# A Knowledge Graph Approach for the Secondary Use of Cancer Registry Data

S.M.Shamimul Hasan[1], Donna Rivera[2], Xiao-Cheng Wu[3], J. Blair Christian[1], Georgia Tourassi[1]
[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]National Cancer Institute, Rockville, MD, USA
[3]Louisiana Tumor Registry, New Orleans, LA, USA
hasans@ornl.gov, donna.rivera@nih.gov, xwu@lsuhsc.edu, christianjb@ornl.gov, tourassig@ornl.gov

***Abstract*—Population-based central cancer registries collect valuable structured and unstructured cancer data primarily for surveillance and reporting. The collected data includes (1) categorization of each cancer case (tumor) at the time of diagnosis, (2) demographic information about the patient such as age, gender, and location at time of diagnosis, (3) first course of treatment information, and (4) survival outcomes when available. While advanced analytical approaches such as SEER*Stat and SAS exist, we provide a knowledge graph approach to organizing cancer registry data for advanced analytics which offers unique advantages over existing approaches. This knowledge graph approach semantically enriches the data and enables straightforward linking capability with third-party data to help understand variation in cancer outcomes. A knowledge graph was developed using Louisiana Tumor Registry data. We present the advantages of the knowledge graph approach by examining: i) scenario-specific queries and ii) linkages with publicly available external datasets. Our results demonstrate this graph based solution can perform complex queries, improve query run-time performance by 81%, and more easily conduct iterative analyses to enhance researchers understanding of cancer registry data.**

***Index Terms*—knowledge graph, cancer registry, treatment**

## I. Introduction

Worldwide, almost 1 out of 6 deaths are due to cancer, a rate which is rising and estimated to increase by approximately 70% in the next 20 years [1]. To tackle this challenge, we must improve our cancer research data infrastructure with new tools to support existing clinical programs that adapt to changing research needs. In the United States and other developed nations, cancer registries systematically collect data on cancer diagnosis and tumor characteristics as well as planned first course of treatment, patient demographics, and outcomes [2], [3]. The information is submitted to centralized national surveillance programs including the National Cancer Institutes (NCI) Surveillance, Epidemiology, and End Results (SEER) program. SEER collects a broad set of clinical data from population based US cancer registries which cover 34.6% of the population. SEER also reports aggregated (non-identifiable) cancer statistics and provides this publicly available dataset to support cancer surveillance and research [2], [4].

Both the centralized SEER datasets and individual cancer registry datasets have potential research benefits, especially when linking to vast third party datasets, augmenting the original clinical data with rich biological and environmental information capable of explaining variation in cancer incidence and outcomes. This resulting use of cancer registry data can not only help gain insights into broader trends for improving care at the population level, while also more precisely understand specific treatment patterns and outcomes at the individual level. To realize the value of this data, a flexible and scalable analytics architecture must be developed. Presently, the core registry datasets are housed in a series of relational database management system (RDBMS), which has a rigid schema structure.

This paper introduces a semantic web-based knowledge graph approach to storing cancer registry data for analytic and research purposes. Our aim is to develop a platform that facilitates both better data representation and fast query execution. In comparison to a RDBMS approach, the knowledge graph approach: i) supports a richer class of queries (e.g., hierarchical), ii) easily links to existing third party datasets, and iii) enables easy schema evolution for iterative analysis. The main contributions of our work are as follows:

- A new analysis architecture for cancer registry data using a knowledge graph. As a proof of concept, we develop a knowledge graph based on the Louisiana Tumor Registry's Cancer/Tumor/Case (CTC) dataset based on the NAACCR data standards [3].
- The development and results of scenario-specific hierarchical queries to understand the population level utilization of cancer treatment sequences.
- The linking of a cancer registry knowledge graph to external datasets. The linked data provides an integrated knowledge base so that researchers can ask more complex queries (e.g., hierarchical, recursive, etc.).

## II. Knowledge Graph Creation

### A. Cancer Registry Data Overview

We used Louisiana Tumor Registry (LTR) data containing data from cancer patients who are Louisiana residents at the time of diagnosis. The registry includes tumor, demographic, and limited treatment information. Each record of our data extract corresponds to a unique cancer, sometimes referred to as a Cancer/Tumor/Case (CTC) as defined by the NAACCR data standard [5]; each patient may occur in the database more than once if they have more than one primary tumor. While the primary data in the database consists of tumor information at

time of diagnosis and first course of treatment, it also contains follow up vital status and date of last contact information. Our data is a CSV file containing 171 columns, 371,915 unique tumor records, and is 164 MB in size, containing data for diagnoses from 2000-2016 [3].

### B. Approach

*1) Loading the CTC dataset into a relational database:* We leverage existing tools to convert the original CSV data into our graph database via an RBDMS. The first step is to load the CSV file into the RBDMS. A schema is required to load the data into a relational database which describes the column names, column datatypes, and primary and foreign key information. We use CTC as the table name and create database column names using the column headers found in the CTC extract, and inherit the column datatypes from the raw data. In the CTC CSV file, null values are presented in numerous ways (e.g., space, NA, NR), and are all set to the databases null value. We used PostgreSQL 9.5.9 as a relational database.

TABLE I
THE SIZE, NUMBER OF TRIPLES, AND CREATION TIME OF THE MAPPING FILE USED TO GENERATE THE CTC KNOWLEDGE GRAPH.

| | |
|---|---|
| Size (KB) | 60 |
| Number of Triples | 1,209 |
| Creation Time (minutes) | <1 |

*2) Creating the relational to knowledge graph mapping file:* We used the resource description framework (RDF) data model to represent our knowledge graph. Numerous RDBMS to RDF conversion tools were available, and the D2RQ tool was selected to create RDF conversion mapping files from the RDBMS, creating mapping files in the RDF format. The D2RQ tool maps the relational database table name to the RDF class name, and maps table attribute names to the RDF property names. We created a CTC mapping file from the PostgreSQL database by using the D2RQ *generate-mapping* service with D2RQ 0.8.1. The mapping file size, number of triples in the mapping file, and mapping generation time are shown in Table I.

*3) Knowledge graph generation:* We apply the D2RQ mapping file to the PostgreSQL CTC table to generate the materialized CTC knowledge graph. We employ the D2RQ *dump-rdf* service for RDF graph creation. We provide knowledge graph size, number of triples, and graph generation time in Table II.

TABLE II
KNOWLEDGE GRAPH SIZE, NUMBER OF TRIPLES, AND CREATION TIME.

| | |
|---|---|
| Size (GB) | 8.4 |
| Number of Triples | 43,501,480 |
| Creation Time (minutes) | 75 |

*4) Loading the knowledge graph into a triplestore:* Next, we loaded our knowledge graph into a triplestore. The triplestore provides a SPARQL endpoint that provides graph-based query execution facilities. We used Virtuoso Open-Source Edition 7.2.4 as our triplestore using a virtual machine running CentOS 7.4 with a 2.0 GHz Intel Xeon E7 4850 CPU, 128 GB of memory and 1 TB of local disk storage.

## III. APPLICATIONS

### A. Application 1: Finding Population Level Treatment Sequence Outcomes in Breast Cancer

Scenario specific queries represent about 60% of physician directed clinical queries, and are usually hierarchical in structure [6]. However, the same structure can be organized in many ways based on the requirements of the domain researchers. Knowledge graphs provide a more flexible data structure for efficiently querying and exploring data from different perspectives.

One class of hierarchical query of interest in cancer surveillance is explaining the variation in breast cancer (Fig. 1). This query groups patients by age, gender, cancer type (specified pre-query as breast cancer), cancer topography, initial treatments, and survival outcome. Breast cancer subsites are defined by ICD-O-3 topography codes (C50.0-C50.9) [7]. This query yielded sequences of first date of treatment for surgery, chemotherapy, radiation, and hormone therapy. In this study, we only consider the registry collected binary data available in the CTC: surgery, chemotherapy, radiation therapy, and hormone therapy treatments. One current limitation of the registry data is that only the date on which the first course of each treatment began was recorded, even if a patient had one or more, or various durations of treatment. Hence, we consider treatment paths of length one treatment (only surgery, etc.), length two (surgery-chemotherapy, etc.), length three (surgery-chemotherapy-radiationtherapy, etc.), or length four (surgery-chemotherapy-radiationtherapy-hormonetherapy, etc.). For instance, if an individual is in the surgery-chemotherapy sequence, it implies that up to the date of last contact, the treatment of the patient started with surgery and was then followed by chemotherapy. For the example scenario, we assume a patient has only one, full, treatment sequence and do not double count their shorter sequences. Our example scenario in Fig. 1 includes patients aged 40-64 to demonstrate the hierarchy.

We used the LTR knowledge graph for this example (Section II). In this work, no survival analysis or logistic regression is performed on the original data collected from LTR, however the knowledge graph approach enables the easy integration of datasets for survival analysis or other statistical modeling. Four results from the scenario represented in Fig. 1 are shown in Figs. 2(a), 2(b), 2(c), and 2(d). The query results show that a large number (4,574 patients) of breast cancers occur in the upper-outer quadrant (C50.4) and most of these patients received a treatment sequence of surgery-chemotherapy-radiation therapy as opposed to only surgery-chemotherapy sequence. Four queries performed on average 81% faster on the knowledge graphs than using an RDBMS. Note that PostgreSQL database and Virtuoso triplestore's internal indexing algorithms play a significant role on the

query performances. In this experiment, we only considered treatment types with valid date information available (surgery, chemotherapy, radiation, and hormone therapy).

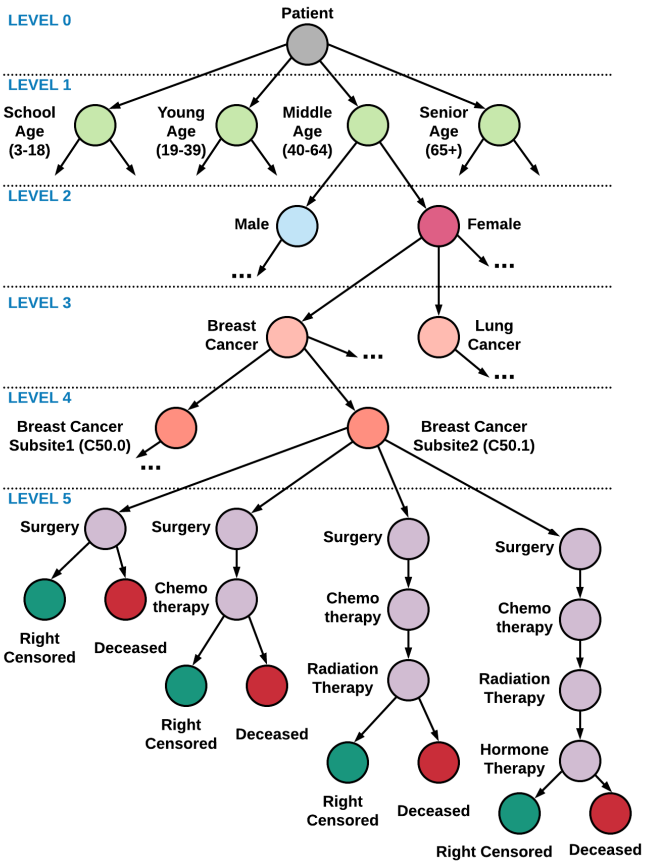



Fig. 1. Hierarchical breast cancer treatment sequence scenarios.

### B. Application 2: Linking External Datasets

Another advantage of using a graph database approach to cancer registry data is that it is flexible and capable of linking to many third party datasets sharing common keys to enable deeper understanding of the causes of variation in treatment and survival outcomes. For example, variables such as environmental exposure, education, income, and occupation are associated with patient location. The linking of the cancer registry knowledge graph and additional datasets allows the execution of advanced queries using multiple datasets to explain the variation associated with socioeconomic status or other factors influencing exposures or outcomes. To demonstrate, we began by linking relevant county level data to show spatial patterns in cancer rates or outcomes.

We downloaded the Rural-Urban Continuum Codes dataset from [8], which includes USA states, county names, and county federal information processing standards (FIPS) codes. In addition, the dataset has a classification scheme specifying metropolitan and non-metropolitan counties.

There are 8 columns and 3,222 records in the Rural-Urban Continuum Codes dataset. We followed the approach outlined in Section II to generate a knowledge graph from this dataset using name and column names as knowledge graph vocabulary. Tables III and IV summarize the rural-urban continuum codes mapping file and knowledge graph information.

TABLE III
RURAL-URBAN CONTINUUM CODES MAPPING FILE SIZE, NUMBER OF TRIPLES, AND CREATION TIME.

| | |
|---|---|
| Size (KB) | 8 |
| Number of Triples | 82 |
| Creation Time (minutes) | <1 |

TABLE IV
RURAL-URBAN CONTINUUM CODES KNOWLEDGE GRAPH SIZE, NUMBER OF TRIPLES, AND CREATION TIME.

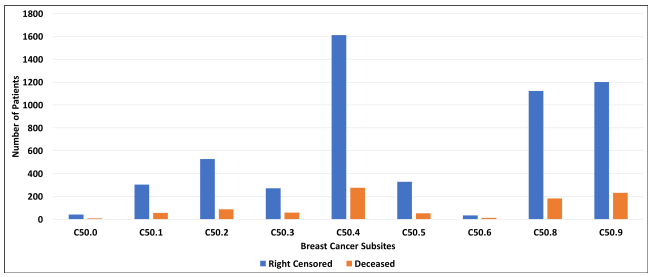
| | |
|---|---|
| Size (MB) | 7.9 |
| Number of Triples | 32,239 |
| Creation Time (seconds) | ∼11 |

TABLE V
CANCER PATIENTS IN THE VARIOUS COUNTIES IN THE STATE OF LOUISIANA.

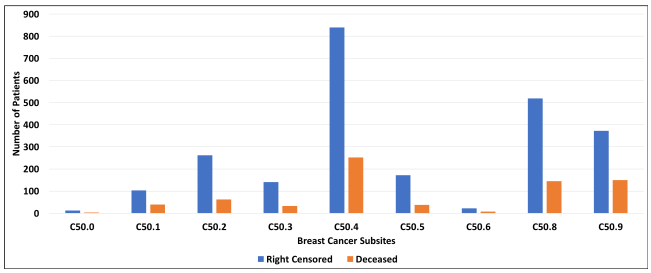
| Rural-Urban Continuum Code 2013 | Code Definitions | Total Counties (with cancer patient) | Total Cancer Patients |
|---|---|---|---|
| 1 | Counties in metro areas of 1 million population or more | 8 | 91,732 |
| 2 | Counties in metro areas of 250,000 to 1 million population | 18 | 141,873 |
| 3 | Counties in metro areas of fewer than 250,000 population | 9 | 70,835 |
| 4 | Urban population of 20,000 or more, adjacent to a metro area | 3 | 16,105 |
| 5 | Urban population of 20,000 or more, not adjacent to a metro area | 1 | 3,470 |
| 6 | Urban population of 2,500 to 19,999, adjacent to a metro area | 16 | 38,325 |
| 7 | Urban population of 2,500 to 19,999, not adjacent to a metro area | 4 | 5,207 |
| 8 | Completely rural or less than 2,500 urban population, adjacent to a metro area | 2 | 1,644 |
| 9 | Completely rural or less than 2,500 urban population, not adjacent to a metro area | 3 | 2,708 |
| 99 | Unknown/not official USDA Rural-Urban Continuum code | 1 | 16 |

The LTR knowledge graph is based on the state of Louisiana which has 64 counties and has the FIPS codes for county of residence at time of diagnosis. The LTR knowledge graph was linked with the Rural-Urban Continuum Codes knowledge graph using these county and state FIPS codes. Next, graph-based queries are developed to quantify the total number of patients residing in different metropolitan and non-metropolitan counties using the rural-urban continuum code for 2013.

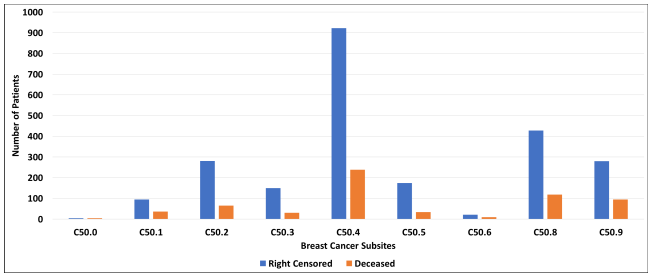
The goal is to identify if there are significant differences in cancer rates between metropolitan and non-metropolitan counties. The results are presented in Table V. The first column of the Table V represents the 2013 Rural-Urban Continuum



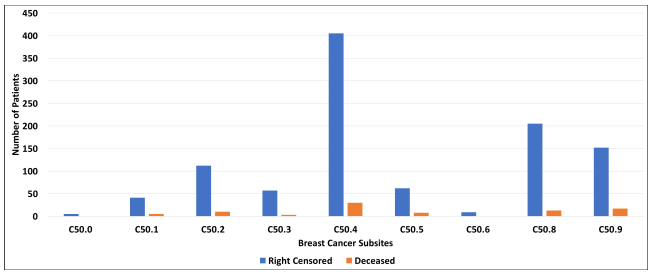

(a) treatment sequence: surgery




(b) treatment sequence: surgery-chemotherapy




(c) treatment sequence: surgery-chemotherapy-radiation therapy




(d) treatment sequence: surgery-chemotherapy-radiation therapy-hormone therapy.

Fig. 2. In this figure, we present results of the following query: find treatment sequence(s) outcomes where age group is middle age, gender is female, and cancer type is breast cancer.

Code, the second column provides code descriptions, the third column shows the total number of Louisiana counties that have cancer patients with the county code description, and the fourth column presents the total number of cancer patients available in the counties. While our results show a significant number of cancer patients reside in the metropolitan counties, as would be expected, the data needs to be normalized by population. Thus, statistical analyses such as tests of equality of proportions can be conducted to identify differences in cancer rates between rural and urban areas, and queue up additional iterations of questions requiring more county level data such as aggregate cancer prevention measures, survey responses, health behaviors, or socioeconomic factors that might help explain any discrepancies. Moreover, researchers should consider patients' residence history and corresponding environmental background information for individual-level data analysis. It is easy to integrate numerous datasets in knowledge graphs because of their flexible structure.

## IV. Conclusion and Future Work

We described a knowledge graph-based approach to the secondary use of cancer registry data. In contrast to RDBMS, the knowledge graph approach makes it easy to handle scenario specific queries and link third-party data. This study demonstrates how cancer registry data management and analysis using the knowledge graph approach receives significant advantages. In future studies, we plan to apply graph pattern mining algorithms to generate clinical hypotheses derived from this knowledge graph that enhance understanding of the patient care trajectory.

## V. Acknowledgment

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of National Institutes of Health. This work was performed under the auspices of the U.S. DOE by ANL under Contract DE-AC02-06-CH11357, LLNL under Contract DE-AC52-07NA27344, LANL under Contract DE-AC5206NA25396, and ORNL under Contract DE-AC05-00OR22725.

## References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] "National Cancer Institute's Surveillance, Epidemiology, and End Results Program," URL: https://seer.cancer.gov, 2018.

[3] "Center for Disease Control's National Program of Cancer Registries," URL: https://www.cdc.gov/cancer/npcr, 2018.

[4] "Surveillance, Epidemiology, and End Results (SEER) Linked Databases," URL: https://seer.cancer.gov/data-software/linked_databases.html, 2018.

[5] "NAACCR DATA DICTIONARY," URL: http://datadictionary.naaccr.org/?c=10, 2018.

[6] W. W. Chu, Z. Liu, W. Mao, and Q. Zou, "Kmex: A knowledge-based digital library for retrieving scenario-specific medical text documents," in *Biomedical Information Technology*. Elsevier, 2008, pp. 307–341.

[7] "ICD-O-3 SITE CODES," URL: https://training.seer.cancer.gov/breast/abstract-code-stage/codes.html, 2018.

[8] "Rural-Urban Continuum Codes," URL: https://seer.cancer.gov/seerstat/variables/countyattribs/ruralurban.html, 2018.