Documenting Variable Comparability with DDI-Lifecycle

IASSIST Virtual Meeting May 20, 2021

ICPSR Documentation Specialist Sanda Ionescu
NACDA Data Project Manager Kathryn Lavender
Today’s Presentation

• Project Background
• About NSHAP and NHATS
• Creating DDI-L Metadata for the NSHAP Series
• DDI-L Metadata for Cross-Series Comparisons
• The NACDA-ICPSR Colectica Portal
• Challenges
• Next Steps
Project Background

- Fall 2018 - March 2019
- The National Archive of Computerized Data on Aging (NACDA) was interested in enhancing longitudinal data availability for data users and began seeking opportunities with DDI Lifecycle
- A Colectica portal was considered value-added to our current process
DDI-L for Longitudinal Data at ICPSR

• Aimed to create and display DDI Lifecycle documentation for the National Social Life, Health, and Aging Project (NSHAP), chosen for
  • Longitudinal approach – 3 rounds of data between 2005-2016 (fourth round pending)
  • U.S. Nationally representative study with cognition and health indicators of Alzheimer's and dementia
  • Data materials were already curated, preserved, and distributed by the NACDA topical archive at ICPSR, as well as DDI-Codebook materials and some existing variable crosswalk efforts
  • Support by the National Institute on Aging
  • Served as proof of concept for extending the use of DDI-L to other NACDA studies

Ultimately, we expanded to other projects 😊

NACDA
About the National Social Life, Health and Aging Project (NSHAP)

• Conducted by NORC, and a team of investigators at the University of Chicago
• Longitudinal study of community-dwelling older people, aged 57 - 85 at the time of recruitment
• Study of health and social factors - Examines the interaction among physical health and illness, use of medication, emotional health, cognitive and sensory function, social connectedness, sex life, and quality of relationships
• Data collected via In-home face-to-face interviews; In-home collected biomeasures; Leave-behind questionnaires (respondent-administered)
• Core NSHAP data, public version available for access through the portal
• Approx. 2,100 variables across the 3 rounds
About the National Health and Aging Trends Study (NHATS)

- Began in 2011, ongoing
- Conducts annual in-person interviews with a nationally representative sample of Medicare beneficiaries ages 65 or older
- 8 Rounds of Data, approx. 11,800 variables
- NHATS has their own Colectica portal, allowing for comparison even though ICPSR does not distribute the NHATS data
- Chosen for similar reasons as NSHAP, and would be compared to NSHAP
Creating DDI-L Metadata for Cross-Study Concordances

• The Data Documentation Initiative (DDI)
  • Widely used metadata standard for describing data from the social, economic and behavioral sciences
  • Two development lines
    • **DDI-Codebook (DDI-C)**
      • Best used to describe individual, standalone datasets
      • Routinely generated at ICPSR for study, and variable descriptions
    • **DDI-Lifecycle (DDI-L)**
      • Includes support for grouping studies and documenting variable comparability across datasets.
      • Currently not supported by ICPSR’s automated curation processes
Creating DDI-L for the NSHAP Series

• Source metadata for creating DDI-Lifecycle
  • DDI-C (V 2.5) variable descriptions for each individual wave
    • Already available - generated in an automated way by ICPSR’s data processing systems
    • Clean, verified documentation
    • Included question text, notes and variable groups
  • Variable lists in the SPSS setups
  • Original codebook information regarding cross-wave variable comparability
    • Valuable, but inconsistent or incomplete throughout the original documentation
• The Colectica suite tools (Designer, Repository and Portal) were used for completing this project
Creating DDI-L for the NSHAP Series

Steps to DDI-Lifecycle:

• Convert pre-existing DDI-C (V2.5) for individual waves to DDI-L using Colectica Designer

Imported (i.e., converted) fields:

- Variable groups
- Variable name
- Variable label
- Summary statistics
- Values
- Value labels
- Category statistics
- Missing data information
- Question text
- (Descriptive text)

Variable notes, if present in DDI-C, are *not* imported (they are not part of the variable description in DDI-L)
Creating DDI-L for the NSHAP Series

Complete study description fields in Colectica Designer

- Easy and convenient editing
- Entries follow the DDI-L structure
- Allows content reuse:
  - Searches pre-existing metadata created in the Designer
  - Creates and saves new content in the Designer
Creating DDI-L for the NSHAP Series

Create cross-wave variable comparison metadata in DDI-L

- DDI-L allows grouping and linking comparable items, adding value for cross-dataset comparison and harmonization

**Variables**

- ATTEND
- SOCIAL
- FAMRELY
- FAMOPEN
- SAMEBED
- HEALTH

**Series: National Social Life, Health, and Aging Project (NSHAP)**

- NSHAP Wave 1 2005-2006
- NSHAP Wave 2 2010-2011
- NSHAP Wave 3 2015-106
- Wave 4 ...

- ATTEND
- SOCIAL
- FAMRELY2
- FAMOPEN2
- SAMEBED2
- SAMEBED3
Creating DDI-L for the NSHAP Series

• Challenge: create a cross-wave variable mapping consistent with DDI-L structure

• Example: variables FAMOPEN and FAMOPEN2

  **Question:** How often can you open up to members of your family if you need to talk about your worries?

<table>
<thead>
<tr>
<th>Wave 1: FAMOPEN</th>
<th>Wave 2: FAMOPEN2</th>
<th>Wave 3: FAMOPEN2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0   if volunteered – no family</td>
<td>0   never</td>
<td>0   never</td>
</tr>
<tr>
<td>1   hardly ever (or never)</td>
<td>1   hardly ever or rarely</td>
<td>1   hardly ever or rarely</td>
</tr>
<tr>
<td>2   some of the time</td>
<td>2   some of the time</td>
<td>2   some of the time</td>
</tr>
<tr>
<td>3   often</td>
<td>3   often</td>
<td>3   often</td>
</tr>
<tr>
<td>4   if volunteered – no family</td>
<td>4   if volunteered – no family</td>
<td></td>
</tr>
</tbody>
</table>
Creating DDI-L for the NSHAP Series

Mapping comparable items in DDI-L

- Traditional crosswalk: suggests comparability
  - All comparable items entered on the same row

<table>
<thead>
<tr>
<th>Wave 1</th>
<th>Wave 2</th>
<th>Wave 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOLUNTEER</td>
<td>VOLUNTEER</td>
<td>VOLUNTEER</td>
</tr>
<tr>
<td>SONS</td>
<td>SONS</td>
<td></td>
</tr>
<tr>
<td>FAMOPEN</td>
<td>FAMOPEN2</td>
<td>FAMOPEN2</td>
</tr>
<tr>
<td>SAMEBED</td>
<td>SAMEBED2</td>
<td>SAMEBED3</td>
</tr>
<tr>
<td>BREAKBONE</td>
<td>BREAKBONE3</td>
<td></td>
</tr>
</tbody>
</table>
Creating DDI-L for the NSHAP Series

Mapping comparable items in DDI-L

• DDI-L variable cascade: documents degree of comparability:

   DDI-L Conceptual Variable
   Concept: How often can R open up to family

   **DDI-L Represented Variable**
   Representation A:
   0    if volunteered – no family
   1    hardly ever (or never)
   2    some of the time
   3    often

   **DDI-L Represented Variable**
   Representation B:
   0    never
   1    hardly ever or rarely
   2    some of the time
   3    often
   4    if volunteered – no family

   **DDI-L Instance Variable**
   W1: FAMOPEN
   **DDI-L Instance Variable**
   W2: FAMOPEN2
   **DDI-L Instance Variable**
   W3: FAMOPEN2

   Data harmonization needed
   Are directly comparable
Creating DDI-L for the NSHAP Series

- Document comparability across waves in DDI-L
  - Compare variable lists to identify which questions were repeated, and in which wave(s)
  - Identify changes in variables representation across waves
  - Create represented variables for different representations
  - Create conceptual variables to link comparable represented variables
  - Create conceptual variable groups to assist in browsing by topic

DDI-L variable comparison: a cascading structure

<table>
<thead>
<tr>
<th>Conceptual Var. Group</th>
<th>Conceptual Variable</th>
<th>Represented Variable</th>
<th>Instance Variable W1</th>
<th>Instance Variable W2</th>
<th>Instance Variable W3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support from Family and Friends</td>
<td>C_FAMOPEN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R_FAMOPEN</td>
<td>FAMOPEN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R_FAMOPEN2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FAMOPEN2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FAMOPEN2</td>
</tr>
</tbody>
</table>
Creating DDI-L for the NSHAP Series

- NSHAP Waves 1-3 now publicly available on the portal
- “Browse by concept” view and initial crosswalk view
Creating DDI-L for the NSHAP Series

Correspondence Tree

- CON_FAMOPEN
  - how often can R open up to family?
  - REP_FAMOPEN
    - National Social Life, Health, and Aging Project (NSHAP): Wave 1, [United States], July 2005-March 2006 - DS1: Core Data - FAMOPEN
  - REP_FAMOPEN2
    - National Social Life, Health, and Aging Project (NSHAP): Wave 3, [United States], 2015-2016 - DS1: Core Data - FAMOPEN2
    - National Social Life, Health, and Aging Project (NSHAP): Wave 2 and Partner Data Collection, [United States], 2010-2011 - DS1: Core Data - FAMOPEN2
Creating DDI-L for the NSHAP Series

### Code Comparison View

<table>
<thead>
<tr>
<th></th>
<th>NSHAP Wave 1</th>
<th>NSHAP Wave 2</th>
<th>NSHAP Wave 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>If volunteered – no family</strong></td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Never</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Hardly ever (or never)</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Hardly ever or rarely</strong></td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Often</strong></td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

### Statistics View

<table>
<thead>
<tr>
<th></th>
<th>NSHAP Wave 1</th>
<th>NSHAP Wave 2</th>
<th>NSHAP Wave 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>If volunteered – no family</strong></td>
<td>0.64%</td>
<td>1.10%</td>
<td>0.0%</td>
</tr>
<tr>
<td><strong>Never</strong></td>
<td>4.90%</td>
<td>3.80%</td>
<td></td>
</tr>
<tr>
<td><strong>Hardly ever (or never)</strong></td>
<td>14.21%</td>
<td>12.02%</td>
<td>10.64%</td>
</tr>
<tr>
<td><strong>Hardly ever or rarely</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Some of the time</strong></td>
<td>38.47%</td>
<td>34.52%</td>
<td>38.83%</td>
</tr>
<tr>
<td><strong>Often</strong></td>
<td>46.68%</td>
<td>47.36%</td>
<td>46.73%</td>
</tr>
</tbody>
</table>

The three types of concordance views are also available from the individual variable description page.
DDI-L Metadata for Cross-Series Comparison
The NSHAP-NHATS Crosswalk

Steps

- Starting point: DDI-L documentation for each individual series, including variable cascades
  - Comparable variables from each series are already linked by a common conceptual variable

- Challenge: find comparability among variables from 11 individual waves (3 from NSHAP and 8 from NHATS)

  DDI-enabled solution:
  - Examine, and find matches between conceptual variables from each series
  - Create new, unifying concept (DDI conceptual variable) that links the conceptual variables from the individual series
    - Project-created and assigned
    - Allows pulling together all of the comparable (instance) variables from the individual waves.
DDI-L Metadata for Cross-Series Comparison

Matching DDI-L concepts to compare across series

Cross-series conceptual variable
HYPERTENSION

NSHAP Conceptual Variable
EVER HAD HYPERTENSION

NHATS Conceptual Variable
HAS HIGH BLOOD PRESSURE

NSHAP Represented Variable
HYPERTENSION

NHATS Represented Variable
High BP

NSHAP Wave 1
Hypertension

NSHAP Wave 2
Hypertension

NSHAP Wave 3
Hypertension

NHATS Wave 1
High BP

NHATS Wave 2
High BP

NHATS Wave 3
High BP

NHATS Wave 8
High BP
## DDI-L Metadata for Cross-Series Comparison

<table>
<thead>
<tr>
<th>NSHAP Variables</th>
<th>NHATS Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Round 1 Core Data</strong></td>
<td><strong>Round 2 Core Data</strong></td>
</tr>
<tr>
<td>Self-rated physical health</td>
<td>PHYSHLTH</td>
</tr>
<tr>
<td>Arthritis</td>
<td>CONDITNS_1</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>OSTEOP</td>
</tr>
<tr>
<td>Emphysema or COPD</td>
<td>CONDITNS_3</td>
</tr>
<tr>
<td>Stroke</td>
<td>CONDITNS_5</td>
</tr>
<tr>
<td>Hypertension</td>
<td>CONDITNS_6</td>
</tr>
</tbody>
</table>

Series-specific conceptual and represented variables and comparison views available when selecting the common concept.
DDI-L Metadata for Cross-Series Comparison

- Four types of variable comparability
  - Directly comparable
  - Need harmonization
  - Related concepts
  - One-to-many or many-to-one matches
    - Example:

<table>
<thead>
<tr>
<th>NSHAP</th>
<th>NHATS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visual acuity</strong></td>
<td><strong>EYESIGHT</strong></td>
</tr>
<tr>
<td>Can see across the street</td>
<td>ss1seewellst ss2seewellst ss3seewellst ss4seewellst ss5seewellst</td>
</tr>
<tr>
<td>Can watch TV from across room</td>
<td>ss1seestvglss2seestvglss3seestvglss4seestvglss5seestvglss</td>
</tr>
<tr>
<td>Can read newspaper print</td>
<td>ss1glrednewp ss2glrednewp ss3glrednewp ss4glrednewp ss5glrednewp</td>
</tr>
</tbody>
</table>
NACDA Portal – Explore Our Comparisons

- NHATS Concept System
- NSHAP-MIDUS Cross Series Comparison
- MIDUS Concept System
- NSHAP Concept System
- NSHAP-NHATS Cross Series Comparison
Cross Series Comparisons – Defined in the Portal

Cross Series Comparisons Information

Where present, the "i" symbol in the black circle indicates the comparability notes for each conceptual variable in the Cross Series Comparisons.

Match Categories:

**Directly Comparable** 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.

**Need Harmonization** 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.

**One to Many or Many to One** matches are those which 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. The alternating shading in the table view of the conceptual variables this portal, in addition to the comparability notes, is used to differentiate between the single matches (one-to-one) and one-to-many, or many-to-one matches.

**Related Concepts** 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.

For more information about our efforts to compare the NSHAP and NHATS collection, please refer to our working paper, "Identifying Cross Series Cognitive Data Similarities Across NSHAP and NHATS".

Please note that the comparability notes in the NACDA portal Cross Series Comparisons differ from the comparability notes the MIDUS portal utilizes. The Cross Series Comparison comparability notes reflect the comparability across variables across the different data series to highlight the potential use of variables from different series in the same data analysis.
Challenges

• Thousands of variables to review

• Creating new conceptual groups instead of using those currently existing within NSHAP and NHATS

• Determining comparability effectively

• Visualizing the results for secondary users
Identifying Cross Series Cognitive Data Similarities Across NSHAP and NHATS

McNally, James; Lavender, Kathryn; Ionescu, Sanda; Koteles, Brendan
Next Steps

• Expand Cross Series Comparison efforts through collaboration within ICPSR and with other data projects

• Assist in standardizing concept groups, where applicable, such as at the project level with agencies like CLOSER and MIDUS

• Increase training and guidance for longitudinal data development
Thank You!

Questions? Email us: sandai@umich.edu, kfrania@umich.edu