*Public Health and Informatics*
*J. Mantas et al. (Eds.)*

*doi:10.3233/SHTI210267*

# Building a Knowledge Graph Representing Causal Associations Between Risk Factors and Incidence of Breast Cancer

Ali DAOWD[a,1], Michael BARRETT[a], Samina ABIDI[b], and Syed Sibte Raza ABIDI[a]
[a]*NICHE Research Group, Faculty of Computer Science, Dalhousie University, Canada*
[b]*Medical Informatics, Department of Community Health and Epidemiology, Dalhousie University, Canada*

**Abstract.** This paper explores the use of semantic- and evidence-based biomedical knowledge to build the *RiskExplorer knowledge graph* that outlines causal associations between risk factors and chronic disease or cancers. The intent of this work is to offer an interactive knowledge synthesis platform to empower health-information-seeking individuals to learn about and mitigate modifiable risk factors. Our approach analyzes biomedical text (from PubMed abstracts), Semantic Medline database, evidence-based semantic associations, literature-based discovery, and graph database to discover associations between risk factors and breast cancer. Our methodological framework involves (a) identifying relevant literature on specified chronic diseases or cancers, (b) extracting semantic associations via knowledge mining tool, (c) building rich semantic graph by transforming semantic associations to nodes and edges, (d) applying frequency-based methods and using semantic edge properties to traverse the graph and identify meaningful multi-node *NCD risk paths*. Generated multi-node risk paths consist of a source node (representing the source risk factor), one or more intermediate nodes (representing biomedical phenotypes), a target node (representing a chronic disease or cancer), and edges between nodes representing meaningful semantic associations. The results demonstrate that our methodology is capable of generating biomedically valid knowledge related to causal risk and protective factors related to breast cancer.

**Keywords.** Knowledge graph, literature-based discovery, disease risk mitigation

## 1. Introduction

Non-communicable Chronic Diseases (NCDs) are most prevalent worldwide and have long been the leading cause of death and disability. The COVID-19 pandemic has further underlined the growing NCD crisis and the cumulative failure of public health measures to handle the complications of NCD due to COVID-19. Data from COVID-19 literature show that patients with NCD are more vulnerable to a severe course of the illness and associated complications [1]. Therefore, there is an urgent need to re-focus preventive health measures for NCD by offering individuals access to novel digital health solutions that support chronic disease prevention and empower individuals to adopt healthier lifestyles in order to mitigate the risk factors leading to the onset of NCD. While

[1] Corresponding Author: Ali Daowd, NICHE Research Group, Faculty of Computer Science, Dalhousie University, Canada, E-mail: ali.daowd@dal.ca.

evidence-based literature on NCD risk factors is growing exponentially, it is not organized in a manner that is accessible to health consumers. With recent technological advances, large amount of literature can be mined to organize validated knowledge such that consumers can apply it to assess and reduce their health risks.

In this paper, we present a novel Literature-Based Discovery (LBD) and knowledge graph based health knowledge synthesis and discovery approach to develop the *RiskExplorer* graph-based knowledge bases to capture and represent the multi-causal association between NCDs and their underlying risk factors. The overarching objective of this work is to provide individuals with an interactive knowledge base to help them understand the underlying associations between lifestyle related risk factors and their contributions to NCD. The PRISM project [2] incorporates the risk factor knowledge-base to help individuals self-assess, and self-manage potential NCD risk factors. Our work is focused on breast cancer, however, we posit that the presented methodology can be extended and adapted to develop similar knowledge graph for other NCDs.

## 2. Background

### *2.1. Literature-Based Discovery (LBD)*

LBD is a mature text-mining approach to extract evidence-based knowledge by uncovering interesting, and often hidden, relationships from existing disparate pieces of published literature [3]. LBD is premised on the ABC model; that is, given concepts of knowledge A-B-C in disjoint articles if A is associated with B, and B is associated with C, then A is associated with C via the intermediate concept B. LBD has two approaches: open-discovery and closed-discovery. In the former, the main objective is to find potential relationships between a known A and an unknown C via an intermediate B. In the latter, A and C are already known and the goal is find relationships between A-B and B-C, thus facilitating implicit associations between A and C. In biomedical research, LBD has been applied to discover disease candidate genes, determine adverse drug reactions, detect drug-drug interactions , and find cancer treatment pathways [3]. In this research, we apply open-discovery LBD and represent the knowledge extracted from the literature as a knowledge graph, such that the graph's nodes represent biomedical concepts and edges represent the semantically-defined association between the concepts.

### *2.2. Semantic MEDLINE Database*

To apply LBD, Natural Language Processing (NLP) techniques are often used to mine textual knowledge sources from existing literature in order to extract meaningful semantic associations between two biomedical concepts from selected knowledge sources. A well-known NLP tool used in LBD research is SemRep [4]; a rule-based system that identifies and extracts semantic association from biomedical text in PubMed. The semantic association takes the form of a subject-PREDICATE-object triplet, where the subject and object are UMLS concepts representing phrases or words in biomedical text, and the predicate is an annotated association between the concepts. For example, SemRep extracts the association [*Physical Activity – PREVENTS – Breast Cancer*] from the sentence: "Physical activity reduces the risk of breast cancer risk". The results of this process are stored in the SemMedDB database which is used in this research—SemMedDB contains >94 million associations from 30 million PubMed articles.

## 3. Methods

Our LBD based knowledge synthesis and discovery methodology comprises the following steps: literature search, semantic association extraction, building a semantic graph, aggregation of semantic associations to build *NCD risk paths*, and ranking of generated paths to outline meaningful multi-node causal associations between risk factors and a target NCD. The NCD risk paths consist of UMLS concepts as nodes, and the semantic associations between them as edges. The nodes represent a source risk factor, one or more intermediate biomedical phenotypes, and a target NCD, while the edges represent the causal associations. The ranking algorithm identifies direct associations (i.e. source-target pair) and indirect associations (i.e. source-intermediate-target triple) via causal UMLS predicates to identify meaningful causal associations.

### *3.1. Literature Search and Synthesis*

We searched PubMed using specified queries consisting of MeSH terms to retrieve a list of PubMed IDs (PMID) of articles relevant to our topic (i.e. breast cancer and risk factors). The query, constructed using the work in [5], returned 44,231 PubMed abstracts.

### *3.2. Semantic Association Extraction*

The list of PMIDs from 3.1 were used as input to extract semantic associations (subject-PREDICATE-object) from SemMedDB. Our search returned 128,632 unique associations with the Concept Unique Identifiers (CUI) and semantic type for each subject and object. We also retrieved all source sentences from which each association was extracted to keep track of the provenance of knowledge. Further, to work with meaningful associations, we filtered out associations that involved uninformative semantic types – e.g., *law or regulation, professional society.*

### *3.3. Building the RiskExplorer Semantic Graph*

Semantic associations can be represented as a graph of nodes and edges, where the nodes represent UMLS concepts as subjects or objects and edges represent the associations (i.e. predicates) between a given subject and object. We used Neo4j graph database to build the RiskExplorer graph. Each node was associated with several properties, such as the subject or object CUI, and concept semantic type. Edges had the following properties: semantic relation type, source PMID, and sentence from which the association was extracted. Further, we employed UMLS Semantic Network to categorize edges as causal or non-causal—an edge is considered causal if its semantic relation type is a child of the *funtionaly_related_to* predicate in UMLS semantic network (e.g. *disrupts, causes, prevents, treats,* etc.). All other non-causal edges (e.g. *part_of, location_of, higher_than* etc.*)* are also retained in the graph and are considered by the path ranking algorithm.

### *3.4. Semantic Association Aggregation and Ranking of Identified Risk Paths*

We use the semantic associations in the RiskExplorer graph to generate meaningful *NCD risk paths.* Here we propose the concept of association patterns that satisfy a set of conditions for the selection and subsequent aggregation of semantic associations to yield

NCD risk paths. We developed a two-step method to generate multi-node discovery patterns as follows: first, target node (i.e., NCD or cancer) is identified and the graph is traversed to look for and retrieve set of source nodes directly linked to the target node. The graph is traversed again to look for and retrieve a set of intermediate nodes that are linked to the source and target nodes, thereby generating aggregated semantic associations with the following pattern: source-intermediate-target (e.g., Obesity – CAUSES – Diabetes– PREDISPOSES – Breast Cancer), where the source is considered a risk factor that indirectly influences the risk of a target cancer via intermediate biomedical phenotypes.

Next, our path ranking algorithm is applied to score and rank the aggregated semantic associations and generate meaningful causal NCD risk paths. For each pair of nodes in source-intermediate-target associations, the ranking algorithm calculates frequency of causal edges between the nodes and subtracts it from the frequency of non-causal edges between same pair of nodes. The resulting difference is then divided by total frequency of causal edges to compute a score for each pair of nodes – i.e., (source, target), (source, intermediate) and (intermediate, target). The final score is the sum of all node pair scores. Our path ranking process prioritizes causal relations, such that a pair of nodes with more non-causal edges than causal edges are assigned negative scores.

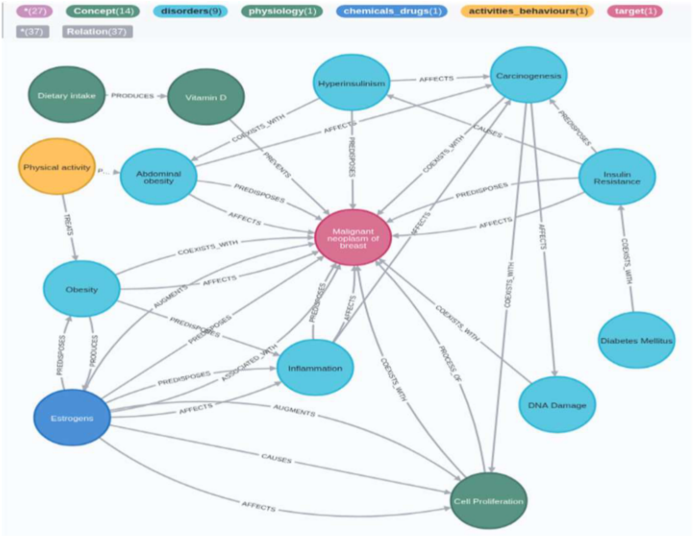

**Figure 1.** A sub-graph of *RiskExplorer* Graph

## 4. Results

The application of our LBD based knowledge synthesis and discovery approach (as described above) yielded a knowledge graph consisting of 3,784 unique nodes and 21,571 unique edges extracted. The ranking algorithm generated 1,034 unique NCD risk paths. The risk paths were assigned a semantic group based on the semantic type of the source node and we retained top 95th percentile score from each group; this allowed us to view highly ranked risk paths based on various causal contexts (e.g., chemicals and drugs, diseases, activities and behaviours, and physiology).

NCD risk paths in **Figure 1**. represent a sub-sample of top ranking biomedically valid risk factors leading to breast cancer. For example, research [6] has indicated obesity as associated with a chronic inflammatory state which results in activation of pathophysiological processes that predispose to carcinogenesis and, thus, cancer. Likewise, early menarche, high levels estrogen (due to obesity) and soy products are well-established risk factors for breast cancer [6]. Interestingly, the subgraph also shows protective risk associations as demonstrated by the path: [Dietary intake–*PRODUCES*–Vitamin D–*PREVENTS*–Breast cancer] in **Figure 1**. Some high ranking risk paths (not in Figure 1) are given in **Table 1**.

**Table 1.** Other high ranking risk paths for breast cancer in the *RiskExplorer* knowledge graph

| **Path (Mechanistic Association)** | **Path Score** |
|---|---|
| [Hormone replacement therapy-*AFFECTS*→Menopause-*AFFECTS*→Breast Cancer] | 2.00 |
| [Anti-inflammatory agents -*PREVENTS*→Tumor growth-*COEXISTS_WITH*→Breast Cancer] | 1.98 |
| [Alcohol consumption-*PREDISPOSES*→Breast Diseases-*PREDISPOSES*→Breast Cancer] | 1.93 |
| [Retinoids-*DISRUPTS*→Cell Proliferation-*ASSOCIATED_WITH*→Breast Cancer] | 1.87 |
| [Estrogens-*COEXITS_WITH*→C reactive protein-*ASSOCIATED_WITH*→Breast Cancer | 1.83 |

## 5. Conclusion and Future Directions

In In this paper, we presented a methodology to build an evidence-based *RiskExplorer* knowledge graph that elucidates multi-causal associations between risk factors and breast cancer. The primary contribution of our work is the utilization of causal semantic associations and frequency-based measures to automatically organize and structure meaningful knowledge from biomedical literature for consumer consumption. The overarching goal of this work is to generate a knowledge synthesis framework that dynamically provides and organizes validated information mined directly from scientific literature. Future direction of this project involves expert-based evaluation of generated NCD risk paths to verify the correctness of the discovered knowledge, whilst improving the ranking algorithm by introducing word embedding models to capture high semantically related associations.

## References

[1] Kmetik KS, Skoufalos A, Nash DB. Pandemic Makes Chronic Disease Prevention a Priority. Popul Health Manag. 2021 Feb;24(1):1-2.

[2] Daowd A, Faizan S, Abidi S, Abusharekh A, Shehzad A, Abidi SSR. Towards Personalized Lifetime Health: A Platform for Early Multimorbid Chronic Disease Risk Assessment and Mitigation. Stud Health Technol Inform. 2019 Aug 21;264:935-939.

[3] Thilakaratne M, Falkner K, Atapattu T. A systematic review on literature-based discovery workflow. PeerJ Computer Science 2019;5:e235.

[4] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.

[5] Bramer, W. M., de Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc. 2018 Oct; 106(4): 531–541.

[6] World Cancer Research Fund International. (2018). Diet, nutrition, physical activity and cancer: a global perspective: a summary of the Third Expert Report. Available at https://www.wcrf.org/dietandcancer. Accessed Jan 2021.