Knowledge Graph-Enabled Cancer Data Analytics

S.M.Shamimul Hasan, *Member, IEEE*, Donna Rivera, Xiao-Cheng Wu, Eric B. Durbin, J. Blair Christian, and Georgia Tourassi, *Senior Member, IEEE*

***Abstract*—Cancer registries collect unstructured and structured cancer data for surveillance purposes which provide important insights regarding cancer characteristics, treatments, and outcomes. Cancer registry data typically (1) categorize each reportable cancer case or tumor at the time of diagnosis, (2) contain demographic information about the patient such as age, gender, and location at time of diagnosis, (3) include planned and completed primary treatment information, and (4) may contain survival outcomes. As structured data is being extracted from various unstructured sources, such as pathology reports, radiology reports, medical records, and stored for reporting and other needs, the associated information representing a reportable cancer is constantly expanding and evolving. While some popular analytic approaches including SEER*Stat and SAS exist, we provide a knowledge graph approach to organizing cancer registry data. Our approach offers unique advantages for timely data analysis and presentation and visualization of valuable information. This knowledge graph approach semantically enriches the data, and easily enables linking with third-party data which can help explain variation in cancer incidence patterns, disparities, and outcomes. We developed a prototype knowledge graph based on the Louisiana Tumor Registry dataset. We present the advantages of the knowledge graph approach by examining: i) scenario-specific queries, ii) links with openly available external datasets, iii) schema evolution for iterative analysis, and iv) data visualization. Our results demonstrate that this graph based solution can perform complex queries, improve query run-time performance by up to 76%, 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 strategically conquer this challenge, we must improve our cancer research data infrastructure to support existing clinical and control programs and adapt to changing research needs. In the United States and other developed nations, cancer registries systematically collect data on diagnosed cancer cases such as tumor characteristics, first course of treatment, patient demographics, and outcomes [2], [3]. The information is abstracted and coded according to the national standards and submitted to a central cancer registry, such as the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program in the United States. 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 public-used case-level dataset to support both population-based cancer research and cancer control programs [2], [4].

A. Major Cancer Data Organization Challenges

Both the centralized SEER datasets and individual cancer registry datasets from which they are extracted have potential research advantages, especially when linking to vast third-party datasets, augmenting the original clinical data with behavioral and environmental information capable of explaining the variation in cancer incidence and outcomes. However, we are facing numerous data-related challenges for the secondary use of cancer registry data. Some challenges include:

- *Heterogeneous data:* An enormous volume of third-party data is available on the web. Moreover, datasets are available in various locations. Furthermore, third-party datasets are available in various formats. For example, neighborhood concentrated disadvantage index (CDI) is available in comma-separated values (CSV) format [5], Wikipedia data is available in RDF format [6], and climate data is available in JavaScript Object Notation (JSON) format [7], to name a few. Customized software development

Manuscript received November 15, 2019; revised March 12, 2020; accepted April 17, 2020. Date of publication May 4, 2020; date of current version July 2, 2020. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). *(Corresponding author: S. M. Shamimul Hasan.)*

S.M.Shamimul Hasan, J. Blair Christian, and Georgia Tourassi are with Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA (e-mail: hasans@ornl.gov; christianjb@ornl.gov; tourassig@ornl.gov).

Donna Rivera is with National Cancer Institute, Bethesda, MD 20892 USA (e-mail: donna.rivera@nih.gov).

Xiao-Cheng Wu is with Louisiana Tumor Registry, New Orleans, LA 70112 USA (e-mail: xwu@lsuhsc.edu).

Eric B. Durbin is with Kentucky Cancer Registry, University of Kentucky, Lexington, KY 40506 USA (e-mail: ericd@cri.uky.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2020.2990797

is needed to analyze these datasets. However, if a new data format comes in, then a substantial code change is required to add a new data format to the customized software.

- *Schema change:* Presently, the core cancer registry datasets are housed in a series of relational database management system (RDBMS), which has a rigid schema structure. Newer versions of the datasets sometimes come with their own new schema. For example, in this paper, we used the Louisiana Tumor Registry's (LTR) Cancer/Tumor/Case (CTC) dataset. The current version of the Louisiana Tumor Registry's dataset contains a column that provides the patients' death location information (U.S. state). However, an earlier version of the dataset does not contain such information. It is not easy to handle schema change in RDBMS. The smallest change in the database requires quite a lot of graphical user interface (GUI)-level code change.
- *Not linked:* Linking a large volume of heterogeneous data sources is quite challenging because heterogeneous data sources do not follow any common data-storing standard.
- *Difficult to execute complex queries:* Data linking is crucial because it is not possible to answer complex cancer questions based on data from a single source. The complex queries (e.g., tree query, recursive query, transitive query) require joining, which is a costly operation in the RDBMS.

B. What is a Knowledge Graph and Why Do We Need It?

A *knowledge graph* denotes a graph of entities with one or more properties that maintain defined semantic relationships to other entities. In a knowledge graph, the entities are represented as nodes, and the relationships are represented as edges [8], [9]. Many academic institutions (e.g., YAGO, DBPedia) and industries (e.g., Google, Microsoft, Facebook) are creating and taking advantage of the knowledge graph technology [10], [11]. This paper introduces a semantic web-based [12] knowledge graph approach to storing cancer registry data for analytic and research purposes. In the following, we briefly discuss the reason for our choice.

- *Standard data model:* A major component of semantic web technology is the Resource Description Framework (RDF). We employed the RDF data model for our knowledge graph creation. The RDF data model is considered a standard model according to the World Wide Web Consortium (W3C) [13]. The RDF is developed on top of web infrastructure, is considered a directed labeled graph in mathematical settings and has been known to represent a graph of things. Furthermore, RDF data is stored in a triple format of subject, predicate, and object. It is easy to map different data models like relational, tree, key value store, and graph to RDF triples. The implication of this is that the RDF is actually a generic data model [9], [14].
- *Dynamic data growth and flexible schema:* Given that it is a graph-based data model, the RDF's structure is considered more flexible than the RDBMS. This is what makes the RDF a better option for dynamic data integration. Furthermore, the RDF does not need code changes or software redesigns to handle schema evolution [9].
- *Easy to incorporate semantics and links:* A large collection of ontologies [15] and vocabularies [16] are available, which model the concepts of a domain. The RDF supports easy ontology and vocabulary integration compared to the RDBMS [17]. It is easy to dynamically link multiple data sources in the RDF, which is tedious in the RDBMS [9], [18].
- *Easy to execute complex queries:* Complex cancer queries follow graph patterns like tree, long path traversal, and transitivity. The above types of queries require numerous joins, which is an expensive operation in the RDBMS's Structured Query Language (SQL). However, joining is easy in SPARQL Protocol and RDF Query Language (SPARQL) because SPARQL is particularly designed to query graph structures [9].
- *Reasoning:* An RDF-based knowledge graph supports dynamically finding new facts that are not explicitly stated in the knowledge graph (through description logic) [19], [20]. It is hard to develop a dynamic reasoner on top of an RDBMS system.
- *Affordable:* An openly available link data browser, semantic search, link discovery, and many visualization tools can be used for RDF data analysis. Furthermore, hardly any programming effort is needed when using these tools. For instance, rather than browsing with a tailored graphical user interface (GUI), which requires customized software development, researchers can use the existing faceted browser for quick data exploration with ontologies or vocabularies. The GUI will automatically reflect any change in the ontology or data. Many of these tools are freely available on the web. These tools have constraints, but they offer a quick data analysis platform for free [9].
- *Query expressive power:* Although SQL is older than SPARQL, Renzo Angles and Claudio Gutierrez proved that the expressive power of SPARQL is equivalent to relational algebra, which is a theoretical foundation for RDBMS and SQL [21]. Hence, along with other benefits mentioned in this section, the expressive power of SPARQL makes the knowledge graph as a better choice for the type of research problem we studied in this paper.

Our aim is to develop a knowledge graph-based scientific digital library platform for advanced cancer data analytics. The main contributions of our work are as follows:

- We propose a scientific digital library framework for the secondary use of cancer registry data using a knowledge graph structure. As a proof of concept, we develop a knowledge graph based on the LTR's CTC dataset abstracted according to 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 execution of complex queries over multiple datasets by linking the cancer registry knowledge graph to external datasets to provide an integrated knowledge base for researchers.

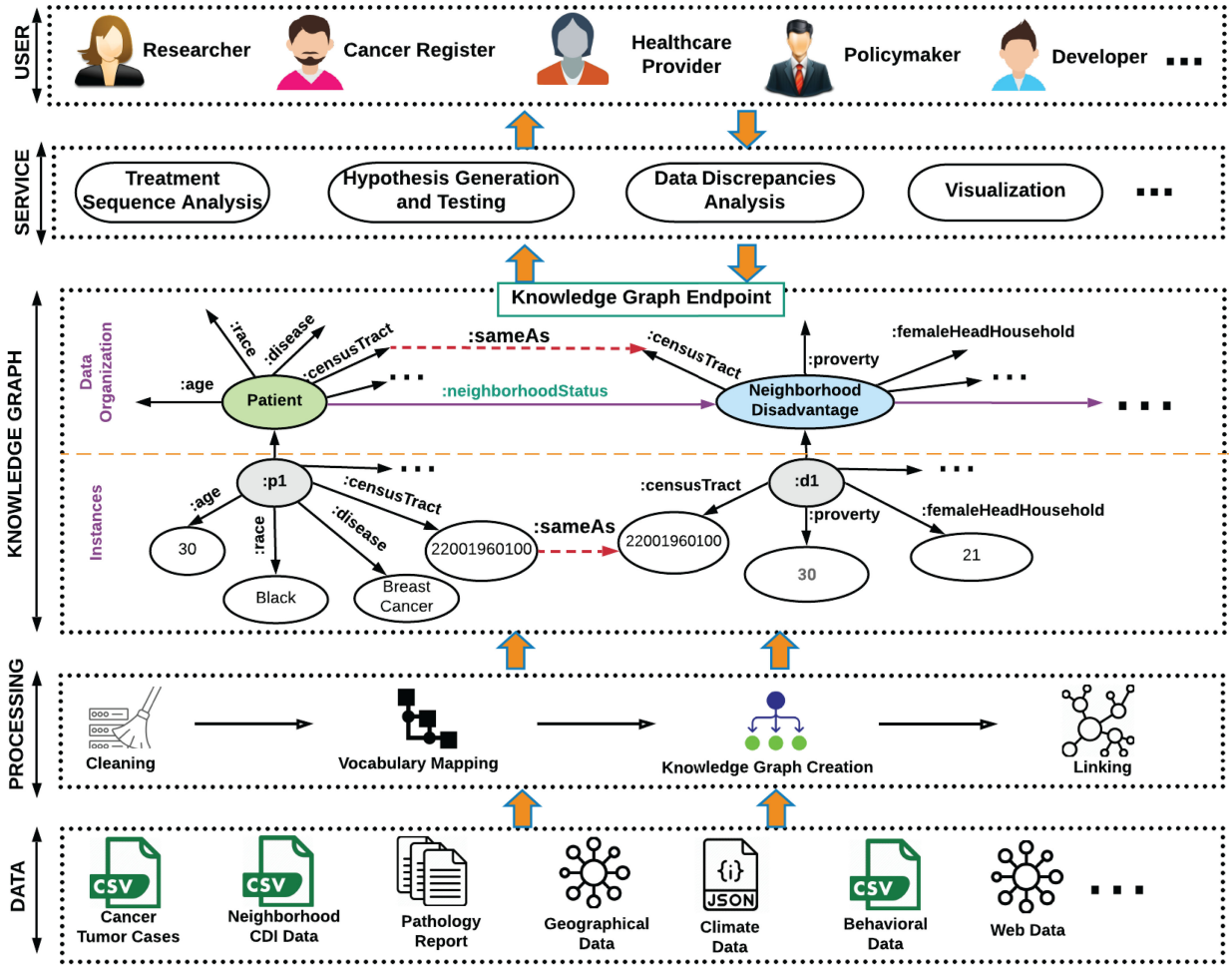

Fig. 1. In this figure, we present a high-level overview of the cancer scientific digital library framework.

- A demonstration of iterative schema evolution using anomaly detection as an example case.
- Present knowledge graph visualization showing the usefulness of visual pattern recognition.

A preliminary version of this work has been reported [22].

II. Related Work

People are suffering from more than 100 types of cancer, which include pancreatic cancer, breast cancer, prostate cancer, etc. Pourshams *et al.* found that worldwide pancreatic cancer cases, death, and disability-adjusted life-years (DALYs) have greater than doubled from 1990-2017 [23]. Atieh Vafajoo, Reza Salarian, and Navid Rabiee proposed a suspension microbead arrays based method for early breast cancer diagnosis [24]. People of any age can have cancers. In [25], Force *et al.* discussed the global burden of childhood cancer disproportionately affects populations with the fewest resources. Although results presented in [23], [25], and [24] are interesting. However, authors did not propose any graph-based dynamic system creation to support advanced cancer analytics. There have been several attempts to formalize the secondary use of healthcare data for research and other uses [26]–[28]. There have been a number of interesting findings regarding the secondary use of electronic health records [29]. Early work in this area showed the application of Web Ontology Language (OWL) based representation of Electronic Health Records (EHR) data [30] but did not create cancer patient-level RDF graphs. Richesson *et al.* discussed the SHARPn framework that enables data normalization, secure transport, and common phenotyping facilities on various EHR data [31], but did not use graph based queries. A relatively small number of studies explored semantic web-based knowledge graph approach for the secondary use of the cancer registry datasets. Esteban-Gil, Ferndez-Breis, and Boeker proposed a semantic web based platform for cancer registries for data analysis and visualization [32]. However, they used simulated cancer registry data, not actual cancer patient data.

III. Proposed Cancer Scientific Digital Library Framework

Cancer researchers need a system to satisfy the information needs of users (societies), provide information services (scenarios), organize information in usable ways (structure), present information in useful ways (spaces), and communicate information with users (streams). A scientific digital library can fulfill the above-mentioned important requirements through the 5S (Streams, Structures, Spaces, Scenarios, and Societies) digital library framework [33]. We propose herein a scientific digital library framework for cancer research. Fig. 1 presents a

high-level overview of the framework. Data, processing, knowledge graph, service, and user are five layers of the framework. In the following, we briefly discuss each layer.

A. Data

First, the data layer shows that we are handling a wide range of heterogeneous datasets to develop a comprehensive knowledge graph. The datasets to be stored include patient-level cancer tumor data, socioeconomic data, pathology reports, geographical data, climate data, behavioral data, web data, and many more. The above-mentioned datasets are stored in numerous formats (e.g., CSV file, Extensible Markup Language (XML) file, graph file, JSON file).

In the above, we present some dataset category examples. ***The knowledge graph that is employed to perform this paper's experimentation contains CTC data, neighborhood CDI data, and Rural-Urban continuum codes data.***

B. Processing

Second, in the processing layer, we perform various cleaning on datasets to make sure that our linked dataset is clean, correct, and useful. We have several types of datasets, and different datasets need different types of cleaning. Hence, we have prepared multiple scripts for data cleaning. In the following, we describe some of the checks we have performed through our cleaning scripts.

- *Data type check:* We checked whether the column values were stored in the correct data types (e.g., integer, string, float) or not [34]. For example, in the CTC file, "Patient ID Number" should use integers. We checked the data types of the CTC, neighborhood CDI, and Rural-Urban Continuum Codes data.
- *Range check:* This check tests whether the values are within the permissible range. For example, the "Marital Status" column in CTC data should contains values within the range of 1 to 9. Another example is that the "Percent Black" column in the neighborhood CDI file should not contain any value of more than 100. If we observe a discrepancy in the data, then we store that information (see Section V – "Application 3: Easy Schema Evolution for Iterative Analysis" for more details).
- *Cross-field consistency check:* This means whether a condition needs other column values to validate a column [34]. For example, in the case of the CTC dataset, the "Date of Chemotherapy" cannot be earlier than the "Date of Diagnosis." We conducted a cross-field validation check.
- *Mandatory field check:* Some columns in the input data file cannot be empty [34]–for example, "Patient ID Number" and "Tumor Record Number" in the CTC dataset. We performed the mandatory field check.
- *Uniqueness check:* One or multiple field values should be unique in the dataset. Hence, we employed a uniqueness check–for example, the Federal Information Processing Standards (FIPS) code in the neighborhood CDI and Rural-Urban Continuum Codes datasets.
- *Format check:* We checked the format of the data values. For example, the CTC file contains many dates (e.g., "Date of Surgery," "Date of Chemotherapy," "Date of Radiation Therapy"). Date values should contain 8 digits and be stored in year-month-day (YYYYMMDD) format. We checked the dates format. We found that month and day information is missing for some of the dates. We added zeros for missing month and day information and made all the dates consist of 8 digits. Moreover, in the different datasets, null values are presented in various ways (e.g., space, NA, NR, etc.). We set all of them to database null value (see Section IV – "Approach" for more details).

After cleaning, we incorporate ontologies or vocabularies to organize the datasets. Next, by using our graph engine (D2RQ [35]) we create a knowledge graph for different datasets. Finally, we link various datasets in the knowledge graph.

C. Knowledge Graph

The knowledge graph is the core of our cancer scientific digital library. It serves as a knowledge base. Third, the "Knowledge Graph" layer is showing a pictorial view of our graph. We have two layers in the graph. The first layer is the data organization layer, which mainly represents various classes (or entities). The second layer represents class instances. In the picture, we are showing an example of a patient and neighborhood disadvantage datasets. We stored our knowledge graph in a graph repository (Virtuoso [36]). We have an endpoint (SPARQL endpoint) available on top of our knowledge graph repository.

D. Service

Fourth, in the "Service" layer, we are showing that different types of services (or apps) can be developed on top of our knowledge graph. Numerous services like treatment sequence analysis, hypothesis generation and testing, data discrepancy analysis, and visualization can be developed on top of our knowledge graph. We discuss services more in section V (Applications).

E. User

Finally, the user layer is showing numerous stakeholders (e.g., Researcher, Cancer Registrar, Healthcare Provider, Policymaker) can derive benefits from our scientific digital library.

Our cancer scientific digital library follows the 5S theory. Various types of input and output data are the *stream* of our digital library. A knowledge graph organizes our data, thus serving as a *structure* of the digital library. We are using indexing and information retrieval algorithms from our graph repository that serves as a *space*. We are providing various services (e.g., treatment sequence analysis, visualization) that fulfills the 5S theory's *scenario* requirement. Finally, we are connecting various *societies* (e.g., Researcher, Cancer Registrar) through our digital library.

IV. Knowledge Graph Creation

A. Cancer Registry Data Overview

We used Louisiana Tumor Registry (LTR) data from cancer patients who were Louisiana residents at the time of diagnosis.

TABLE I
THE SIZE, NUMBER OF TRIPLES, AND CREATION TIME OF THE LOUISIANA MAPPING FILE AND KNOWLEDGE GRAPH

Variables	Louisiana Mapping File	Louisiana Knowledge Graph
Size (KB)	56	16,000,000
Number of Triples	1,467	90,673,527
Creation Time (minutes)	<1	~75

Each record of our data extract corresponds to a unique cancer, sometimes referred to as a CTC as defined by the NAACCR data standard [37]; 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 demographic information, vital status, and date of last contact. Our data is a CSV file containing 240 columns, 374,682 unique tumor records, and is 207 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 used CTC as the table name and created 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 database's null value. We used PostgreSQL 9.5.9 as a relational database.

2) Creating the Relational to Knowledge Graph Mapping File: We used the RDF data model to represent our knowledge graph. Numerous RDBMS to RDF conversion tools were available, and the D2RQ tool [35] was selected to create RDF conversion mapping files from the RDBMS. The D2RQ tool maps the relational database table name to the RDF class name, and table attribute names to the RDF property names. We created a CTC mapping file from the PostgreSQL database [38] 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's column 2 (Louisiana Mapping File).

3) Knowledge Graph Generation: We applied the D2RQ mapping file to the PostgreSQL CTC table to generate the materialized CTC knowledge graph. We employed the D2RQ *dump-rdf* service for RDF graph creation. We provide knowledge graph size, number of triples, and graph generation time in Table I's column 3 (Louisiana Knowledge Graph).

4) Loading the Knowledge Graph Into a Triplestore: Next, we loaded our knowledge graph into a triplestore. The triplestore provides a SPARQL endpoint that supports graph-based query execution facilities. We used Virtuoso Open-Source Edition 7.2.4 [36] 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.

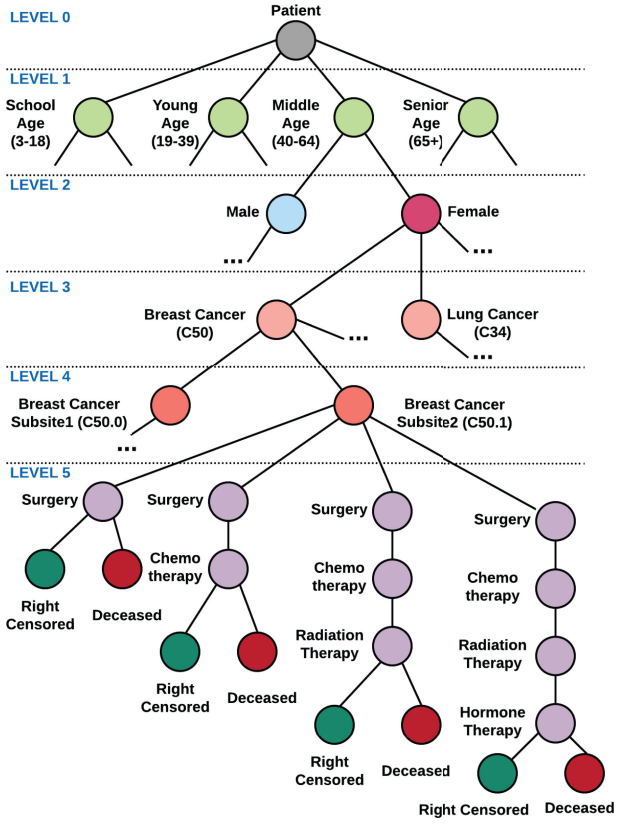

Fig. 2. Hierarchical breast cancer treatment sequence scenarios.

V. APPLICATIONS

A. Application 1: Finding Population Level Treatment Sequences in Breast Cancer

Scenario specific queries represent about 60% of physician directed clinical queries, and are usually hierarchical in structure [39]. 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 treatment sequence (Fig. 2). This query groups patients by age, gender, cancer site (specified pre-query as breast cancer), initial treatments, and survival status. Breast cancer subsites are defined by ICD-O-3 topography codes (C50.0-C50.9) [40]. This query yielded sequences of dates of first treatments for surgery, chemotherapy, radiation, and hormone therapy. In this study, we only consider the registry collected binary treatment data available in the CTC: surgery, chemotherapy, radiation therapy, and hormone therapy. 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 therapies, or various durations of treatment. Hence, we consider treatment paths of length one treatment (only surgery, etc.),

TABLE II

IN THIS TABLE, WE PROVIDE OUR SCENARIO RESULTS FOR TREATMENT SEQUENCE LENGTH ONE. THE FIRST COLUMN REPRESENTS THE QUERY ID. THE SECOND COLUMN SHOWS THE SCENARIO PATH (MIDDLE AGE, FEMALE, BREAST CANCER, AND TREATMENT SEQUENCE LENGTH ONE). IN THE SECOND COLUMN, **S** MEANS SURGERY, **C** MEANS CHEMOTHERAPY, **R** MEANS RADIATION THERAPY, AND **H** MEANS HORMONE THERAPY. WE PRESENT OUTCOMES IN THE THIRD COLUMN, WHICH IS DIVIDED INTO BREAST CANCER SUBSITES. THE BREAST CANCER SUBSITE COLUMNS ARE FURTHER DIVIDED INTO **RC** (RIGHT CENSORED) AND **DE** (DECEASED) COLUMNS. THE "∗" SYMBOL MEANS "WE CAN NOT REPORT THE VALUE BECAUSE OF OUR DATA REPORTING RESTRICTION." EMPTY TABLE CELL MEANS "NO VALUE IS AVAILABLE"

Query ID	Scenario Path	Outcome																	
		C50.0		C50.1		C50.2		C50.3		C50.4		C50.5		C50.6		C50.8		C50.9	
		RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE
Q1.1	S	42	∗	303	55	521	92	271	60	1,583	290	324	55	32	∗	1,123	191	1,189	244
Q1.2	C			∗	∗	22	19	∗	∗	64	59	∗	∗	∗	∗	56	49	45	131
Q1.3	R	∗		∗	∗	∗	∗	∗	∗	∗	∗	∗	∗			∗	∗	∗	18
Q1.4	H						∗	∗		∗	∗	∗	∗	∗			∗	∗	17

TABLE III

IN THIS TABLE, WE PROVIDE OUR SCENARIO RESULTS FOR TREATMENT SEQUENCE LENGTH TWO. THE FIRST COLUMN REPRESENTS THE QUERY ID. THE SECOND COLUMN SHOWS THE SCENARIO PATH (MIDDLE AGE, FEMALE, BREAST CANCER, AND TREATMENT SEQUENCE LENGTH TWO). IN THE SECOND COLUMN, **S** MEANS SURGERY, **C** MEANS CHEMOTHERAPY, **R** MEANS RADIATION THERAPY, AND **H** MEANS HORMONE THERAPY. HERE, TREATMENT PATHS ARE MENTIONED IN THE ORDERED INITIAL FORMAT. FOR EXAMPLE, **S-C** MEANS THE PATIENT FIRST USED SURGERY AND THEN USED CHEMOTHERAPY. WE PRESENT OUTCOMES IN THE THIRD COLUMN, WHICH IS DIVIDED INTO BREAST CANCER SUBSITES. THE BREAST CANCER SUBSITE COLUMNS ARE FURTHER DIVIDED INTO **RC** (RIGHT CENSORED) AND **DE** (DECEASED) COLUMNS. THE "∗" SYMBOL MEANS "WE CAN NOT REPORT THE VALUE BECAUSE OF OUR DATA REPORTING RESTRICTION." EMPTY TABLE CELL MEANS "NO VALUE IS AVAILABLE"

Query ID	Scenario Path	Outcomes																	
		C50.0		C50.1		C50.2		C50.3		C50.4		C50.5		C50.6		C50.8		C50.9	
		RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE
Q2.1	S-C	∗	∗	100	41	249	64	138	35	820	263	164	42	20	∗	502	153	365	152
Q2.2	S-R	∗	∗	156	19	338	53	159	17	1,014	121	217	24	21	∗	547	63	361	53
Q2.3	S-H	16	∗	116	∗	180	16	99	∗	483	45	117	∗	∗	∗	357	35	277	32
Q2.4	C-S	∗		24	∗	34	∗	18	∗	147	78	30	∗	∗	∗	82	45	63	70
Q2.5	C-R			∗	∗		∗	∗	∗	∗	∗	∗	∗		∗	∗	∗	∗	23
Q2.6	C-H			∗	∗	∗	∗			17	∗	∗	∗	∗		∗	∗	∗	17
Q2.7	R-S	∗		∗		∗		∗	∗	∗	∗	∗		∗		∗	∗	16	∗
Q2.8	R-C				∗		∗		∗	∗	∗		∗				∗	∗	23
Q2.9	R-H		∗							∗	∗					∗	∗	∗	∗
Q2.10	H-S	∗		∗	∗	∗	∗	∗	∗	17	∗	∗	∗			∗	∗	17	∗
Q2.11	H-C			∗	∗	∗	∗			∗	∗	∗				∗	∗	∗	∗
Q2.12	H-R						∗			∗	∗						∗	∗	∗

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. 2 includes patients aged 40-64 to demonstrate the hierarchy.

We implemented the hierarchy (Fig. 2) and used the LTR knowledge graph for this example (Section IV). 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 creation and/or integration of datasets for survival analysis or other statistical modeling. We created 64 different combinations of treatment sequences with four treatment types (surgery, chemotherapy, radiation therapy, and hormone therapy). We present our results in Tables II, III, IV, and V. The query results show that a large number (10,262 primary tumors) of breast cancers occur in the upper-outer quadrant (C50.4), and at date of last contact the majority of these patients have received a treatment sequence of length three as opposed to length one, two, and four. The 64 queries performed on average 76% faster on the knowledge graphs than using an RDBMS (see Appendix VI). Note that PostgreSQL database and Virtuoso triplestore's internal indexing algorithms play a significant role in query performances. In this experiment, we only considered treatment types with valid date information available (surgery, chemotherapy, radiation, and hormone therapy).

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 advanced queries using multiple datasets to explain the variation associated with socioeconomic status or other factors influencing exposures or outcomes. To demonstrate, we linked neighborhood concentrated disadvantage index (CDI) dataset with cancer registry dataset to show associations between socioeconomic factors and cancer cases.

Triple negative breast cancer (TNBC) is an aggressive form of breast cancer [5]. In TNBC, three receptors (estrogen, progesterone, and human epidermal growth factor receptor 2)

TABLE IV

THIS TABLE PRESENTS TREATMENT SEQUENCE LENGTH THREE SCENARIO RESULTS. THE FIRST COLUMN REPRESENTS THE QUERY ID. THE SECOND COLUMN SHOWS THE SCENARIO PATH (MIDDLE AGE, FEMALE, BREAST CANCER, AND TREATMENT SEQUENCE LENGTH THREE). IN THE SECOND COLUMN, **S** MEANS SURGERY, **C** MEANS CHEMOTHERAPY, **R** MEANS RADIATION THERAPY, AND **H** MEANS HORMONE THERAPY. HERE, TREATMENT PATHS ARE MENTIONED IN THE ORDERED INITIAL FORMAT. FOR EXAMPLE, **S-C-R** MEANS THE PATIENT FIRST USED SURGERY, THEN USED CHEMOTHERAPY, AND NEXT USED RADIATION THERAPY. WE PRESENT OUTCOMES IN THE THIRD COLUMN, WHICH IS DIVIDED INTO BREAST CANCER SUBSITES. THE BREAST CANCER SUBSITE COLUMNS ARE FURTHER DIVIDED INTO **RC** (RIGHT CENSORED) AND **DE** (DECEASED) COLUMNS. THE "∗" SYMBOL MEANS "WE CAN NOT REPORT THE VALUE BECAUSE OF OUR DATA REPORTING RESTRICTION." EMPTY TABLE CELL MEANS "NO VALUE IS AVAILABLE"

Query ID	Scenario Path	Outcomes																	
		C50.0		C50.1		C50.2		C50.3		C50.4		C50.5		C50.6		C50.8		C50.9	
		RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE
Q3.1	S-C-R	∗	∗	94	38	281	69	151	33	923	255	175	35	20	∗	435	122	276	101
Q3.2	S-C-H	∗	∗	77	∗	114	∗	63	∗	330	58	68	∗	∗		254	40	173	40
Q3.3	S-R-C	∗		∗	∗	16	∗	∗	∗	40	33	∗	∗	∗		25	∗	23	∗
Q3.4	S-R-H	∗	∗	130	∗	330	∗	177	∗	939	39	176	∗	∗		534	17	224	∗
Q3.5	S-H-C	∗		22	∗	33	∗	21	∗	97	37	17	∗	∗		73	28	52	26
Q3.6	S-H-R	∗		51	∗	134	∗	72	∗	425	44	62	∗	∗	∗	198	∗	115	17
Q3.7	C-S-R		∗	32	21	43	19	24	∗	213	103	42	17	∗	∗	106	61	86	83
Q3.8	C-S-H			∗		∗		∗		26	∗	∗	∗			∗	∗	∗	∗
Q3.9	C-R-S			∗	∗	∗	∗	∗		∗	∗			∗		∗	∗	∗	∗
Q3.10	C-R-H										∗					∗	∗		∗
Q3.11	C-H-S			∗		∗	∗	∗		16	∗	∗		∗		∗	∗	∗	∗
Q3.12	C-H-R				∗	∗	∗	∗	∗	∗	∗					∗	∗	∗	∗
Q3.13	R-S-C				∗					∗	∗					∗	∗	∗	∗
Q3.14	R-S-H					∗		∗		∗						∗		∗	∗
Q3.15	R-C-S				∗		∗	∗		∗	∗					∗	∗	∗	∗
Q3.16	R-C-H									∗	∗					∗	∗		∗
Q3.17	R-H-S									∗			∗				∗		∗
Q3.18	R-H-C			∗						∗	∗							∗	∗
Q3.19	H-S-C			∗		∗		∗		∗	∗					∗			∗
Q3.20	H-S-R			∗		16		∗	∗	31	∗	∗		∗		17	∗	∗	∗
Q3.21	H-C-S			∗		∗	∗	∗		∗	∗	∗				∗	∗	∗	∗
Q3.22	H-C-R				∗	∗	∗		∗	∗	∗					∗	∗	∗	∗
Q3.23	H-R-S			∗				∗		∗	∗		∗			∗	∗	∗	∗
Q3.24	H-R-C										∗								∗

TABLE V

THIS TABLE PRESENTS TREATMENT SEQUENCE LENGTH FOUR SCENARIO RESULTS. THE FIRST COLUMN REPRESENTS THE QUERY ID. THE SECOND COLUMN SHOWS THE SCENARIO PATH (MIDDLE AGE, FEMALE, BREAST CANCER, AND TREATMENT SEQUENCE LENGTH FOUR). IN THE SECOND COLUMN, **S** MEANS SURGERY, **C** MEANS CHEMOTHERAPY, **R** MEANS RADIATION THERAPY, AND **H** MEANS HORMONE THERAPY. HERE, TREATMENT PATHS ARE MENTIONED IN THE ORDERED INITIAL FORMAT. FOR EXAMPLE, **S-C-R-H** MEANS THE PATIENT FIRST USED SURGERY, FOLLOWED BY CHEMOTHERAPY, RADIATION THERAPY, AND HORMONE THERAPY. WE PRESENT OUTCOMES IN THE THIRD COLUMN, WHICH IS DIVIDED INTO BREAST CANCER SUBSITES. THE BREAST CANCER SUBSITE COLUMNS ARE FURTHER DIVIDED INTO **RC** (RIGHT CENSORED) AND **DE** (DECEASED) COLUMNS. THE "∗" SYMBOL MEANS "WE CAN NOT REPORT THE VALUE BECAUSE OF OUR DATA REPORTING RESTRICTION." EMPTY TABLE CELL MEANS "NO VALUE IS AVAILABLE"

Query ID	Scenario Path	Outcomes																	
		C50.0		C50.1		C50.2		C50.3		C50.4		C50.5		C50.6		C50.8		C50.9	
		RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE	RC	DE
Q4.1	S-C-R-H	∗		40	∗	120	∗	57	∗	421	35	67	∗	∗		213	16	156	18
Q4.2	S-C-H-R	∗		45	17	104	19	53	∗	333	75	60	∗	∗		171	41	90	32
Q4.3	S-R-C-H	∗		∗	∗	∗	∗	∗		29	∗	∗	∗			21	∗	∗	∗
Q4.4	S-R-H-C	∗		∗	∗		∗	∗		∗	∗	∗	∗			∗	∗	∗	∗
Q4.5	S-H-C-R	∗		22	∗	64	∗	32	∗	220	58	46	∗	∗		103	30	49	23
Q4.6	S-H-R-C	∗				∗	∗	∗		∗	∗					∗			∗
Q4.7	C-S-R-H	∗		∗	∗	21	∗	∗	∗	71	∗	∗	∗	∗		47	∗	41	∗
Q4.8	C-S-H-R			∗		∗	∗	∗		40	∗	∗	∗	∗		16	∗	15	∗
Q4.9	C-R-S-H									∗	∗					∗	∗	∗	
Q4.10	C-R-H-S																	∗	
Q4.11	C-H-S-R	∗	∗	∗	∗	∗	∗	∗		34	∗	∗	∗			22	∗	17	∗
Q4.12	C-H-R-S									∗								∗	∗
Q4.13	R-S-C-H				∗	∗				∗			∗						
Q4.14	R-S-H-C					∗				∗			∗						
Q4.15	R-C-S-H			∗						∗						∗		∗	∗
Q4.16	R-C-H-S					∗											∗		∗
Q4.17	R-H-S-C										∗								
Q4.18	R-H-C-S					∗											∗		
Q4.19	H-S-C-R					∗				∗	∗	∗				∗	∗	∗	∗
Q4.20	H-S-R-C									∗						∗			
Q4.21	H-C-S-R	∗		∗	∗	∗	∗	∗		23	∗	∗	∗			∗	∗	∗	∗
Q4.22	H-C-R-S									∗									∗
Q4.23	H-R-S-C										∗								
Q4.24	H-R-C-S					∗													

TABLE VI
THE SIZE, NUMBER OF TRIPLES, AND CREATION TIME OF THE LOUISIANA CDI MAPPING FILE AND KNOWLEDGE GRAPH

Variables	Louisiana CDI Mapping File	Louisiana CDI Knowledge Graph
Size (KB)	4	864
Number of Triples	54	6,767
Creation Time (minutes)	<1	<1

responsible for breast cancer growth are absent in the cancer tumor [41], [42]. TNBC is aggressive and difficult to treat [5], [43]. Studies show that African American (AA) women are more likely than European American (EA) women to be diagnosed with TNBC [5], [44]–[47]. Hence, it is important to understand racial disparities in TNBC cases to design targeted interventions. A study about neighborhood social determinants of TNBC among AA (or black) and EA (or white) women is presented in [5], which uses the LTR data in our knowledge graph. We replicated some experiments discussed in [5]. Data preparation and experimentation results are explained below.

Dataset: We prepared the CDI dataset for the state of Louisiana. We calculated the CDI scores for the Louisiana census tracts by following the PhenX Toolkit Protocol, as mentioned in the "Measuring Disadvantage" subsection in [5]. CDI was derived from six census variables: i) percent of individuals below the poverty line, ii) percent of households receiving public assistance, iii) percent of female-headed families, iv) percent of unemployed, v) percent of individuals less than age 18, and vi) percent black (AA). We also used the factor loading presented in [5]. We used R's American Community Survey (ACS) API for the preparation of CDI datasets [48]. Our Louisiana CDI dataset contains 7 columns (census tract FIPS and the six variables mentioned earlier) and 1,149 records. We followed the approach mentioned in Section IV of this paper to generate a knowledge graph from the Louisiana CDI dataset using column names as knowledge graph vocabulary. Table VI summarizes the Louisiana CDI mapping file and knowledge graph information.

Results: Hossain *et al.* present the demographic and tumor stage characteristics of TNBC patients in Louisiana from 2010 to 2012 in **TABLE 1** in [5]. We replicated several experiments in **TABLE 1** in [5], and our results are available in Table VII along with results mentioned in [5]. We found a total of 1,299 TNBC cases in the LTR dataset, although **TABLE 1** in [5] presented 1,216 TNBC cases. Similar to **TABLE 1** in [5], we also observed that greater than 25% of TNBC cases were available in women less than 50 years old. We identified that AA women had ∼47% of TNBC cases (similar to **TABLE 1** in [5]). Furthermore, for SEER SUMMARY STAGE 2000, similar to **TABLE 1** in [5], we found the percentage of localized tumors was higher than other types.

We conducted the two-sample *t*-test to find whether our results and results mentioned in [5] deviated significantly or not. Our results and results mentioned in [5] are available in Table VII, which shows that for AGE, YEARS (%) group "<30" and SEER SUMMARY STAGE 2000(%) "In situ" category, Hossain *et al.* [5] does not provide any value. Therefore, we are not considering the above mentioned rows in our statistical test, because it could lead us to misleading insights. We imported datasets in R and performed a two-sample *t*-test. For AGE, RACE, and SEER SUMMARY STAGE 2000 variables, we considered:

- μ_1: mean of our result percentages (Table VII - column 3)
- μ_2: mean of percentage mentioned in [5] (Table VII - column 4)
- Null Hypothesis $H_0 : \mu_1 - \mu_2 = 0$
- Alternative Hypothesis $H_1 : \mu_1 - \mu_2 \neq 0$

We performed two-sample *t*-test for AGE, RACE, and SEER SUMMARY STAGE 2000. We present our test results in Table VIII. We provide descriptive statistics (sample, sample size, and mean), estimation for difference (95% confidence interval for difference), and test information (t-value, degrees of freedom (df), and p-value) in Table VIII. Our significance level was 0.05. We considered that if p-value >0.05 the difference between the samples is not statistically significant, if p-value is < than 0.05 then the converse is true: The samples are statistically different. Table VIII shows that for AGE, RACE, and SEER SUMMARY STAGE 2000 variables p-values are > 0.05, which means that the difference between the samples is not statistically significant. Hence, we can not reject the null Hypothesis (H_0). Therefore, our results and the results mentioned in [5] are conveying similar information on average.

Hossain *et al.* show the age-specific unadjusted incidence of TNBC by race in Louisiana from 2010 to 2012 in **FIGURE 1** in [5]. We replicated the experiment, and our results are available

TABLE VII
IN THIS TABLE, WE PRESENT OUR EXPERIMENTAL RESULTS AND ALSO RESULTS MENTIONED IN [5] ABOUT DEMOGRAPHICS AND TUMOR STAGE CHARACTERISTICS OF TRIPLE NEGATIVE BREAST CANCER PATIENTS, LOUISIANA 2010–2012. COUNT INFORMATION IS NOT AVAILABLE IN [5]. HOWEVER, IN [5], THE AUTHORS PROVIDE PERCENTAGE INFORMATION. HENCE, THIS TABLE DOES NOT CONTAIN A COUNT COLUMN FOR [5]

Total TNBC Cases (Female) - Our Result: 1,299			
Total TNBC Cases (Female) - Mentioned in [5]: 1,216			
AGE, YEARS (%)			
Age	**Count (Our Result)**	**Percentage (Our Result)**	**Percentage (Mentioned in [5])**
<30	15	1.2	Not Available
30-39	95	7.3	7.3
40-49	233	17.9	18.6
50-59	376	29	29.3
60-69	303	23.3	23.5
70>	277	21.3	21.3
RACE (%)			
Race	**Count (Our Result)**	**Percentage (Our Result)**	**Percentage (Mentioned in [5])**
White	683	52.6	51.8
Black	616	47.4	47.5
SEER SUMMARY STAGE 2000(%)			
Derived SEER Summary Stage 2000	**Count (Our Result)**	**Percentage (Our Result)**	**Percentage (Mentioned in [5])**
In situ	26	2	Not Available
Localized	737	56.7	57.8
Regional	424	32.6	33.3
Distant	107	8.2	8.4
Unknown	5	0.4	0.5

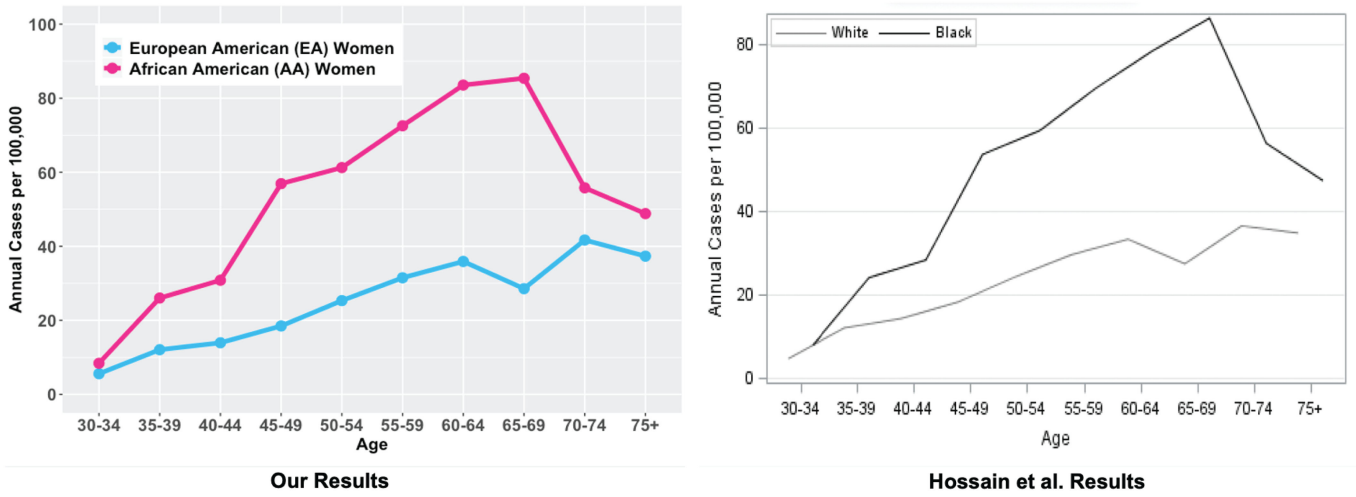

Fig. 3. In this figure, we present, age-specific TNBC incidence rates (Female, Louisiana 2010–2012). The left figure shows our results and the right figure presents Hossain *et al.* results [5] (right figure's credits go to Hossain *et al.* [5]). According to the 2010's census in Louisiana, the total number of EA women is 1,463,058 and the total number of AA women is 766,756 (the ratio is almost 2:1), which includes the age group of <30 years. "The estimates are based on the 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions" [49].

TABLE VIII
IN THIS TABLE WE PRESENT TWO-SAMPLE *T*-TEST RESULTS PERFORMED ON TABLE VII DATA

AGE		
DESCRIPTIVE STATISTICS		
Sample	Sample Size	Mean
Our Result	5	19.76
Result Mentioned in [5]	5	20
ESTIMATION FOR DIFFERENCE		
95% CI for Difference	(-12.0287, 11.5487)	
TEST		
t-value	df	p-value
-0.046947	7.9994	0.9637
RACE		
DESCRIPTIVE STATISTICS		
Sample	Sample Size	Mean
Our Result	2	50
Result Mentioned in [5]	2	49.65
ESTIMATION FOR DIFFERENCE		
95% CI for Difference	(-14.66804, 15.36804)	
TEST		
t-value	df	p-value
0.10374	1.9319	0.9271
SEER SUMMARY STAGE 2000		
DESCRIPTIVE STATISTICS		
Sample	Sample Size	Mean
Our Result	4	24.475
Result Mentioned in [5]	4	25
ESTIMATION FOR DIFFERENCE		
95% CI for Difference	(-45.03209, 43.98209)	
TEST		
t-value	df	p-value
-0.028866	5.998	0.9779

in Fig. 3 with Hossain et al.'s results (**FIGURE 1** in [5]). Similar to Hossain *et al.*, we found that as opposed to EA women, AA women had greater age-specific TNBC incidence (given as cases per 100,000). Furthermore, the age-specific TNBC incidence difference between EA and AA women was higher for the age group 65-69 and lower for the age group 30-34. **FIGURE 2** in [5] shows TNBC CDI distribution by race in Louisiana from 2010 to 2012. Our experimental results are available in Fig. 4 along with Hossain et al.'s results (**FIGURE 2** in [5]). In Fig. 4, our results, the blue bars represent the EA population (white women) and the pink bars depict the AA population (black women), while the purple bars represent the overlapping of blue and pink bars. In Fig 4, both our results and Hossain *et al.* results shows that a significant number of AA women are living in neighborhoods with a high disadvantaged index compared with EA women. Neighborhood CDI is normalized to have mean 0 and standard deviation of 1. So, a census tract with a value of 4 indicates it is four standard deviations more disadvantage relative to the mean. Negative neighborhood CDI means less disadvantage relative to the mean. Because EA and AA populations have different CDI distributions, this could be associated with health outcomes.

Although our results are similar to those in [5], they are not identical. In the following, we summarize the differences.

- We found 1,299 TNBC cases from our experimentation. However, Hossain *et al.* mentioned 1,216 TNBC cases.
- Table VII shows that for the 40–49 age group, we found a lower percentage (17.9) compared with Hossain *et al.* (18.6) [5]. As Table VII shows that, we have a higher white race percentage (52.6) than in Hossain *et al.* (51.8) [5]. Furthermore, Table VII shows that our results (56.7) for the SEER Summary Stage Localized percentage is lower than the result of Hossain et al.'s result (57.8) [5].
- In Fig. 3 (age-specific TNBC incidence), for the age group 75+, there is a shorter gap between EA and AA women in our results compared with those of Hossain *et al.* [5].
- Fig. 4 shows that our maximum Percentage of the Study Population value (y-axis of the plot) for us is less than 10. However, for Hossain *et al.* [5], the maximum Percentage of Study Population value is more than 15. Moreover, in our results, the various Concentrated Disadvantage Index (x-axis) values, the Percentage of Study

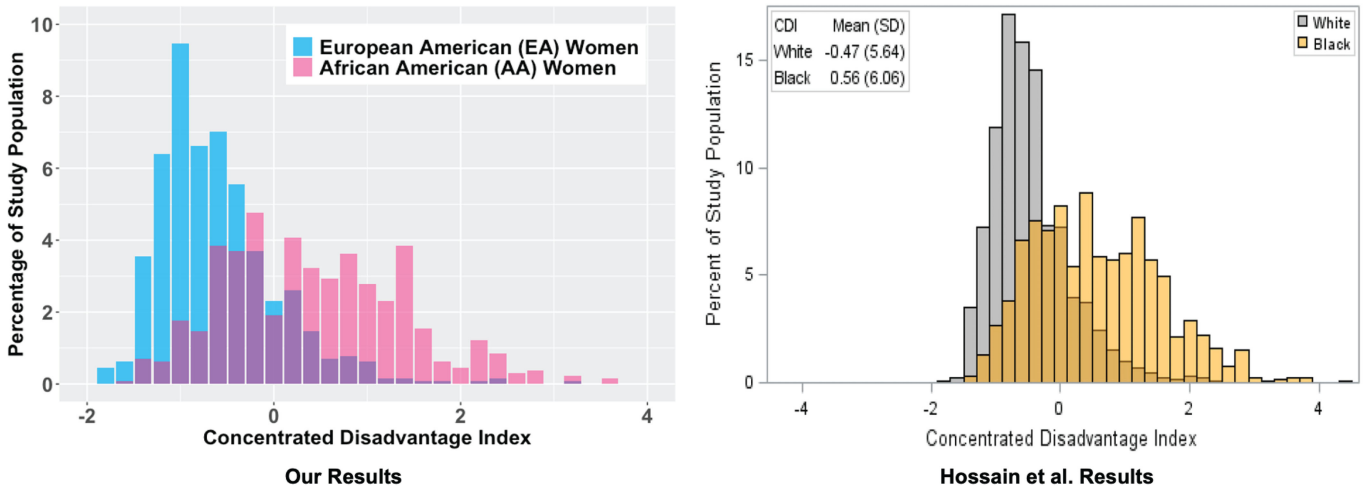

Fig. 4. In this figure, we present TNBC CDI distribution (Female, Louisiana 2010–2012). The left figure shows our results and the right figure presents Hossain *et al.* results [5] (right figure's credits go to Hossain *et al.* [5]).

Population (y-axis) value is less than that of Hossain et al.'s findings [5].

There are several reasons for the small discrepancies, including that Hossain *et al.* [5] excluded some of the census tracts from the experimentation (see the "Census Tract Exclusions" subsection in [5]), and that records had changed in the interval between our analyses. Furthermore, some of the differences maybe attributed to lack of lower-level details (e.g., data normalization strategy) of the experimentation in [5] for full replication. Also, we used an updated version of the dataset (compared to the dataset used by Hossain *et al.* [5]) for our experimentation.

We developed our knowledge graph queries in generic form. We created Python APIs using RDFLib [50], which can communicate with SPARQL Endpoints and can execute SPARQL queries on top of the RDF knowledge graph. The SPARQL queries were written as a template. We provide an API example in Listing 1 that shows the API name, input parameter description, and example input and output. We show a SPARQL query template for the **getCDI** API in Listing 2. In Listing 2, bold texts (stateName, diseaseName, stageSiteFactor, race, sex, diagnosisStartYear, diagnosisEndYear, cdiLowerRange, cdiUpperRange) are placeholders that can be filled with the appropriate values. The query templates allow the creation of an large set of queries. We used our **getCDI** API with various *cdiLowerRange* and *cdiUpperRange* values to create Fig. 4 (our results). In Listing 1, the example output value is 25. We divide the example output value by the total TNBC cases and multiply it by 100 to compute the percentage of the study population value ([25/1,299]*100). We found the percentage of the study population value of 1.9 for Listing 1's input parameters. We plot this value for our results in Fig. 4 (purple bar for x-axis range −0.1 to 0.1). We used another API called **getCaseCount** to get the total TNBC cases. *Here, we omit the implementation detail about other APIs due to space constraints.*

Our knowledge graph supports dynamic data integration. To show the benefits of dynamic data integration and query

Listing 1: getCDI API Example.

```
API:
---
getCDI (stateName, diseaseName, stageSiteFactor,
race, sex, diagnosisStartYear, diagnosisEndYear,
cdiLowerRange, cdiUpperRange)

Parameter Description:
----------------------
stateName = Name of the state
diseaseName = Name of the cancer
stageSiteFactor = Collaborative Stage
Site-Specific Factors
race = Patient race
sex = Patient sex
diagnosisStartYear = Start year of diagnosis
diagnosisEndYear = End year of diagnosis
cdiLowerRange = CDI lower range value
cdiUpperRange = CDI upper range value

Example Input:
--------------
stateName = 'Louisiana'
diseaseName = 'Breast Cancer'
stageSiteFactor = 'Triple Negative'
race = 'African American'
sex = 'Woman'
diagnosisStartYear = 2010
diagnosisEndYear = 2012
cdiLowerRange = -0.1
cdiUpperRange = 0.1

Example Output:
---------------
25
```

generalization, we created another knowledge graph with Kentucky Cancer Registry's (KCR) CTC data. The KCR data is available in CSV format, contains 232 columns, and 207,766 unique tumor records, is 118 MB in size, and has data from 2010 to 2016. We followed the approach mentioned in Section IV for knowledge graph creation. We present the KCR mapping file and knowledge graph size, number of triples, and graph generation time in Table IX. We stored our KCR knowledge graph in our

Listing 2: getCDI API implementation example.

```
SELECT COUNT(?breastCancerPatient)
WHERE{
?breastCancerPatient rdf:type :ctc_stateName.
?breastCancerPatient :disease ?dName.
FILTER(?dName='diseaseName')
?breastCancerPatient :siteSpecificFactor ?sFactor.
FILTER(?sFactor='stageSiteFactor')
?breastCancerPatient :censusTract ?ctcFips.
?breastCancerPatient :race ?rInfo.
FILTER(?rInfo='race')
?breastCancerPatient :sex ?sInfo.
FILTER(?sInfo='sex')
?breastCancerPatient :dateOfDiagnosis ?dDate.
FILTER(?dDate>=?diagnosisStartYear &&
?dDate<=diagnosisEndYear)
?cdi rdf:type :cdi_stateName.
?cdi :tractFips ?cdiFips.FILTER (?cdiFips=?ctcFips)
?cdi :cdiValue ?cVal.
FILTER(?cVal>=?cdiLowerRange &&
?cVal<=?cdiUpperRange)
}
```

TABLE IX
THE SIZE, NUMBER OF TRIPLES, AND CREATION TIME OF THE KENTUCKY MAPPING FILE AND KNOWLEDGE GRAPH

Variables	Kentucky Mapping File	Kentucky Knowledge Graph
Size (KB)	56	8,100,000
Number of Triples	1,420	48,409,945
Creation Time (minutes)	<1	~52

TABLE X
DEMOGRAPHICS AND TUMOR STAGE CHARACTERISTICS OF TRIPLE NEGATIVE BREAST CANCER PATIENTS, KENTUCKY 2010–2012

Total triple negative breast cancer cases (female): 1,096		
AGE, YEARS (%)		
Age	**Count**	**Percentage**
<30	10	0.9
30-39	59	5.4
40-49	200	18.3
50-59	293	26.7
60-69	301	27.5
70+	233	21.3
RACE (%)		
Race	**Count**	**Percentage**
White	961	87.7
Black	135	12.3
SEER SUMMARY STAGE 2000(%)		
Derived SEER Summary Stage 2000	**Count**	**Percentage**
In situ	29	2.6
Localized	665	60.7
Regional	333	30.4
Distant	68	6.2
Unknown	1	0.1

Virtuoso triplestore. We used our graph query APIs to study the TNBC cases for the state of Kentucky. *We received results for Kentucky without changing any code in the APIs (examples are provided in Listings 1 and 2).*

The Kentucky TNBC study results are available in Table X and Figs. 5 and 6. We found 1,096 TNBC cases in Kentucky. Less than 25% of the cases were women below 50 years old, AA women had ~12% of the TNBC cases, and the percentage of localized tumors was higher than other types (see Table X).

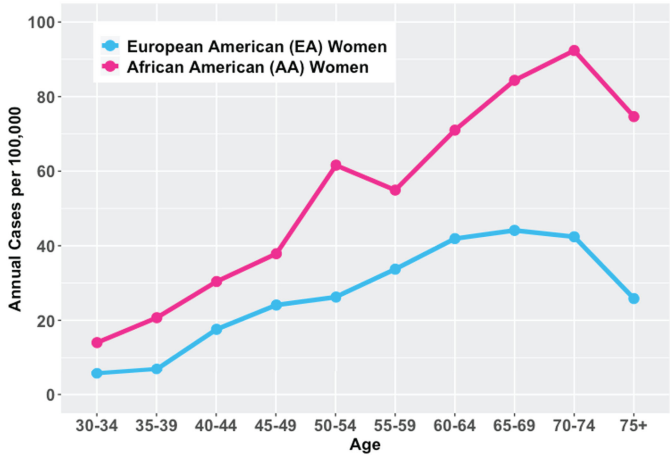

Fig. 5. Age-specific TNBC incidence rates (Female, Kentucky 2010–2012). According to the 2010's census in Kentucky, the total number of EA women is 1,963,670 and the total number of AA women is 173,032 (the ratio is almost 12:1), which includes the age group of <30 years. "The estimates are based on the 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions" [49].

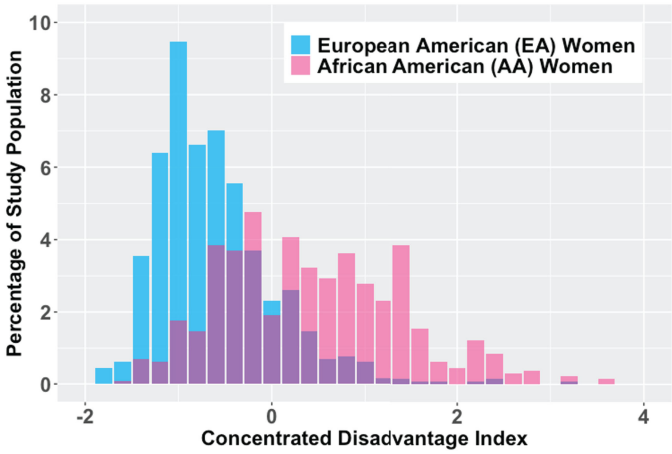

Fig. 6. TNBC CDI Distribution (Female, Kentucky 2010–2012).

Fig. 5 shows, per 100,000, AA women had a higher age-specific incidence of TNBC than EA women. Fig. 6 presents that the CDI distribution between Kentucky and Louisiana are different. Although some of the AA cancer population live in extreme highly disadvantaged neighborhoods, a portion of the EA population also lives in these disadvantaged neighborhoods. Please note that Kentucky has a lower AA population than Louisiana. Hossain *et al.* [5] did not provide any experimentation with KCR data. Hence, for KCR data, we are not providing any comparison of our results with Hossain et al.'s results [5].

We proposed a cancer scientific digital library framework in Fig. 1. "Hypothesis Generation and Testing" is a service mentioned in Fig. 1. The API example that we provide in Listing 1 and 2 can be employed to implement "Hypothesis Generation and Testing" service. For example, based on LTR data researchers can have following hypothesis "Neighborhood

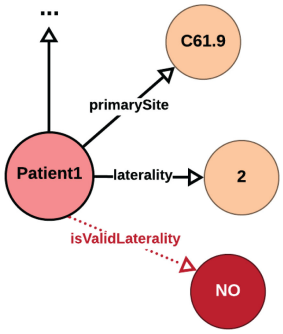

Fig. 7. This figure illustrates an example of the knowledge graph's easy schema evolution. Here we show a patient with prostate cancer (code: C61.9) and a laterality value of 2 (value for a paired organ). To identify invalid laterality codes, we add an "isValidLaterality" flag dynamically to the graph without substantially changing the software code.

disadvantage has an impact on disparities in the incidence of TNBC." By using KCR data researchers can validate the hypothesis. Moreover, our knowledge graph supports dynamic data integration. Hence, we can quickly integrate data from many registries (e.g., New Jersey, Utah) for a robust "Hypothesis Generation and Testing" service.

C. Application 3: Easy Schema Evolution for Iterative Analysis

Exploratory data analysis is naturally iterative, so it is imperative to use a database which enables the schema or design to be updated quickly and easily. One example of schema evolution (modification of the schema) is identifying anomalies and updating the schema to flag them with an indicator variable. While there may not be a large number of discrepancies, it is important to identify them and analyze them appropriately to avoid bias and to examine outliers. We checked the Louisiana CTC dataset for deviations in prostate cancer laterality and breast cancer histology, and found various deviations from the SEER coding standards [51]. For example, the prostate is a single organ, so there are not left and right prostate, however we identified 118 prostate tumors with laterality values for paired organs. We also discovered 88 breast tumors with histology values in the range of 9590-9992. However, breast cancer patient histology reporting values within the range of 9590-9992 are outliers and merit further scrutiny [51]. It is important to identify outliers so they can be excluded from analyses where appropriate and analyzed independently as needed. Identifying data discrepancies for further investigation requires the addition of an outlier "flag" column in the dataset. This flag triggers a schema update. In an RDBMS, it is difficult to manage schema updates for continuously updated data. However, because of the flexible structure of knowledge graphs, it is easier to handle schema changes and evolution. To address the data discrepancy problem, we added new nodes and edges to the graph without substantially changing the core software code that uses the knowledge graph (see Fig. 7).

D. Application 4: Knowledge Graph Visualization

It is difficult for a cancer researcher to graphically understand relationships among many entities and recognize patterns from a massive volume of data. However, data visualization makes it easy for cancer researchers to read and understand those various relationships and patterns. It is possible to develop a visualization on top of an RDBMS tool; however, customized software development is required. Moreover, as we mentioned earlier, expensive join operations are necessary to connect numerous heterogeneous datasets. Furthermore, it is difficult for an RDBMS-based visualization tool to handle continuous data growth and schema change, which is common in cancer research. Conversely, the knowledge graph is developed on top of the graph data structure. Hence, it is easy to visualize a knowledge graph using various open-source tools to find important nodes, edges, and patterns from the graph.

We developed a knowledge graph prototype visualization service with the use of the open-source Gephi graph visualization tool [52]. We employed Gephi's Semantic Web Importer plugin, which queries our Virtuoso SPARQL endpoint and provides a graph visualization [53]. Gephi supports numerous graph layout options (e.g., ForceAtlas 2, Yifan Hu) and coloring scheme options for sophisticated visualization creation. Moreover, Gephi freely provides several graph algorithms such as PageRank, connected components, average degree, modularity, graph density, and diameter. The above mentioned algorithms help us to discover patterns from the massive scale knowledge graph easily. However, in the RDBMS world, a significant amount of programming is required to develop a visualization service with the algorithms mentioned earlier. Gephi is an example of a visualization tool. Many free tools (e.g., D3 [54], Cytoscape [55]) are available for knowledge graph visualization. *To the best of our knowledge, this study presents the first attempt to create an interactive visualization service on top of a knowledge graph that contains cancer registries data and numerous third-party data.*

In this paper, we present Gephi-based high-level (Fig. 8) and a low-level (Fig. 9) visualization examples of our knowledge graph. Fig. 8 presents our high-level visualization, which is a partial visualization of the Rural-Urban Continuum Codes data that we have in our knowledge graph (see the preliminary version of this work for details [22]). The visualization is based on node degree information (means the number of edges connected to the node). The red nodes and some of the gray nodes represent the U.S. state's (e.g., Kentucky, Texas) information. Many counties are connected to state nodes. Hence, state nodes have more connections. The degree of the red nodes is more than 100 and the degree of the gray nodes is 80-100. The green, black, blue, and orange nodes and some pink nodes represent the counties information (e.g., East Baton Rouge Parish in Louisiana). In our knowledge graph, all the counties contain 5 Rural-Urban Continuum codes (codes for 1974, 1983, 1993, 2003, and 2013). Hence, the county level nodes' degree is 5. However, in Fig. 8, we show a partial visualization. Thus, some of the county degrees are less than 5. Therefore, the degrees of the green, black, blue, and orange nodes are 5-2, and they are clustered in the middle

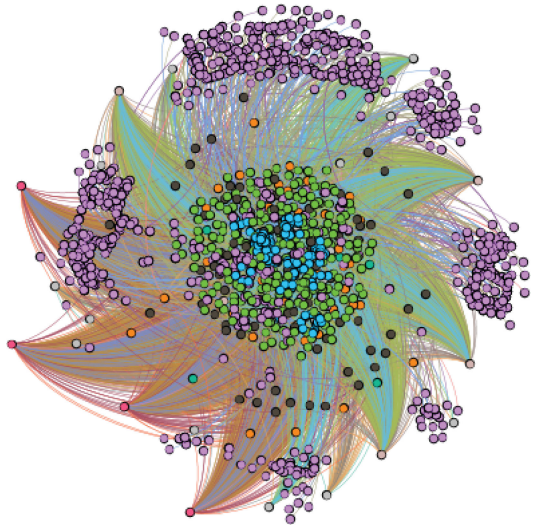

Fig. 8. In this figure we present, a partial high level visualization of LTR knowledge graph (Rural-Urban Continuum Codes portion).

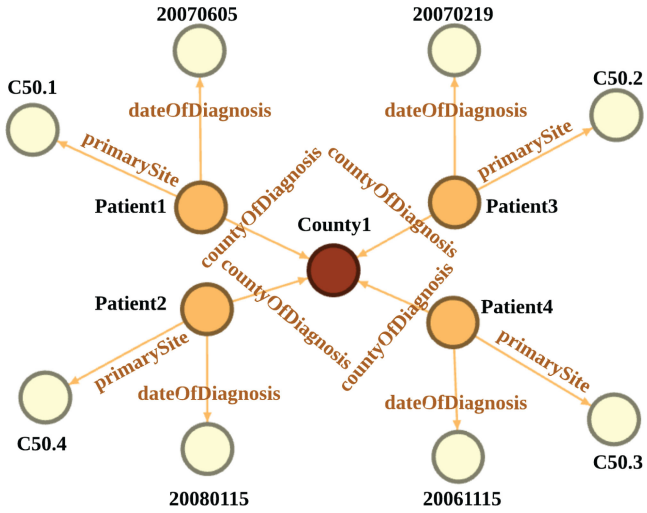

Fig. 9. Visualization of the four hypothetical patients information. Although the patient information was simulated to avoid revealing protected health information (PHI) data, the underlying schema is consistent with the actual CTC data.

of the visualization. The pink nodes degree is 1. The edge colors are based on the connections among the nodes and their colors. For example, if a green node points to a gray node, then the visualization presents a green line from the green node to the gray node. We used the "Yifan Hu" layout algorithm [56] for visualization, which is responsible for node aggregation and scattering. The degree of pink nodes is less than those of the green, black, blue, and orange nodes. However, some of the pink nodes exist in the middle of the cluster. The "Yifan Hu" layout algorithm presents the visualization in this form.

We present a low-level example of our knowledge graph visualization in Fig. 9, where we show the relationships between four hypothetical patients with their primary site, date of diagnosis, and county of diagnosis properties from the Louisiana CTC knowledge graph. Fig. 9 shows that all four patients diagnosed are in the same county. This approach demonstrates the visual link between counties with unusual trends and can help to identify hypotheses that are ripe for future investigation.

VI. Conclusion and Future Work

We described a knowledge graph-based approach to the secondary use of cancer registry data. In contrast to RDBMS, a knowledge graph approach makes it easy to handle scenario-specific queries, link third-party data, manage schema changes, and create data visualization. Our study demonstrated that cancer registry data management and analysis may receive significant advantages from a knowledge graph approach. We developed a prototype based on patient-level data from LTR and KCR. To exhibit the benefits of the knowledge graph approach, we used scenario specific queries that determine the relationships between cancer treatment sequences and survival outcomes. For treatment sequence length one (Table II), our results show the following: (i) for the surgery option, we found 5,388 right censored (not verified dead) patients and 1,008 deceased patients; (ii) for the chemotherapy option, we discovered 215 right censored patients and 283 deceased patients; (iii) for the radiation therapy option, we found 37 right censored patients and 46 deceased patients; and (iv) for the hormone therapy option, we identified 32 right censored patients and 36 deceased patients. To calculate the above results, for each treatment sequence (surgery, chemotherapy, radiation therapy, and hormone therapy), we aggregated all the right censored and deceased columns (for all subsites: C50.0-C50.9) available in Table II. Our results shows that surgery is most the popular option for treatment sequence length one. We present the right censored and deceased information for treatment sequence length two, three, and four, in Tables III, IV, and V respectively. To illustrate the knowledge graph's ease of use in iterative analysis, we linked the knowledge graph to external datasets for performing complex queries using multiple datasets. In addition, we showed the advantage of using knowledge graphs to identify discrepancies in the data and handling schema changes. Finally, we presented the benefits of knowledge graph visualization for pattern discovery. The human brain can process visual information faster than written text. The visualization platform combines human knowledge with machines to interact with data, recognizing patterns that help to understand the context of the research problem better [57]. We present two visualizations (high-level and low-level) examples in Subsection V-D (Application 4: Knowledge Graph Visualization). High-level visualization (Fig. 8) helps us to quickly understand the structure of the graph (e.g., important nodes, edges, clusters). On the other hand, low-level visualization provides detail information. For example, Fig. 9 shows that multiple cancer patients belong to a particular county (*County1*). Policymakers can use this information to design targeted intervention.

In the future, we plan to incorporate different external datasets, including BioPortal, GeoNames, DrugBank, LinkedCT, DBpedia, Bio2RDF, SemMed, etc. Incorporation of these datasets will open a new window of opportunity for advanced analytics including eligibility identification for clinical trials, evaluation of drug treatment, analysis of patient geographic variation and risk factors, and more. We are interested in

applying graph pattern mining algorithms to generate clinically meaningful hypotheses from our knowledge graph. This technique could also be applied to evaluate the treatment sequences and outcomes using additional linked data. Finally, we also plan to integrate relevant ontologies within our knowledge graph to develop a domain specific language which is valuable for cancer researchers.

APPENDIX A
ABBREVIATIONS

In the Table XI, we present list of abbreviations used in this paper.

TABLE XI
LIST OF ABBREVIATIONS

Abbbreviation	Full Form
5S	Streams, Structures, Spaces, Scenarios, and Societies
AA	African American
ACS	American Community Survey
API	Application Programming Interface
CDI	Concentrated Disadvantage Index
CSV	Comma-Separated Values
CTC	Cancer/Tumor/Case
DALY	Disability Adjusted Life Year
DE	Deceased
DF	Degree of Freedom
EA	European American
EHR	Electronic Health Record
FIPS	Federal Information Processing Standards
GUI	Graphical User Interface
JSON	JavaScript Object Notation
KB	Kilobyte
KCR	Kentucky Cancer Registry
LTR	Louisiana Tumor Registry
NA	Not Available
NCI	National Cancer Institute
NR	Not Reporting
OWL	Web Ontology Language
PHI	Protected Health Information
RC	Right Censored
RDBMS	Relational Database Management System
RDF	Resource Description Framework
SEER	Surveillance, Epidemiology, and End Results
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TNBC	Triple Negative Breast Cancer
W3C	World Wide Web Consortium
XML	Extensible Markup Language

APPENDIX B
QUERY TIME

Here we present SQL and SPARQL run-time of the queries mention in Subsection V-A- (Application 1: Finding Population Level Treatment Sequences in Breast Cancer).

TABLE XII
IN THIS TABLE, WE PRESENT, SQL AND SPARQL QUERY RUN-TIME, AND IMPROVEMENT INFORMATION FOR THE QUERIES MENTIONED IN TABLE II

Query ID	Scenario Path	SQL Time (Millisecond)	SPARQL Time (Millisecond)	Improvement (Percentage)
Q1.1	S	340	74	78.2
Q1.2	C	320	79	75.3
Q1.3	R	312	95	69.6
Q1.4	H	308	87	71.8

TABLE XIII
IN THIS TABLE, WE PRESENT, SQL AND SPARQL QUERY RUN-TIME, AND IMPROVEMENT INFORMATION FOR THE QUERIES MENTIONED IN TABLE III

Query ID	Scenario Path	SQL Time (Millisecond)	SPARQL Time (Millisecond)	Improvement (Percentage)
Q2.1	S-C	342	82	76.0
Q2.2	S-R	344	75	78.2
Q2.3	S-H	334	83	75.1
Q2.4	C-S	328	84	74.4
Q2.5	C-R	311	75	75.9
Q2.6	C-H	310	79	74.5
Q2.7	R-S	331	78	76.4
Q2.8	R-C	309	74	76.1
Q2.9	R-H	314	76	75.8
Q2.10	H-S	324	75	76.9
Q2.11	H-C	315	77	75.6
Q2.12	H-R	311	78	74.9

TABLE XIV
IN THIS TABLE, WE SHOW, SQL AND SPARQL QUERY RUN-TIME, AND IMPROVEMENT INFORMATION FOR THE QUERIES MENTIONED IN TABLE IV

Query ID	Scenario Path	SQL Time (Millisecond)	SPARQL Time (Millisecond)	Improvement (Percentage)
Q3.1	S-C-R	356	79	77.8
Q3.2	S-C-H	342	76	77.8
Q3.3	S-R-C	351	77	78.1
Q3.4	S-R-H	360	73	79.7
Q3.5	S-H-C	338	75	77.8
Q3.6	S-H-R	354	78	78.0
Q3.7	C-S-R	344	75	78.2
Q3.8	C-S-H	337	75	77.7
Q3.9	C-R-S	343	74	78.4
Q3.10	C-R-H	308	89	71.1
Q3.11	C-H-S	335	74	77.9
Q3.12	C-H-R	307	76	75.2
Q3.13	R-S-C	343	78	77.3
Q3.14	R-S-H	341	76	77.7
Q3.15	R-C-S	334	75	77.5
Q3.16	R-C-H	308	105	65.9
Q3.17	R-H-S	337	94	72.1
Q3.18	R-H-C	312	84	73.1
Q3.19	H-S-C	331	90	72.8
Q3.20	H-S-R	337	73	78.3
Q3.21	H-C-S	329	76	76.9
Q3.22	H-C-R	306	72	76.5
Q3.23	H-R-S	331	74	77.6
Q3.24	H-R-C	307	88	71.3

TABLE XV
IN THIS TABLE, WE PRESENT, SQL AND SPARQL QUERY RUN-TIME, AND IMPROVEMENT INFORMATION FOR THE QUERIES MENTIONED IN TABLE V

Query ID	Scenario Path	SQL Time (Millisecond)	SPARQL Time (Millisecond)	Improvement (Percentage)
Q4.1	S-C-R-H	345	75	78.3
Q4.2	S-C-H-R	344	82	76.2
Q4.3	S-R-C-H	339	75	77.9
Q4.4	S-R-H-C	343	76	77.8
Q4.5	S-H-C-R	357	77	78.4
Q4.6	S-H-R-C	342	89	74.0
Q4.7	C-S-R-H	328	76	76.8
Q4.8	C-S-H-R	327	79	75.8
Q4.9	C-R-S-H	333	95	71.5
Q4.10	C-R-H-S	330	86	73.9
Q4.11	C-H-S-R	324	77	76.2
Q4.12	C-H-R-S	332	79	76.2
Q4.13	R-S-C-H	330	75	77.3
Q4.14	R-S-H-C	326	74	77.3
Q4.15	R-C-S-H	325	75	76.9
Q4.16	R-C-H-S	324	102	68.5
Q4.17	R-H-S-C	325	75	76.9
Q4.18	R-H-C-S	328	81	75.3
Q4.19	H-S-C-R	326	75	77.0
Q4.20	H-S-R-C	322	73	77.3
Q4.21	H-C-S-R	322	78	75.8
Q4.22	H-C-R-S	324	92	71.6
Q4.23	H-R-S-C	324	75	76.9
Q4.24	H-R-C-S	322	80	75.2

APPENDIX C
LOUISIANA AND KENTUCKY FEMALE POPULATION

TABLE XVI
HERE WE PRESENT TOTAL EA AND AA WOMEN AVAILABLE IN LOUISIANAN AND KENTUCKY BY AGE GROUPS [49]

	Louisiana		Kentucky	
Age	EA Women	AA Women	EA Women	AA Women
30-34	89,554	51,535	122,215	11,886
35-39	88,387	44,813	125,870	11,272
40-44	93,277	47,579	130,795	10,973
45-49	108,226	52,710	148,031	12,323
50-54	110,552	52,786	147,354	12,442
55-59	100,534	45,944	135,413	10,318
60-64	88,181	34,711	120,089	7,513
65-69	68,860	24,203	91,424	5,137
70-74	53,553	18,523	70,765	3,970
75+	116,048	33,449	147,045	8,040

ACKNOWLEDGMENT

This work was 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 J. Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] "National Cancer Institute's surveillance, epidemiology, and end results program," 2018. [Online]. Available: https://seer.cancer.gov

[3] "Center for Disease Control's national program of cancer registries," 2018. [Online]. Available: https://www.cdc.gov/cancer/npcr

[4] "Surveillance, epidemiology, and end results (SEER) linked databases," 2018. [Online]. Available: https://seer.cancer.gov/data-software/linked_databases.html

[5] F. Hossain *et al.*, "Neighborhood social determinants of triple negative breast cancer," *Frontiers Public Health*, vol. 7, pp. 1–8, 2019.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web*. Berlin, Germany: Springer, 2007, pp. 722–735.

[7] "Climate JSON," 2020. [Online]. Available: https://github.com/michaelx/climate, Accessed: Feb. 20, 2020.

[8] "What is the difference between Knowledge graph and structured graph?" 2018. [Online]. Available: https://www.quora.com/What-is-the-difference-between-Knowledge-graph-an d-structured-graph, Accessed: Feb. 11, 2018.

[9] S. S. Hasan, E. A. Fox, K. Bisset, and M. V. Marathe, "Epik: A knowledge base for epidemiological modeling and analytics of infectious diseases," *J. Healthcare Informat. Res.*, vol. 1, no. 2, pp. 260–303, 2017.

[10] J. Yan, C. Wang, W. Cheng, M. Gao, and A. Zhou, "A retrospective of knowledge graphs," *Frontiers Comput. Sci.*, vol. 12, no. 1, pp. 55–74, 2018.

[11] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, "Industry-scale knowledge graphs: Lessons and challenges," *Queue*, vol. 17, no. 2, pp. 48–75, 2019.

[12] G. Antoniou and F. Van Harmelen, *A Semantic Web Primer*. Cambridge, MA, USA: MIT Press, 2004.

[13] E. Miller, "An introduction to the resource description framework," *Bull. Amer. Soc. Inf. Sci. Technol.*, vol. 25, no. 1, pp. 15–19, 1998.

[14] "Statistical Linked Dataspaces," 2012. [Online]. Available: https://csarven.ca/statistical-linked-dataspaces, Accessed: Feb. 18, 2020.

[15] B. Smith *et al.*, "The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechno.*, vol. 25, no. 11, pp. 1251–1255, 2007.

[16] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[17] D. L. McGuinness *et al.*, "Owl web ontology language overview," *W3C Recommendation*, vol. 10, no. 10, pp. 1–22, 2004.

[18] "Ontology mapping with owl:sameAs property," 2020. [Online]. Available: http://graphdb.ontotext.com/documentation/free/sameAs-backgroung-inform ation.html, Accessed: Feb, 18 2020.

[19] R. B. Mishra and S. Kumar, "Semantic web reasoners and languages," *Artif. Intell. Rev.*, vol. 35, no. 4, pp. 339–368, 2011.

[20] Z. Pan and I. Horrocks, *Description Logics: Reasoning Support for the Semantic Web*. Manchester, U.K.: Univ. Manchester, 2004.

[21] R. Angles and C. Gutierrez, "The expressive power of SPARQL," in *Proc. Int. Semantic Web Conf.*, 2008, pp. 114–129.

[22] S. S. Hasan, D. Rivera, X.-C. Wu, J. B. Christian, and G. Tourassi, "A knowledge graph approach for the secondary use of cancer registry data," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2019, pp. 1–4.

[23] A. Pourshams *et al.*, "The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990–2017: A systematic analysis for the global burden of disease study 2017," *Lancet Gastroenterol. Hepatol.*, vol. 4, no. 12, pp. 934–947, 2019.

[24] A. Vafajoo, R. Salarian, and N. Rabiee, "Biofunctionalized microbead arrays for early diagnosis of breast cancer," *Biomed. Phys. Eng. Express*, vol. 4, no. 6, 2018, Art. no. 065028.

[25] L. M. Force *et al.*, "The global burden of childhood and adolescent cancer in 2017: An analysis of the global burden of disease study 2017," *Lancet Oncol.*, vol. 20, no. 9, pp. 1211–1225, 2019.

[26] C. Safran *et al.*, "Toward a national framework for the secondary use of health data: An American medical informatics association white paper," *J. Amer. Med. Informat. Assoc.*, vol. 14, no. 1, pp. 1–9, 2007.

[27] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[28] H. A. Piwowar *et al.*, "Towards a data sharing culture: Recommendations for leadership from academic health centers," *PLoS Medicine*, vol. 5, no. 9, 2008, Paper e183.

[29] W. R. Hersh, "Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance," *Clin. Pharmacol. Ther.*, vol. 81, pp. 126–128, 2007.

[30] C. Tao *et al.*, "A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data," *J. Amer. Med. Informat. Assoc.*, vol. 20, no. 3, pp. 554–562, 2012.

[31] S. Rea *et al.*, "Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project," *J. Biomed. Informat.*, vol. 45, no. 4, pp. 763–771, 2012.

[32] A. Esteban-Gil, J. T. Fernández-Breis, and M. Boeker, "Analysis and visualization of disease courses in a semantically-enabled cancer registry," *J. Biomed. Semantics*, vol. 8, no. 1, pp. 1–16, 2017.

[33] E. A. Fox, M. A. Goncalves, and R. Shen, "Theoretical foundations for digital libraries: The 5S (societies, scenarios, spaces, structures, streams) approach," *Synthesis Lectures Inf. Concepts, Retrieval, Services*, vol. 4, no. 2, pp. 1–180, 2012.

[34] "Data cleansing," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Data_cleansing, Accessed: Feb, 20 2020.

[35] C. Bizer and R. Cyganiak, "The D2RQ plattform," 2009. [Online]. Available: http://d2rq.org/

[36] "Virtuoso Open-Source Edition," 2018. [Online]. Available: http://vos.openlinksw.com/owiki/wiki/VOS, Accessed: Oct. 27, 2019.

[37] "NAACCR DATA DICTIONARY," 2018. [Online]. Available: http://datadictionary.naaccr.org/?c=10.

[38] F. Milicchio and W. A. Gehrke, "Postgresql database," *Distrib. Services OpenAFS: For Enterprise Educ.*, pp. 275–286, 2007.

[39] 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*. New York, NY, USA: Elsevier, 2008, pp. 307–341.

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

[41] R. Dent *et al.*, "Triple-negative breast cancer: Clinical features and patterns of recurrence," *Clin. Cancer Res.*, vol. 13, no. 15, pp. 4429–4434, 2007.

[42] "Triple Negative Breast Cancer," 2019. [Online]. Available: http://www.hipxchange.org/ADI, Accessed: Oct. 16, 2019.

[43] E. A. Rakha, M. E. El-Sayed, A. R. Green, A. H. Lee, J. F. Robertson, and I. O. Ellis, "Prognostic markers in triple-negative breast cancer," *Cancer*, vol. 109, no. 1, pp. 25–32, 2007.

[44] M. J. Lund *et al.*, "Race and triple negative threats to breast cancer survival: A population-based study in Atlanta, GA," *Breast Cancer Res. Treatment*, vol. 113, no. 2, pp. 357–370, 2009.

[45] M. L. Plasilova, B. Hayse, B. K. Killelea, N. R. Horowitz, A. B. Chagpar, and D. R. Lannin, "Features of triple-negative breast cancer: Analysis of 38,813 cases from the national cancer database," *Medicine*, vol. 95, no. 35, pp. 1–6, 2016.

[46] H. M. Sineshaw *et al.*, "Association of race/ethnicity, socioeconomic status, and breast cancer subtypes in the national cancer data base (2010–2011)," *Breast Cancer Res. Treatment*, vol. 145, no. 3, pp. 753–763, 2014.

[47] K. A. Vallega, N. Liu, J. S. Myers, K. Yu, and Q.-X. A. Sang, "Elevated resistin gene expression in African American estrogen and progesterone receptor negative breast cancer," *PloS One*, vol. 11, no. 6, 2016, Paper e0157741.

[48] E. H. Glenn, *acs: Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census*, 2019, r package version 2.1.4. [Online]. Available: https://CRAN.R-project.org/package=acs

[49] "United States Census Bureau - American FactFinder (AFF) - Louisiana," 2020. [Online]. Available: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml, Accessed: Feb. 20 2020.

[50] D. Krech, "RDFLib: A Python library for working with RDF," 2006. [Online]. Available: https://rdflib.readthedocs.io/en/stable/gettingstarted.html

[51] "Site Recode ICD-O-3/WHO 2008 Definition," 2018. [Online]. Available: https://seer.cancer.gov/siterecode/icdo3_dwhoheme/, Accessed: Feb. 11, 2018.

[52] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Proc. 3rd Int. AAAI Conf. Weblogs Social Media*, 2009, pp. 361–362.

[53] "Gephi—Semantic Web Importer," 2018. [Online]. Available: https://seinecle.github.io/gephi-tutorials/generated-html/semantic-web-importer-en.html, Accessed: Oct. 27, 2019.

[54] N. Q. Zhu, *Data Visualization With D3. js Cookbook*. Birmingham, U.K.: Packt Publishing Ltd, 2013.

[55] "Cytoscape," 2020. [Online]. Available: https://cytoscape.org, Accessed: Feb. 20, 2020.

[56] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica J.*, vol. 10, no. 1, pp. 37–71, 2005.

[57] D. White, "Visualization: Set your analytics users free," Tech. rep. Aberdeen Group, Boston, Massachusetts, USA, Tech. Rep., 2013.