

# Towards Machine-Assisted Biomedical Data Preparation: A Use Case on Disparity in Access to Health Care

Paulo Pinheiro<sup>1</sup>, Henrique Santos<sup>2</sup>, Miao Qi<sup>2</sup>, Kristin P. Bennett<sup>2</sup> and Deborah L. McGuinness<sup>2</sup>

<sup>1</sup>*Parcela Semântica, Funchal, Portugal*

<sup>2</sup>*Rensselaer Polytechnic Institute, Troy NY, United States*

## Abstract

Data preparation is a time-consuming task required for data analytics. In the biomedical field, we observe that datasets tend to have a large number of diversified variables, especially when we consider data coming from healthcare facilities. When data analytics depends on variables from several studies, one approach is to use semantics to annotate and support the alignment and combination of variables. We propose a novel use of semantics to support biomedical data preparation, specifically the use of semantic variable normalization in support of machine-assisted biomedical data preparation. To illustrate our approach, we present a use case in disparity in access to health care using data from the U.S. National Health and Nutrition Examination Surveys (NHANES), one of the most studied biomedical datasets in the U.S. This use case is a multi-cycle study of disparities in access to needed care that requires the semantic combination of data from three survey cycles. We demonstrate that NHANES data can be normalized and accessed regardless of cycle by the use of a semantic representation of study variables and a semantically-enabled faceted search. This approach can reduce the time required for data understanding and preparation, especially in settings like NHANES where it is common to combine data from several cycles.

## Keywords

data preparation, NHANES, semantic variable normalization, health informatics, Health Disparity

## 1. Introduction

The typical input for data analysis activities is a dataset rather than the raw data from data files [1]. *Data preparation* is commonly used to describe time-consuming processes that combine data manipulation operations and culminate in the generation of a dataset out of data file content [2]. A *data preparation criterion*, which specifies the variables from one or more sources that are required to perform a data analysis activity, guides the execution of data preparation activities. As an example, an (over-simplified) description such as “all known demographic data

---


*SeWebMeDa-2023: 6th International Workshop on Semantic Web solutions for large-scale biomedical data analytics, May 29, 2023, Hersonissos, Greece*

✉ paulo@psemantica.com (P. Pinheiro); oliveh@rpi.edu (H. Santos); qimiaorpi@gmail.com (M. Qi); bennek@rpi.edu (K. P. Bennett); dlm@rpi.edu (D. L. McGuinness)

🆔 0000-0001-8469-4043 (P. Pinheiro); 0000-0002-2110-6416 (H. Santos); 00000-0002-2917-0965 (M. Qi); 0000-0002-8782-105X (K. P. Bennett); 0000-0001-7037-4567 (D. L. McGuinness)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

in a given study” can be understood as a data preparation criterion that is expected to identify the required elements for generating a dataset out of a collection of data files. The assumption is that the content of the dataset resulting from successful data preparation, i.e., the “prepared data”, is made of data values that meet the inclusion statement in the data preparation criterion.

In the biomedical context, we observe the complexity in analyzing data is often related to a set of variables that is much larger than we can observe in non-biomedical data [3, 4]. For example, in the context of analyzing financial and marketing data, we observe companies performing data analysis in extremely large datasets, e.g., tabular data with literally billions of rows, although with a very reduced number of variables. Nevertheless, this large set of variables behind biomedical data tends to include a complex network of relationships between the entities behind the variables containing the biomedical data.

NHANES is the main national weighted survey for the United States that was started in 1971 and continues to be conducted. Continuous implementation of NHANES has occurred since 1999. NHANES gave origin to thousands of derived studies, and many more studies will be derived from it in the future. Naturally, it is overwhelming for a person who is new to NHANES data, policies, and documentation to confidently perform data preparation. In addition, NHANES data preparation is usually performed to support a single study (or a constrained set of studies). Thus, this effort is rarely translated into knowledge and tools that can simplify the future task of someone preparing NHANES data for new data analysis activities. Because NHANES data has resulted from a careful and complex survey design, conclusions from NHANES can be very accurate if properly combined and analyzed. Otherwise, results may be biased potentially leading to inaccurate conclusions.

In this paper, we explore the problem of using data from the National Health and Nutrition Examination Survey (NHANES) [5] to perform data analysis to assess disparity in access to health care in the U.S. More specifically, we discuss semantic challenges that may arise when accessing the original data from NHANES. We present an approach based on semantic web technologies for facilitating biomedical data preparation, more specifically for the task of uniformly selecting variables from different studies. We propose the abstraction of variables as semantic variables that use properties of entities that are shared among several variables. We operationalize this approach in a use-case for quantifying disparity in access to antidiabetic medication and immunization using NHANES data from three cycles. Further, we discuss that many of the semantic challenges one may experience while using NHANES can be partially mitigated by using a data ingestion framework that is capable of creating semantic normalized and annotated variables from NHANES’ original variables.

## **2. NHANES Background**

NHANES is a list of cross-sectional studies starting in 1971. Since 1999, a new study is added to the list of studies every two years, and each one of these studies done since 1999 has approximately ten thousand subjects. NHANES is based on a sampling design that is used to select participants representative of the civilian, non-institutionalized US population. Each of these periods of two years is called a “cycle.” For each cycle, NHANES uses a complex survey design including oversampling, survey non-response, and post-stratification adjustment to

match total population counts from the Census Bureau [6]. A statistical weight is associated with each survey participant to identify how many members of the US population that participant represents. Therefore, the sampling weights must be used for all subsequent analyses to make valid conclusions based on the NHANES data. A further complication is that the variables included in each cycle evolve over time. Variables may be added and removed between different cycles, and/or the names and/or definitions of the variables may change (e.g. the variable *RIDRETH3 - Race/Hispanic origin w/ NH Asian* introduced in 2011 to include the Non-Hispanic Asian category). Considerable analyst expertise is needed to identify the variables available for analysis in each cycle, to appropriately combine data from different cycles for analysis, to calculate the correct survey weights, and to perform the appropriate survey-weighted design.

NHANES data is acquired for properties of any subject in any cycle just once. The cross-sectional nature of each cycle implies that every observation of a subject occurs once within the entire project. We note that other than the weights associated with each subject, the overall collection of subjects of all cycles are providing the same kind of information. Thus, it is common practice to combine multiple adjacent cycles to create a larger sample size leading to more robust conclusions. This means that if one wants to analyze NHANES data with a very large sample population, one needs to keep aggregating cycles to the pool of subjects to be analyzed since approximately ten thousand new subjects are added for each cycle that is aggregated.

## 2.1. NHANES Cycle Aggregation Challenges

According to NHANES documentation, the only concern one needs to be aware of while aggregating cycles is the fact that the cycles need to be adjacent (i.e., consecutive years with no gaps), and that any weighted variable needs to be averaged by the total number of cycles being aggregated. In practical terms, however, cycle aggregation is a semantic challenge when it comes to understanding how variables of multiple studies are harmonized – reminding us that each cycle is an individual study. Below we list six of these challenges.

- For variables available in one cycle, it is not assured that corresponding variables are available in the next cycle. For example, if one decides to aggregate three consecutive cycles, it is possible that a variable available in the first and third cycles is not available in the second cycle (e.g. Generalized Anxiety Disorder scores only exist in the 1999-2004 cycles).
- Codebooks of categorical variables can change over time, meaning that a manual process of reading the NHANES documentation, understanding codebook changes over time, and developing code to harmonize those variables for a given selection of cycles may be required (e.g. removal of Spanish-speaking countries codes from the country of birth variable in 2011).
- NHANES data contain some “split-categorical variables” where a set of distinct variables represents a single property. To understand the meaning of these variables, one is required to carefully read the documentation to realize these split variables are indeed values of a common property and that they are often used when the study may accept multiple values for the property behind these split variables (e.g. several variables to characterize health insurance coverage).

- NHANES data contain some “merged variables.” We consider a variable to be “merged” when the values of the variable are from two or more distinct properties but are put together as a single variable. An example of a merged variable in NHANES is the variable about “drug usage”. We consider this variable to be merged because some of its values are about “drug usage for disease treatment” while the other values are for “drug usage for disease prevention.”
- None of the knowledge required to address the previous challenges can be obtained from NHANES’ original data. Instead, this is knowledge provided as documentation that needs to be interpreted by humans and translated into data preparation solutions based on code.

An approach to mitigate the challenges above is for scientists to share their data preparation code, hoping it will be very similar for someone doing data preparation with the same set of selected variables. But, as explained above, simply using another cycle’s data and performing the same analysis that was done in the past is likely to lead to incorrect results. There are significant chances that the code may run and produce some results that are apparently correct but biased. Thus, we do consider a risk for scientific accuracy the strategy of reusing existing data analytic code without revisiting the overall NHANES documentation, understanding what are the variables of interest in a selected set of cycles, and understanding changes that may have occurred during selected cycles, every time a new data preparation criterion is established.

### 3. Semantic Solutions

Before we dive into our semantic infrastructure that enables semantic variable normalization, we need to revisit some definitions and introduce the notion of *semantic variable* used in this paper.

#### 3.1. Variables and Semantic Variables

A *Variable*, from the point of view of a tabular data file, is a column in the file. Each variable value, i.e., a value in a column of the table corresponding to our variable of concern, is the measured, elicited, or simulated value of an entity’s attribute. For example, for the cycle 2017-2018 of NHANES there is a variable named RIDAGEYR that corresponds to the attribute “age” of an entity of type “human subject”, and it is stated in “years”. We consider all information about the variable (such as attribute, entity, and unit, in this example) as properties of the variable. For example, “age” is the property *Attribute* of the variable, “human subject” is the property *Entity* associated with the variable, and “years” is the property *Unit* of the variable. It is important to mention that the US population in 2017-2018 is the property *Population* of the RIDAGEYR variable for the 2018-2018 cycle.

A *Variable Specification* is the description of the properties of a variable. The population “US Population in 2017-18”, the attribute “age”, the entity “human subject”, and the unit “years” are all part of the specification of the RIDAGEYR variable for the 2017-2018 cycle. We observe that some properties are present in some variable specifications while others may be not. For instance, not all variables required a property “Unit”, especially when these variables are categorical like “Biological sex”. A comprehensive discussion about variable specification

formalization is beyond the scope of this paper. However, we would like to particularly stress that variable specifications may be missing essential content if their properties *Entity*, *Attribute*, and *Population* are not provided, or are provided as empty definitions.

From our definition of Variable Specification, we define a *Semantic Variable* as a variable specification that does not include a population property. From the variable definition above, each variable is bound to a given population. When the only distinction between the set of properties of any two variables is their populations, we would say that the two variables have the same semantic variable, i.e., the two variables share a common semantic variable. In this case, we can say that RIGAGEYR for the 2015-2016 cycle and RIDAGEYR for the 2017-2018 cycle are two variables with the same semantic variable (“age of the participant in years”) reference. In fact, the only distinction between these two variables is their populations: the population of the first variable is the US population in 2015-2016 while the second is the US population in 2017-2018. The reuse of variable names across cycles is an informal way of NHANES handling the notion that many variables share the same semantic variable. This however can quickly become confusing for a new NHANES user since tools like the use of NHANES variable search<sup>1</sup> to look for RIDAGEYR would return multiple entries. For example, as shown in Figure 1, the search for RIDAGEYR returns 11 entries, one entry for each cycle of NHANES’s continuous period.

Number of variables found: 11

Variable Name	SAS Label	Variable Description	Data File Name	Component Link
RIDAGEYR	Age at Screening Adjudicated - Recode	Best age in years of the sample person at time of HH screening. Individuals 85 and over are topcoded at 85 years of age.	Demographic Variables & Sample Weights (DEMO)	<a href="#">1999-2000 Demographics</a>
RIDAGEYR	Age at Screening Adjudicated - Recode	Best age in years of the sample person at time of HH screening. Individuals 85 and over are topcoded at 85 years of age.	Demographic Variables & Sample Weights (DEMO_B)	<a href="#">2001-2002 Demographics</a>

**Figure 1:** First two entries of the result of 11 entries when performing a variable search for “RIDAGEYR” in NHANES.

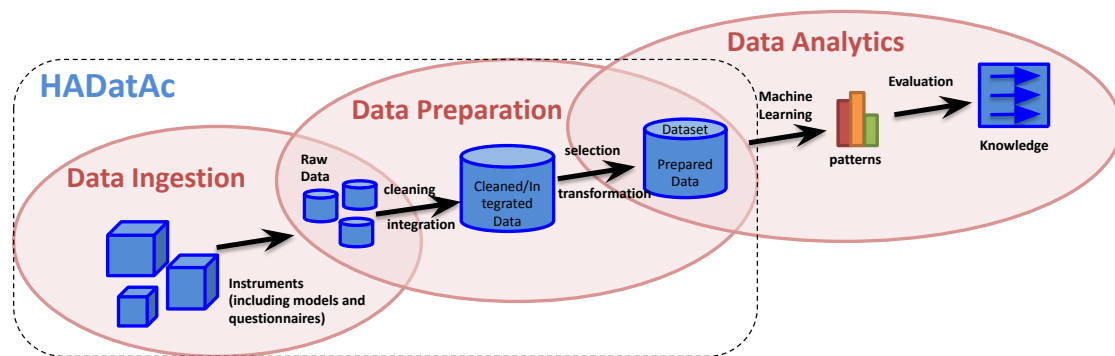
### 3.2. HADatAc: A Data Ingestion Solution for Data Preparation

A *data ingestion* is an activity within studies that, like data preparation, also manipulates the content of data files. However, in contrast with data preparation that focuses on dataset generation, data ingestion focuses on persisting study content in data stores (i.e. databases, search engines, graphs, etc.) to support data analysis activities [7, 8, 9], as well as data harmonization activities across studies [10, 11]. A data ingestion activity, by allowing each data value to be retrieved from a single source, mitigates potential time-consuming tasks such as handling distinct data file formats, unstructured data, data harmonization, missing values, provenance, and more.

<sup>1</sup><https://wwwn.cdc.gov/nchs/nhanes/search/>

We use the Human-Aware Data Acquisition Framework (HADatAc) [12] to support a data ingestion approach centered around the construction of a knowledge graph that comprehensively describes a collection of scientific studies. HADatAc employs an extensive set of concepts and associated terms used to represent studies' components (i.e. activities, subjects, samples, etc.), while logically connecting each data value to its related KG entities. Our data ingestion activity follows a systematic approach for acquiring knowledge from ontologies, semantic documents, and data files to build its knowledge graphs. With the use of data preparation criteria as described in Section 4, datasets ready for data analysis are automatically generated from data ingestion-generated knowledge graphs. The HADatAc data ingestion process also provides a systematic, normalized, and reusable way of organizing variables and variable data than is not an expected output of traditional extract-transform-load (ETL) tools, used in support of data preparation.

Figure 2 shows the role of a data ingestion process in the context of acquiring new knowledge from a data file to use the data in support of machine learning. The entire process starts with the acquisition of raw data from sources like physical instruments, questionnaires, and computer models. From the raw data and through the use of several operations, a knowledge graph is built with the use of data ingestion. Figure 2 assumes the use of data ingestion since its output is a knowledge graph that is later used to generate datasets from data preparation requests. HADatAc covers several aspects of data ingestion and data preparation as outlined above.



**Figure 2:** Data preparation in the context of data ingestion and data analytics.

### 3.3. Semantic Data Normalization

*Semantic data normalization* is the process of transforming a dataset based on a set of original variables of a given study into a corresponding set of normalized variables. The variable normalization process consists of two steps: the first step is a manual analysis of any available documentation and metadata of each original variable identifying its semantic variable properties; the second step is the process of encoding the semantic variable properties as annotation of the data. The list below shows essential semantic variables properties identified during the normalization process:

- type of the variable's entity of interest;

- type of the attribute that characterized the property of the variable's entity of interest;
- in the case of continuous variables, and optionally for some categorical values, the variable's unit;
- in the case of categorical variables, their codebooks;
- any spatial restriction related to the variable, E.g., a location where the variable was acquired;
- any temporal restriction related to the variable. E.g., when the variable was acquired.

The information above about each variable is described in the NHANES documentation through its data dictionaries, codebooks, file descriptions, interview descriptions, questionnaire descriptions, and many other auxiliary pieces of documentation.

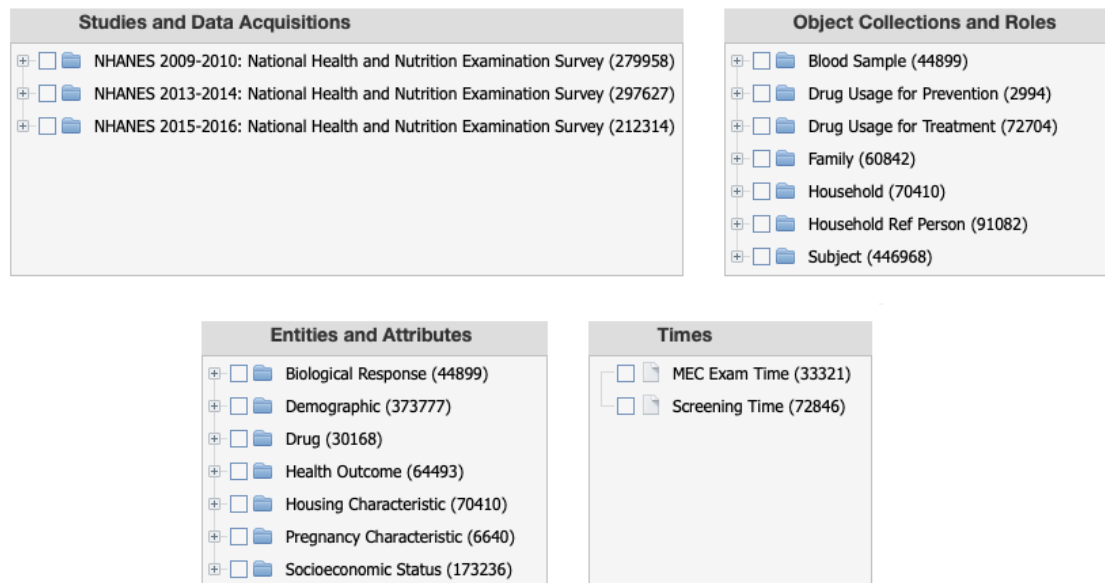
After the properties of variables are annotated, the normalization process is considered completed after the adjustment of original variables regarding *split-variables* and *merged-variables*. Two or more original variables are said to be *split-variables* if they share a common semantic variable, i.e., all the properties of their semantic variables are the same. NHANES data about insurance coverage from survey participants is an example of merged variables. The Insurance datasets contain about seven variables (depending on the cycle) to fully characterize insurance coverage. Each variable contains the participation status of the survey participant in one specific type of insurance (such as Medicaid, Medicare, Private insurance, etc.). However, we understand insurance coverage as not the value of a single variable but the combination of several variables insurance-related variables, all contributing to the insurance coverage attribute of the participant. For example, we can only infer if a person does not have insurance coverage if all variables contain the information of not being covered.

Merged-variable adjustment occurs when one original variable in NHANES cannot be represented by a single semantic variable i.e., the values of the original variable may require to be separated into two or more distinct semantic variables. One example of a merged variable is the Prescription Drug Usage dataset. This dataset's contents convey information on drugs being taken by survey participants to treat and/or prevent some diseases. The diseases and drugs are identified by codes (such as ICD10-CM). However, the dataset organizes the differentiation between treatment and prevention by modifying the original ICD-10CM codes to append a 'P' when the disease is being prevented (they remain unmodified for treatment).

### 3.4. Semantic Faceted Data Search

HADatAc provides a user interface where all the variables and studies are shown at once. Variable normalization as described in Section 3.3 is a key enabler for a uniform faceted data search for NHANES: semantic variable properties are indexed and treated as facets and used to facilitate variable selection; variable availability can be explored by executing data search. Therefore, through the use of a semantic faceted data search, one can browse and select available studies, study data files, hierarchies of entities, hierarchies of attributes, codebooks, time restrictions, space restrictions, and all of the above together.

Once a user of the faceted search selects the desirable values in each facet, the user can press the search button to verify if there is actual data matching the search request. If just a fraction of the requested data is returned or if no value is returned, that means that not all requested



**Figure 3:** HADatAc's semantic variable faceted-search using NHANES data.

data is available. The search process can be done gradually since one can further perform a search over the result of a previous search. By performing many searches, one is capable of probing NHANES content for many combinations of cycles and variables.

The machine's capability of processing the search request described above and showing what is available is the result of two features from HADatAc: the indexing of available context, and the fact that available content is data coming from normalized variables.

#### 4. Data Preparation Use-case: Demographic Determinants of Access to Care

The use-case used as a running data analytics example is based on survey-weighted logistic regression models and an equity-focused approach used to identify demographic and socioeconomic factors associated with patients' health care. A full description of the use case with the complete set of data analysis results can be found at [13]. We first focus on the survey-weighted logistic regression analysis. The task of selecting and preparing NHANES data for logistic regression analysis in the R code requires a deep understanding of the internals of NHANES.

To support this use-case, we produced several Semantic Data Dictionaries (SDDs) [14] (as well as additional metadata templates required by HADatAc) to describe NHANES datasets<sup>2</sup>. SDD is a specification that formalizes the assignment of a semantic representation of data, which can enable standardization and harmonization across diverse datasets. Specifically, SDDs allow the characterization of columns in tabular data using objects, attributes, and units defined

<sup>2</sup><https://github.com/tetherless-world/nhanes-hadatac>



Column	Attribute	attributeOf	Unit
RIDAGEYR	sio:SIO_001013	??participant	sio:SIO_000428
RIDRETH1	hhear:00609	??participant	

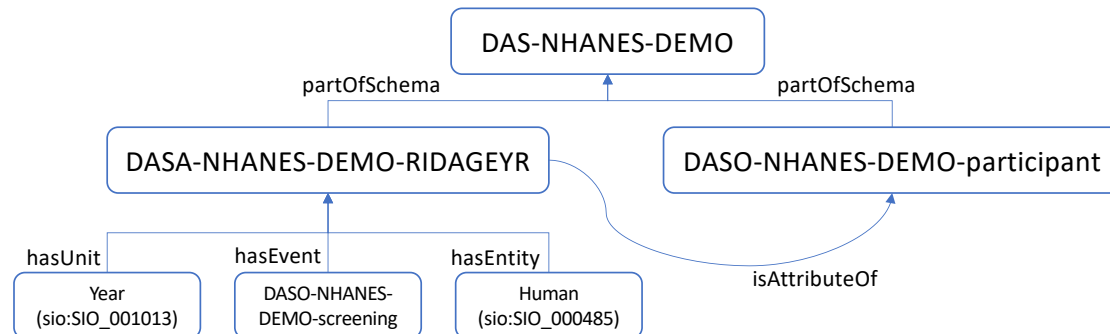
**Table 1**

Part of the Semantic Data Dictionary for the NHANES Demographics dataset, displaying the modeling of the *Age in years at screening* and *Race/Hispanic origin* variables.

in existing ontologies. In addition, for columns containing categorical values, SDDs support the representation of codebooks that resolve literal values to resources in ontologies.

The SDDs covered four survey cycles and about 40 datasets. Table 1 shows part of the SDD that models the RIDAGEYR and RIDETH1 variables (from the Demographics dataset). The RIDAGEYR variable is defined as the attribute `sio:SIO_001013` (“age”) of the survey participant (denoted by SDD’s object notation `??participant`), measured in the `sio:SIO_000428` (“year”) unit. The RIDETH1 variable is defined as the attribute `hhear:00609` (“Race or Ethnicity Combined”) of the survey participant. This variable does not have a unit because it is a categorical variable with an associated codebook.

We then used HADatAc to process all the produced SDDs and associated NHANES datasets to bootstrap a knowledge graph using the metadata. As an example, Figure 4 shows the “Age in years at screening” semantic variable RDF representation. Semantic variables are represented as RDF resources that compose an SDD (using the `partOfSchema` predicate). We represent semantic variable properties using specific predicates. In this case, we use `hasUnit`, `hasEvent`, and `hasEntity` to assert unit, time, and entity properties. We also show how a semantic variable is related to an object by using the `isAttributeOf` predicate. In this case, “Age in years at screening” is an attribute of the survey participant.



**Figure 4:** Graph representation of the “Age in years at screening” semantic variable. We highlight properties such as unit, time, and entity.

Once the NHANES Knowledge Graph was created within HADatAc, we utilized the semantic faceted data search to select the desired NHANES cycles. Then, we selected the semantic variables of interest, which included the race/ethnicity of survey participants, prescription drug usage (only antidiabetic drugs), drug classification, and immunization. Based on this selection, HADatAc generated a tabular dataset containing these semantic variables for analysis. More details about the NHANES knowledge graph creation in HADatAc can be found in [15].

## 5. Results

Utilizing the generated dataset, we examined equity of access to needed care with respect to race/ethnicity in the United States with data prepared using the proposed approach. To illustrate the flexibility of the approach, we looked at two problems: access of adult subjects with Type-2 Diabetes (T2D) to anti-diabetic drugs, and access to vaccines for hepatitis A (HAV), the hepatitis B (HBV), and the human papillomavirus (HPV). A logistic regression model was constructed as a function of race/ethnicity, age, gender, educational attainment, insurance type, poverty level, comorbidity severity based on CCI, and HbA1c condition. The model calculates the odds ratio (OR), which is the odds of drug/vaccination access and utility of a racial-ethnic subgroup divided by the odds of the same healthcare source access and utility in a reference non-Hispanic White group. The reader should consult [13] for full details of the analysis as well as a more extensive analysis of equity of access with respect to other social determinants of health.

Vaccine	HAV	HBV	HPV
NH Black	<b>1.09 (1.03, 1.16)*</b>	1.05 (1.00, 1.10)	0.99 (0.94, 1.04)
NH Asian	<b>1.17 (1.08, 1.26)*</b>	<b>1.10 (1.04, 1.17)*</b>	<i>0.94 (0.89, 0.99)*</i>
Hispanic	<b>1.13 (1.08, 1.18)*</b>	1.03 (0.97, 1.08)	0.99 (0.94, 1.05)

**Table 2**

Associations between population groups and vaccination by race/ethnicity in the U.S. with reference group: race/ethnicity = non-Hispanic White. **Bold\*** (*italics\**) indicates statistically significant increased (*decreased*) odds of vaccination than reference with  $p \leq 0.05$ .

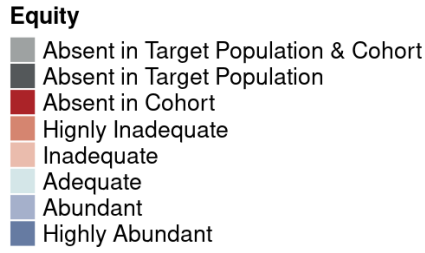
Table 2 shows the logistic regression analysis of vaccination coverage for minorities compared to the non-Hispanic White population. For example, minorities were more likely to be vaccinated against HAV/HBV while less likely for HPV. The statistics show that non-Hispanic Asian subjects experience an increase of 17% in the odds of getting the HAV vaccine, an increase of 10% in the odds of getting HBV, and a decrease of 6% in the odds of getting the HPV vaccine compared to non-Hispanic White subjects.

Additional equity analysis is performed using the approach developed in [13]. We use NHANES to estimate the target rate for each subgroup (e.g., percentage non-Hispanic Blacks with T2D in US) in the US and the rate of each subgroup that receives the treatment (e.g., percentage of non-Hispanic Blacks with T2D who received meglitinide). Then the log disparity metric and the associated statistical test are used to measure and visualize the disparity [16]. The results are color coded for easy interpretation as shown in Figure 5.

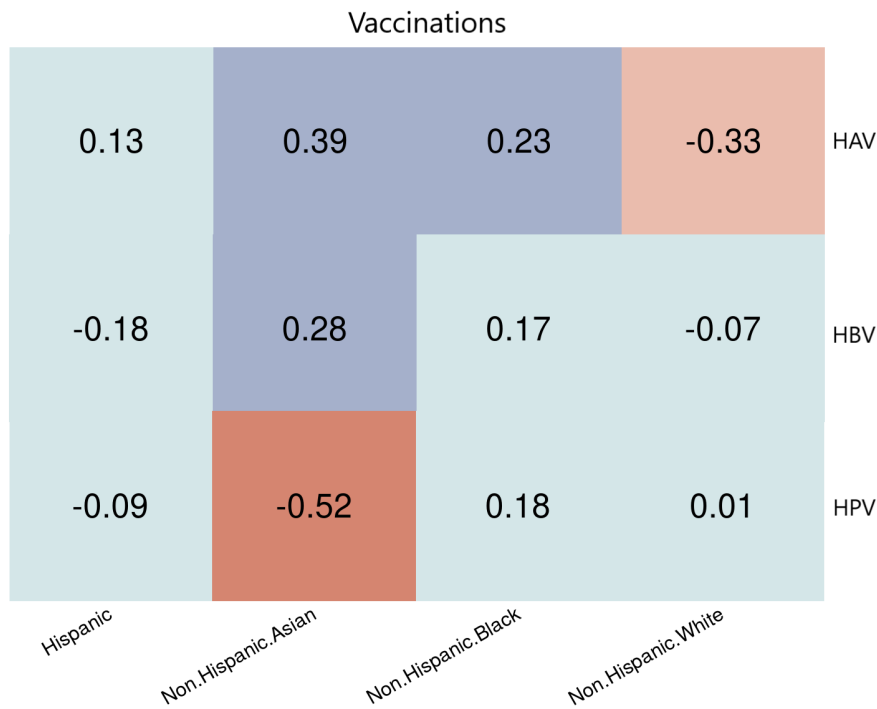
The general trend of the equity analysis of vaccinations is similar to the results shown in Table 2. For instance, Figure 6 shows that the non-Hispanic Asian population was more likely to receive HAV and HBV vaccinations but less likely for HPV while the non-Hispanic White population was less likely to get HAV vaccination.

Table 3 displays findings from the logistic regression model on the hyperglycemic medication utilization among U.S. patients with diabetes using non-Hispanic White as reference. For example, non-Hispanic Black patients had a 10% decrease in the odds ratio in biguanides in the odds ratio of prescribing rate compared to the non-Hispanic White group.

Figure 7 displays results from the log disparity equity metric for vaccinations. Though the



**Figure 5:** Description of colors used in health equity evaluation comparing subgroup cohort to target population using a color scheme in [16]. Red means inadequately represents and Blue means abundant with darkness indicating degree. Teal means adequate or no significant difference between the cohort and target populations.

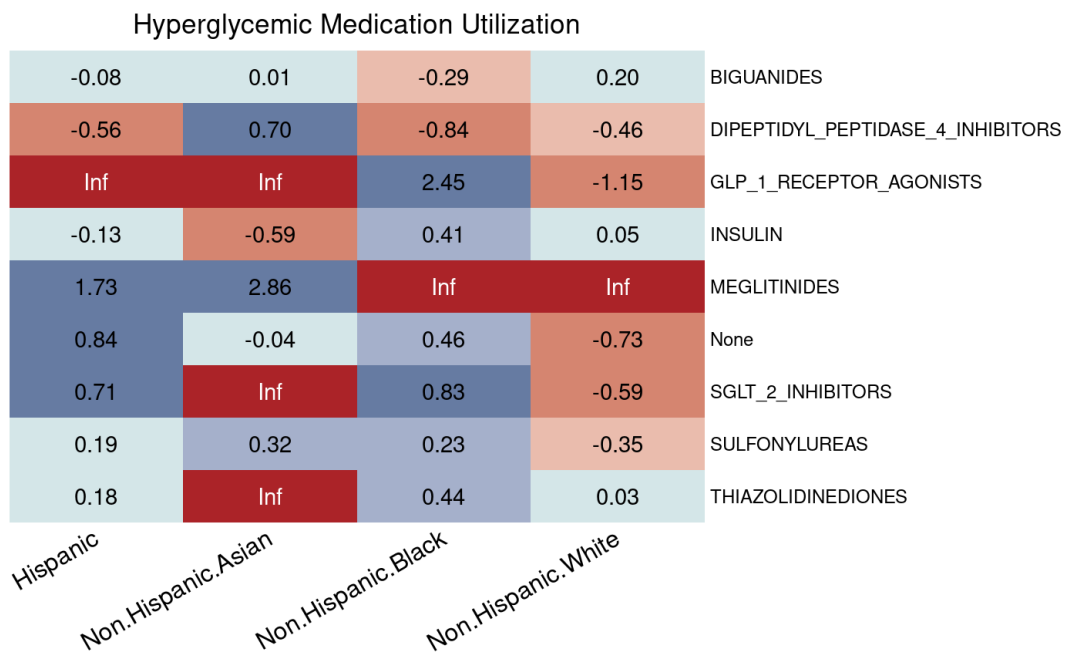


**Figure 6:** Equity evaluation heatmaps of racial/ethnic disparities on HAV, HBV, and HPV vaccinations.

logistic regression model shows that, non-Hispanic Asians had a similar prescribing rate for insulin compared to White populations, they had a disproportionate utilization based on the overall U.S. racial/ethnic distribution among the T2DM population.

Medication Class	NH Black	NH Asian	Hispanic
Biguanides	0.90 (0.81,0.99)*	1.02 (0.89,1.17)	1.00 (0.91,1.11)
DPP-4is	0.97 (0.92,1.03)	1.04 (0.96,1.13)	1.02 (0.95,1.10)
GLP-1RAs	0.99 (0.94,1.04)	0.94 (0.88,1.00)*	0.97 (0.93,1.01)
Insulin	1.03 (0.91,1.16)	0.89 (0.79,1.01)	0.95 (0.87,1.05)
Meglitinides	1.00 (0.99,1.02)	1.00 (0.99,1.02)	1.00 (0.99,1.01)
SGLT-2is	1.02 (0.98,1.06)	1.01 (0.96,1.06)	1.00 (0.97,1.03)
SUs	1.01 (0.91,1.11)	<b>1.14 (1.02,1.27)*</b>	0.98 (0.89,1.08)
TZDs	1.01 (0.97,1.06)	0.96 (0.93,1.00)*	1.02 (0.98,1.06)
None	0.99 (0.94,1.05)	1.00 (0.95,1.05)	1.03 (0.95,1.11)

**Table 3** Associations between population groups and diabetes drug use by race/ethnicity in the U.S. **Bold\*** (*italics\**) indicates statistically **increased** (*decreased*) chance of group receiving drug compared to non-Hispanic White with  $p \leq 0.05$ .



**Figure 7:** Disparities of hyperglycemic medication utilization by race/ethnicity.

## 6. Related Work

The problem of systematically accessing variables in a setting like NHANES is not new and has been somewhat explored in the literature. The NHANES Unified Dataset [17] was an effort to integrate several NHANES datasets in a unified way using an API. One of the contributions of this research was to support the examination beyond a few variables (usually constrained to a single survey cycle), using a method that can combine multiple variables across several NHANES survey cycles.

The earlier work in [18] performed a similar, but constrained to a subset of variables, method to preprocess NHANES' datasets to classify variables in three categories (environmental chemicals, health biomarkers, and questionnaire responses). In addition, this work performed several fine-grained variable normalizations based on the requirements of the study (such as the identification of cardiovascular disease and diabetes). The normalized variables across the 1999-2010 cycles were used in conjunction with their data analysis pipeline.

In terms of facilitating data analysis from R environments, a few packages for working with NHANES data exist. RNHANES [19] provides simple search capabilities for retrieving datasets and variables from specific NHANES components (such as lab results) and cycles. This package allows data to be downloaded and used directly. nhanesA [20] is a package that provides similar features, while also allowing access to some of the associated metadata such as codebooks. This metadata can be used in conjunction with data to resolve values, such as replacing codes with their natural language values in the codebook.

These previous efforts have provided solutions for specific parts of the challenge we are tackling. Our work expands these earlier accomplishments and starts to provide a more systematic method for formalizing semantic variables using semantic web technologies, as a basis for building data preparation pipelines that have an increased level of automation.

Going forward, the review in [21] suggests that "validation through independent replication will be critical in data-driven studies", in response to the problem that exposure measurement errors are common. Our approach aims to support the consistent use of variables that can lead to reproducibility.

## 7. Conclusion

We presented an approach for facilitating biomedical data preparation based on the notion of semantic variables. In this work, we abstract the meaning of a variable and represent it as a semantic variable, using semantic web technologies. A semantic variable is a human understanding of a property of an entity (and related aspects such as a unit of measurement) that can be shared among several variables. In a setting like NHANES, in which variables are being revised and evolved in each cycle, leading to new variables being created, semantic variables support data users in consistently finding and using relevant variables in their studies.

We have demonstrated this approach in a cross-sectional subgroup disparity analysis of 2013-2018 NHANES (3 cohorts) on U.S. adults for receipt of diabetes treatments and vaccines against Hepatitis A, Hepatitis B, and Human Papilloma. The results show that race/ethnicity is a determinant in access to certain diabetes medication classes and certain vaccines. While our current scenarios are based on NHANES's cycle aggregation (which is very specific to NHANES), the description of variable normalization and variable harmonization are generalizable to most studies.

In future work, we plan to continue applying this data preparation method to more analysis pipelines. Specifically, we are currently developing an experiment intended to reproduce existing studies based on NHANES and formally compare the obtained results in terms of the amount of time reduced in data preparation and error mitigation. For this, we are seeking to develop metrics that could quantify these aspects. This work could be directly generalized to the Health

Examination Survey Methods that are similar to NHANES in other nations including Brazil, Chile, Colombia, Mexico, England, and Scotland [22]. By removing data preparation barriers and supporting appropriate survey-weighted analyses of NHANES and related datasets, the proposed machine-assisted approach can potentially help conduct and accelerate many future public health studies leading to a better understanding of and improvements in human health.

## Acknowledgments

This work was partially supported by IBM Research AI Horizons Network and NIEHS' Human Health Exposure Analysis Resource (HHEAR), project number 5U2CES026555-05. Further thanks to Marcello P. Bax, Ph.D. who collaborated in the creation of Figure 2.

## References

- [1] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufmann, San Francisco, Calif, 1999.
- [2] A. Ruiz, *The 80/20 data science dilemma*, 2017. URL: <https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html>.
- [3] W. Raghupathi, V. Raghupathi, *Big data analytics in healthcare: promise and potential*, *Health Information Science and Systems* 2 (2014) 3.
- [4] Y. Wang, L. Kung, T. A. Byrd, *Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations*, *Technological Forecasting and Social Change* 126 (2018) 3–13.
- [5] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), *National Health and Nutrition Examination Survey Data, 2023*. URL: <https://www.cdc.gov/nchs/nhanes/>.
- [6] *NHANES Tutorials - Weighting Module*, 2023. URL: <https://wwwn.cdc.gov/nchs/nhanes/tutorials/weighting.aspx>.
- [7] C. H. Yu, *Exploratory Data Analysis*, in: *Psychology*, Oxford University Press, 2017.
- [8] T. G. Dietterich, *Machine-Learning Research*, *AI Magazine* 18 (1997) 97–97. Number: 4.
- [9] P. Domingos, *A few useful things to know about machine learning*, *Communications of the ACM* 55 (2012) 78–87.
- [10] I. Fortier, P. R. Burton, P. J. Robson, V. Ferretti, J. Little, F. L'Heureux, M. Deschênes, B. M. Knoppers, D. Doiron, J. C. Keers, P. Linksted, J. R. Harris, G. Lachance, C. Boileau, N. L. Pedersen, C. M. Hamilton, K. Hveem, M. J. Borugian, R. P. Gallagher, J. McLaughlin, L. Parker, J. D. Potter, J. Gallacher, R. Kaaks, B. Liu, T. Sprosen, A. Vilain, S. A. Atkinson, A. Rengifo, R. Morton, A. Metspalu, H. E. Wichmann, M. Tremblay, R. L. Chisholm, A. Garcia-Montero, H. Hillege, J.-E. Litton, L. J. Palmer, M. Perola, B. H. Wolffenbuttel, L. Peltonen, T. J. Hudson, *Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies*, *International Journal of Epidemiology* 39 (2010) 1383–1393.
- [11] J. Kalter, M. G. Sweegers, I. M. Verdonck-de Leeuw, J. Brug, L. M. Buffart, *Development and use of a flexible data harmonization platform to facilitate the harmonization of individual patient data for meta-analyses*, *BMC Research Notes* 12 (2019) 164.

- [12] P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, M. P. Bax, HADatAc: A Framework for Scientific Data Integration using Ontologies, in: The 17th International Semantic Web Conference Posters & Demonstrations Track, Monterey, CA, 2018.
- [13] M. Qi, H. Santos, P. Pinheiro, D. L. McGuinness, K. P. Bennett, Demographic and socioeconomic determinants of access to care: A subgroup disparity analysis using new equity-focused measurements, Under review (2023).
- [14] S. M. Rashid, J. P. McCusker, P. Pinheiro, M. P. Bax, H. Santos, J. A. Stingone, A. K. Das, D. L. McGuinness, The Semantic Data Dictionary – An Approach for Describing and Annotating Data, *Data Intelligence* 2 (2020) 443–486.
- [15] H. Santos, P. Pinheiro, D. L. McGuinness, Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the National Health and Nutrition Examination Surveys Use Case, in: 4th U.S. Semantic Technologies Symposium, Michigan State University, East Lansing, MI, 2022.
- [16] M. Qi, O. Cahan, M. A. Foreman, D. M. Gruen, A. K. Das, K. P. Bennett, Quantifying representativeness in randomized clinical trials using machine learning fairness metrics, *JAMIA Open* 4 (2021) ooab077.
- [17] C. J. Patel, N. Pho, M. McDuffie, J. Easton-Marks, C. Kothari, I. S. Kohane, P. Avillach, A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey, *Scientific Data* 3 (2016) 160096.
- [18] S. M. Bell, S. W. Edwards, Identification and Prioritization of Relationships between Environmental Stressors and Adverse Human Health Impacts, *Environmental Health Perspectives* 123 (2015) 1193–1199.
- [19] H. Susmann, RNHANES: Facilitates Analysis of CDC NHANES Data, 2016. URL: <https://CRAN.R-project.org/package=RNHANES>, R package version 1.1.0.
- [20] C. Endres, nhanesA: NHANES Data Retrieval, 2023. URL: <https://CRAN.R-project.org/package=nhanesA>, R package version 0.7.2.
- [21] A. K. Manrai, Y. Cui, P. R. Bushel, M. Hall, S. Karakitsios, C. J. Mattingly, M. Ritchie, C. Schmitt, D. A. Sarigiannis, D. C. Thomas, D. Wishart, D. M. Balshaw, C. J. Patel, Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health, *Annual Review of Public Health* 38 (2017) 279–294.
- [22] J. S. Mindell, A. Moody, A. I. Vecino-Ortiz, T. Alfaro, P. Frenz, S. Scholes, S. A. Gonzalez, P. Margozzini, C. de Oliveira, L. M. Sanchez Romero, A. Alvarado, S. Cabrera, O. L. Sarmiento, C. A. Triana, S. Barquera, Comparison of Health Examination Survey Methods in Brazil, Chile, Colombia, Mexico, England, Scotland, and the United States, *American Journal of Epidemiology* 186 (2017) 648–658.