Ontology for Documentation of Variable and Data Source Selection Process to Support Integrative Data Analysis in Cancer Outcomes Research

Hansi Zhang[1], Yi Guo[1], Jiang Bian[1]

[1] University of Florida, Gainesville FL 08544, USA

Abstract. To improve cancer prognosis and survival, it is crucial to gain a comprehensive view of potential risk factors (RFs) associated with cancer outcomes (e.g., stage of diagnosis, cancer survival). Guided by the National Institute on Minority Health and Health Disparities (NIMHD) Research Framework, cancer outcomes are influenced by RFs from multiple levels (e.g., individual, inter-personal) and multiple domains (e.g., biological, behavioral). Prior research on RFs of cancer survival, however, has primarily focused on RFs from the individual level (e.g., tumor characteristics) due to the lack of integrated datasets that contain multi-level, multi-domain RFs. It is important to pool RFs from heterogeneous data sources, so that we can examine as many RFs as possible in a multi-level integrative data analysis (IDA). However, RF selection and data integration are inconsistently performed and poorly documented in current cancer research, which threatens scientific reproducibility. Therefore, in this paper, we developed a preliminary reporting protocol for RF variable and data source selection based on our previous cancer survival research. Our protocol is informed by NIMHD framework that provides guidance and promotes structural thinking on identifying multi-level cancer RFs. Further, we propose an ontology-based approach to document RF variable and data source selection so that it is (1) explicitly modeled with a shared, controlled vocabulary, (2) understandable to humans and executable by computers, and (3) adaptive to changes when the process being refined.

Keywords: Ontology, Integrative Data Analysis, Cancer Outcomes Research

1 Introduction

In the United States (US), as the 2nd leading cause of death, cancer is responsible for 1 in every 4 deaths [1]. The lifetime probability of being diagnosed with cancer is 39.7% and 37.6% for men and women, respectively [2]. To improve prognosis and survival, the first and most crucial step is to gain a comprehensive view of potential risk factors (RFs) associated with various cancer outcomes such as the stage of diagnosis (the most important prognostic factor [3, 4]) and survival.

Recognized by the National Institute on Minority Health and Health Disparities (NIMHD) Research Framework [5], individuals are embedded within the larger social system and constrained by the physical environment they lived in. Within this framework, cancer outcomes are influenced by RFs from multiple levels (i.e., individual, interpersonal, community, and societal) and multiple domains (i.e., biological, behavioral, physical/built environment, sociocultural environment, and healthcare system). Prior research on RFs of cancer outcomes, however, has primarily focused on factors from the individual level (e.g., tumor characteristics) due to the lack of integrated datasets that contain multi-level, multi-domain RFs. Very few studies have explored contextual-level RFs (e.g., access to health care services); and certainly no study has comprehensively explored all possible RFs together. To do so, it is important to pool RFs from heterogeneous data sources through data integration, so that we can examine as many RFs as possible in a multi-level integrative data analysis (IDA).

However, RF selection and data integration are inconsistently performed and poorly documented, threatening transparency and reproducibility. When reporting research, it is critical to document the steps that were followed to select, integrate, and process data; so that others can repeat the same steps and reproduce the findings. In this paper, based on our previous experience with multi-level IDAs [6], we developed a preliminary reporting protocol for RF variable and data source selection. Our protocol is informed by the NIMHD framework that provides guidance on identifying multi-level RFs. Further, we propose an ontology-based approach so that the selection process is (1) explicitly modeled with a shared, controlled vocabulary, (2) understandable to humans and executable to computers, and (3) adaptive to changes when the process being refined.

2 Method

2.1 A reporting protocol for cancer risk factor selection, data source selection, and data integration informed by a multi-level IDA case study

In a previous study, we assessed the effect of data integration on predictive ability of cancer survival models [6] and created a semantic data integration (SDI) framework [7] to pool multi-level RFs from heterogenous data sources to support IDA. Table 1 lists the selected RFs and their data sources. Through the case study, a number of variable selection and data integration steps need to be clearly documented. For example, area-rurality status of an individual's residency has different representations based on the choice of using either the rural-urban commuting area definition [8] (i.e., 10 levels from rural to metropolitan) or the NCHS urban-rural classification [9] (i.e., a hierarchal schema with 7 categories). It is important to document how rurality was defined as different representations of the same variable or concept have differential impacts on the predictive ability of the survival model. Further, a number of data assumptions were made as the different datasets were collected at different time periods and on different populations. For example, the Florida Cancer Data System data include cancer patients from 1996 to 2010, while the US Census data we used were from the general population in 2010. Thus, we made assumptions that the area-level characteristics derived from the Census data were applicable across different time periods. Without a clear documentation of such assumptions and choices, other researchers generally would not have a clear picture of these data integration nuances.

Table 1. Identified cancer risk factors, corresponding levels, and data sources.

Cancer Risk Factor		Data Source
Individual level	Sex; Race; Age at diagnosis; Year of diagnosis; Stage of diagnosis; Treatment; Tobacco use; Marital status; Health insurance	FCDS
Contextual level	Social Vulnerability Index (SVI) Area rurality status[a] County-level: smoking rate; alcohol consumption rate; health status; County density of primary care physicians[b]	US Census; BRFSS; FLHealthCHARTS

FCDS: Florida Cancer Data System; BRFSS: Behavioral Risk Factor Surveillance System

[a] Defined based on rural-urban commuting area (RUCA) codes and the National Center for Health Statistics (NCHS) urban-rural classification scheme

[b] Number of primary care physicians per 1,000 people

Through this multi-level IDA case study, we realized that to ensure the transparency and reproducibility of our study, the documentation of our multi-level RF selection choices (e.g., individual smoking status vs county-level smoking rate), data source selection (e.g., individual-level data from FCDS and contextual-level data from US Census), integration (e.g., data integration strategies and use cases), and processing steps (e.g., the need to calculate body max index [BMI] using weight and height vs using a calculated BMI field that came with the raw data) in the study are the key elements. Through discussions with expert biostatisticians, data analysts and cancer outcomes researchers, we summarized the typical IDA process and developed a prototype reporting protocol for RF variable and data source selection. Further, we propose to inform the multi-level and multi-domain RF selection process with the NIMHD research framework, so that investigators can structurally and comprehensively identify relevant RFs and data sources in their IDA studies.

2.2 Ontology for documentation of variable/data source selection (ODVDS)

Scope. The scope of ODVDS was to standardize and document the selection, integration and processing steps of RF variables and data sources to support IDAs guided by the NIMHD research framework for cancer outcomes research.

Approach. Using the basic formal ontology (BFO) as the top-level ontology, the ODVDS was first developed with a top-down approach, where we started by identifying candidate entities (classes and relations) based on the reporting protocol for RF variable and data source selection. Following best practices, we reviewed existing widely accepted ontologies and reused their classes and relations identified using the NCBO BioPortal [10]. We also took a bottom-up process that started with creating the definition of the most specific classes and then subsequent grouped similar classes into more general concepts. The bottom-up approach helped us determine what new classes and relations are needed to fully represent the IDA process.

3 Result

3.1 A reporting protocol for RF variable and data source selection

Informed by the NIMHD research framework, our preliminary reporting protocol consists of two parts as shown in **Fig 1(a)**, reporting (1) the objective of the study including the background, rationale and hypotheses; and (2) the study design for variable and data source selection. The selection process consists of 5 key steps: (1) set up the outcome variables (i.e., primary and secondary outcomes [if necessary]); (2) for each outcome variable, follow an iterative process (**Fig 1(a). A**) to determine the data sources for each outcome variable according to NIMHD framework. For example, if the outcome of interest is individual's *lung cancer* risks, we shall first identify potential data sources that contain individual-level patient data where lung cancer incidence data are available. Then, based on the cohort criteria and other information such as sample size and data range (e.g., time range and geographic information) of the potential data sources, we could determine all qualified data sources; (3) determine the individual-level predictors and covariates of the study; (4) for each individual-level predictor, follow loop B in **Fig 1(a)** to identify the different levels/domains of predictors according to NIMHD framework. Note that multiple variables often exist for the same predictor variable across different data sources, thus, it is important to contrast and consolidate a new predictor with the existing selected predictors to resolve conflicts. Nevertheless, it is often a difficult choice and these "duplicate" variables might all need to be tested in models before a selection is finalized; and (5) follow loop C in **Fig 1(a)** to identify additional contextual-level predictors and data sources of interest.

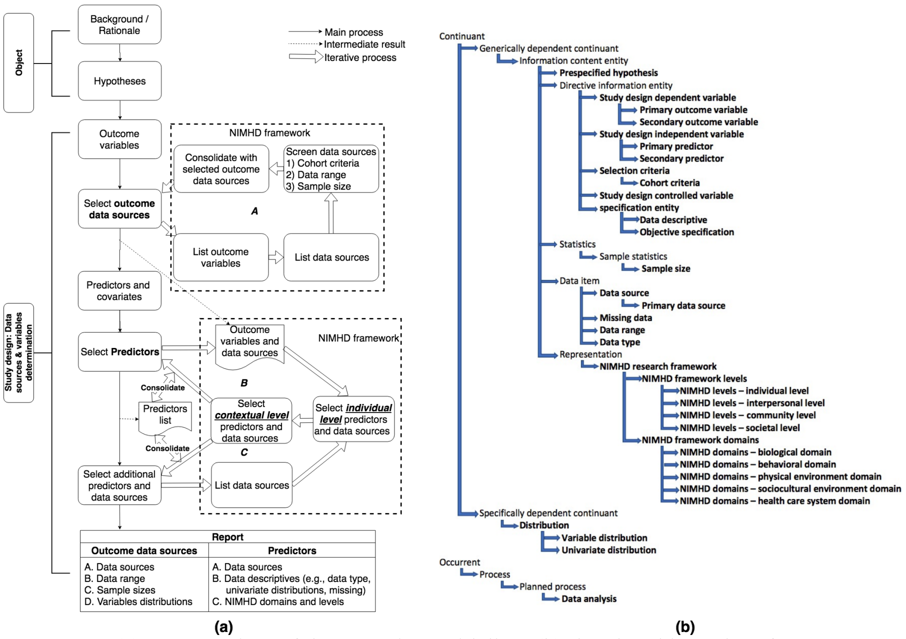
Object
Background / Rationale
Hypotheses
Main process
Intermediate result
Iterative process
Study design: Data sources & variables determination
Outcome variables
NIMHD framework
Consolidate with selected outcome data sources
Screen data sources
1) Cohort criteria
2) Data range
3) Sample size
Select outcome data sources
A
List outcome variables
List data sources
Predictors and covariates
Outcome variables and data sources
NIMHD framework
Select Predictors
B
Consolidate
Predictors list
Select contextual level predictors and data sources
Select individual level predictors and data sources
Consolidate
C
Select additional predictors and data sources
List data sources
Report
Outcome data sources
Predictors
A. Data sources
B. Data range
C. Sample sizes
D. Variables distributions
A. Data sources
B. Data descriptives (e.g., data type, univariate distributions, missing)
C. NIMHD domains and levels
(a)
Continuant
Generically dependent continuant
Information content entity
Prespecified hypothesis
Directive information entity
Study design dependent variable
Primary outcome variable
Secondary outcome variable
Study design independent variable
Primary predictor
Secondary predictor
Selection criteria
Cohort criteria
Study design controlled variable
specification entity
Data descriptive
Objective specification
Statistics
Sample statistics
Sample size
Data item
Data source
Primary data source
Missing data
Data range
Data type
Representation
NIMHD research framework
NIMHD framework levels
NIMHD levels – individual level
NIMHD levels – interpersonal level
NIMHD levels – community level
NIMHD levels – societal level
NIMHD framework domains
NIMHD domains – biological domain
NIMHD domains – behavioral domain
NIMHD domains – physical environment domain
NIMHD domains – sociocultural environment domain
NIMHD domains – health care system domain
Specifically dependent continuant
Distribution
Variable distribution
Univariate distribution
Occurrent
Process
Planned process
Data analysis
(b)

Fig. 1. (a) An overview of the reporting guideline; (b) the class hierarchy of ODVDS

3.2 Development of the ODVDS ontology

Based on the reporting protocol above, we identified the necessary classes and properties to represent the IDA process. We reused classes from the following existing well-known ontologies: Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIt), Data Science Education Ontology (DSEO), and Relations Ontology (RO). We created 20 new classes in ODVDS. Overall, we identified 33 classes and 5 properties. **Fig 1(b)** shows the class hierarchy of ODVDS**.**

4 Discussion, conclusion, and future work

In this work, we developed (1) a reporting protocol for RF variable and data source selection, and then an initial version of the ODVDS ontology for annotating the documentation of the reporting protocol. However, our current work is limited: (1) the protocol is developed based on one case study where the coverage RFs, cancer outcomes and data integration scenarios is limited; (2) we only reviewed a limited number of existing reporting guidelines such as the Checklist for One Health Epidemiological Reporting of Evidence (COHERE) [11] and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [12] statement. A more systematic review of existing reporting guideline for variable and data source selection, data integration process, statistical methods, and analysis plan in health research is warranted to expand the reporting protocol. A good resource of these reporting guidelines is the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network; and (3) current classes and properties in the initial ODVDS only covered RF variable and data source selection. With the expansion of the reporting protocol (e.g., to include data integration scenarios, processing of the data), new classes and properties to fully represent the entire IDA process are needed. Further, tools associated with the reporting protocol and ODVDS are needed as our ultimate goal is to help other investigators to "automatically" reproduce the analytic steps, especially the data integration and processing steps.

Nevertheless, our ontology-based documentation approach provides a good start for researchers to document the RF variable and data source selection process in their multi-level IDAs. Clear documentation is necessary to help researchers communicate their studies to other investigators, assist others to reproduce the analytic datasets, and improve transparency and scientific reproducibility.

References

1. CDC. Statistics for Different Kinds of Cancer. 2017. https://www.cdc.gov/cancer/dcpc/data/types.htm. Accessed 1 Jul 2019.

2. American Cancer Society. Cancer Facts & Figures 2018. Atlanta: American Cancer Society; 2018. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2018/cancer-facts-and-figures-2018.pdf. Accessed 28 Jun 2019.

3. American Cancer Society. Cancer Facts & Figures 2017. 2017. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf. Accessed 1 Jul 2019.

4. Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, et al. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin. 2016;66:271–89.

5. NIMHD. NIMHD Research Framework. https://www.nimhd.nih.gov/about/ overview/research-framework.html. Accessed 28 Jun 2019.

6. Guo Y, Bian J, Modave F, Li Q, George TJ, Prosperi M, et al. Assessing the effect of data integration on predictive ability of cancer survival models. Health Informatics J. 2019;:1460458218824692.

7. Zhang H, Guo Y, Li Q, George TJ, Shenkman E, Modave F, et al. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med Inform Decis Mak. 2018;18. doi:10.1186/s12911-018-0636-4.

8. United Stats Dpartment of Agriculture - Economic Reaserch Service. 2010 Rural-Urban Commuting Area (RUCA) Codes. 2019. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/documentation/. Accessed 8 Jul 2019.

9. National Center for Health Statistics - Office of Analysis and Epidemiology. NCHS Urban-Rural Classification Scheme for Counties. 2017. https://www.cdc.gov/nchs/data_access/urban_rural.htm#2013_Urban-Rural_Classification_Scheme_for_Counties. Accessed 8 Jul 2019.

10. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39 Web Server issue:W541-545.

11. Davis MF, Rankin SC, Schurer JM, Cole S, Conti L, Rabinowitz P, et al. Checklist for One Health Epidemiological Reporting of Evidence (COHERE). One Health Amst Neth. 2017;4:14–21.

12. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. [The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies]. Rev Esp Salud Publica. 2008;82:251–9.