Data and Metadata Harmonization for the RAND Survey Meta Data Repository

Alerk Amin

April 3, 2013

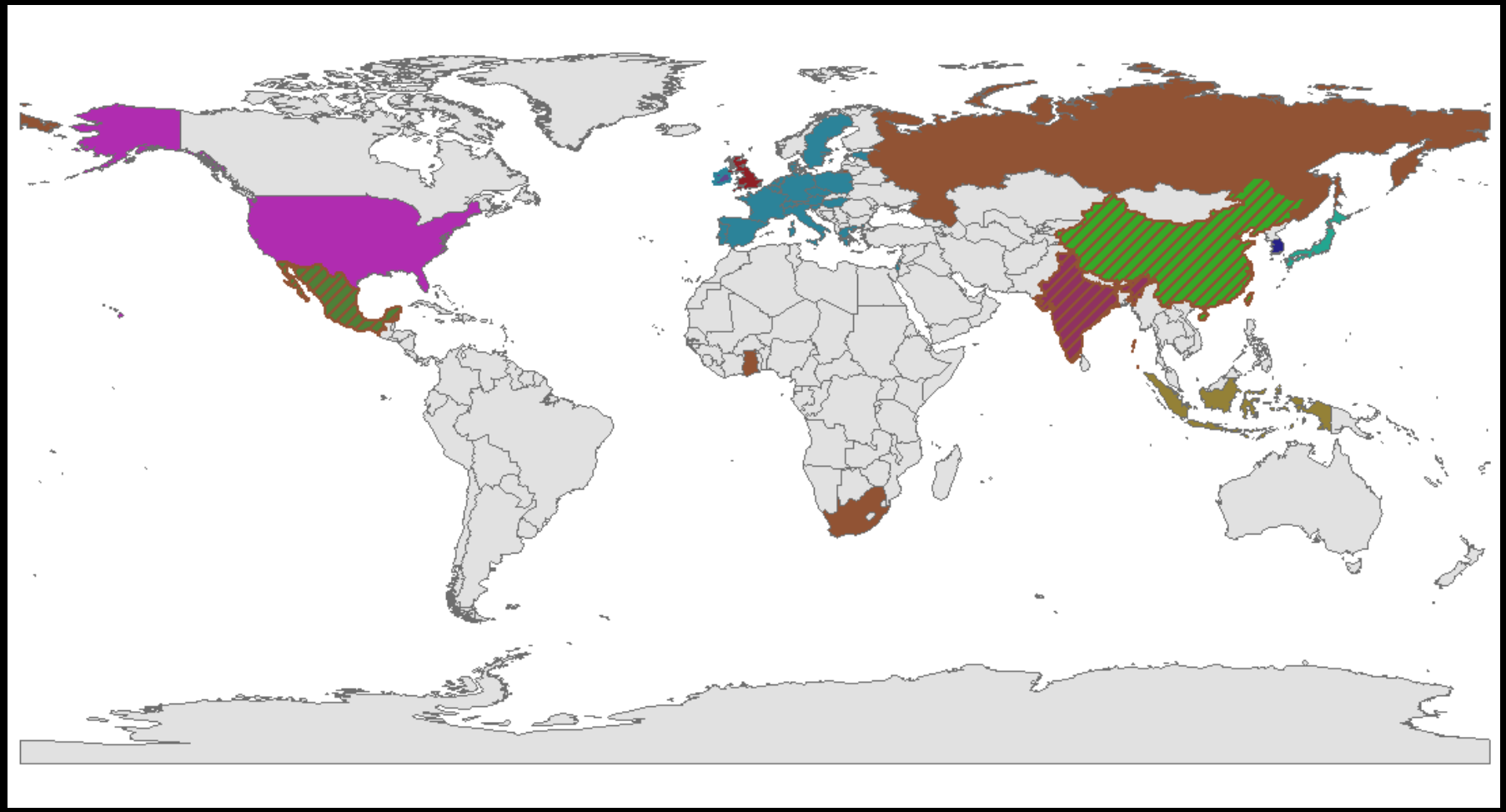
Outline

- About the studies
- RAND work
  - Repository Website
  - Import metadata
  - RAND harmonized datasets
- How DDI could help

Health and Retirement Study

- University of Michigan
- Longitudinal panel study of Americans age 50+
- Biennially since 1992

HRS Worldwide "Family"



HRS "family"

- Health and Retirement Study (HRS) since 1992
- Mexican Health and Aging Study (MHAS) since 2001
- English Longitudinal Study of Ageing (ELSA) since 2002
- Survey of Health, Ageing and Retirement in Europe (SHARE) since 2004
- Korean Longitudinal Study on Aging (KLoSA) since 2006
- Japanese Study of Aging and Retirement (JSTAR) since 2007
- Indonesian Family Life Survey (IFLS), modified in 2007
- China Health and Retirement Study (CHARLS) since 2008
- Irish Longitudinal study on Ageing (TILDA) since 2010
- Longitudinal Aging Study in India (LASI) since 2010
- Study on Global Ageing and Adult Health (SAGE), since 2010

Repository Goals

- **Make it easier for researchers to**
  - **Find appropriate data across the different studies**
  - **Cross-country analysis**

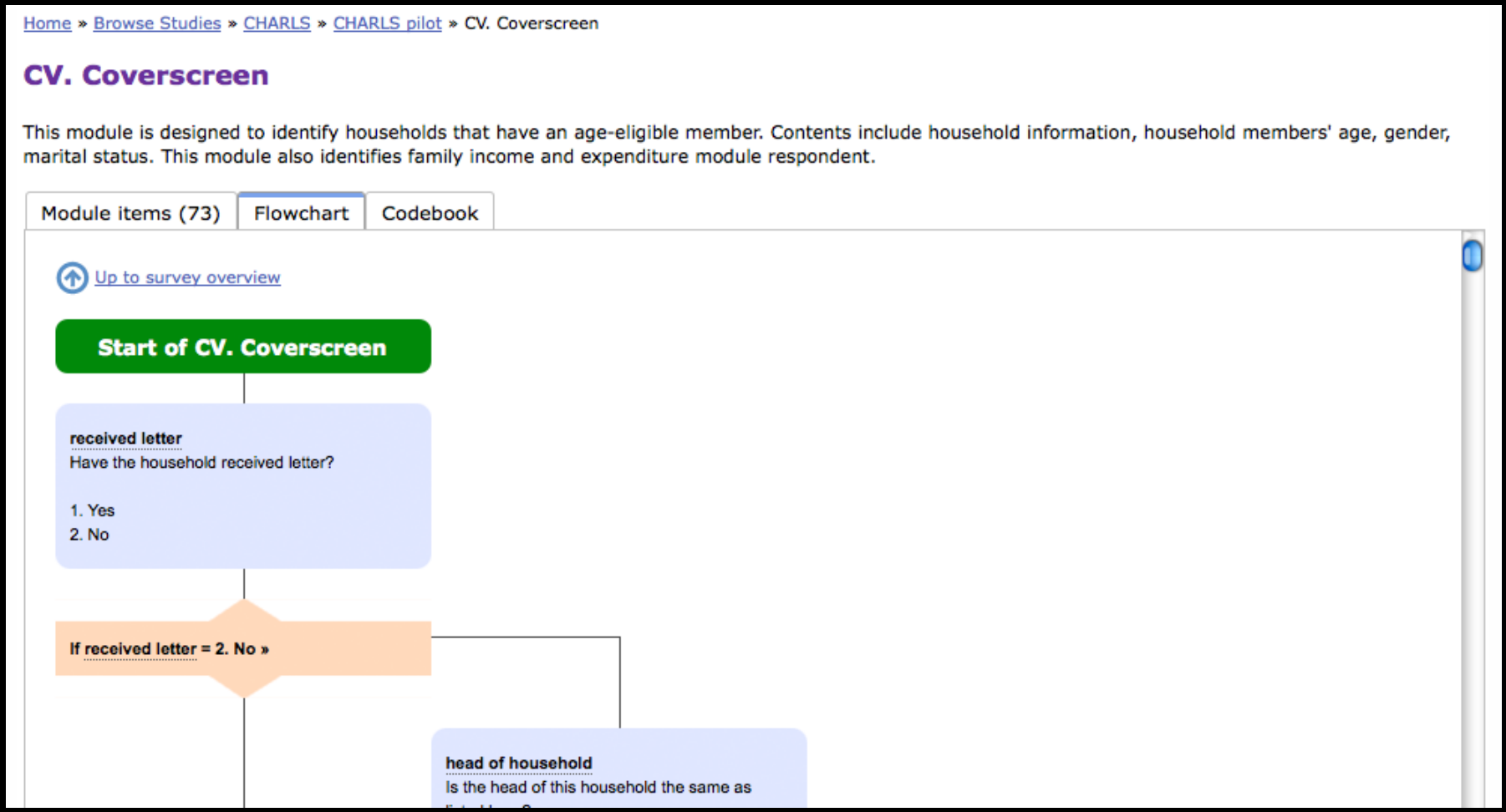
Repository Item

Home » Browse Studies » SHARE » SHARE 2006 » EP-Employment and pension » EP005_CurrentJobSit

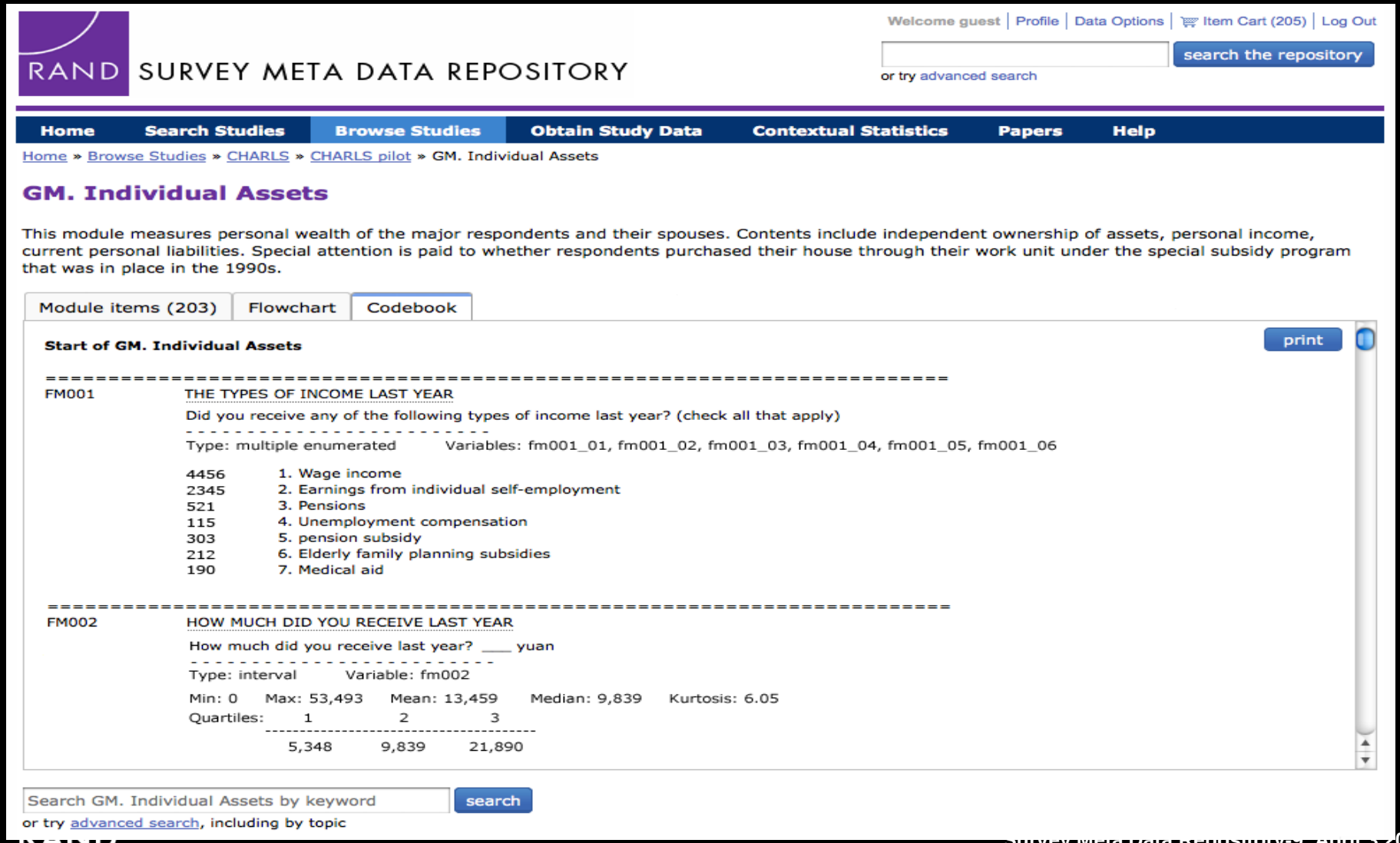
EP005_CurrentJobSit

Location:	SHARE » SHARE 2006 » EP-Employment and pension
Description:	CURRENT JOB SITUATION
Item type:	Question
Question text:	Please look at card 21. In general, how would you describe your **current** situation? IWER:CODE ONLY ONE
Answer type:	Enumerated
Answer choices:	1. Retired 2. Employed or self-employed (including working for family business) 3. Unemployed and looking for work 4. Permanently sick or disabled 5. Homemaker 97. Other (Rentier, Living off own property, Student, Doing voluntary work)
Associated Variable(s):	ep005_ (current job situation)
Topics:	disabled, employed, employee, looking after home or family, retired, retiree, self-employee, work status
Translations:	English (Generic)
Researchers were also interested in:	FM002_a MONEY GM002 WHY DIDN'T YOU PURCHASE GM014 AT WHAT PRICE DID YOU SELL IT ?
User comments:	Be the first to comment!

Repository Questionnaire Routing



Repository Codebook



RAND SURVEY META DATA REPOSITORY

Welcome guest | Profile | Data Options | Item Cart (205) | Log Out

search the repository

or try advanced search

Home | Search Studies | Browse Studies | Obtain Study Data | Contextual Statistics | Papers | Help

Home » Browse Studies » CHARLS » CHARLS pilot » GM. Individual Assets

GM. Individual Assets

This module measures personal wealth of the major respondents and their spouses. Contents include independent ownership of assets, personal income, current personal liabilities. Special attention is paid to whether respondents purchased their house through their work unit under the special subsidy program that was in place in the 1990s.

Module items (203) | Flowchart | Codebook

print

Start of GM. Individual Assets

FM001 THE TYPES OF INCOME LAST YEAR

Did you receive any of the following types of income last year? (check all that apply)

Type: multiple enumerated Variables: fm001_01, fm001_02, fm001_03, fm001_04, fm001_05, fm001_06

4456	1. Wage income
2345	2. Earnings from individual self-employment
521	3. Pensions
115	4. Unemployment compensation
303	5. pension subsidy
212	6. Elderly family planning subsidies
190	7. Medical aid

FM002 HOW MUCH DID YOU RECEIVE LAST YEAR

How much did you receive last year? ____ yuan

Type: interval Variable: fm002

Min: 0 Max: 53,493 Mean: 13,459 Median: 9,839 Kurtosis: 6.05

Quartiles:	1	2	3
	5,348	9,839	21,890

Search GM. Individual Assets by keyword search

or try advanced search, including by topic

Metadata Harmonization

- **Metadata import from each study**
- **Create links across studies to aid researchers**

Importing Survey Metadata

- **Only IFLS provides DDI (v 1.2.2) metadata**
- **Studies using MMIC can be imported via scripts**
- **Other studies involve a lot of work by hand**
  - **Sometimes CSV files are available for import**
  - **Otherwise, cut-and-paste from codebook**
  - **Routing is entered by hand, based on codebook**

Linking Questions to Variables

- **Variables are imported via Stata files**
- **Questions are linked to Variables**
  - **usually via CSV import**

Linking Questions to Concepts

- **RAND Working Paper Series, WR861-1/7**
- **A set of domain-specific user guides on**
  - Chronic medical conditions
  - Financial transfer
  - Expectation
  - Employment and retirement
  - Income
  - Wealth
  - Cognition

Linking Questions to Concepts

- **One wave of concepts is available from working papers**
- **Concept-Questions are entered by hand**
- **More work is done to then link other waves**

How DDI Could Help

- **If studies provided metadata in DDI3 format**
  - **Import Questions, Variables, Routing**

Importing Published DDI

- **Other studies copy concepts, questions from HRS**
- **IDEAL - all studies link Questions to a common ConceptScheme**
- **REALISTIC – all studies link Questions to HRS Questions**

Data Harmonization

- **RAND creates harmonized datasets**
  - **RAND HRS, RAND ELSA, RAND SHARE, RAND KLoSA**
  - **Coming soon – RAND CHARLS, RAND JSTAR**
- **Identical/Comparable set of variables across studies**

RAND HRS, RAND ELSA, RAND SHARE, RAND KLoSA: Harmonized variables for cross-country, longitudinal study

Domains	Variables
Identifiers, weights	Person specific identifier; household identifier; couple identifier; spouse identifier; wave status: response indicator; wave status: interview status; sample cohort; whether eligible for sample; sampling weight; person-level analysis weight; household analysis weight (not available for ELSA); country;
Demographics	Number of household respondents; whether couple household; financial respondent; family respondent; whether proxy interview; interview dates; birth date; age at interview (in months and years); place of birth (not available for ELSA); gender; race; education: years of education; education: categorical summary; current marital status: with partnership; current marital status: without partnership; number of marriage; marital history: never married; marital history: number of times divorced; marital history: number of times widowed; marital history: number of times don't know how marriage ended; length of current marriage; length of longest marriage; religion (not available for ELSA); place of birth (not available for ELSA); parental mortality: mother alive; parental mortality: father alive; parental mortality: mother's current age or age at death; parental mortality: father's current age or age at death
Health	Self-report of health; whether health limits work; activities of daily living (ADLs): some difficulty; instrumental activities of daily living (IADLs): some difficulties; other functional limitations: raw recode; ADL summary: sum ADLs where respondent reports any difficulty; IADL summary: sum IADLs where respondent reports any difficulty; other summary indices: mobility, large muscle, gross fine motor activities; mental health (CESD score); doctor diagnosed health problems: ever have condition; doctor diagnosed health problems: memory-related disease; health behaviors: physical activity or exercise; health behaviors: drinking; health behaviors: smoking (cigarettes); change in health: self-reported health; change in health: functional limitations; change in health: conditions; change in health: memory-related disease;
Financial and Housing Wealth	Net value of business; value of primary residence; value of all mortgage (primary residence); net value of primary residence; net value of real estate; net value of cars; net value of stocks, mutual funds, and investment funds; value of checking, savings, or money market accounts; net value of bonds and bond funds; net value of non-housing financial wealth; total family wealth (respondent & spouse)
Income	Individual earnings; income from employer pension or annuity; individual income from public pension; individual unemployment benefits or workers compensation (not available for ELSA); family capital income; family government transfer income; total family income (respondent & spouse)
Family structure	Number of people living in household; number of children; number of living siblings; number of living parents
Employment history	Currently working for pay; whether self-employed; labor force status; hours of work per week at current job; weeks worked per year at job; wage rate; level of physical effort at current job; years of tenure on current job; occupation code for current job; month and year last job ended

RAND Harmonized Datasets

- One data file per wave
- “fat” format with 1 row per respondent
- Respondent, Spouse and Household variables
- Respondent ID consistent across waves

RAND-enhanced Fat Files For Each Interview Year

HOUSEHOLD LEVEL
Preload
Coverscreen (CsR)
Family Structure (FamR)
Housing (FinR)
Assets and Income (FinR)
Asset Change (FinR)

8 Files

RESPONDENT LEVEL
FAT FILE FOR 1 YEAR

24 Files

RESPONDENT LEVEL
Preload
Coverscreen
Demographics
Physical Health
Cognition
Parents and Siblings
Functional Limitations
Physical Measures
Employment (1-6 sections)
Health Services and Insurance
Expectations
Widowhood and Divorce
Wills and Life Insurance
Experimental Modules
Event History and Social Security
Interviewer Observations

Current Process – New Study

- **The RAND HRS is the “baseline”**
- **For each variable in RAND HRS**
  - **Find the corresponding variable in the new study**
  - **If the variable is exactly equivalent**
    - **Create a variable with the “same” name**
  - **If the variable is not exactly equivalent**
    - **Create a related variable and document the differences**

Current Process

- For each new wave of data
  - All variables are compared with previous waves to see if they are the same or different
  - If variable is the same
    - Create a variable with the “same” name
  - If variable is different
    - Create a related variable and document the differences

Codebook Sample

How Constructed:

RwSHLT is the respondent's self-reported general health status. Codes range from 1 for Excellent to 5 for Poor. SwSHLT is the respondent's spouse or partner's self-reported general health status.

RwSHLT is assigned the value of the raw variable except that missing values for don't know, refused, and other missings are recoded to .D, .R, and .M, respectively.

RwSHLT and SwSHLT are used in contruction of a change in health variable RwSHLTC. Please see "Change in Health" for a description of these measures.

The SwSHLT variables are taken from the Wave 'w' spouse's self-reported RwSHLT variables.

Cross Wave Differences in Original HRS Data

In Wave 1 values for self-reported health status are imputed by HRS if missing. These imputations are used. From Wave 2 forward, values are not imputed by HRS.

If Studies Distributed DDI…

- **Harmonization would be much faster**
- **Studies are based on HRS**
  - **Their Questions would already reference HRS, and describe how their questions are different**
  - **Reference HRS Concepts**
- **Would make it much easier for RAND to evaluate variables and determine if data is comparable**

If RAND Distributed DDI…

- **Currently**
  - **distribute a paper codebook**
  - **use variable names to show identical/related variables**
- **With DDI**
  - **RAND distributes structured metadata describing the differences in variables across studies**
  - **To researchers or back to original studies**

Conclusion

- **If studies published DDI**
  - **Metadata and data harmonization would be faster and easier**
- **If RAND published DDI**
  - **Other researchers/agencies could make better use of RAND's knowledge of the HRS family of studies**


RAND
CORPORATION