


RESEARCH

Open Access



# An ontology-based approach for harmonization and cross-cohort query of Alzheimer's disease data resources

Xubing Hao<sup>1†</sup>, Xiaojin Li<sup>2,3†</sup>, Guo-Qiang Zhang<sup>1,2,3</sup>, Cui Tao<sup>1</sup>, Paul E. Schulz<sup>2</sup>, The Alzheimer's Disease Neuroimaging Initiative and Licong Cui<sup>1,3\*</sup> 

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM)  
Virtual. 9-12-December 2021. <https://ieeebibm.org/BIBM2021/>

## Abstract

**Background** In the United States, the National Alzheimer's Coordinating Center (NACC) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) are two major data sharing resources for Alzheimer's Disease (AD) research. NACC and ADNI strive to make their data more FAIR (findable, interoperable, accessible and reusable) for the broader research community. However, there is limited work harmonizing and supporting cross-cohort interoperability of the two resources.

**Method** In this paper, we leverage an ontology-based approach to harmonize data elements in the two resources and develop a web-based query system to search patient cohorts across the two resources. We first mapped data elements across NACC and ADNI, and performed value harmonization for the mapped data elements with inconsistent permissible values. Then we built an Alzheimer's Disease Data Element Ontology (ADEO) to model the mapped data elements in NACC and ADNI. We further developed a prototype cross-cohort query system to search patient cohorts across NACC and ADNI.

**Results** After manual review, we found 172 mappings between NACC and ADNI. These 172 mappings were further used to construct common concepts in ADEO. Our data element mapping and harmonization resulted in five files storing common concepts, variables in NACC and ADNI, mappings between variables and common concepts, permissible values of categorical type data elements, and coding inconsistency harmonization, respectively. Our cross-cohort query system consists of three core architectural elements: a web-based interface, an advanced query engine, and a backend MongoDB database.

**Conclusions** In this work, ADEO has been specifically designed to facilitate data harmonization and cross-cohort query of NACC and ADNI data resources. Although our prototype cross-cohort query system was developed for exploring NACC and ADNI, its backend and frontend framework has been designed and implemented to be generally applicable to other domains for querying patient cohorts from multiple heterogeneous data sources.

<sup>†</sup>Xubing Hao and Xiaojin Li contributed equally to this work.

\*Correspondence:

Licong Cui  
[licong.cui@uth.tmc.edu](mailto:licong.cui@uth.tmc.edu)

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

**Keywords** Alzheimer's disease, Ontology, Data element mapping, Data harmonization, Cross-cohort query

## Background

Alzheimer's disease (AD) is a neurodegenerative disease affecting over 5.5 million Americans with significant economic and social impacts [1]. It has received a great deal of attention from biomedical research community. In the United States, two major data sharing resources for AD research are the National Alzheimer's Coordinating Center (NACC) [2] and the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3]. NACC and ADNI strive to make their data more findable, interoperable, accessible and reusable (FAIR) for the broader research community [4, 5]. They provide valuable resources for discoveries such as AD biomarkers [6], disease progression [7], and cross-cohort model validation [8, 9].

Cross-cohort comparisons allow the research findings obtained from one study to be tested and replicated by another study [10]. Harmonization of heterogeneous data from different resources is essential to enable such cross-cohort comparisons. In addition, harmonizing and integrating data from multiple sources increase the statistical power that an individual dataset would provide. There have been various efforts for cross-cohort data harmonization and integration [11–13] and cross-cohort data exploration [14–17] in different disease domains. For instance, Cui et al. have performed data harmonization on heterogeneous datasets in the the National Sleep Research Resources (NSRR) and developed a cross-cohort search interface called X-search for querying patient cohorts from multiple datasets in NSRR [17].

However, there has been limited work harmonizing AD-related datasets from different resources (such as NACC and ADNI) and supporting cross-cohort data exploration from these AD-related data resources. In a recent preprint [18], Salimi et al. manually harmonized 1,196 variables across 20 AD cohort datasets including NACC and ADNI, and presented a web-based platform called ADataViewer to explore the cohort datasets. However, their data harmonization only mapped a sub-collection of variables (or data elements) from each dataset, and did not harmonize inconsistent codes (or permissible values); in addition, the web-based data exploration was provided in a summarized manner (e.g., pre-computed variable distribution plots), and only supported variable-level queries (e.g., number of patients with variable “Mini-Mental State Examination (MMSE)” captured) rather than value-level queries (e.g., number of patients with MMSE below 12).

In this work, we focused on mapping data elements in NACC and ADNI, and harmonizing inconsistent codes

such as different values ranges or different units of measurement for mapped data elements. For data exploration, we developed a prototype cross-cohort, value-level query system for searching patient cohorts across the two resources. In particular, we created an Alzheimer's Disease Data Element Ontology (ADEO), which not only models harmonized data elements between NACC and ADNI, but also serves as the knowledge source to drive the cross-cohort query system.

## NACC and ADNI

NACC has been collecting data from Alzheimer's Disease Research Centers (ADRCs) funded by the National Institute on Aging since 2005 [2]. The goal is to translate research advances into improved diagnosis and care for AD patients, and find ways to treat and possibly prevent Alzheimer's disease and related dementias (ADRD). It includes participants with cognitive status ranging from normal cognition, to mild cognitive impairment (MCI), and demented. In each participant's annual Uniform Data Set (UDS) visit, standardized clinical data consisting of 16 forms are collected, covering topics including subject demographics, neurological examination findings, and diagnosis. Subsets of UDS subjects have imaging data, and cerebrospinal fluid (CSF) biomarker data, genetic data, and autopsy data.

ADNI began in 2004, and its goal is to detect AD at the earliest possible stage, identify ways to track the disease progression with biomarkers, and support advances in AD intervention, prevention, and treatment [3]. The participants include AD patients, mild cognitive impairment subjects, and elderly controls. The data types that ADNI collects include clinical (demographics, physical examinations, and cognitive assessment data), genetic, MRI image, PET image, and biospecimen (blood, urine, and CSF). ADNI data has been used by AD researchers around the world resulting in over 2,100 publications [19].

## Related work on data element mapping and harmonization

Data element harmonization across cohorts has been an active research area for improving the interoperability across different datasets. For example, Pathak et al. mapped phenotype data elements from five Electronic Medical Records and Genomics (eMERGE) Network sites to the National Cancer Institute (NCI) Cancer Data Standards Registry (caDSR) [20]. Liu et al. mapped data

elements in the Dental Information Model to the caDSR common data elements [21]. Tao et al. developed a web-based interactive mapping interface for users to find mappings of variables from the North American Association of Central Cancer Registries (NAACCR) data dictionary to the National Cancer Institute (NCI) Thesaurus (NCIt) [22]. In a preprint [18], Salimi et al. created a variable mapping catalog that harmonized 1,196 unique variables in 20 AD cohort datasets through meticulous manual curation.

### Related work on cross-cohort data exploration

There have been query systems developed for searching patient cohorts across different data sources [14–17]. For example, Weber et al. [14] developed the Shared Health Research Information Network (SHRINE) based on i2b2 [23] to query patient cohorts from multiple data sources. Zhang et al. [15] designed and implemented VISual AGgregator and Explorer (VISAGE) for querying across disparate databases in clinical research. Bache et al. [16] defined and validated an adaptable architecture for identifying patient cohorts from multiple heterogeneous data sources. Cui et al. [17] developed an open access interface for querying patient cohorts across nine datasets in NSRR.

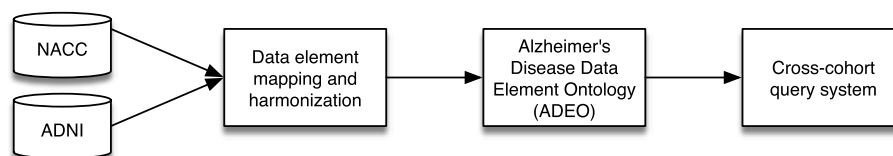
### Methods

Figure 1 shows the overall workflow of our ontology-based approach. After acquiring data from NACC and ADNI, we performed mapping and harmonization of data elements between the two resources. Then we constructed ADEO to formalize the harmonized data elements, which were further leveraged to develop the web-based cross-cohort query system.

### Datasets

We requested study data stored in the format of comma-separated values (CSV) from NACC and ADNI. Each patient may have multiple visits recorded in the study data. NACC stores their study data in a single file, while ADNI separates their study data in different tables. In addition, we downloaded structured data dictionaries (in CSV and PDF) that semantically define the scope and characteristics of data elements (or variables) in the study data. For NACC, the data dictionaries of Uniform Data Set (UDS), Neuropathology (NP) data set, and genetic data are stored in CSV, while the data dictionaries of the imaging and biomarker data sets are stored in PDF. We converted PDF data dictionaries to plain text files using the *pdftotext* utility (part of the *Xpdf* software suite [24]). Then we parsed the plain text files and extracted attributes of data elements and stored them in CSV.

Essential attributes of data elements in NACC's data dictionary include variable name, form, short descriptor, data type and allowable codes. Essential attributes of data elements in ADNI's data dictionary include FLDNAME, TEXT, CRFNAME, TYPE and CODE. Variable name in NACC and FLDNAME in ADNI serve as the column name in the study data. Form in NACC and CRFNAME in ADNI are the broader category of each data element. Short descriptor in NACC and TEXT in ADNI store the full name of the data element and are displayed to users in the query interface. Data type in NACC and TYPE in ADNI demonstrate the data type of the data element such as numerical or categorical. Allowable codes in NACC and CODE in ADNI store the permissible values of the categorical type data element or the range of the numerical type data element. Table 1 shows three examples of data elements in NACC's data dictionary. Table 2



**Fig. 1** Workflow of our ontology-based approach

**Table 1** Examples of data elements in NACC

VariableName	Form	VariableType	ShortDescriptor	DataType	AllowableCodes
EDUC	a1	Original UDS question	Years of education	Numeric cross-sectional	0 - 36; 99 = Unknown
NORMCOG	d1	Original UDS question	Normal cognition and behavior	Numeric longitudinal	0 = No; 1 = Yes
NACCVASC	np	NACC Derived Variable	Ischemic, hemorrhagic, or vascular pathology present	Numeric cross-sectional	0 = No; 1 = One or more vascular pathology; 9 = Unknown

**Table 2** Examples of data elements in ADNI

Phase	FLDNAME	TBLNAME	CRFNAME	TEXT	TYPE	LENGTH	CODE	UNITS
ADNI1	GDAFRAID	GDSCALE	Geriatric Depression Scale	6. Are you afraid that something bad is going to happen to you?	N	1	1=Yes(1); 0=No(0)	
ADNI1	ST127SV	UCSFFRESFR	Longitudinal FreeSurfer	Volume (WM Parcellation) of Third-Ventricle	N	8		mm3
ADNIGO	PTDOBMM	PTDEMOG	Participant Demographics	2a. Participant Month of Birth	N	2	1..12	

shows three examples of data elements in ADNI's data dictionary.

### Data element mapping and harmonization

We conducted manual data element mapping to identify overlaps between NACC and ADNI. We first did some pre-processing on the data dictionaries. In NACC's data dictionaries, the attribute form of the data element is stored using short names (e.g., "a1", "a2", "b6"). We converted the short names to their full names provided on NACC's website for data element mapping. For example, form "a1" has a full name of "Subject Demographics". In ADNI's data dictionary, for some imaging-related data elements, words in phrases describing brain regions are concatenated without spaces. We pre-processed such cases and added spaces between the concatenated words. For example, data element "Cortical Thickness Average of LeftIsthmusCingulate" in ADNI after pre-processing is "Cortical Thickness Average of Left Isthmus Cingulate". The manual data element mapping were initially performed by XH (expertise in biomedical informatics) and further reviewed by LC and CT (with expertise in biomedical data science and ontology) as well as PES (clinical expert in AD). Disagreements were resolved through discussion.

A unique challenge in exploring data in multiple heterogeneous data sources is to address the coding inconsistency issue, which involves the detection and harmonization of inconsistencies among the disparate permissible values (or value domains) for the same data element [17]. Such inconsistencies occur frequently for numerical and categorical variables. For example, the permissible values of concept "Banked postmortem CSF" is inconsistent between NACC and ADNI. Table 3 shows how we handle the harmonization of the inconsistency for this data element. For the inconsistency of numerical concepts such as "Years of education" has a range of "0 - 36" in NACC and "0 - 20" in ADNI. We always kept the wider range, which is "0 - 36" in this case. To ensure accurate results for data exploration, we manually harmonized such heterogeneity.

We further pre-processed the study data to address other types of inconsistencies before importing them to the database of the cross-cohort query system. For

**Table 3** Harmonizing coding inconsistencies for data element "Banked postmortem CSF"

Source	Value	Name	Value - Harmonized	Name - Harmonized
ADNI	1	Yes	1	Yes
ADNI	2	No	0	No
NACC	0	No	0	No
NACC	1	Yes	1	Yes
NACC	9	Missing/unknown	9	Missing/unknown
NACC	-4	Not available	-4	Not available

instance, concept "Segmented right hippocampus volume (cc)" in NACC uses unit "cc" but ADNI uses unit "mm3". Since 1 cc = 1000 mm3, all data points of this data element in ADNI were divided by 1000. Also, for data element "Average number of packs smoked per day", in NACC it has a categorical data type with permissible values "0 = No reported cigarette use; 1 = 1 cigarette to less than  $\frac{1}{2}$  pack; 2 =  $\frac{1}{2}$  pack to less than 1 pack; 3 = 1 pack to  $1\frac{1}{2}$  packs; 4 =  $1\frac{1}{2}$  packs to 2 packs; 5 = More than two packs; 8 = Not applicable; 9 = Unknown; - 4 = Not available", while in ADNI it has a numerical data type with a range of 0 to 10. We grouped the numeric values of this data element in ADNI according to NACC's categories and stored the categorical results as a new column in the study data.

### ADEO construction

We used the Protege OWL editor (Version 5.5.0) [25] and Owlready2 [26] for building ADEO. Owlready2 is a Python package for manipulating ontologies in the format of Web Ontology Language (OWL). It not only provides the functionality of loading, modifying, and saving ontologies, but also supports reasoning via Hermit [26]. Owlready2 allows a transparent access to OWL ontologies. We used the "types", which is a Python module to create classes and subclasses dynamically. Subclasses can be created by inheriting an ontology class. Multiple inheritance is also supported.

We constructed ADEO based on the mapped data elements that we identified from the manual mapping. We defined ADEO classes (or common concepts) to

represent the mapped data elements. We created data property classes “hasCategory” and “hasRange” under class “DataProperty” to model permissible values for categorical concepts and specify the range for numerical concepts, respectively. And we used “some” restriction that Owlready2 provided when defining classes. Permissible values and ranges were leveraged to define the classes. We organized the classes into different sub-hierarchies. Figure 2(a) shows the sub-hierarchies of ADEO and the classes under sub-hierarchy “*Demographics*”, including “*Marital status*”, “*Month of birth*”, “*Year of birth*”, “*Primary language*”, “*Type of Residence*”, and “*Years of education*”. Figure 2(b) shows the range defined for class “*Years of education*”. Figure 2(c) shows the categories of permissible values defined for class “*Primary language*”.

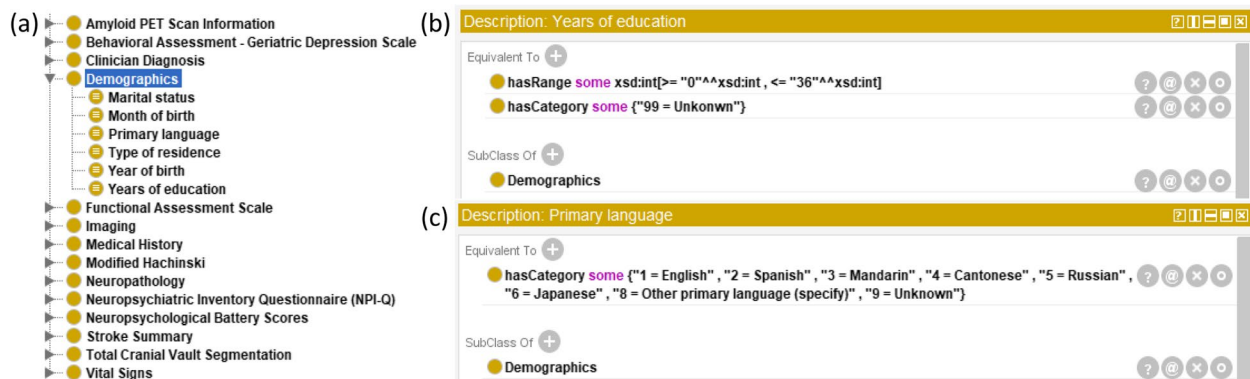
In addition, the ADEO classes (or common concepts) will serve as the core query terms in the query system for users to browse or search. Since an ADEO class may correspond to different variable names in NACC and ADNI, there is a need for mapping data elements from NACC and ADNI to the common concepts. Take the common concept “*Large arterial infarcts present*” as an example,

its mapped data element in NACC is “*Large arterial infarcts present*” and its mapped data element in ADNI is “*Are one or more large artery cerebral infarcts present?*”.

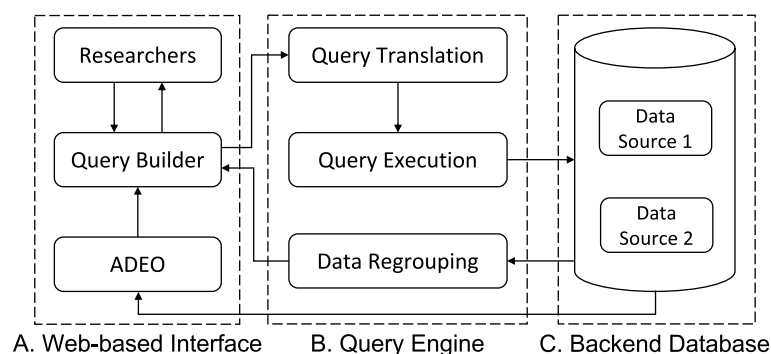
### Cross-cohort query system development

Figure 3 shows the general architecture design of our cross-cohort query system, consisting of 3 core architectural elements:

- 1 a web-based interface (see Fig. 3.A), called query builder, which is a powerful and intuitive interface that has been designed and developed to enable researchers to quickly find the right common concepts and perform an exploratory cross-cohort query;
- 2 an advanced query engine for searching records across different datasets (see Fig. 3.B); such a query engine translates the user queries built from the web-based interface into executable database query languages, and consists of three modules: a query translation module, a query execution module, and data regrouping module; and



**Fig. 2** Examples of ADEO classes



**Fig. 3** The system architecture of our cross-cohort query system



- 3 a backend MongoDB database for storing ADEO and data from different sources (see Fig. 3.C).

We built the query builder interface using React [27], an open-source JavaScript library that is used for building user interfaces specifically for web-based applications. The cross-cohort query engine was implemented using Ruby on Rails [28], which is an agile web development framework.

The query builder consists of four areas, which correspond to four steps to perform cross-cohort queries as follows: 1) dataset selection area, where researchers can select dataset(s) of interest; 2) query term selection area, where users can find and select query terms (i.e., common concepts) and add them to the query construction area; 3) query construction area, where query criteria can be specified for each query term; and 4) query results display area, where the patient counts retrieved from selected dataset(s) satisfying the query criteria are returned to the user.

In this work, the dataset selection area is filled with the names of the two datasets: ADNI and NACC. In the query term selection area, two modes are provided to find query terms of interest: browsing and searching. The browsing mode displays query terms in a hierarchical order, allowing users to explore all accessible query terms level by level. For users with background knowledge, the search mode provides the functionality of directly searching for query terms of interest. Based on the query terms selected by a user, the query builder automatically generates visual query widgets using a dynamic approach, such as generating widgets with checkboxes for specifying permissible values when selecting a categorical term, while creating widgets with sliders for specifying a range of values when selecting a numerical term. The query construction area is designed to be as close to natural language as possible to make the query logic clear and readable to the users. The query results display area is driven by the query criteria specified in the query construction area.

As users add query terms to define queries, the interface creates an array of key-value pairs in JSON objects representing the current state of the user interface and query criteria. Such objects themselves do not contain query language, but instead, contain the query terms as well as additional metadata that describes the query. The query translation module automatically translates the JSON objects into actual MongoDB statements to query the backend database.

The translation relies on the specified query terms and values, as well as the mappings from the data elements in NACC and ADNI to the common concepts representing the query terms. The query statements for disparate data

sources are distinct since these data sources have different variable information mapping to a common concept. We have two mapping files that are specifically designed for query term mapping and query value mapping. For example, the query term “Difficulty or need help with: Playing a game of skill” is mapped to the variable “GAMES” in NACC and variable “FAQGAME” in ADNI. The value of “Requires assistance” of this query term is mapped to “2” in NACC and “4” in ADNI.

For each type of query terms, a general template is predefined and used for dynamically generating the actual MongoDB statement for query translation. For instance, the template for querying a numerical query term with a specified range [min, max] is defined as:

```
db.records_collection.
distinct(<mapped_patient_identifier>,
{"dataset":<mapped_dataset>,
<mapped_variable_name>:{"$gte":min,
"$lte":max}})
```

where <mapped\_patient\_identifier> represents the variable name of the unique patient identifier in the mapped dataset, <mapped\_dataset> represents the name of the mapped dataset, and <mapped\_variable\_name> is the variable name to which the query term is mapped in a dataset. For instance, to query the number of patients with years of education between 5 and 15 years in NACC can be translated to:

```
db.records_collection.
distinct("NACCID", {"dataset":"NACC",
"EDUC":{"$gte":5, "$lte":15}})
```

The query execution module sends the translated MongoDB statements to the backend database to execute the query. The MongoDB backend returns numeric counts of eligible patients satisfying the query criteria. The data regrouping module summarizes and reorganizes the query results to facilitate the user interface display.

## Results

### Data element mapping and harmonization

The data dictionaries that we downloaded contain 1,195 NACC data elements and 13,918 ADNI data elements. After manual review, we found 172 mappings between NACC and ADNI. Among these mappings, 23 of them required numerical harmonization, 26 of them required a categorical harmonization, and 7 of them required a unit harmonization. These 172 mappings were further used to construct common concepts in ADEO. The core concepts capture information regarding Demographics (e.g., Year of birth, Marital status), Medical History (e.g., Average number of packs smoked per day), vital signs (e.g., Seated

Blood Pressure: Systolic), Behavioral Assessment - Geriatric Depression Scale (e.g., Do you feel full of energy?), Modified Hachinski (e.g., Somatic Complaints, Focal Neurologic Signs), Neuropathology (e.g., Banked frozen brain, PRNP codon 129, FTLTD-tau subtype - Pick's (PiD)), Neuropsychological Battery Scores (e.g., Multilingual Naming Test (MINT) - Semantic cues: Number given), Neuropsychiatric Inventory Questionnaire (NPI-Q) (e.g., Anxiety severity, Delusions severity), Functional Assessment Scale (e.g., In the past four weeks, did the subject have any difficulty or need help with: Preparing a balanced meal), Imaging (e.g., Left lingual mean cortical thickness (mm)), Stroke Summary (e.g., Total brain white matter hyperintensity volume (cc)), Amyloid PET Scan Information (e.g., Amyloid Imaging radiotracer used), Total Cranial Vault Segmentation (e.g., Total intracranial volume (cc)), and Clinician Diagnosis (e.g., Normal cognition and behavior). Our data element mapping and harmonization resulted in five files storing common concepts, variables in NACC and ADNI, mappings between variables and common concepts, permissible values of categorical type data elements, and coding inconsistency harmonization, respectively.

#### ADEO and cross-cohort query system

The current version of ADEO contains 186 classes. Figure 4 shows the overall structure of ADEO with the following four sub-hierarchies expanded: “Demographics”, “Modified Hachinski”, “Neuropsychiatric Inventory Questionnaire (NPI-Q)”, and “Vital Signs”.

The query builder interface with the four areas annotated is shown in Fig. 5. In the dataset selection area (Fig. 5.A), both datasets are chosen. All the core concepts from ADEO are listed in the query term selection area (Fig. 5.B), and researchers can enter text in search mode to obtain the query terms of interest. The query construction area (Fig. 5.C) contains two query widgets for “Marital status” (with checkboxes) and “Years of education” (with a slider bar), with specified query criteria: married, and between 5 and 15 years of education. The query results display area (Fig. 5.D) shows the number of patients satisfying the query criteria in each dataset, as well as the total number of patients satisfying the query criteria across all the selected datasets.

To visualize the information for query terms in ADEO, we designed and implemented an interface to display the metadata of each query term. The example of different query term types is shown in Fig. 6, including a categorical query term “Marital status” (Fig. 6(a)) and a numerical query term “Years of education” (Fig. 6(b)). For each type, we generated different interactive visualizations for the distributions of corresponding query term's values in each dataset, with bar charts for the category type and box plots for the numerical type.

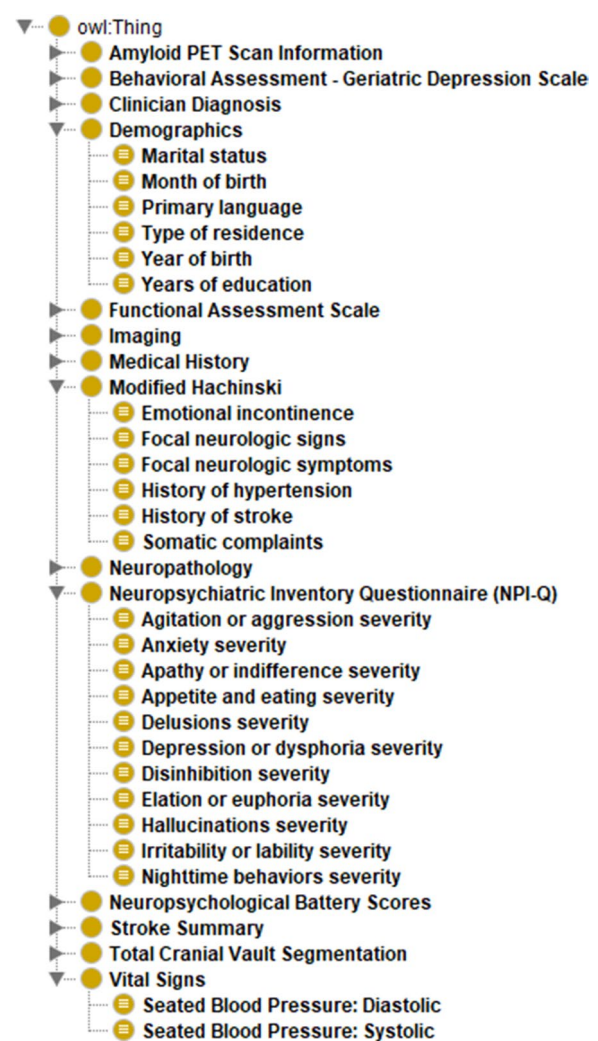


Fig. 4 Overall structure of ADEO

#### Discussion

Although our prototype cross-cohort query system was developed for exploring NACC and ADNI, its backend and frontend framework (Figs. 3 and 5) has been designed and implemented to be generally applicable to other domains for exploring patient cohorts from multiple heterogeneous data sources.

#### Comparison with related work

In previous studies regarding data element mapping and harmonization, the mapped data elements served different roles in downstream research. For example, Salimi et al. leveraged the harmonized variables to build the interactive ADataViewer to semantically and statistically facilitate the scientific community to explore multiple AD cohort datasets [18]. Tao et al. provided an interface for users to find mappings from data dictionaries to ontologies [22]. Different from these works, we leveraged our

The screenshot displays a query builder interface with four main sections: A. Dataset Selection, B. Query Term Selection, C. Query Construction, and D. Query Results Display.

**A. Dataset Selection:** Shows two datasets selected: ADNI and NACC.

**B. Query Term Selection:** A list of query terms with their counts:

Query Term	Count
Demographics	6
Medical History	2
Vital Signs	2
Behavioral Assessment - Geriatric Depression Scale	16
Modified Hachinski	6
Neuropathology	31
Neuropsychological Battery Scores	17
Neuropsychiatric Inventory Questionnaire (NPI-Q)	11
Functional Assessment Scale	10
Imaging	68
Stroke Summary	1
Amyloid PET Scan Information	1
Total Cranial Vault Segmentation	1
Clinician Diagnosis	1

**C. Query Construction:** Shows the query logic. Under "Marital status", "Married" is selected. Under "Years of education", a slider bar is set from 5 to 15.

**D. Query Results Display:** Shows the results of the query. The total number of patients is 11470. The results are displayed in a table:

Dataset	Number of Patients	Query Time(s)
ADNI	951	0.025919
NACC	10519	0.450813

**Fig. 5** Screenshot of the query builder interface. This example queries the number of married patients with 5 to 15 years of education

harmonized data elements to build a web-based query system for users to search patient cohorts across two widely used AD data resources. Our data element mapping was purely through manual creation to ensure the accuracy of mapping results. We not only mapped variables between data resources but also harmonized their permissible values for our query system purpose.

For the cross-cohort query system, differing from previous work X-search [17] that uses MySQL as the backend database, we choose MongoDB as the backend database in this work considering its good query performance with the large-scale dataset and flexible data models. While X-search stores different data sources in separate MySQL tables, this new system stores all data sources in one collection by leveraging the flexible data model in MongoDB. Such design reduces the complexity of data integration and makes it easier to import data from different sources into the database, and it avoids the need to perform queries across multiple tables. In X-search, the data need to be preprocessed before importing into the MySQL database to handle the coding inconsistency. In this work, we apply a different strategy of creating an additional query value mapping file and mapping inconsistent variable values in real-time in the query translation module; therefore, we do not need to map each inconsistent variable value for all records before importing, thus reducing the tasks and time to build the system.

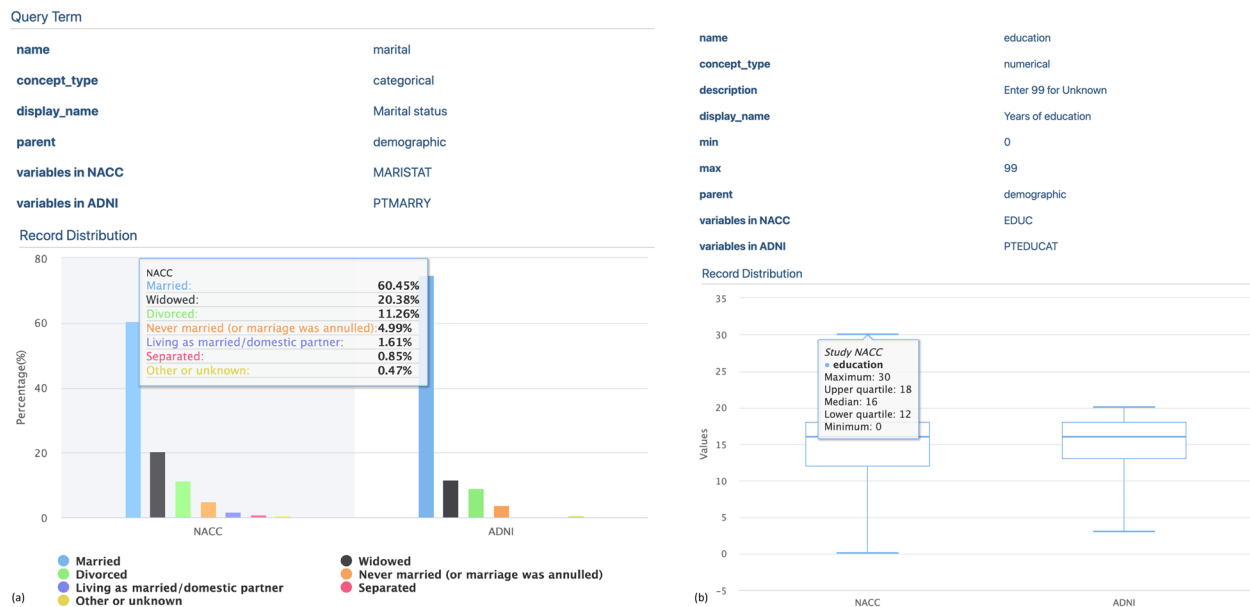
In addition, although there have been efforts to develop AD-related ontologies such as the Common Alzheimer's Disease Research Ontology (CADRO) [29] and the

Alzheimer's disease ontology (ADO) [30], these existing AD-related ontologies were not designed (and thus are not sufficient) to be directly usable for supporting harmonizing and querying data elements across different AD data resources. In this work, ADEO has been specifically designed to facilitate such data harmonization and cross-cohort query among different resources.

#### Limitations and future work

One limitation of this work is that we did not perform a usability evaluation for the prototype cross-cohort query system. We plan to invite AD researchers to evaluate our prototype query system and provide feedback for us to enhance the system's functionality and usability. Another limitation of our work is that there may exist missed mappings between NACC and ADNI, even though manual curation was performed. Comparing to Salimi et al.'s work that identified 170 mapped data elements between ADNI and NACC [18], our work identified 172 mapped data elements. Among these mappings, there are 72 identified by both works, 98 identified by Salimi et al.'s work but not ours, and 100 identified by our work but not Salimi et al.'s. We will further incorporate and harmonize those mappings identified by Salimi et al.'s work but not ours. Automated mapping techniques along with manual validation may also help identify further mappings between the two data resources. Additional future work includes enriching ADEO with synonyms leveraging other AD-related ontologies and the Unified Medical Language System.





**Fig. 6** Screenshots of the query term interface: **a** Marital status; and **b** Years of education

## Conclusions

In this paper, we presented an ontology-based approach to map and harmonize data elements in NACC and ADNI, two widely used data resources for AD research. We also developed a prototype cross-cohort query system to search patient cohort counts across the two data resources. Our prototype query system is generally applicable to other domains for supporting cross-cohort queries.

## Abbreviations

NACC	National Alzheimer's Coordinating Center
ADNI	Alzheimer's Disease Neuroimaging Initiative
AD	Alzheimer's Disease
NSRR	National Sleep Research Resources
MMSE	Mini-Mental State Examination
UDS	Uniform Data Set
CSF	cerebrospinal fluid
NCI	National Cancer Institute
caDSR	Data Standards Registry
eMERGE	Electronic Medical Records and Genomics
NP	Neuropathology
OWL	OntologyWebLanguage
CADRO	Common Alzheimer's Disease Research Ontology
ADO	Alzheimer's disease ontology

## Acknowledgements

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADRCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50

AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG03514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Data collection and sharing for part of this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health ([www.fnih.org](http://www.fnih.org)). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Part of the data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<https://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [https://adni.loni.usc.edu/wp-content/uploads/how\\_to\\_apply/ADNI\\_Acknowledgement\\_List.pdf](https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf).

## About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making Volume 23 Supplement 1, 2023: Quality Assurance and Enrichment of Biological and Biomedical Ontologies and Terminologies. The full contents of the supplement are available online at <https://bmcmmedinformdecismak.biomedcentral.com/articles/supplements/volume-23-supplement-1>.

## Authors' contributions

LC conceptualized and designed this study. XH performed the data element mapping and harmonization, and constructed ADEO. LC, CT, and PES conducted review of mapping results and ADEO. XL and GQZ designed and developed the cross-cohort query system. ADNI prepared part of the data used in this work. LC, XH and XL wrote the manuscript. All the authors read and approved the final manuscript.

## Funding

This work was supported by the National Institutes of Health (NIH) through grants R21AG068994, R01LM013335, R01NS116287, and RF1AG072799. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. Publication costs are funded by R01NS116287.

## Availability of data and materials

The results for mappings between NACC and ADNI as well as the ADEO ontology are available at [https://github.com/XubingHao/BMC2022\\_AD-Query](https://github.com/XubingHao/BMC2022_AD-Query).

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>2</sup>Department of Neurology, McGovern School of Medicine, The University of Texas Health Science Center at Houston, Houston, TX, USA. <sup>3</sup>Texas Institute for Restorative Neurotechnologies, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Received: 31 August 2022 Accepted: 26 July 2023

Published online: 04 August 2023

## References

- Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*. 2013;80(19):1778–83.
- Beekly DL, Ramos EM, van Belle G, Deitrich W, Clark AD, Jacka ME, et al. The national Alzheimer's coordinating center (NACC) database: an Alzheimer disease database. *Alzheimer Dis Assoc Disord*. 2004;18(4):270–7.
- Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15(4):869.
- Kukull WA, Ganguli M. Clinic-based data serving Population Neuroscience: NACC example. *Alzheimers Dement*. 2021;17:e051214.
- Weiner MW, Aisen PS, Jack Jr CR, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A. The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*. 2010;6(3):202–11.
- Banning LC, Ramakers IH, Rosenberg PB, Lyketsos CG, Leoutsakos JMS, Initiative ADN. Alzheimer's disease biomarkers as predictors of trajectories of depression and apathy in cognitively normal individuals, mild cognitive impairment, and Alzheimer's disease dementia. *Int J Geriatr Psychiatry*. 2021;36(1):224–34.
- Ghazi MM, Nielsen M, Pai A, Modat M, Cardoso MJ, Ourselin S, et al. Robust parametric modeling of Alzheimer's disease progression. *NeuroImage*. 2021;225:117460.
- Bron EE, Klein S, Papma JM, Jiskoot LC, Venkatraghavan V, Linders J, et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage Clin*. 2021;31:102712.
- Archetti D, Young AL, Oxtoby NP, Ferreira D, Mårtensson G, Westman E, et al. Inter-cohort validation of SuStain model for Alzheimer's disease. *Front Big Data*. 2021;4:661110.
- Cross-cohort research: Opportunities, challenges and examples. <https://www.closer.ac.uk/event/cross-cohort-research-opportunities-challenges-and-examples-2/>. Accessed 08 Mar 2022.
- Flanagan T, Fortier I, Sing MF, Moore C. An International Cross-cohort Harmonization and Data Integration Initiative towards Achieving Statistical Power and Meaningful Results: IJPDs (2017) Issue 1, Vol 1: 362 Proceedings of the IPDLN Conference (August 2016). *Int J Popul Data Sci*. 2017;1(1).
- Salter A, Stahmann A, Ellenberger D, Fneish F, Rodgers W, Middleton R, et al. Data harmonization for collaborative research among MS registries: a case study in employment. *Mult Sclerosis J*. 2021;27(2):281–9.
- Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28(3):427–43.
- Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc*. 2009;16(5):624–30.
- Zhang GQ, Siegler T, Saxman P, Sandberg N, Mueller R, Johnson N, et al. VISAGE: a query interface for clinical research. *Summit Transl Bioinforma*. 2010;2010:76.
- Bache R, Miles S, Taweel A. An adaptable architecture for patient cohort identification from diverse data sources. *J Am Med Inform Assoc*. 2013;20(e2):e327–33.
- Cui L, Zeng N, Kim M, Mueller R, Hankosky ER, Redline S, et al. X-search: an open access interface for cross-cohort exploration of the National Sleep Research Resource. *BMC medical informatics and decision making*. 2018;18(1):1–10.
- Salimi Y, Domingo-Fernandez D, Bobis-Alvarez C, Hofmann-Apitius M, Vasculature I, Birkenbihl C, et al. ADataViewer: Exploring Semantically Harmonized Alzheimer's Disease Cohort Datasets. *medRxiv*. 2021.
- Alzheimer's Disease Neuroimaging Initiative. ADNI publications. <http://adni.loni.usc.edu/news-publications/publications/>. Accessed 08 Mar 2022.
- Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc*. 2011;18(4):376–86.
- Liu K, Acharya A, Alai S, Schleyer T. Using electronic dental record data for research: a data-mapping study. *J Dent Res*. 2013;92(7 suppl):S90–S96.
- Tao S, Zeng N, Hands I, Hurt-Mueller J, Durbin EB, Cui L, et al. Web-based interactive mapping from data dictionaries to ontologies, with an application to cancer registry. *BMC Med Inform Decis Making*. 2020;20(S10):271.
- Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124–30.
- Glyph L Cog. XPDFReader. 2021. <https://www.xpdfreader.com/about.html>. Accessed 01 Aug 2023.
- Musen MA. The protégé project: a look back and a look forward. *AI Matters*. 2015;1(4):4–12.
- Lamy JB. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artif Intell Med*. 2017;80:11–28.

27. Rawat P, Mahajan AN. ReactJS: A Modern Web Development Framework. *Int J Innov Sci Res Technol*. 2020;5(11):698–702.
28. Bächle M, Kirchberg P. Ruby on rails. *IEEE Softw*. 2007;24(6):105–8.
29. Refolo LM, Snyder H, Liggins C, Ryan L, Silverberg N, Petanceska S, et al. Common Alzheimer's disease research ontology: National Institute on Aging and Alzheimer's Association collaborative project. *Alzheimers Dement*. 2012;8(4):372–5.
30. Malhotra A, Younesi E, Gündel M, Müller B, Heneka MT, Hofmann-Apitius M. ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimers Dement*. 2014;10(2):238–46.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

