)
(2) the provenance (e.g., extraction time, annotator).
(3) associated publication information (e.g., author, title, and published time of the article where the triple is extracted from)

- **Triple Store:** GraphDB - a popular graph database with inference and SPARQL query support.

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

Jiang Bian[1], Hansi Zhang[1], Yi Guo[1] | Department of Health Outcomes & Biomedical Informatics, University of Florida

RESULTS

END SHOW
TITLE SLIDE

Click on a bubble to navigate the e-poster

introduction
methods
results
discussion

Annotation result

Annotation	Count
PubMed Abstract	59 abstracts
Classes	119 concept classes
Relations	44 relations
Triple statements	374 triple statements

Ontology development

Reference ontology: We selected 3 main ontologies, National Cancer Institute Thesaurus (NCIt), Relation Ontology (RO), and Time Event Ontology (TEO) as the foundation for creating CBRFO

Mapping result comparison

Mappings	CBRFO (%)	SNOMED CT (%)
Classes (N = 119)	105 (88.23%)	83 (70%)
Relations (N = 44)	12 (27.27%)	6 (14%)

SNOMET CT mapping details

Mapped class classification

Mapped class category	Coverage (%) N = 83	Examples
Clinical findings	48 (57.82%)	“cancer”, “obesity”
Observable entities	12 (14.46%)	“birth weight”, “body mass index”
Body structures	8 (9.64%)	“polyp”, “meningioma”
Quantifier values	6 (7.23%)	“early stage”, “increased”
Procedures	4 (4.82%)	“chemotherapy”, “radiotherapy”
Social context	3 (3.61%)	“adult”, “woman”
Environment / geographical locations	2 (2.41%)	“United States of America”, “India”

Reasons summary for not mapped class:

- Issues related to the granularity of a concept
  - We mapped the concept “biochemically recurrent prostate carcinoma” a subclass of “recurrent prostate carcinoma” in NCIt, while SNOMED CT only contains “recurrent prostate carcinoma”
- The corresponding concept does not exist in SNOMED CT likely because that the concepts do not fit the scope (clinical terms) of SNOMED CT
  - e.g., “cancer incidence”, “cancer related death”, not clinical terms

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

Jiang Bian[1], Hansi Zhang[1], Yi Guo[1] | Department of Health Outcomes & Biomedical Informatics, University of Florida

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS SPONSORED BY:
tpp

DISCUSSION

END SHOW
TITLE SLIDE

Click on a bubble to navigate the e-poster

introduction
methods
results
discussion

Relation mapping summary

- For the relation terms in CBRF-KB, the majority of them cannot be mapped to the relations in SNOMED CT.
- **Reason**: granularity issue of the relation classes in SNOMED CT
  - e.g., “associated with” **VS.** “significantly associated with”, “not significantly associated with”, “positively associated with”, and “inversely associated with”
- **Solution:** group 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”)

Conclusion and future work

- We curated a CBRF-KB to better organize high-quality evidence extracted from scientific literature on the relationships between various behavioral risk factors and cancer.
- To build CBRF-KB, we created the CBRFO ontology and compared it with SNOMED CT to standardize the terms and relations used across different articles.

Conclusion and future work (continue.)

- Base 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
- As the most comprehensive clinical 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.

References

1. Institute of Medicine (U.S.), F. A. Sloan, and H. Gelband, Eds., Cancer control opportunities in low- and middle-income countries. Washington, DC: National Academies Press, 2007.
2. National Cancer Institute, “Risk Factors for Cancer,” 23-Dec-2015. [Online]. Available: https://www.cancer.gov/about-cancer/causes-prevention/risk. [Accessed: 20-Jun-2019].
3. Susannah Fox, “The social life of health information,” 15-Jan-2014. [Online]. Available: https://www.pewresearch.org/fact-tank/2014/01/15/the-social-lifeof-health-information/. [Accessed: 25-Apr-2019].