How to Use the Health and Retirement Study

A simple researcher's primer

Part 1

Ryan D. Edwards

redwards@qc.cuny.edu

Queens College, City University of New York & NBER
Visiting UC Berkeley Demography 2012-2013

March 22, 2013

Broad outline

- The Health and Retirement Study:
  - What is it?
  - What's in it?

} Tell

  - How do you get access to it?
  - How do you work it?

} Show

Thanks

- Many researchers far more wise than I have spent far more time and energy on HRS to make it what it is

- I have drawn heavily from the HRS's online summary, *Growing Older in America: The Health & Retirement Study*

- The most thanks are owed to the over 30,000 respondents who have donated their time, earnings histories, biological samples, and much more to HRS

- Thanks to the RAND team for their efforts to clean, distribute, and document user-friendly versions of HRS

- My own special thanks to Alice Zulkarnain, David Weir, Heidi Guyer, Cindi Leacock, Luis Rosero-Bixby, Will Dow, and Amal Harrati for many insights

Warning label

- I'm just a researcher who uses HRS, not a statistician nor a true HRS expert

- I first started using HRS here at Berkeley in 2000 for a dissertation paper that was published in 2008, and I've used it since

- I sincerely hope I'm not leading anyone astray; if I am, it's purely an accident, but at least you'll be making the same mistakes I'm making

- Any mistakes in this presentation and hands-on are my own

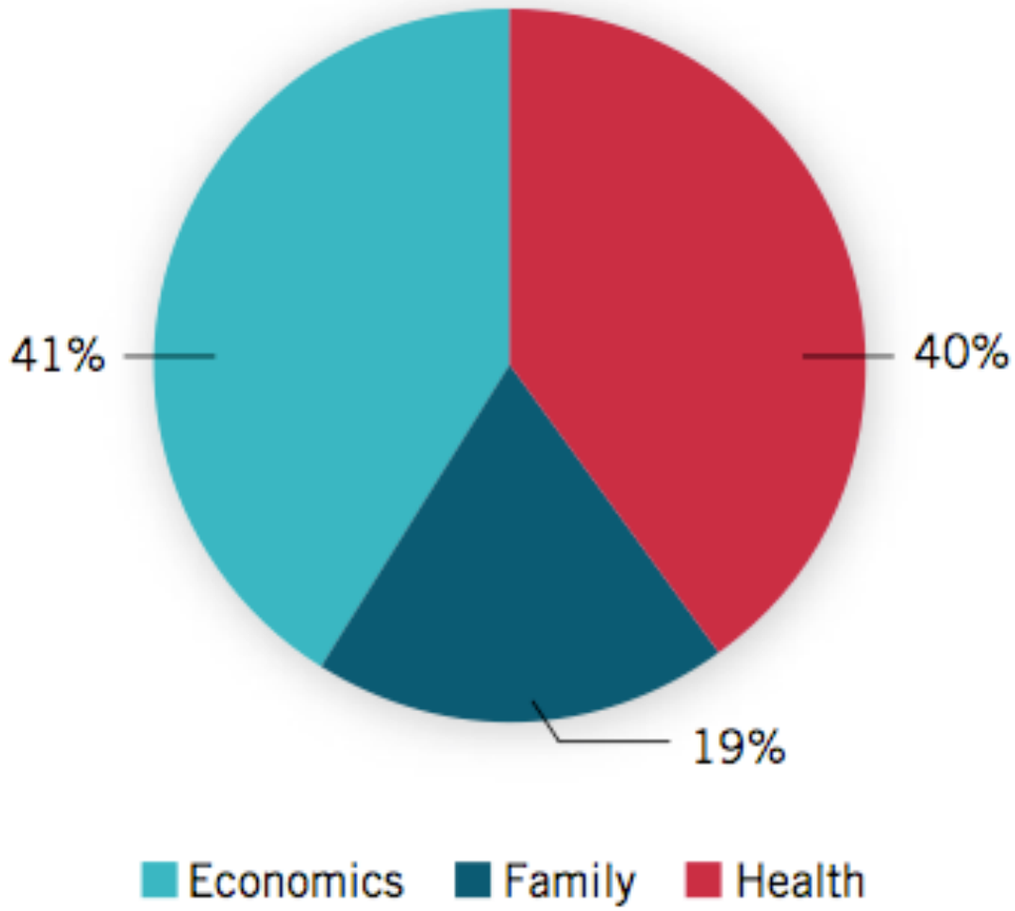
- I encourage you to use HRS with an appropriate degree of careful skepticism about how you have constructed the data

HRS: What is it and what's in it?

- A biennial (every other year) panel survey that is nationally representative of Americans aged 50+, since 1992 or 1998
- Each even-year core survey wave includes about 18,000 people in perhaps 11,000 households, primarily interviewed by telephone ...
- Conducted by the Institute for Social Research at the University of Michigan, funded in large part by the National Institute on Aging
- Intended to be an interdisciplinary study for economists, sociologists, psychologists, epidemiologists, demographers, biomedical researchers
  - Causes and consequences of retirement
  - Health, income, and wealth dynamics and interrelationships
  - Life-cycle patterns of saving, disability, family, etc.

HRS: What is it and what's in it?

FIG. A-2
THE ALLOCATION OF HRS INTERVIEW TIME BY BROAD TOPIC

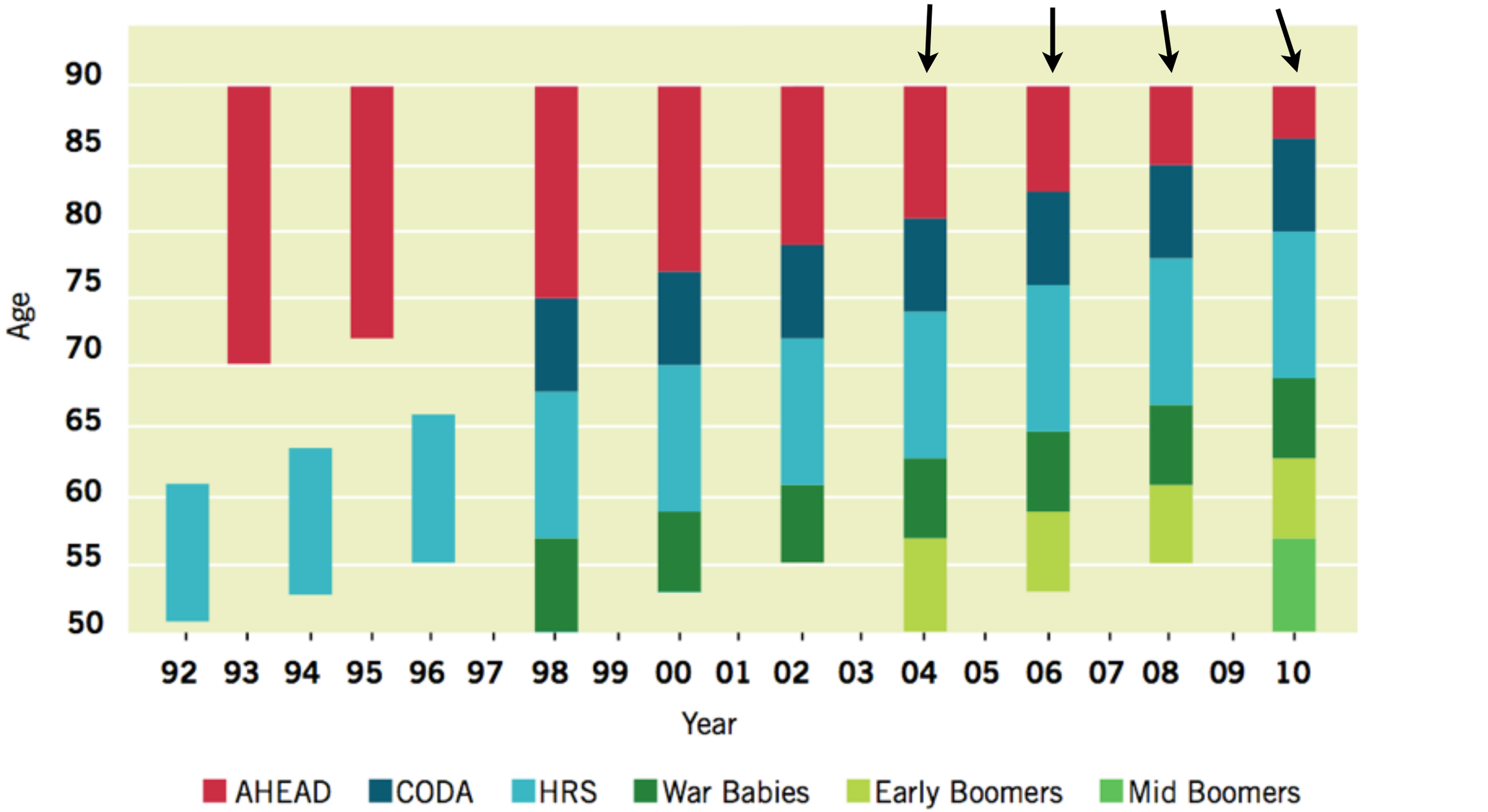


- Economic circumstances
- Occupations and employment
- Health and health care
- Cognition
- Living and housing arrangements
- Demographics and family relationships

A fragmented sample structure prior to 1998, with introductions of new cohorts every fourth wave since

FIG. A-3

THE HRS LONGITUDINAL SAMPLE DESIGN



Representativeness, oversampling

- HRS oversamples African Americans, Hispanics, and Floridians
- Like other surveys, HRS initiates contact *only with non-institutionalized individuals* not in prisons, jails, nursing homes, long-term care facilities
- Once in the study, HRS follows respondents in and out of nursing care. (Jail? The country? Apparently no offenders; 10% are foreign born ...)
- So over time, the sample will become representative of folks in and out of nursing homes
- Weir (2010 PAA paper), Adams et al. (J Econometics 2003): mortality surveillance is "essentially complete" after the first few years of the study

Sample weights

- Starting with the redesign in 1998, HRS has produced weights for all waves by post-stratifying each wave's weights to the March CPS using birth year, sex, and race/ethnicity

- Two types of weights for each wave: household and respondent

- Universe is civilian noninstitutionalized; nursing home residents, the dead, and nonresponders all get 0 weights

- A third type of weight, for nursing home residents, only available in 2000 and 2002 waves; unclear but probably not post-stratified

- Weights for longitudinal analysis? Choice of starting, terminal, other

Imputations

- For questions about financial wealth, HRS asked about (a) ownership, (b) values, then (c) unfolding brackets of value if needed

(a) "Do you have any shares of stock or stock mutual funds?"

(b) "If you sold all those ... about how much would you have?"

(c) "Would it amount to less than $X, more than $Y, or what?"

X and Y ranges are preset randomly across respondents

- HRS has used these brackets to impute values for each wave
- Since the 2006 wave, RAND has released imputations for all waves using a consistent method
- Some other types of responses are also imputed; be cautious

Unit of observation

- The HRS sampling unit is the household (variable name: "**hhid**")
  - Within the household are age-eligible respondents and any spouses, who are measured whether age-eligible or not
- In the RAND file and in many of the HRS data files, the unit of observation — i.e., each row — is the individual
  - The respondent and the spouse each gets his or her own row in the data file, with an individual identifier **hhidpn**, a concatenation of **hhid** and a person number **pn**
  - Beware: within such data files, household-level variables like assets and wealth are often duplicated across spouses

The RAND HRS Data File

- With funding from NIA, the RAND Corporation has produced a consistently measured longitudinal file with much core HRS content

  - Benefits: Consistent measures; Missing data from "don't know" or "refused" etc. are coded conveniently; Data labels help clarify values

  - Shortcomings: Does not cover all variables; Their cleanup and recodes may or may not be reasonable for all applications

- Use it with some caution; to make it consistent, judgment calls were made that may or may not be 100% right for your specific use

- But I recommend starting with the RAND file and merging in other data as you find necessary

Contents of the RAND HRS Data File

- In the current version L there are 30,671 observations with 8,920 variables for a dataset of 408 MB in Stata
- Arrayed loosely like the HRS questionnaire

Section	Topic	# of variable categories
A	Demographics	46
B	Health, disability, and cognition	48
C	Financial and housing wealth	23
D	Income	10
E	Social Security and disability benefits	13

Section	Topic	# of variable categories
F	Pensions	7
G	Health insurance	9
H	Family structure	5
I	Retirement plans, expectations	21
J	Employment history	20

Overview of some special content areas beyond the core data

1. Mortality

2. Exit interviews and bequests

3. Family structure and the RAND family file

There are many others that I won't discuss!

4. Consumption and Activities Mail Survey (CAMS)

5. Childhood health retrospectives

6. Restricted access files: (a) Social Security, (b) Medicare, (c) geocodes

7. Biomarkers and Genetic Data

1.a Mortality

- HRS documents mortality two ways: (a) tracking each ever-respondent, and (b) matching to the National Death Index (NDI)

- Tracking involves contacting each respondent and arranging an exit interview with next-of-kin for the deceased

- Panel attrition is the clear issue here, but HRS attrition rates are low compared to other surveys like ELSA (Banks, Muriel, and Smith, 2010)

- Through 2006, David Weir (2010) found:
  - Tracking had identified 97.4% of deaths, NDI identified 95.6%
  - Comparisons to life tables suggest that "mortality surveillance is essentially complete in HRS"
  - With 8,000 deaths over 300,000 person-years for 30,000 people, there was enough power to examine SES differentials

1.b Mortality, representativeness, weights

- Each "new" HRS cohort, when first interviewed, is noninstitutionalized
- Over short periods of time, each new cohort's mortality experience will not be representative b/c nursing home residents are excluded
- HRS sample weights are for the civilian noninstitutionalized population, no quick fix (but there are nursing home weights for 2000 and 2002)

- Because life tables vary dramatically by age, sex, and race/ethnicity, use caution when comparing HRS to life tables without some kind of weights
- HRS David Weir has used two methods:
  - Reweighting U.S. life tables to HRS — match age/sex/race life tables to HRS respondents by age/sex/race
  - Reweighting HRS to U.S. — generate new age/sex/race weights using e.g. 2000 population weights from Census or Human Mortality Database

2.a Exit interviews and bequests

- For respondents identified as deceased, HRS conducts exit interviews of proxy respondents, typically widow(er) or next of kin

  - Content is similar to core interviews for living respondents

  - 1,446 deceased respondents covered in 2010 Exit Final

- Post-exit telephone interviews of respondents interviewed in prior exit waves and who had unresolved financial situations (wills, trusts, real estate)

  - 134 deceased respondents in 2010 Post-Exit Proxy Final

2.b Exit interviews and bequests

- Section T of the questionnaire asks about wills, insurance, trusts
  - Value and fate of primary residence, of secondary residence
  - Death expenses
  - Fate of assets and possessions, excluding life insurance
  - Value of assets and possessions, excluding life insurance, whether some is in a trust, who is the trustee
  - Beneficiaries of life insurance
  - Value of life insurance
- These data do not appear to have been filtered and harmonized by RAND or anybody else, but publications exist that use them

3. RAND Family File

- Unlike PSID, children of respondents do not become HRS respondents, but some of their characteristics are measured; same for parents of respondents

- Now in version B, the RAND family file consists of two datasets:

1. Respondent-child file with data on parent-child pairs, where HRS respondents are the parents, their children are the observations (rows)

2. Respondent file with data on each HRS respondent's parents, siblings, and children, where HRS respondents are the observations (rows)

- RAND personnel collected and cleaned these data from a variety of sources in the core and modules & produced these longitudinal files

- I think I may have found some panel inconsistencies with respondents' siblings

4. Consumption and Activities Mail Survey (CAMS)

- Mail survey sent out biennially during off-years to a subset of about 5,000 core respondents, one randomly chosen per household

- RAND file v. B combines data from 5 waves: 2001, '03, '05, '07, '09

- Panel consistency: A total of 5,407 observations in total, of which 2,458 are present both in 2001 and 2009

- Inspired by the U.S. Consumer Expenditure Survey (CEX), with comparable questions

- CAMS also asks about time use by & labor force status of the respondent (randomly chosen if in a couple household), and some questions about spending around retirement, either pro- or retrospective

5. Childhood health retrospective questions

- Starting in 2008, the core survey asks a larger set of retrospective questions about childhood conditions before age 16

  - Primarily focused on childhood health conditions: measles, mumps, diabetes, allergies, etc.

  - Also asks about parents' smoking, own smoking; core has always asked about parental education and other basic characteristics

  - Also asks about learning problems in school, special training

- These data are only in the core files, not the RAND dataset yet

6. HRS restricted and sensitive files

- Social Security earnings history, benefits
- Medicare beneficiary records
- Geocodes for each wave down to ZIP code
- Detailed industry-occupation
- ★ Biomarkers: blood composition (“biomarkers”) & genes (“genetic”)
- Aging, Demographics, and Memory Study (ADAMS)
- 2003 Diabetes Study, 2005 Prescription Drug Study ... and more

Obtaining access to HRS restricted-use files

- Some of these files — SSA earnings, e.g., — require that the PI have federal research funds prior to data use agreement

  - The idea has been that the PI faced the risk of losing future research funds, and that would help insure data security

  - Some PI's have data use agreements that permit research assistants and other collaborators to use the data under specified conditions

- For other files — biomarkers and other "sensitive health" files — requirements are less restrictive but still stringent

  - Application requires a data protection plan; HRS likes to see a standalone, encrypted workstation in a locked single-user office

7.a HRS Biomarkers and Genetic Data

- Starting in 2006, HRS has asked rotating halves of the sample to submit physical measures each wave
  - "You're in HRS? And you can submit biomarkers? Great, flip a coin:
  - "Heads, we ask you to submit biomarkers in 2006, 2010,
    Tails, we ask you to submit in 2008, 2012,

- Physical measures consisted of:
  - Physical capabilities & metrics:
    Balance, walking, breath, grip strength, objective height and weight
  - Blood pressure, pulse, and blood composition analysis:
    A lot like what your doctor measures in your annual physical
  - Saliva sample collection leading to genetic analysis:
    Not at all like what your doctor measures!

7.b Physical capabilities/metrics and Biomarkers

- Many measures are in the HRS core files (but not in the RAND file)
  - Blood pressure, pulse, and all of the physical capabilities & metrics
- 2006 Biomarkers data, which are restricted and require an application, includes 3 measures of blood characteristics for about 6,000 respondents:
  - Hemoglobin A1C, a measure of average blood sugar over several months; high A1C is an indicator of diabetes
  - Total or "bad" cholesterol, which is linked to heart disease & stroke
  - HDL or "good" cholesterol, protective against heart disease & stroke
  - All of these are valid measures even when the respondent hasn't been fasting
  - Other interesting measures, like cortisol, are absent at least for now, maybe because their validity is questionable, I'm not sure

7.c Genetic Data

- Many thanks to Amal Harrati, who is writing her dissertation with these
- 2006-08 Genetic Data, which are restricted and require an application, cover about 12,500 respondents measured in 2006 and 2008
- The dataset is beyond "rich," it's enormous
  - For each respondent, 2.5 million pieces of genetic information called single nucleotide polymorphisms (SNPs)
  - SNPs are specific places along the human genome where variation in humans occurs
  - Per Amal: most SNPs don't do anything "exciting"
  - But in principle, that's up to 2.5 million variables by 12,500 observations, and a dataset of 1.2 terabytes (1,200 gigabytes) that needs to sit on a secure workstation