ICPSR Working Paper 4:

Identifying Cross Series Cognitive Data Similarities Across NSHAP and NHATS

AUGUST 27, 2020

The mission of the NACDA Program on Aging

The National Archive of Computerized Data on Aging (NACDA), located within ICPSR, is funded by the National Institute on Aging. NACDA's mission is to advance research on aging by helping researchers to profit from the under-exploited potential of a broad range of datasets. NACDA acquires and preserves data relevant to gerontological research, processing as needed to promote effective research use, disseminates them to researchers, and facilitates their use. By preserving and making available the largest library of electronic data on aging in the United States, NACDA offers opportunities for secondary analysis on major issues of scientific and policy relevance.

ICPSR Working Paper 4: Identifying Cross Series Cognitive Data Similarities Across NSHAP and NHATS

JAMES W. MCNALLY

SANDA IONESCU

KATHRYN LAVENDER

BRENDAN KOTELES

UNIVERSITY OF MICHIGAN

This work was funded by Factors in Aging: Best Pactices in Archiving and Sharing Longetudinal Data Resources on Aging –Alzheimer's Supplement (NIA 3U24AG056918-01S1)

(McNally ORCID: https://orcid.org/0000-0002-6807-4538)

Identifying Cross Series Cognitive Data Similarities Across NSHAP and NHATS

Research Goals of the NACDA Program on Aging

The National Archive of Computerized Data on Aging (NACDA) seeks to create a dynamic and flexible data infrastructure to stimulate health research and advance knowledge as it relates to the gerontological lifecourse. Through the development and delivery of research resources and data services, NACDA alerts researchers to opportunities for secondary data analysis, provides tools to locate and access relevant materials, and enhances the availability of gerontological data. To fulfill this mission, NACDA strives to accomplish four specific goals: 1) Enhance and expand the longitudinal collections maintained by NACDA; 2) Advise and assist in the documentation and archiving of data and metadata for researchers who are producing longitudinal, repeat measure and linked data; 3) Distribute enhanced data and documentation to researchers in a form that will facilitate their use; and 4) Facilitate secondary analysis by providing user support, access to data, training, and consultation. By carefully tracking all aspects of user support requests and data use by the community we will more effectively identify and target emerging research trends and training needs so we can continue to evolve our programs and support services.

Abstract

This paper provides an overview of a methodology used to identify and organize health questions and measures related to Alzheimer's and other cognitive impairments using data maintained or supported by NACDA. This project specifically used the National Social Life, Health and Aging Project (NSHAP) and the National Health and Aging Trends Study (NHATS) as our comparison proof of concept. The methodology used in this process identifies variables that measure Alzheimer's disease (A.D.) and other cognitive impairments within NSHAP and NHATS as well as sociodemographic, and comorbidity data commonly associated with increased risk of A.D. and other cognitive impairments. As both NSHAP and NHATS represent multiple waves of longitudinal follow-up information we created longitudinal metadata files that allow for the comparison of A.D. and other cognitive impairments risk across time using these two studies. The project generated enhanced metadata using DDI Lifecycle software to make the discovery of A.D. and other cognitive impairments variables more straightforward and increase the user-friendly elements of these studies. Finally, the proposed supplement included the creation of a customized bibliography (see Appendix) of the use of NSHAP and NHATS data in the analysis of A.D. and other cognitive impairments research, allowing researchers to more easily review the existing body of literature using these data resources. This report describes NACDA's effort to increase the availability, usability, and discoverability of A.D. and other cognitive impairments information in these studies, encouraging use of NSHAP and NHATS for Alzheimer's related research and adding to our understanding of how cognitive issues change across time.

Introduction

The rapid growth of Alzheimer's disease diagnosis, treatment, and related mortality represent an ongoing source of financial and emotional stress for individuals, caregivers, and the U.S. healthcare system. Current estimates suggest approximately 5.7 million Americans have Alzheimer's disease, the vast majority over the age of 65. By 2025, some estimates show this number could reach 7.1 million, and increase to almost 14 million by 2050. After cancer and coronary heart disease; A.D. is the third most expensive disorder in the United States. In 2017, the estimated cost of formal care for Americans with Alzheimer's and other dementias was estimated to be $277, with almost $200 billion paid through Medicare and Medicaid, and out-of-pocket costs reaching as much as $60 billion. By 2050, the total costs of care for people with Alzheimer's and other dementias could be over $1 trillion. Recently, the Alzheimer's Association estimated a cost of approximately $85,775 per year for an individual to reside in a semi-private room in a nursing home. While Medicaid contributions offset this cost, nursing home care still represents a significant burden on federal and state governments, families, and individuals. A central concern for these cost and care estimates is that they generally address formal types of care provision, tracked through direct payment, insurance, Medicare, or Medicaid, but fail to capture the indirect cost of caregiving. While some estimates for informal costs of care exist, they have varied widely. Hurd et al., for example, estimated indirect costs of care in the U.S. at $109 billion for 2010, while a recent review article estimated the aggregate cost for informal caregiving during the same period to between $159 billion and $215 billion depending on the estimation approach.

This concern has translated into persistent calls to significantly enhance our understanding of Alzheimer's Disease (A.D.) and other dementias and their impacts across broad population groups. Collaborations such as the National Alzheimer's Project Act, the 2012 Alzheimer's Disease Research Summit, the health disparities session of the 2013 Alzheimer's Disease and Related Dementias Meeting[5], and the 2015 Alzheimer's Disease Research Summit[6] have specifically emphasized the need for more research across cohorts and improvements in the methods and tools to measure health disparities related to A.D. There is a clear need for increased enrollment of all impacted populations and expanding the use of existing cohorts captured by existing data to create of robust estimates of A.D. The number of longitudinal studies that capture Alzheimer's populations is also limited, and studies used to generate current estimates of A.D. prevalence are often drawn from cross-sectional measurements or a single wave of a longitudinal study used in a cross-sectional manner.

Longitudinal studies funded by the National Institute on Aging (NIA) provide unique insights into the lives of the U.S. population. By following individuals over extended periods of time and collecting multidisciplinary data on them, the resulting datasets provide researchers the opportunity to measure change and stability in these individuals and investigate the phenomena of aging from an integrated theoretical perspective. While these publicly-available datasets contain a wealth of data, they have several limitations: (1) similar data collected over multiple time periods are vulnerable to changing question structure or respondent consistencies which are not adequately documented; (2) data are often published in stand-alone data files, organized by wave or round; (3) and within studies themselves, naming conventions of variables are often not consistent across study periods. These challenges pose obstacles to researchers--especially early career and student--in the time-intensive search for related longitudinal data that are relevant to their research questions.

The current project illustrates best practices for improving metadata and descriptive documentation that facilitate the creation of user-friendly datasets and codebooks for the longitudinal analysis of Alzheimer's and other dementias. This approach created variable matches across longitudinal files and codebooks for the National Social Life, Health, and Aging Project (NSHAP) and the National Health and Aging Trends Study (NHATS) as our proof of concept. We reviewed each study individually and identified a subset of variables that share features across both studies.

This report describes work to enhance our growing ability to address gaps in our measurement of A.D. and other dementias through the use of information drawn from multiple studies measuring related activities. The NACDA project organized detailed metadata from NSHAP and NHATS in lieu of merged longitudinal analysis files for those studies to illustrate the benefits of structured longitudinal analysis files. This approach represents an alternative to the more common individual wave dataset format, which, at present, represents the delivery standard for most longitudinal studies. The project identified specific measures that capture the risk or presence of Alzheimer's and other dementias and make them the focus for an analysis data set that can be created by the researcher. In support of direct measures of cognition, the project identified common comorbidities such as obesity, heart disease, diabetes, stroke, high blood pressure, and high cholesterol, as well as socioeconomic and demographic variables. Collectively these data will allow researchers to look at changes in the expression or risk of Alzheimer's across time using these longitudinal analysis files and compare this change across two independent studies.

The measurement instruments and variables employed in these resources were structured in a searchable format consistent with facilitated comparative analysis. This approach increases discoverability and provides direct information on how A.D. and other dementias are measured in independent surveys. Secondly, the project identified and organized available research literature and studies into a discoverable bibliographic resource to identify existing measurements and findings from these studies and make them available to the research community in a systematic manner. The data resources emerging from this project have been made available to the research community through the NACDA website (nacda-aging.org) as well as the Colectica (colectica.com), NSHAP (www.norc.org/Research/Projects/Pages/national-social-life-health-and-aging-project.aspx) and NHATS (nhats.org) websites.

This process increases the value of both the NSHAP and NHATS studies by illustrating commonalities and variations in the measurement of Alzheimer's and other dementias. The ultimate goal is to improve opportunities across multiple studies for researchers to engage in guided analysis of cognitive change across time in the U.S. population. By increasing discoverability and tools that facilitate the use of longitudinal resources as a research tool this work is the first step in a more ambitious plan to simplify data management tasks associated with the measurement of change across time for Alzheimer's and other dementias. This data management approach can also facilitate the study of health conditions that represent barriers to successful aging.

ICPSR History with DDI; DDI Lifecycle; DDI Codebook; Colectica as a Resource

The metadata created for this project used the DDI-Lifecycle (DDI-L) standard, one of the development lines of the Data Documentation Initiative (DDI) specification. The Data Documentation Initiative (DDI) is an international metadata standard for describing data from the social, behavioral, economic, and health sciences. ICPSR has been a driving force in DDI development since its creation. The standard was created through a collaborative effort initiated and led by ICPSR, and supported by a National Science Foundation grant (SBR-9617813) between 1997 and 2000.

First published in 2000, the initial DDI metadata specifications quickly gained acceptance, being adopted by a significant number of data centers, projects, and archives worldwide. Its XML expression and structured design enable task automation and machine-processing of metadata, facilitating the creation of browsing and search tools that ultimately enhance the data users' experience through improved discovery and access. The use of DDI as a documentation standard

also encourages data sharing by ensuring interoperability. The cross-series cognitive comparison project with NSHAP and NHATS described in this paper represents an excellent example of the potential of DDI, allowing us to seamlessly bring together metadata created by two different organizations for direct comparison.

Supported by the DDI Alliance, a membership-based organization founded in 2005 under the directorship of ICPSR, there is an ongoing development of more sophisticated DDI standards and tools. Currently, the DDI standard supports two major development lines. DDI-Codebook (DDI-C), which is best suited to documenting individual, survey-type studies, and DDI-Lifecycle (DDI-L), which offers additional tools for describing related studies, longitudinal or time-series data, and variable comparability across series. Most recently, the DDI Alliance has released its evaluation for the DDI – Cross Domain Integration (DDI – CDI) specification. This new specification seeks to provide a model for working with a wide variety of research data across many scientific and policy domains. It provides a level of detail that supports machine-actionable processing of data, both within and between systems, and could dramatically enhance existing DDI capacities. This product is still under development but reflects the ongoing efforts of the DDI Allance to improve the functionality of the DDI instance.

ICPSR was among the first data archives to become DDI compliant, its metadata at the study, and the variable level fully aligned with the DDI-Codebook (DDI-C) specification since 2005. The organizational use of DDI-Codebook versus DDI-Lifecycle is driven by the preponderance of cross-sectional data across the broader ICPSR collections, ongoing investment in refinements to the Codebook model, and the cost of computer infrastructure migration. The new emphasis on longitudinal data central to NACDA's FACTORS IN AGING U24 (AG056918) award has offered the opportunity to expand our project into new areas of DDI compliance and take advantage of the longitudinal design features of DDI Lifecycle.

NACDA has aggressively explored the increased potential of DDI-Lifecycle for describing longitudinal, or time-series collections since initiating a relationship with Colectica, a leading software innovator in DDI Lifecycle tools. In our first project, we successfully created DDI-Lifecycle documentation for the three waves of the NSHAP study. In addition to describing individual waves in a longitudinal project, DDI-Lifecycle facilitates grouping them in a single series and creating series-level metadata that supports comparison and harmonization across waves. The NHATS research team has also documented the NHATS series using DDI-Lifecycle; therefore, we

had a good starting foundation with both collections being already compliant to the same standard. The novelty of the project described in this paper was to create cross-series comparison metadata of independent studies, something not previously attempted with DDI-Lifecycle. Based on a detailed analysis of each survey's content, we were able to develop a harmonization structure that identifies comparable variables with their value domains across all of the waves from both longitudinal studies (explicitly focused on topics of health, cognition, and A.D.).

This project used Colectica software to accomplish its goals. Colectica is a suite of tools designed to assist in creating and publishing data documentation compatible with the DDI-Lifecycle standard. Colectica Designer is a DDI-L metadata editor with multiple formats conversion features, useful for importing statistical data or questionnaire formats and translating them to DDI. The Colectica Repository may be used in conjunction with the Designer to store and manage the metadata in a shared environment. From the repository, the DDI-L documentation may be published and displayed on the Colectica Portal, which is a user-friendly Web application with browse and search features as well as data and metadata downloads.

NSHAP and NHATS and the Process of Identifying and Defining Matches

NSHAP and NHATS are both Tier 1 NIA funded, longitudinal, and U.S. National series of studies, both focusing on respondents in a similar age group (NSHAP - age 60 and up; NHATS - ages 65 and older and enrolled in Medicare). The timeframes are relatively close - NSHAP began in 2005, and NHATS began in 2011; both studies are ongoing. Both NSHAP and NHATS contain measures related to patterns in aging, incorporating topics of cognition, social engagement, and physical health measures. The process of identifying similarities across NSHAP and NHATS meant reviewing variable and collection-level metadata across three waves of NSHAP and eight rounds of NHATS. Additionally, both NSHAP and NHATS were already in Colectica portals through NACDA and the NHATS project, respectively.

In planning this project, another goal motivating our team was the desire to link longitudinal data measuring similar cognition related issues within aging populations. In this sense, we are using "link" quite loosely; the work in this project illustrates steps associated with identifying, organizing, and matching information across two longitudinal data series. There is no full harmonization (recoding) across variables, nor is this seen as appropriate. Underlying the desire to "link" these longitudinal collections, we provide solid examples of the consistency within the social sciences

across survey methods, design, and data formatting, which in turn allows for valid comparative analysis across samples.

Beyond the importance of these studies to NIA, a key consideration for using NSHAP and NHATS in this project is that both collections are generated by investigators and research teams that take significant care in creating the documentation supporting these data. As we are exploring new applications for relating independent metadata elements, it is reasonable to start with studies created with secondary data users in mind as part of the research process. The NSHAP and NHATS studies reflect examples of the active employment of data best practices, before and during data collection, and as a integral element of primary analysis. This effort results in data collections that encourage independent research, replication studies, and the exploration of more sophisticated research models. Deeply invested in the identification of common data elements across studies supported by NIA's Division of Behavioral and Social Research (BSR) data collections, NACDA saw these reliable data resources as an excellent starting point.

NACDA and our ICPSR colleagues familiarized themselves with the data and supporting documentation for each series to determine the best approach to develop a comparative structure. Each series offers extensive documentation to support the analytic data files, and each series also had preexisting structured variable level metadata available on independent Colectica portals. Drawing information from the existing portal metadata files for both NSHAP and NHATS proved to be the most efficient approach as it allowed us to maintain consistency with the independent study portals. Linking the new cross-series comparison crosswalk to the existing projects allows the users to move from the metadata to the analysis files without the need for moving through mutiples portals conveniently. As the NHATS portal is still under development, there are a few question text, descriptive text, and variable labels still extant for some variables. To clarify their content, in addition to the portal documentation, we consulted the original NHATS documentation (codebooks, questionnaires) available on their dedicated website.

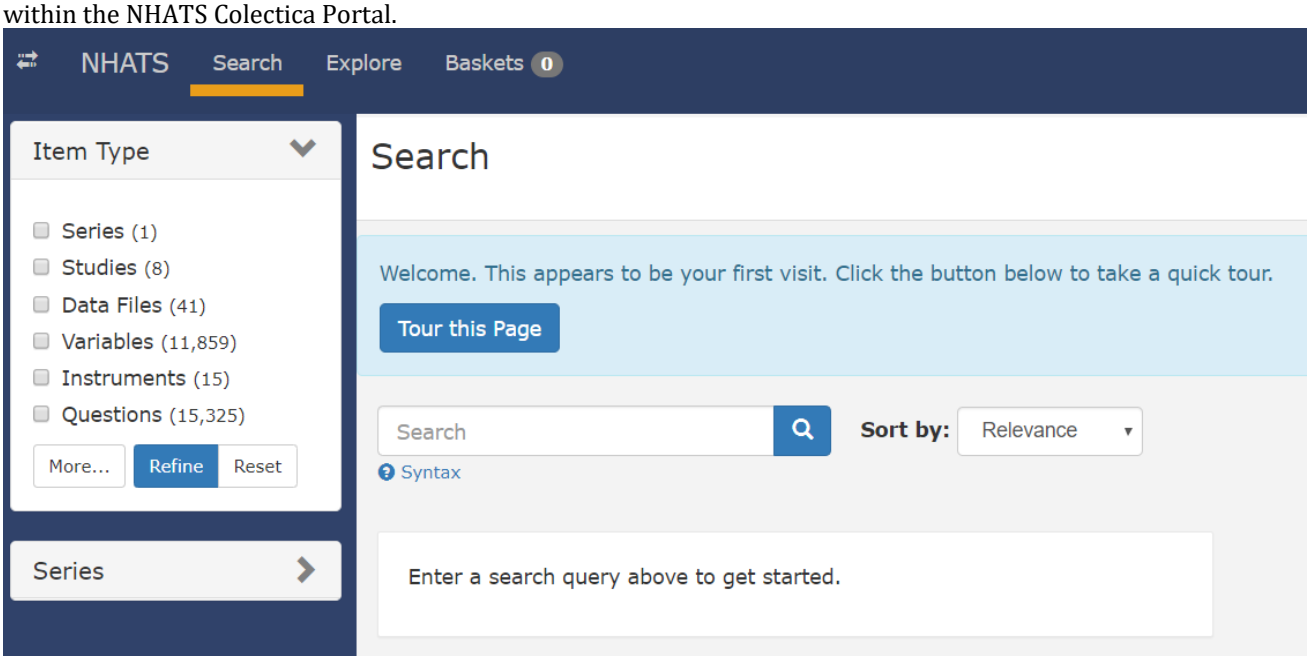
Using multiple Excel sheets to organize the initial work, we started with an exhaustive list of the NSHAP variables from the DDI-L documentation already created for this series. As the variable list, complete with topical groups and variable labels, was already available, it reduced the person time costs considerably for this proof of concept project. The NACDA-ICPSR team then made two passes through the variables. First, we reviewed the list of NHATS variables in order to identify any potential matches with NSHAP variables; any potential matches were flagged as possible matches.

We then did this same search in reverse - exploring the NSHAP variable list to pinpoint potential matches with NHATS variables. These two passes resulted in a first draft of a crosswalk with potentially comparable variables. Second, we analyzed the detailed variable descriptions (question text and values with value labels) to determine the type of comparability. In this step, we eliminated some variables were not actually comparable, and determined variables which were either directly comparable or would require harmonization to be comparable, and variables which were only related at the conceptual level. This review also revealed a fourth type of comparability (one that we did not anticipate), where a single variable from one of the series could be compared to a group of variables from the other series, measuring the same concept. We defined these as "one-to-many" or "many-to-one" types of matches.

To determine the degree of comparability between the matching pairs across the two series, we added question text (where available), as well as values and value labels for the NHATS variables to the spreadsheet, replicating the process we performed with the NSHAP metadata. This process created a document that contained all of the variable-level detail required to evaluated and categorize the matches. This master spreadsheet was then subsetted to create a file that exclusively identified the matches and classified them according to their degree of comparability.

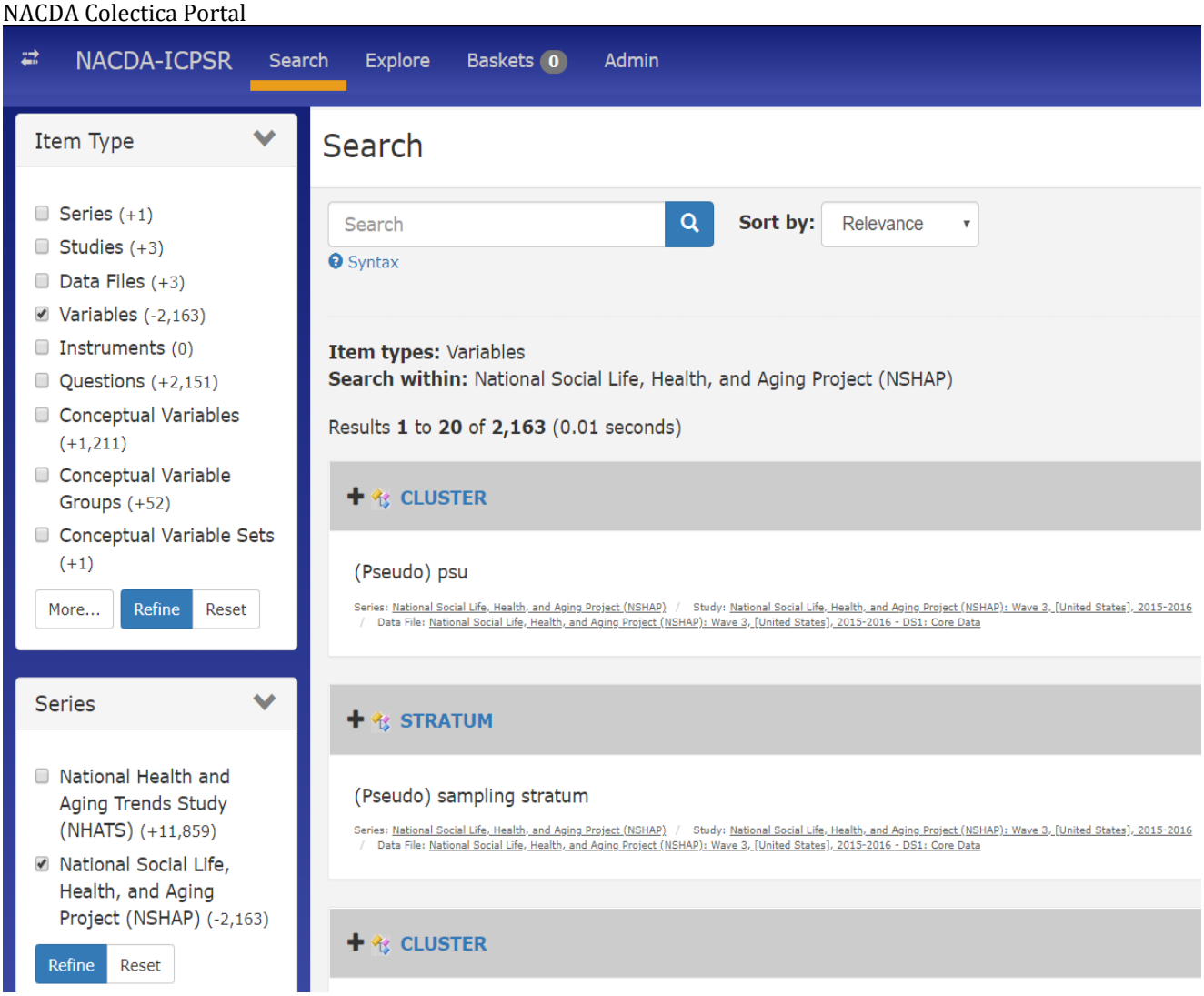
NHATS COLECTICA PORTAL FIG. 1

Fig.1 (below) NHATS portal image from Search page; NHATS has 11,859 variables in the across the series within the NHATS Colectica Portal.



NACDA PORTAL FIG. 2

Fig. 2 (below) NACDA Portal image from Search Page; NSHAP has 2,163 variables across the series within the NACDA Colectica Portal



To identify comparable variables across both series, we took advantage of a specific DDI-Lifecycle structure that connects similar variables to a common concept within a "variable cascade." This function helped to identify cross-series commonalities by examining a single item from each series, i.e., the unique conceptual variable that represents the link among the actual measures used in each dataset. These links are intrinsic to the DDI-Lifecycle design, further allowing the user to retrieve all of the comparable variables across individual waves from each conceptual match. This incorporates these matches into a complete cross-series concordance.

We identified matching conceptual variables by selecting similar topics and then narrowing down the search to an evaluation of the question text and response domains for the variables that

appeared comparable. Ultimately a total of 252 conceptual variable matches were identified across the two series. A "new cross-series conceptual item" was then created (column B in the image below, fig. 3) to flag each of these matches, and further group these concepts into topical groups and subgroups.

FIG. 3 NACDA EXCEL CONCORDANCE SHEET FOR NSHAP-NHATS COMPARISON (COLUMNS A-B)

	A	B
1	Topical Groups and Subgroups	NEW Cross-Series Conceptual Items
256	Cognitive function	Self-rated mental health
257	Cognitive function	Self-rated memory
258	Cognitive function	Cognitive impairment (Alzheimer's, dementia)
259	Cognitive function	Dementia or Alzheimer's
260	Cognitive function	Alzheimer's
261	Cognitive function	Dementia
262		
263	Standard memory question: today's date	Today's month correct (NSHAP SPMSQ)
264	Standard memory question: today's date	Today's month correct (NSHAP MOCA)
265	Standard memory question: today's date	Today's day correct (NSHAP SPMSQ)
266	Standard memory question: today's date	Today's year correct (NSHAP SPMSQ)
267	Standard memory question: today's date	Today's date correct (NSHAP SPMSQ combined)
268	Standard memory question: today's date	Today's month correct (NHATS)
269	Standard memory question: today's date	Today's day correct (NHATS)
270	Standard memory question: today's date	Today's year correct (NHATS)
271	Standard memory question: today's date	Today's date correct (MOCA combined)
272	Standard memory question: today's date	Today's month correct (NHATS)
273	Standard memory question: today's date	Today's day correct (NHATS)
274	Standard memory question: today's date	Today's year correct (NHATS)
275	Standard memory question: today's date	Day of week correct?
276		
277	Standard memory question: president's name	President's name correct/incorrect
278	Standard memory question: president's name	President's last name correct?
279	Standard memory question: president's name	President's last name incorrect?
280	Standard memory question: president's name	President's first name correct?
281	Standard memory question: president's name	President's first name incorrect?
282		
283	Standard memory question: clock drawing	Attempted clock drawing
284	Standard memory question: clock drawing	Clock drawing score
285	Standard memory question: clock drawing	Clock contour
286	Standard memory question: clock drawing	Clock numbers

Excel spreadsheets manage the variable organization work, allowing the importation of these files into the Colectica software. The spreadsheet workflow reflects our internal organization system to create the DDI-L necessary for the Colectica portal; the Colectica system is flexible enough to allow

for different preparation and importation approaches. The "NEW Cross-Series Conceptual Items" in our spreadsheet became new conceptual variables for the project (across both the NSHAP and NHATS series). These are called "concepts" because they apply to unique pairs, one concept-one pair. With this in mind, the groupings (on the left-hand side in the image above, column A) are not definitively concepts. They are conceptual variable groups created by topic; calling them topical groups would be more accurate as they indicate the broader topic covered. In DDI, there is a direct relationship between concepts and variables, but a group cannot be called "concept." It's either a concept group or a variable group; in this project, these are variable groups.

In 2018-2019, NACDA worked with Colectica to add the NSHAP series to the NACDA Portal. The Colectica Designer software was used to create the DDI-Lifecycle for both the individual waves and the entire series. Initially, the DDI metadata did not have active links between individual variables across the waves of NSHAP. These links could have been created manually in the Designer, but it would have been a time-consuming activity. Instead, a separate Excel spreadsheet was used to indicate these links, and in turn used by the Colectica support team with an automated script that incorporated all of the variable links in the DDI-L documentation.

For the NHATS-NSHAP comparison project, NACDA created the crosswalk in spreadsheet format, which Colectica translated (also using a script that they put together according to the specs our teams agreed on) into a new DDI-L instance that explicitly documented the comparison project. This linked the new cross-series conceptual items created for the project to the conceptual variables from each series, and the topical groups and subgroups (left-hand side column A in the spreadsheet image above, fig. 3). Note that the new cross-series conceptual items for the series are artifacts that serve to link the actual variables across all waves and both series.

- 59 unique NSHAP-NHATS common topics ("Topical Groups and Subgroups," column A in the image of the sheet below, fig. 4)
- 503 Cross-Series Conceptual Items (column B in the image below, fig. 4)
- 28 unique one-to-many matches where a single NSHAP variable matches with multiple NHATS variables taken together (column D in the image of the sheet below, corresponding to "one-to-many" in column C in images below, fig. 5)
- 34 unique many-to-one matches where a single NHATS variable matches with multiple NSHAP variables taken together (column E in the image of the sheet below, corresponding to "many-to-one" in column C in images below, fig. 5)

- 11 matches between variables that are directly comparable
- 88 matches between variables that need harmonization
- 91 matches between variables that measure related concepts
- Therefore, 28+34+11+88+91 = 252 unique matches of some type across the two series

FIG. 4 NACDA EXCEL CONCORDANCE SHEET FOR NSHAP-NHATS COMPARISON (COLUMNS A-C)

	A	B	C
1	Topical Groups and Subgroups	NEW Cross-Series Conceptual Items	Comparability note
26	Respondent demographics	Number of stepchildren	Comparability: one to many
27	Respondent demographics	Number of sons	**Comparability: directly comparable**
28	Respondent demographics	Number of daughters	**Comparability: directly comparable**
29	Respondent demographics	Born in the US	Comparability: need harmonization
30	Respondent demographics	Financial status in childhood	Comparability: need harmonization
31	Respondent demographics	Lived with both parents in childhood	Comparability: related concepts
32	Respondent demographics	Health status in childhood	Comparability: need harmonization
33	Respondent demographics	Gender of partner	**Comparability: directly comparable**
34	Respondent demographics	Education of spouse or partner	Comparability: need harmonization
35	Respondent demographics	Internet use	Comparability: related concepts
36			
37	Employment and income	Currently employed	Comparability: need harmonization
38	Employment and income	Retired	Comparability: need harmonization
39	Employment and income	Unemployed	Comparability: related concepts
40	Employment and income	Worked for pay past week	Comparability: need harmonization
41	Employment and income	Hours worked per week	Comparability: need harmonization
42	Employment and income	Household income past year	Comparability: need harmonization
43	Employment and income	Household income bracket	Comparability: many to one
44	Employment and income	Household income more/less than 50k	Comparability: many to one
45	Employment and income	Household income more/less than 25k	Comparability: many to one
46	Employment and income	Household income more/less than 100k	Comparability: many to one
47			
48	Social life	Size of social network	Comparability: need harmonization
49	Social life	Socializing with friends or family	Comparability: related concepts
50	Social life	Volunteer work	Comparability: related concepts
51	Social life	Attending organized group activities	Comparability: related concepts
52	Social life	Attending religious services	Comparability: related concepts
53	Social life	Total years lived in current neighborhood	**Comparability: directly comparable**
54			
55	Biomeasures	Current weight	**Comparability: directly comparable**
56	Biomeasures	Waist measurement final status	Comparability: related concepts
57	Biomeasures	Waist measurement (inches)	**Comparability: directly comparable**
58	Biomeasures	Height	Comparability: one to many
59	Biomeasures	Height (feet)	Comparability: one to many
60	Biomeasures	Height (inches)	Comparability: one to many

FIG. 5 NACDA CONCORDANCE SHEET FOR NSHAP-NHATS COMPARISON (COLUMNS B-E)

B	C	D	E
NEW Cross-Series Conceptual Items	Comparability note	NSHAP C. variable	NHATS C variable
Number of stepchildren	Comparability: one to many		dnmstpchd
Number of sons	**Comparability: directly comparable**	CON_SONS	dnumson
Number of daughters	**Comparability: directly comparable**	CON_DAUGHTER	dnumdaugh
Born in the US	Comparability: need harmonization	CON_BORN_US	borninus
Financial status in childhood	Comparability: need harmonization	CON_FAMFIN	fingrowup
Lived with both parents in childhood	Comparability: related concepts	CON_LIVEPARENT	lvbhpar15
Health status in childhood	Comparability: need harmonization	CON_CHLDHLTH	hlthchild
Gender of partner	**Comparability: directly comparable**	CON_PGENDER	spgender
Education of spouse or partner	Comparability: need harmonization	CON_PEDUC	spouseduc
Internet use	Comparability: related concepts	CON_INTERNET	online
Currently employed	Comparability: need harmonization	CON_JOBSTAT_1	doccup
Retired	Comparability: need harmonization	CON_JOBSTAT_2	abstlstwk
Unemployed	Comparability: related concepts	CON_JOBSTAT_4	doccup
Worked for pay past week	Comparability: need harmonization	CON_WEEKPAY	workfpay
Hours worked per week	Comparability: need harmonization	CON_HRSCJOB	hrswkwork
Household income past year	Comparability: need harmonization	CON_HEARN_RECODE	totinc
Household income bracket	Comparability: many to one		toincesjt
Household income more/less than 50k	Comparability: many to one	CON_IML50K	
Household income more/less than 25k	Comparability: many to one	CON_IML25K	
Household income more/less than 100k	Comparability: many to one	CON_IML100K	
Size of social network	Comparability: need harmonization	CON_ALTERS	dnumsn
Socializing with friends or family	Comparability: related concepts	CON_SOCIAL	vistfrfam
Volunteer work	Comparability: related concepts	CON_VOLUNTEER	voltrwork
Attending organized group activities	Comparability: related concepts	CON_ATTEND	clbmtgrac
Attending religious services	Comparability: related concepts	CON_ATNDSERV	attrelser
Total years lived in current neighborhood	**Comparability: directly comparable**	CON_RESIDEY	yrslived
Current weight	**Comparability: directly comparable**	CON_WEIGHT	currweigh
Waist measurement final status	Comparability: related concepts	CON_WAIST	waistrslt
Waist measurement (inches)	**Comparability: directly comparable**	CON_WAISTM	wstmsrinc

Examples of Matches

In total, the project identified **252** unique matches at the variable level containing cross-study counterpart or counterparts, determined by comparing the question text and responses from the candidate variables. There are five different categories of matches, each with different criteria: directly comparable; need harmonization; many-to-one; one-to-many; related concepts.

Initially, matches were grouped into three initial categories: exact matches, near matches, and topic only matches. These matches are defined in the "Comparability notes," as seen in the images above (fig. 3-4). The NACDA-ICPSR team determined the match/comparability types after a careful examination of how to compare different conceptual variable pairs. This stage was the initial contribution to the project, adding a necessary layer to indicate to users that matches are not all of the same types, representing value-added to the metadata.

"Exact matches" indicate directly comparable variables, measuring the same concept with similar or identical questions, and require no manipulation of the data for comparison. An exact match means the variables reflect the same information, were collected in the same or nearly the same way (as far as to question phrasing and timing) and contains the same range of values. Typically seen within the biometrics and respondent demographics topics, exact matches are relatively uncommon across broader conceptual topics. We found only eleven matches of this type: gender, marital status (2), number of sons and daughters, the gender of partner, total years in the current neighborhood, current weight, waist measurement, number of cigarettes smoked per day, number of hours slept (fig. 6).

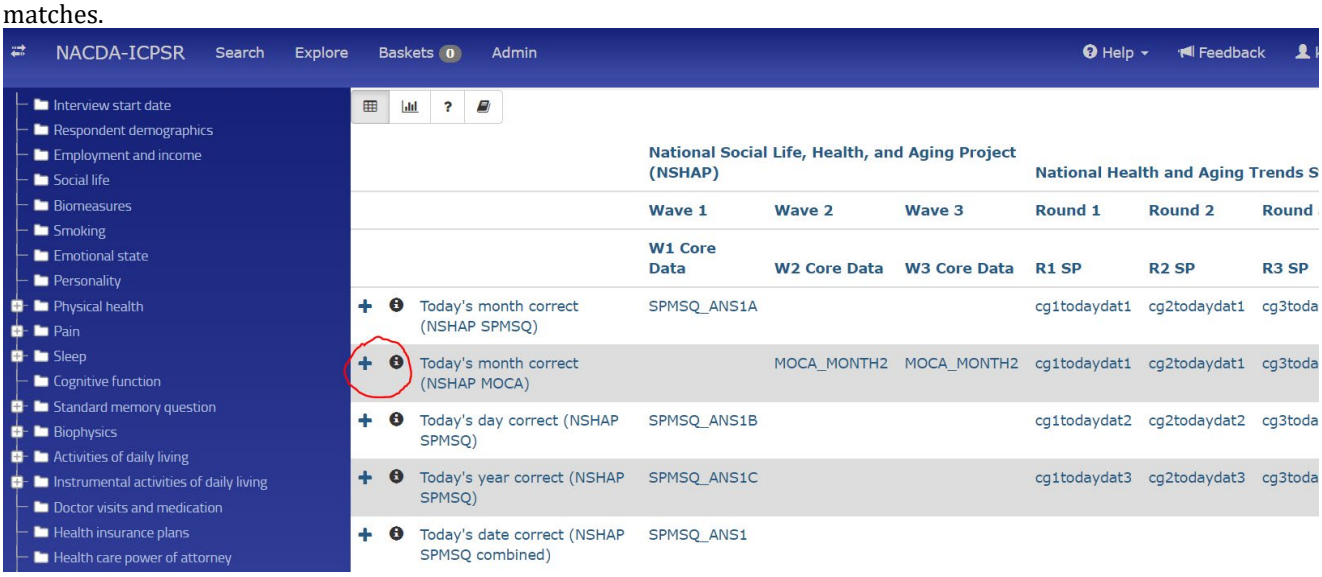
FIG. 6 NACDA CONCORDANCE SHEET FOR NSHAP-NHATS COMPARISON (COLUMNS A-E)

A	B	C	D	E
Topical Groups and Subgroups	NEW Cross-Series Conceptual Items	Comparability note	NSHAP C. variable	NHATS C variable
Respondent demographics	Gender	**Comparability: directly comparable**	CON_GENDER	dgender
Respondent demographics	Marital status	**Comparability: directly comparable**	CON_MARITLST	martlstat
Respondent demographics	Marital status: derived	**Comparability: directly comparable**	CON_MARITLST	dmarstat
Respondent demographics	Number of sons	**Comparability: directly comparable**	CON_SONS	dnumson
Respondent demographics	Number of daughters	**Comparability: directly comparable**	CON_DAUGHTER	dnumdaugh
Respondent demographics	Gender of partner	**Comparability: directly comparable**	CON_PGENDER	spgender
Social life	Total years lived in current neighborhood	**Comparability: directly comparable**	CON_RESIDEY	yrslived
Biomeasures	Current weight	**Comparability: directly comparable**	CON_WEIGHT	currweigh
Biomeasures	Waist measurement (inches)	**Comparability: directly comparable**	CON_WAISTM	wstmsrinc
Smoking	Number of cigarettes smoked per day	**Comparability: directly comparable**	CON_AVECIG	numcigday
Sleep: night-time sleep	Number of hours slept at night	**Comparability: directly comparable**	CON_HRSSLEEP	sleephour

Near matches reflect variables that carry similar content, measuring the same idea or collecting the same kind of information, but may reflect different ranges of values or timeframes in their measurement. To be directly compared or used in statistical analysis, these variables across NSHAP and NHATS may need to be adjusted or "harmonized" to obtain a more exact match. The standard memory questions, such as asking respondents "what is today's date" reflect examples of variables across NSHAP and NHATS in which the respondent response was recorded as correct or not; they differ in the values used ("correct" and "incorrect" vs "yes" and "no"). There are 88 instances total in the Excel spreadsheet, including duplicates due to one-to-many and many-to-one matches. The "need harmonization" label and flag generated for these variables will help researchers identify those variable matches across NSHAP and NHATS that can be analyzed together after recoding the data instances to reflect the same set of values. This kind of manipulation can also include transforming a continuous set of values into a set of ranges, or a set of ranges into a yes or no response.

FIG. 7 NACDA COLECTICA PORTAL – EXAMPLE OF ICON INDICATING COMPARABILITY NOTES IN THE PORTAL

The "i" symbol in the black circle in the image below indicates the comparability notes for each conceptual variable; variables across this conceptual variable grouping would need to be harmonized in order to be exact matches.



The alternating shading in the Colectica portal — as illustrated in the image below, fig. 8 — is used to differentiate between the single matches (one-to-one) and another type of matches — **one-to-many, or many-to-one**. Technically the one-to-many and many-to-one could be double counted as "need harmonization" or "concept-only matches." Similarly, one-to-many and many-to-one could require multiple variables from one study to build a match with a single variable from the other study. Thus, a single NSHAP or NHATS variable may be comparable to two or more similar NSHAP or NHATS variables. An example of this kind of match would be comparing a question in one series that asks for the respondent's "age" with another where three questions asking for the respondent's day of birth, month of birth, and year of birth are required to construct an age variable. One-to-many and many-to-one matches can also be classified as directly comparable or needing harmonization. In this case, we decided to apply just one type of comparability note to each pair for the sake of simplicity.

FIG. 8 NACDA COLECTICA PORTAL – EXAMPLE OF ALTERNATING COLORS

Image below represents an example of a many-to-one match - see the "Cognitive Impairment (Alzheimer's, dementia)" conceptual variables, in the explore view.

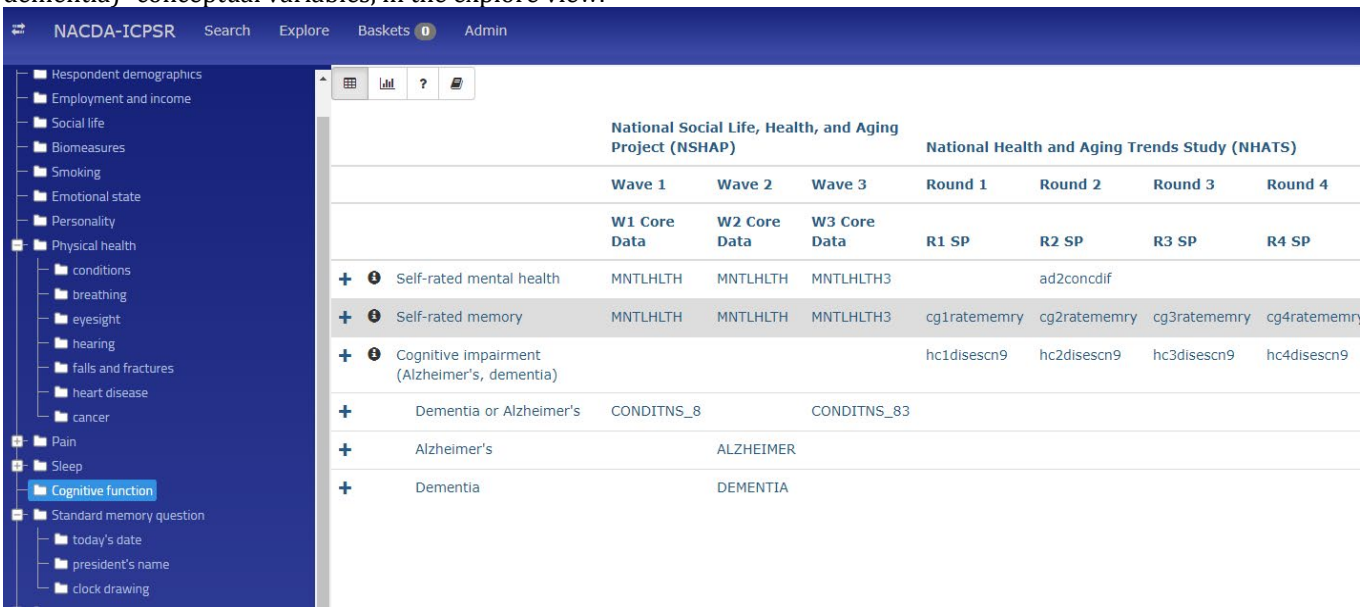


FIG. 9 NACDA CONCORDANCE SHEET FOR NSHAP-NHATS COMPARISON (COLUMNS C-F)

Closeup screenshot of one-to-many matches listed in spreadsheet

C	D	E	F
Comparability note	NSHAP C. variable	NHATS C variable	Multiplicity
Comparability: one to many	CON_INT_START_NEW		
Comparability: one to many		spstat	CON_INT_START_NEW
Comparability: one to many		spstatdtmt	CON_INT_START_NEW
Comparability: one to many		spstatdtyr	CON_INT_START_NEW

Closeup screenshot of many-to-one matches listed in spreadsheet

C	D	E	F
Comparability note	NSHAP C. variable	NHATS C variable	Multiplicity
Comparability: many to one		toincesjt	
Comparability: many to one	CON_IML50K		toincesjt
Comparability: many to one	CON_IML25K		toincesjt
Comparability: many to one	CON_IML100K		toincesjt

"Topic only" or "related concept" matches are one-to-one matches between questions with related subject matter, but which are not comparable because they do not precisely measure the same

concept, only a related one. The project identified 91 cases of this type of match. Examples include comparing questions asking about the frequency of socializing with peers, which while similar, are not conceptually comparable with questions asking about the frequency with which the respondent visits peers. Both questions regard the respondents' social lives in similar ways, but there is no way to manipulate the data so that the responses give "one to one" comparable information.

Challenges throughout the Process

As this was a proof of concept project with no template to guide our initial activities, it required some trial and error early in the process. Initially, NHATS conceptual variables were directly matched to NSHAP conceptual variables by topic. In review, the need for complete variable descriptions, including question text and values with value labels, became apparent. This more detailed information allowed us to assess whether the variables were comparable, and to which degree. We did not preserve the topical groupings from the individual NSHAP and NHATS portals since the focus of the project was on identifying related cognition and A.D. items. The goal was to identify common topics across series, instead of within each series. For these reasons, it made more sense to create new topical groupings specifically for this project, representing an additional value added to the metadata, expanding the research use of both data collections.

The variable review to determine the degree of comparability identified some anomalies. For example, a single variable from one of the series could capture a broader concept that was accounted for by a group of variables from the other series. This issue is not an uncommon occurrence, arising from the approach study authors/designers choose to measure a concept. For this particular use case, we chose to indicate that the type of comparability was "one-to-many" or "many-to-one." We used the one-to-many identification approach, for example, in instances where an NSHAP variable reflected values based on multiple-choice questions. In contrast, the matching NHATS variables often measured the same idea but had recoded the multiple-choice values into a series of yes/no values for each possible answer.

The significant number of variables that had to be reviewed across each series made it difficult to identify all potential matches while ensuring consistency in the approach. Each variable's detailed description was examined to determine whether the pair was directly comparable (the original survey questions measured the same concept, and the values were the same) or needed harmonization (same concept, different values).

An important decision in determining the degree of comparability among variables was whether to take the missing values into account when comparing response categories. We found that variables from the two series could have identical valid values, while the missing values were different. These pairs could be interpreted as "directly comparable" if we only looked at the valid values, or as "needing harmonization" if we considered the missing values as well. Our decision for this project was to ignore the differences in missing values as the adjustment of these values occurs as part of analysis preparation.

Visualizing the expected outcome of the portal view represented a challenge. Previous work has successfully represented independent longitudinal studies in the Colectica portals (NSHAP, MIDUS, NHATS, and others). However, cross-study comparisons did not previously reside in the Colectica portal at the variable level. The U.K. longitudinal studies portal, CLOSER, does this in a sense, presenting variables within a particular topic across studies listwise. Working with the Colectica team, we chose to display the variables side by side in a quasi calendar format. The implementation of direct "one to one" variable matches representing direct comparisons proved straightforward, but other types of matches represented a challenge. We needed to be able to show how one variable in NSHAP could be comparable to multiple variables in NHATS, and vice versa; this lead to implementing the greyscale within the portal and the "i" bubble/toolkit widgets (see images above). The one-to-many and many-to-one comparisons reflected in the Excel sheet delivered to Colectica specified the match type, but by adding the "Multiplicity" column clarified variables and their relationship further as to the level of match achieved. We expect to continue to work with the Colectica team to improve the display of the cross-series cognitive comparisons, as well as hope to add more longitudinal series to the portal.

Lastly, there are administrative challenges to every project, and this was no exception. Administrative considerations include time, budget, roles, specific software requirements (licenses, and online tools), versioning, and communication. Experience in archival and repository costs allowed us to generate budget estimates that reasonably captured the staff and equipment costs for this project and the amount of time it would take to complete. The NACDA and Colectica team estimates for the time needed to communicate the project goals and accommodate revisions to the Excel sheets and the portal display fell within the budget. NACDA and Colectica maintained open and ongoing communication and thus were able to adjust the metadata work to ensure the optimal web presentation.

The early discussions that established what we wanted to see in the final product allowed both teams to collaborate more efficiently throughout the entire lifetime of the project. Careful consideration of the roles and the number of people required to execute this project, as well as the need to purchase additional software licenses, will offer additional guidance in designing new projects to expand this initiative. The planning process also established protocols for identifying staff to play an "administrative" role within the Colectica portal online tool. The "admin" person has the authority to edit elements of the hosted portal without the need to contact the Colectica team. This role also allows the admin designate to monitor registered users and portal activity. One issue that we did not initially account for was version control; this led to backtracking later after the portal was live and became part of "lessons learned," which, again, will allow us to expand the project across additional studies more efficiently. Some lessons learned related to versioning also include utilizing the Colectica "zen desk"/service desk system, and creating internal tracking outside of email communications.

FIG. 10 NACDA COLECTICA PORTAL CODE COMPARISON VIEW

Code comparison view of "Presence of pain" items (classified as "needs harmonization"); as illustrated below, the conceptual variables related to a respondent having pain in NSHAP and in NHATS have similar value ranges. There is a challenge here with these slight differences in the way the different studies labeled their variable values.

Statistics | Code Comparison | Correspondence Tree

	National Social Life, Health, and Aging Project (NSHAP)		National Health and Aging Trends Study (NHATS)				
	Wave 2	Wave 3	Round 1	Round 2	Round 3	Round 4	Round 5
	W2 Core Data HADPAIN	W3 Core Data HADPAIN	R1 SP ss1painbothr	R2 SP ss2painbothr	R3 SP ss3painbothr	R4 SP ss4painbothr	R5 SP ss5painbot
-9 Missing			-9	-9	-9	-9	-9
-8 DK			-8	-8	-8	-8	-8
incomplete interview	-8	-8					
-7 RF			-7	-7	-7	-7	-7
missing in error	-6	-6					
not returned	-5	-5					
no answer	-4	-4					
not applicable	-3	-3					
don't know	-2	-2					
-1 Inapplicable			-1	-1	-1	-1	-1
refused	-1	-1					
no	0	0					
1 YES			1	1	1	1	1
yes	1	1					
2 NO			2	2	2	2	2
7 PREVIOUSLY REPORTED							

Summary/Conclusion:

By allowing users the ability to review NSHAP and NHATS side by side, we are increasing the opportunities that users will have in the discovery and exploration of data related to Alzheimer's and other dementias. This work described in this project reflects the amount of effort involved in searching for across independent studies and identifying shared attributes and differences. By performing these activities proactively and making them readily available to the research community, this project shows the ongoing value of archival best practices. This enhanced comparative structure makes it easier for researchers to select the variables they are interested in, freeing time for data analysis and drawing inferences that would have previously been dedicated to data preparation. The project also offers a standardized approach to data comparison. Researchers will not "reinvent the wheel," and by using the same data, the likelihood that comparable results will emerge from the independent analysis increases.

FIG. 11 NACDA COLECTICA PORTAL HOME PAGE

Screenshot of NACDA Portal home page; "Cross Series Cognitive Comparisons" directs users to view NSHAP and NHATS variables side by side by topic.

Appendix

Types of Matches:

Types of matches were defined, and then used to label variables according to the degree of comparability:

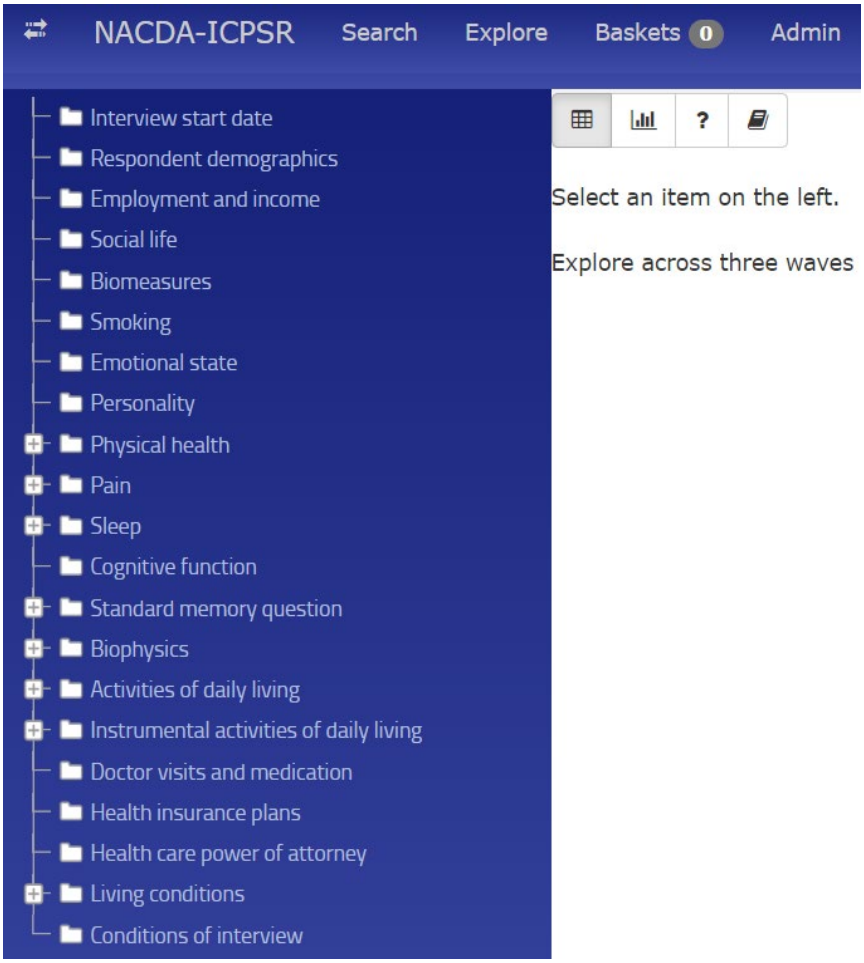
- Directly comparable: measuring the same concept, with the same data values and value labels
- Need harmonization: variables need to be harmonized before they can be compared using statistical procedures; with alteration to value labels could be made directly comparable
- Related concepts: are not measuring the exact same concept, but a related one. Cannot be harmonized, but the topics are close enough that viewing together could be of use
- One-to-many: two or more NHATS variables taken together are comparable to a single NSHAP variable
- Many-to-one: two or more NSHAP variables taken together are comparable to a single NHATS variable

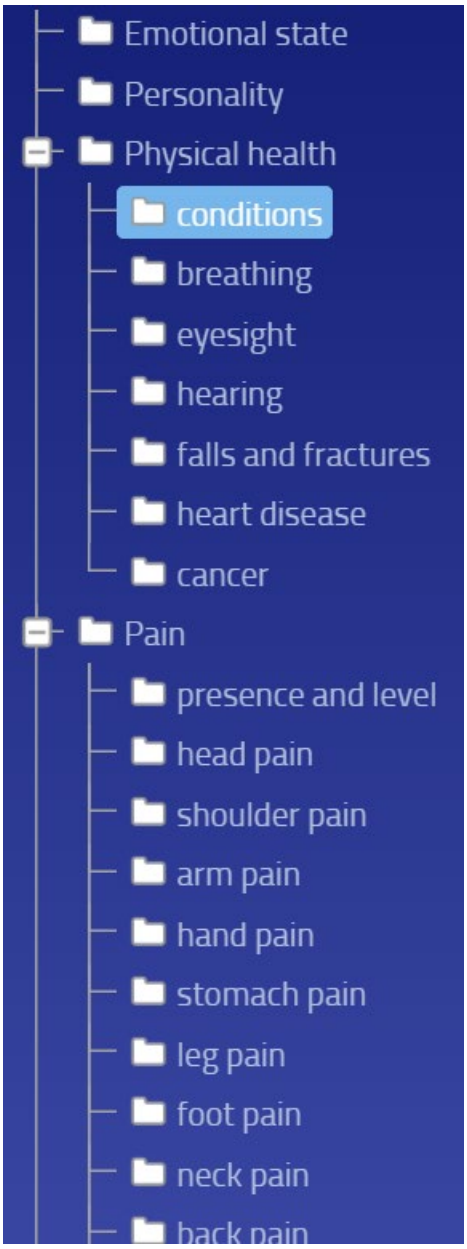
Topics:

1. Interview start date
2. Respondent demographics
3. Employment and income
4. Social life
5. Biomeasures
6. Smoking
7. Emotional state
8. Personality
9. Physical health: conditions
10. Physical health: breathing
11. Physical health: eyesight
12. Physical health: hearing
13. Physical health: falls and fractures
14. Physical health: heart disease
15. Physical health: cancer
16. Pain
17. Head pain
18. Shoulder pain
19. Arm pain
20. Hand pain
21. Stomach pain
22. Leg pain
23. Foot pain
24. Neck pain
25. Back pain
26. Hip pain
27. Sleep: naps
28. Sleep quality
29. Sleep: night-time
30. Cognitive function
31. Standard memory question: today's date
32. Standard memory question: president's name
33. Standard memory question: clock drawing
34. Biophysics: timed walk
35. Biophysics: chair stands
36. Biophysics: balance measure, side by side
37. Biophysics: balance measure, one foot ahead
38. Biophysics: balance measure, heel to toe
39. Activities of daily living: getting dressed
40. Activities of daily living: bathing
41. Activities of daily living: eating
42. Activities of daily living: getting in and out of bed
43. Activities of daily living: toileting
44. Activities of daily living: walking
45. Activities of daily living: use of help
46. Instrumental activities of daily living: managing medications
47. Instrumental activities of daily living: preparing meals
48. Instrumental activities of daily living: shopping
49. Instrumental activities of daily living: housework
50. Instrumental activities of daily living: driving
51. Instrumental activities of daily living: use of telephone
52. Instrumental activities of daily living: managing finances
53. Doctor visits and medication
54. Health insurance plans
55. Health care power of attorney
56. Residence type
57. Residence condition
58. Neighborhood description
59. Conditions of interview

On the Portal, the topics are organized in groups and subgroups. The main groups (topics) are represented by the text before the colon in the list above, while the text following the colon has become a subgroup.

Images below represent the NACDA Colectica Portal; the left image is the initial view and list of topics (main groups) when a user selects the “cross series cognitive comparisons” on the welcome page or the link to “NACDA Concepts (NSHAP-NHATS [variable] comparison)” from the Explore page. The right image is an up-close view of those topics with some items expanded to show the subgroups.





Cross-Series Conceptual Items used more than once (duplicates):

- Skin cancer (2)
- Taking medications to sleep (2)
- Today's month correct (NHATS) (2)
- Today's day correct (NHATS) (2)
- Today's year correct (NHATS) (2)
- Gets help dressing (2)
- Gets help bathing (2)

In general, the subconcepts should be completely unique since these are considered like variables in a database; however, these items were overlooked and do not currently cause any issue.

Tagged Related Publications

DISTRIBUTION OF CITATION RECORDS WITH SUBJECT TERMS ("COGNITION", "DEMENTIA", AND/OR "ALZHEIMER'S"), BY SERIES AND STUDIES	
Series/Study name (acronym)	**# of records added & tagged during project term**
NHATS	145
SALSA	46
CLHLS	40
MIDUS	24
TILDA	17
NSHAP	14
WHO SAGE	13
H-EPESE	5
REACH II	4
HAALSI	4
NHANES	3
Americans' Changing Lives	3
National Poll on Healthy Aging	2
CRELES	2
SWAN	1
CogUSA	1
Australian [Adelaide] Longitudinal Study of Aging	1
Images of Aging	1
Americans' Changing Lives	1
EPESE	1
Survey of Long-Term Care Awareness and Planning	1
National Nursing Home Survey Series	1
National Long-Term Care Survey	1
National Hospital Discharge Survey Series	1