Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the National Health and Nutrition Examination Surveys Use Case

Henrique Santos, Paulo Pinheiro, and Deborah L. McGuinness
To study the relationship between diet, nutrition, and health and their roles in designated population subgroups with select diseases and risk factors

- Occurs every 2 years
  - Interview, Examination, and Laboratory results
  - Tens of datasets per survey cycle

- Challenges in understanding the data
  - Implicit objects in the data (e.g. participant, household, household reference person)
  - Questionnaires vs. Examination (e.g. “Do you have high blood pressure?” vs. Blood pressure readings)
  - Not trivial to quickly identify which types of data are there (e.g. “What are the diabetes-related data?”)
NHANES Knowledge Graph Infrastructure

Data, Documentation, Codebooks
- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

Contents in Detail
- Questionnaire Instruments
- Laboratory Methods
- Procedure Manuals
- Brochures and Consent Documents

Semantic Data Dictionaries

Metadata Templates

Supporting ontologies
- SIO, CHEBI, NCIT, HHEAR, NHANES, NHANES-Drug, …

HADatAc Framework

Search

Dataset generation

Graph browsing API
Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the NHANES Use Case

National Health and Nutrition Examination Survey
2017-2018 Data Documentation, Codebook, and Frequencies

Demographic Variables and Sample Weights (DEMO_J)

Data File: DEMO_J.xpt

Data Processing and Editing

Frequency counts were checked, "skip" patterns were verified, and the reasonableness of question responses was reviewed. Edits were made to some variables to ensure the completeness, consistency, and analytic usefulness of the data. Edits were also made, when necessary, to address data disclosure concerns.

SDDSRVYR: This variable represents the two-year data release cycle number. A value of "10" denotes NHANES 2017-2018.

RIDSTATR: This variable is used to identify whether a participant was both interviewed at home and examined in the mobile examination center (MEC) or was only interviewed in the home but never went through the examination.

RIDAGEYR: Age in years at the time of the screening interview, is reported for survey participants between the ages of 1 and 79 years of age. All responses of participants aged 80 years and older are coded as '80.' The reporting of age in single years for adults 80 years and older was determined to be a disclosure risk. In NHANES 2017-2018, the weighted mean age for participants 80 years and older is 85 years.

RIDAGEYR was calculated based on the participant's date of birth. In rare cases, if the actual date of birth was missing but the participant's age in years was provided, then the reported age was used.

Community developed ontologies

NHANES Demographics Semantic Data Dictionary
Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the NHANES Use Case

The NHANES Ontology and NHANES-Drug Ontology

- 9–11th grade (includes 12th grade with no diploma) (NHANES:00049)
  - 10th grade (NHANES:00041)
  - 11th grade (NHANES:00042)
  - 12th grade, no diploma (NHANES:00043)
  - 9th grade (NHANES:00040)
- College graduate or above (NHANES:00052)
- High school graduate/GED or equivalent (NHANES:00050)
  - GED or equivalent (NHANES:00045)
  - High school graduate (NHANES:00044)
- Less than 9th grade (NHANES:00048)
  - 5th grade (NHANES:00036)
  - 6th grade (NHANES:00037)
  - 7th grade (NHANES:00038)
  - 8th grade (NHANES:00039)
- Less than 5th grade (NHANES:00047)
  - 1st grade (NHANES:00032)
  - 2nd grade (NHANES:00033)
  - 3rd grade (NHANES:00034)
  - 4th grade (NHANES:00035)
  - Never attended / kindergarten only (NHANES:00031)
- More than high school (NHANES:00046)
- Some college or AA degree (NHANES:00051)
Further survey data understanding: Treatment and Prevention of diseases

- Modified ICD10-CM codes in NHANES – problem!
- We formalized in the graph the notion of Prescription Drug Usage (for treatment and for prevention)
- Normalization of ICD10-CM codes using the formal codes

<table>
<thead>
<tr>
<th>SEQN</th>
<th>RXDUSE</th>
<th>RXDDRUG</th>
<th>RXDDRGIID</th>
<th>RXQSEEN</th>
<th>RXDAYS</th>
<th>RXDRSC</th>
<th>RXDRSD</th>
<th>RXDCOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>93759</td>
<td>1</td>
<td>ASPIRIN</td>
<td>d00170</td>
<td>1</td>
<td>35</td>
<td>Z79.0</td>
<td></td>
<td>Long term (current) use of anticoagulants and</td>
</tr>
<tr>
<td>93759</td>
<td>1</td>
<td>ATORVASTATIN</td>
<td>d04165</td>
<td>1</td>
<td>1825</td>
<td>E78.0</td>
<td></td>
<td>Pure hypercholesterolemia</td>
</tr>
<tr>
<td>93759</td>
<td>1</td>
<td>GLIFIZIDE</td>
<td>d00246</td>
<td>1</td>
<td>1825</td>
<td>E11</td>
<td></td>
<td>Type 2 diabetes mellitus</td>
</tr>
<tr>
<td>93759</td>
<td>1</td>
<td>HYDROCHLOROTHIAZIDE</td>
<td>d00253</td>
<td>1</td>
<td>1825</td>
<td>R60.9</td>
<td></td>
<td>Edema, unspecified</td>
</tr>
<tr>
<td>93759</td>
<td>1</td>
<td>METFORMIN</td>
<td>d03807</td>
<td>1</td>
<td>1825</td>
<td>E11</td>
<td></td>
<td>Type 2 diabetes mellitus</td>
</tr>
</tbody>
</table>
Metadata templates: Including survey knowledge in the graph

- HADatAc’s metadata templates help incorporate knowledge about scientific activities and resources related to the surveys

- The DPL (Deployment) template allows the representation of data acquisition instruments (e.g. questionnaires, sample analyzers)

<table>
<thead>
<tr>
<th>hasURI</th>
<th>a rdfs:label</th>
<th>owl:sameAs</th>
</tr>
</thead>
</table>
Human-Aware Data Acquisition Framework (HADatAc-NHANES)

Study View of NHANES-2013-2014

SOC Structure
(select nodes to browse their objects)
The NHANES Knowledge Graph in use & ongoing efforts

- Supporting cross-sectional data analysis for quantifying the representativeness of subjects in randomized clinical trials
  - Determining equitable access to antidiabetic medications and vaccinations

- Representation of statistical components of weighted surveys
  - Weights and variances
  - Goal: Support the determination of correct weights to use in data analysis based on variables of interest
Takeaway messages…

- Scientific studies and surveys have knowledge that are scattered around documentation, codebooks, and data
  - We are formalizing this knowledge as a Knowledge Graph, facilitating several tasks by the different roles of data users

- Importance of reducing the barriers between new users and the data

- All produced assets are freely available (and being evolved) at:
  - https://github.com/tetherless-world/nhanes-hadatac

- Cited works
Thank you!

Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the National Health and Nutrition Examination Surveys Use Case

Henrique Santos, Paulo Pinheiro, and Deborah L. McGuinness

email: oliveh@rpi.edu