

# Aero: An Evidence-based Semantic Web Knowledge Base of Cancer Behavioral Risk Factors

Hansi Zhang<sup>1</sup>, Xing He<sup>1</sup>, Tyler Harrison<sup>1</sup>, and Jiang Bian<sup>1\*</sup>

<sup>1</sup> University of Florida, Gainesville FL 08544, USA

\*Corresponding author: bianjiang@ufl.edu

**Abstract.** 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. However, existing CBRF information online is poorly organized, not evidenced-based, and confusing to health information consumers. In this paper, we present a prototype semantic web cAncer bEhavioral Risk knOwledgebase—**Aero** to (1) better organize and provide evidence-based CBRF knowledge extracted from scientific literature (i.e., PubMed), and (2) provide users with access to high-quality scientific knowledge, yet easy to understand answers for their frequently encountered CBRF questions. Our current prototype focuses on the top 4 types of CBRFs: smoking, alcohol drinking, physical activity, and overweight. We manually annotated 59 high-quality PubMed abstracts (i.e., review articles with impact factor  $\geq 8$ ) and created a preliminary version of Aero with 787 triples. We built an interactive user interface with graph-based visualization of the KB, where users can explore answers to commonly asked CBRF questions according to the cancer risk factor fact sheet of National Cancer Institute. A preliminary evaluation of Aero was also conducted.

**Keywords:** Ontology, Semantic Web Knowledge Base, Cancer Behavioral Risk Factors, Question Answering, Interactive Graph-based Visualization

## 1 Introduction

Cancer is the second leading cause of death worldwide and responsible for an estimated 9.6 million deaths in 2018 [1]. An immense amount of evidence from research studies has linked the development of cancer to a wide range of risk factors [2]. Many of these factors cannot be altered, such as age, sex and family history; while risky health behaviors (e.g., smoking and overweight) can be avoided and managed [3]. Recognized by the integrated behavioral model (IBM) [4], an individual's health behavior is determined by her intension, while intention is directly influenced by her knowledge, attitudes, among many other factors. Nevertheless, research has shown that the public's awareness of these cancer behavioral risk factors (CBRFs) is poor, and the public lacks the necessary knowledge towards a healthy lifestyle [5]. Given that 72% adults internet users in the United States (US) searched online for health information [6], the Internet is a great communication venue to disseminate CBRF-related health information. However, existing online information about CBRFs is not well-organized, not evidenced-based, and of poor quality. Much of this online information consists of personal opinion, salesmanship, testimonials, and claims that are not evidence-based (i.e., supported by high-quality scientific literature and/or scientific consensus). Even in scientific literature, evidence describing the relationships between various cancers and CBRFs are heterogenous ranging from pre-clinical models and case studies to mere hypothesis-based arguments. In short, current wealth of online health information on CBRFs is overwhelming and disorganized. We believe that a formal knowledge representation model (e.g., ontology) along with associated Semantic Web technology stack can help

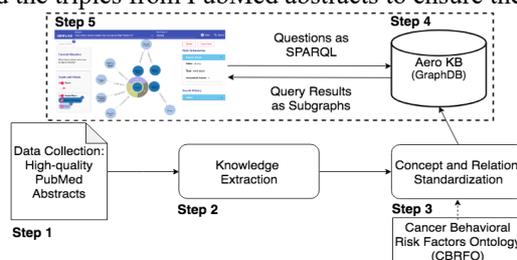
Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

organize and present quality health information to the public.

Previously, Lossio-Ventura *et al.* created a natural language processing (NLP)-based system to construct an obesity and cancer knowledge base [7–9]. However, the system is limited by the performance of the NLP methods and it only focused on one risk factor. In this paper, we present a prototype semantic web **c**Ancer **b**Ehavioral **R**isk **kn**Owledgebase—Aero to (1) better organize evidence-based CBRF-related knowledge extracted from free-text scientific literature (i.e., PubMed abstract), and (2) provide users with high-quality, yet easy to understand answers to their commonly asked health questions relevant to CBRFs. Our preliminary work of Aero consists of 787 triples extracted from 59 annotated high-quality PubMed abstracts (i.e., impact factor  $\geq 8$ ). We constructed semantic queries based on commonly asked questions in the National Cancer Institute (NCI)’s cancer risk factor (CRF) fact sheet [10]; and evaluated Aero by comparing its query results with the answers from the NCI fact sheet. We have also built an prototype interactive user interface (UI) for Aero.

## 2 Methods

**Fig. 1** illustrates our process of creating Aero. Rather than using an NLP system, we manually extracted the triples from PubMed abstracts to ensure the quality of the KB.



**Fig. 1.** The process of creating Aero.

**Step 1: Data collection.** The initial Aero KB focused on 4 types of CBRFs: smoking, alcohol drinking, physical activity, and overweight. 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. To ensure the quality of evidence, we only considered articles that were published in high-quality journals (i.e., impact factors  $> 8$ ). Two annotators screened each abstract to filter out articles whose study results are not related to CBRFs. Note that an article that discussed CBRF (e.g., in the introduction) but does not itself generate results or evidence indicating the relationships between the CBRF and cancer is excluded.

**Step 2: Knowledge extraction.** Two annotators reviewed each abstract and extracted information relevant to either cancer or CBRFs, paying more attention to the direct relationships between the two. The extracted knowledge is represented as semantic triples (i.e., “*subject-predicate-object*”). Our process and annotation guideline are as follows: (1) for each abstract, identify all terms and sentences that related to cancer and the CBRF of interest; (2) the terms and sentences must be identified as study results or conclusion; (3) identify the relations between the extracted terms within the sentence; and (4) construct individual triples using the extracted terms and relations.

**Step 3: Concept and relation standardization.** 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*”). Following best practices in ontology

engineering, we first considered reusing concept 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 (widely used, regularly maintained) candidate ontologies related to the 4 CBRFs of interest and cancer using the National Center for Biomedical Ontology BioPortal. An ontology is considered as a candidate if it contains the terms relevant to the 4 CBRFs or cancer. The same concept (or relation) may exist in multiple ontologies; thus, we used an ontology alignment tool (i.e., LogMap [11]) to link 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.

**Step 4: Triple and associated provenance data management.** We organized extracted semantic triples and corresponding provenance data in the form of a nanopublication [12]. Provenance data of the extracted triples are important and can facilitate consumers of the KB to form assessments of its quality. A nanopublication has three basic elements: (1) an assertion; (2) the provenance (e.g., extraction time, annotator); and (3) associated publication information (e.g., author, title, and published time of the article where the triple is extracted from). We then used a python library, RDFLib, to serialize all nanopublications into Resource Description Framework (RDF) using TriG syntax [13]. We stored all serialized RDF triples in GraphDB—a popular graph database with inference and SPARQL query support.

**Step 5: User interface (UI) design.** We created a prototype UI with interactive graph-based visualizations. Previously, we have shown that graph-based visualizations stimulating visual thinking and help end-users better comprehend the presented information [14, 15]. The UI consists of two main parts: (1) a top bar for users to select a set of predefined question templates (i.e., common questions related to CBRF and cancer, summarized from NCI’s CRF fact sheet), and (2) a canvas for graph-based interactive visualization of the query results. We have also implemented a number of other convenient functions (e.g., visualization options such as zooming and filtering).

## 3 Results

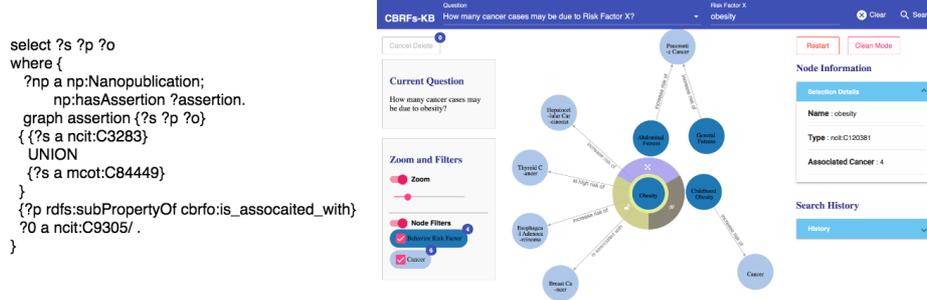
### 3.1 An Aero prototype

We first identified 169 articles published in journals with impact factors equal or greater than 8. Two annotators reviewed these articles and retained 59 articles that are relevant based on the inclusion criteria (i.e., inter-rater agreement: 0.8421). The two annotators further extracted 126 concept classes, 53 relations, and 787 triple statements (i.e., inter-rater agreement: 0.7241). Out of the 787 triples, 374 are assertions of CBRFs, 118 are associated provenance data, and 295 are used to describe the publication information. Out of the 126 concept classes and 53 relations, we obtained 119 unique classes and 44 unique relations. The selected 3 ontologies (i.e., NCIt, RO, TEO) for creating the CBRFO cover 88.23% of the 119 concept classes and 27.27% of the 44 relations. New classes and relations were created in CBRFO to provide full coverage. Then, for each article, we represented the extracted triple statements and associated provenance data in the form of a nanopublication and imported into GraphDB.

### 3.2 Question answering with graph-based interactive visualization in Aero

We extracted 53 questions related to the 4 CBRFs from the NCI CRF fact sheet and summarized them into 3 categories: (1) “*What is known about the relationship between X and cancer?*”; (2) “*Does X cause cancer and other disease?*”; and (3) “*What research being done related to X and cancer?*”, where the X refers to a specific CBRF

and “cancer” can refer to cancer in general or a specific type of cancer (e.g., lung cancer, oral cancer). We then created 3 SPARQL query templates for these 3 categories of questions. **Fig 2** shows an example SPARQL query for the question “*What is known about the relationship between obesity and cancer?*”. We simply used the parent class “cancer” (i.e., `ncit:C9305`) in the query and the reasoner will automatically consider all subclasses of cancer associated with obesity (**Fig 2**).



**Fig. 2.** An example SPARQL query (left) and the interactive visualization of the query results (right) for question “*What is known about the relationship between obesity and cancer?*” \*In SPARQL queries, variables are prefixed with “?”, where “*?s*” represents a CBRF (i.e., obesity [`ncit:C3283`] and childhood obesity [`ncit:C84449`]), “*?o*” represents cancer (i.e., NCI id `ncit:C9305`), and “*?p*” represents the relation.

We manually evaluated the query results of Aero in terms of whether the retrieved triples accurately (i.e., precision) and comprehensively (recall) cover the answers on the NCI risk factor fact sheet, focusing on one CBRF (i.e., obesity).

**Table 1.** Aero query result performance in terms of its ability to answer consumer questions.

Question	Precision	Recall	F-score
Question type 1: “ <i>What is known about the relationship between obesity and cancer?</i> ”	1	0.29	0.45
Question type 2: “ <i>Do smoking cause cancer and other diseases?</i> ”	0.5	0.43	0.46
Question type 3: “ <i>What research being done related to obesity and cancer?</i> ”	0.1	0.33	0.15

#### 4 Discussion and conclusion

We curated a semantic web KB (i.e., Aero) to better organize high-quality evidence extracted from scientific literature on the relationships between various behavioral risk factors and cancer. To build Aero, we created the CBRFO ontology to standardize the terms and relations used across different articles. Further, we experimented with interactive graph-based visualizations to provide consumers with an easy to understand visual representation of the answers to commonly asked CBRF questions, stimulating their visual thinking. Given how frequent that the general public searches online for health information, our ultimate goal for Aero is to provide evidence-based CRBF information that can lead to behavioral change towards a healthy lifestyle.

Our current study is still limited. Only 59 articles were annotated limiting the coverage of the KB. Manual annotation is labor-intensive and time-consuming. Thus, we are actively investigating a crowdsourcing solution that can improve the efficiency of the KB curation process at scale. Further, the usability of the Aero UI needs to be assessed, and any usability issues raised should be addressed with inputs from stakeholders especially the lay consumers following a user-centered design process.

## References

1. World Health Organization. Cancer - key facts. 2018. <https://www.who.int/news-room/fact-sheets/detail/cancer>. Accessed 3 Jul 2019.
2. National Cancer Institute. Risk Factors for Cancer. 2015. <https://www.cancer.gov/about-cancer/causes-prevention/risk>. Accessed 20 Jun 2019.
3. Cancer Research UK. The causes of cancer you can control. 2011. <https://science-blog.cancerresearchuk.org/2011/12/07/the-causes-of-cancer-you-can-control/>. Accessed 3 Jul 2019.
4. Glanz K, Rimer BK, Viswanath K, editors. Health behavior and health education: theory, research, and practice. 4th ed. San Francisco, CA: Jossey-Bass; 2008.
5. Ryan AM, Cushen S, Schellekens H, Bhuachalla EN, Burns L, Kenny U, et al. Poor Awareness of Risk Factors for Cancer in Irish Adults: Results of a Large Survey and Review of the Literature. *The Oncologist*. 2015;20:372–8.
6. Susannah Fox. The social life of health information. 2014. <https://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/>. Accessed 25 Apr 2019.
7. Lossio-Ventura JA, Hogan W, Modave F, Hicks A, Hanna J, Guo Y, et al. Towards an Obesity-Cancer Knowledge Base: Biomedical Entity Identification and Relation Detection. *Proc IEEE Int Conf Bioinforma Biomed*. 2016;2016:1081–8.
8. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, Hicks A, et al. OC-2-KB: A software pipeline to build an evidence-based obesity and cancer knowledge base. *Proc IEEE Int Conf Bioinforma Biomed*. 2017;2017:1284–7.
9. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, Yang X, et al. OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system. *BMC Med Inform Decis Mak*. 2018;18 Suppl 2:55.
10. National Cancer Institute. NCI Fact Sheets - Risk Factors and Possible Causes. 2019. <https://www.cancer.gov/publications/fact-sheets>. Accessed 3 Jul 2019.
11. Jiménez-Ruiz E, Cuenca Grau B. LogMap: Logic-Based and Scalable Ontology Matching. In: Aroyo L, Welty C, Alani H, Taylor J, Bernstein A, Kagal L, et al., editors. *The Semantic Web – ISWC 2011*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 273–288.
12. Paul Groth, Erik Schultes, Mark Thompson, Zuotian Tatum, Tobias Kuhn, Christine Chichester. Nanopublication Guidelines. 2018. [http://nanopub.org/guidelines/working\\_draft/](http://nanopub.org/guidelines/working_draft/). Accessed 3 Jul 2019.
13. Chris Bizer, Richard Cyganiak. RDF 1.1 TriG. 2014. <https://www.w3.org/TR/trig/>. Accessed 3 Jul 2019.
14. Bian J, Xie M, Hudson TJ, Eswaran H, Brochhausen M, Hanna J, et al. CollaborationViz: interactive visual exploration of biomedical research collaboration networks. *PloS One*. 2014;9:e111928.
15. He X, Zhang R, Rizvi R, Vasilakes J, Yang X, Guo Y, et al. Prototyping an Interactive Visualization of Dietary Supplement Knowledge Graph. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE; 2018. p. 1649–52. doi:10.1109/BIBM.2018.8621340.