Harnessing the U.S. Health and Retirement Study for Research in the Social Sciences: A Technical Workshop

Ryan D. Edwards

Division of Data Sciences and
Berkeley Population Center

UC Berkeley Demography Department
September 14, 2018

Thanks

- National Institute on Aging (NIA)

  Center grant 5P30AG012839

  Center on the Economics and Demography of Aging (CEDA)

- Many HRS folks over the years, especially current HRS Director David Weir

- HRS background publications *Aging in the 21st Century* and *The Health and Retirement Study: An Introduction*

Introductions

- Name
- Field or program
- Why HRS is interesting to you
- Have you ever been asked to participate in a survey? Which one? Did you agree

The U.S. Health and Retirement Study

- Not just about health and retirement!

health conditions	marital history	children and grandchild	time use
biomarkers	household structure	friends and social networks	consumption
retirement (expectations)	neighborhood characteristics	own childhood conditions	expectations and beliefs
work history	genetic data	military info and VA records	psychosocial questionnaire
income and wealth	Social Security earnings histories	Medicare records	

Applications in which I’ve used HRS

Expectations of health status and holding stocks vs. bonds

Military service, combat exposure, and later-life health

The “shape” of survivorship expectations and information

Advent of biomarker collection and responses to new information

Smoking and weight over the panel

Pet ownership and physical activity

Implications of Hispanic first name & nativity for health

Socioeconomic gradients in health among Hispanics

Some of these have explicitly used the **panel nature** of the data: repeated observation of individuals

Expectations of health status and holding stocks vs. bonds

The "shape" of survivorship expectations and information

Smoking and weight over the panel

[redacted]

Military service, combat exposure, and later-life health

Advent of biomarker collection and responses to new information

[redacted]

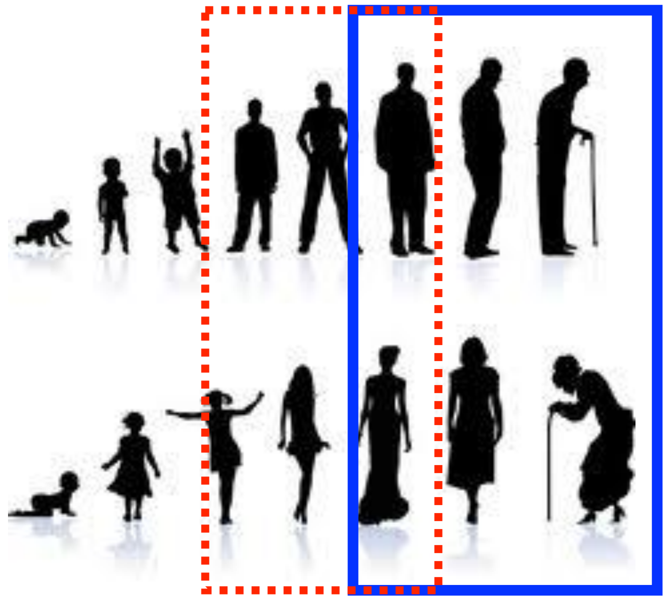
[redacted]

Agenda today

- Brief overview of HRS & variable naming, practical tips
- Tips about finding Topic X in the HRS
- More on contents and scope of HRS
- More practical tips about working with the variables
- A few slides about statistical languages
- A brief tutorial in Stata
- Overview of some recent work on biomarkers and Hispanic names

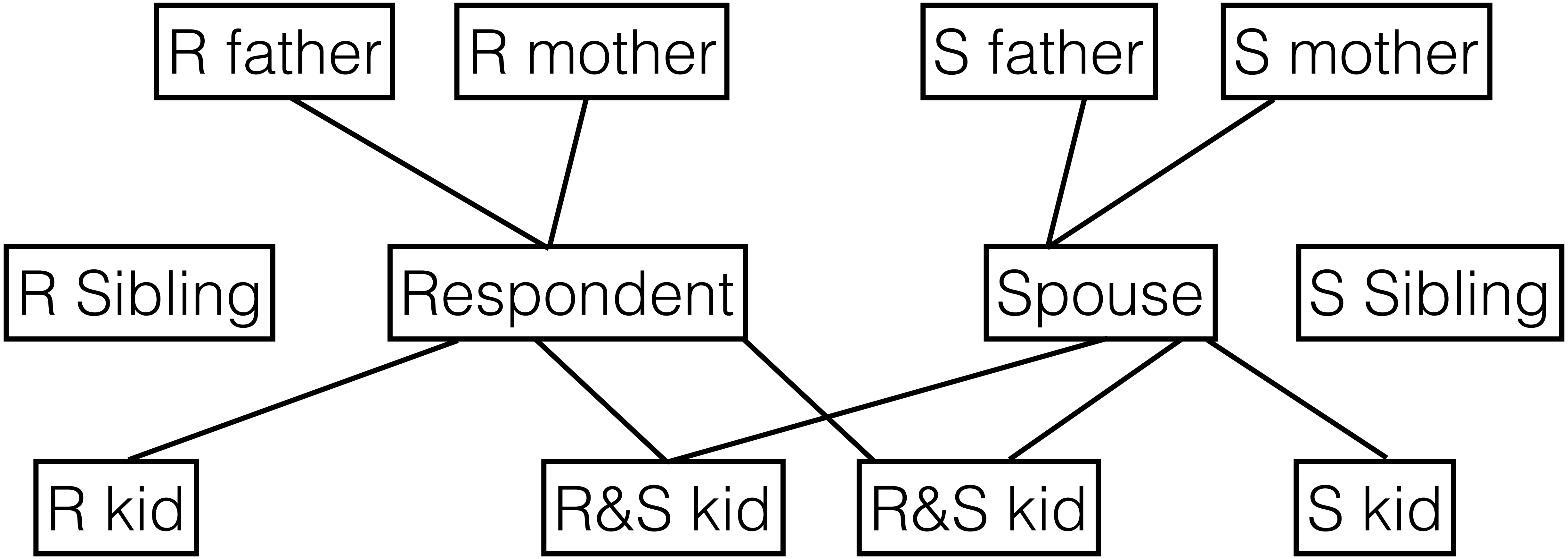
HRS interviews Americans aged 50+ in households

Some data on earlier periods in life are collected via **retrospective questions**



Social Security earnings data are collected via **administrative linkages**

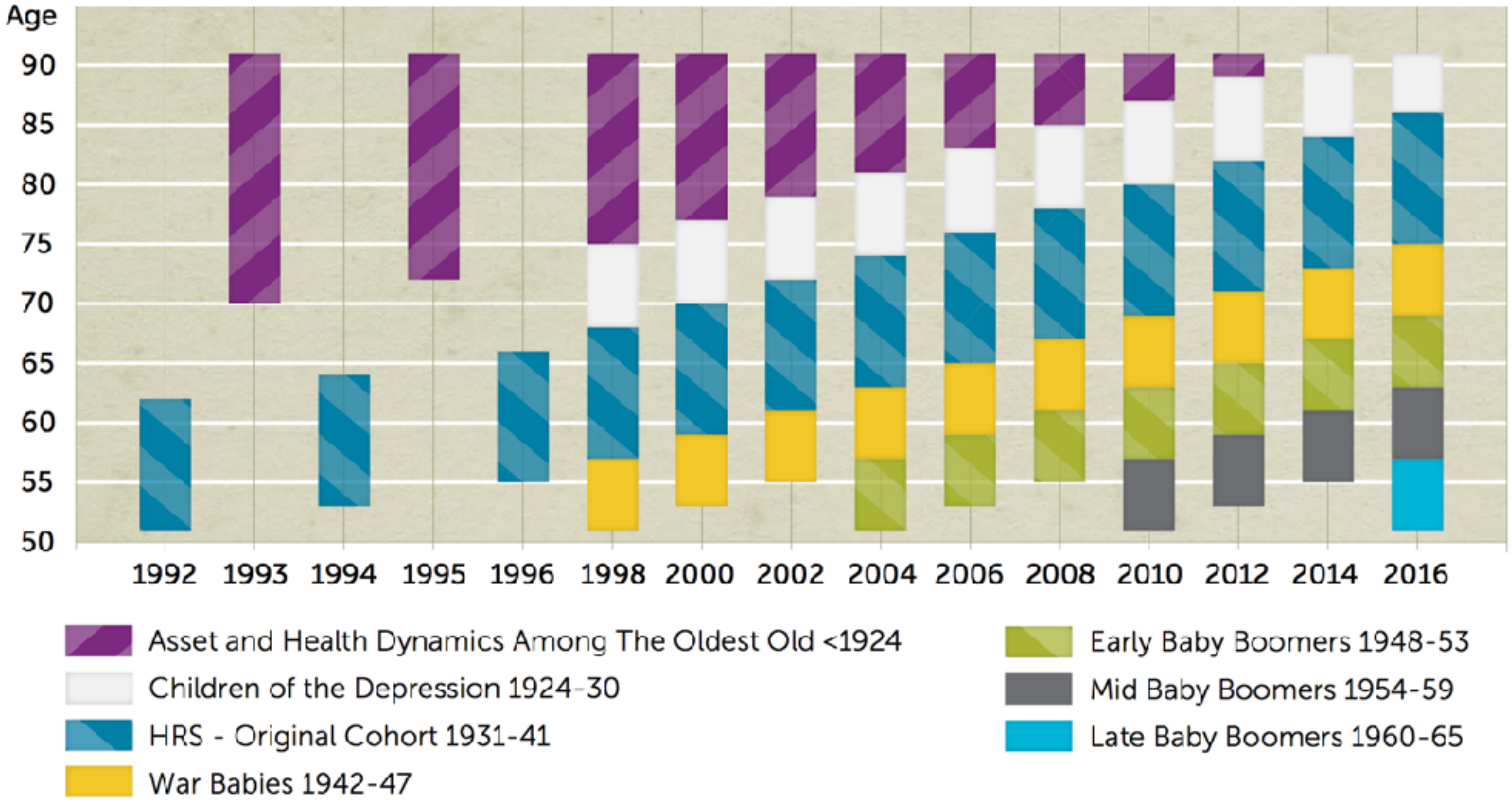
Respondents are Americans aged 50+ *and their spouses*, HRS has some information about their parents and children



Mostly data like age, age at death, education

Most other panel datasets follow "one cohort" through age, but HRS follows many

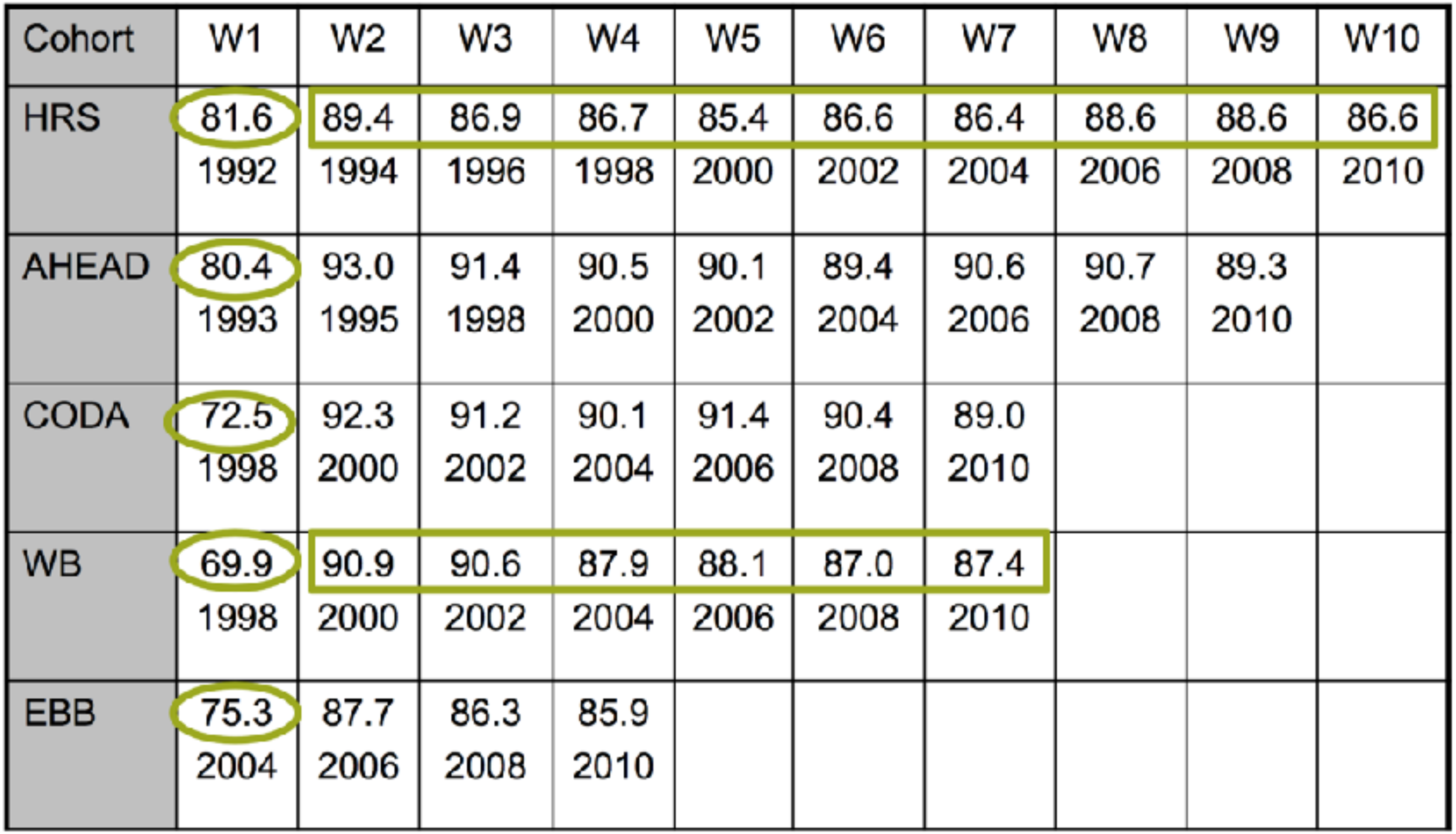
FIGURE A-4 **Longitudinal cohort design of the HRS**



Quick upshots and fine print

- A household-based survey refreshed every 6 years with new cohorts, remaining nationally representative of ages 50+
  - Accomplished via grit and toil and (for users) survey weights
  - Fine print: Hispanics and African-Americans are oversampled 2x
- HRS works hard to follow respondents wherever they may go: nursing homes, abroad, afterlife
  - Current residents nursing homes get a zero sample weight in the dataset because the universe is civilian noninstitutionalized
  - HRS is very serious about good mortality followup, but there also is some amount of nonresponse

BASELINE AND RE-INTERVIEW RESPONSE RATES



Cohort	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
HRS	81.6 1992	89.4 1994	86.9 1996	86.7 1998	85.4 2000	86.6 2002	86.4 2004	88.6 2006	88.6 2008	86.6 2010
AHEAD	80.4 1993	93.0 1995	91.4 1998	90.5 2000	90.1 2002	89.4 2004	90.6 2006	90.7 2008	89.3 2010	
CODA	72.5 1998	92.3 2000	91.2 2002	90.1 2004	91.4 2006	90.4 2008	89.0 2010			
WB	69.9 1998	90.9 2000	90.6 2002	87.9 2004	88.1 2006	87.0 2008	87.4 2010			
EBB	75.3 2004	87.7 2006	86.3 2008	85.9 2010						

Mortality

- HRS documents mortality two ways: (1) tracking each ever-respondent, and (2) 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

Mode of interview

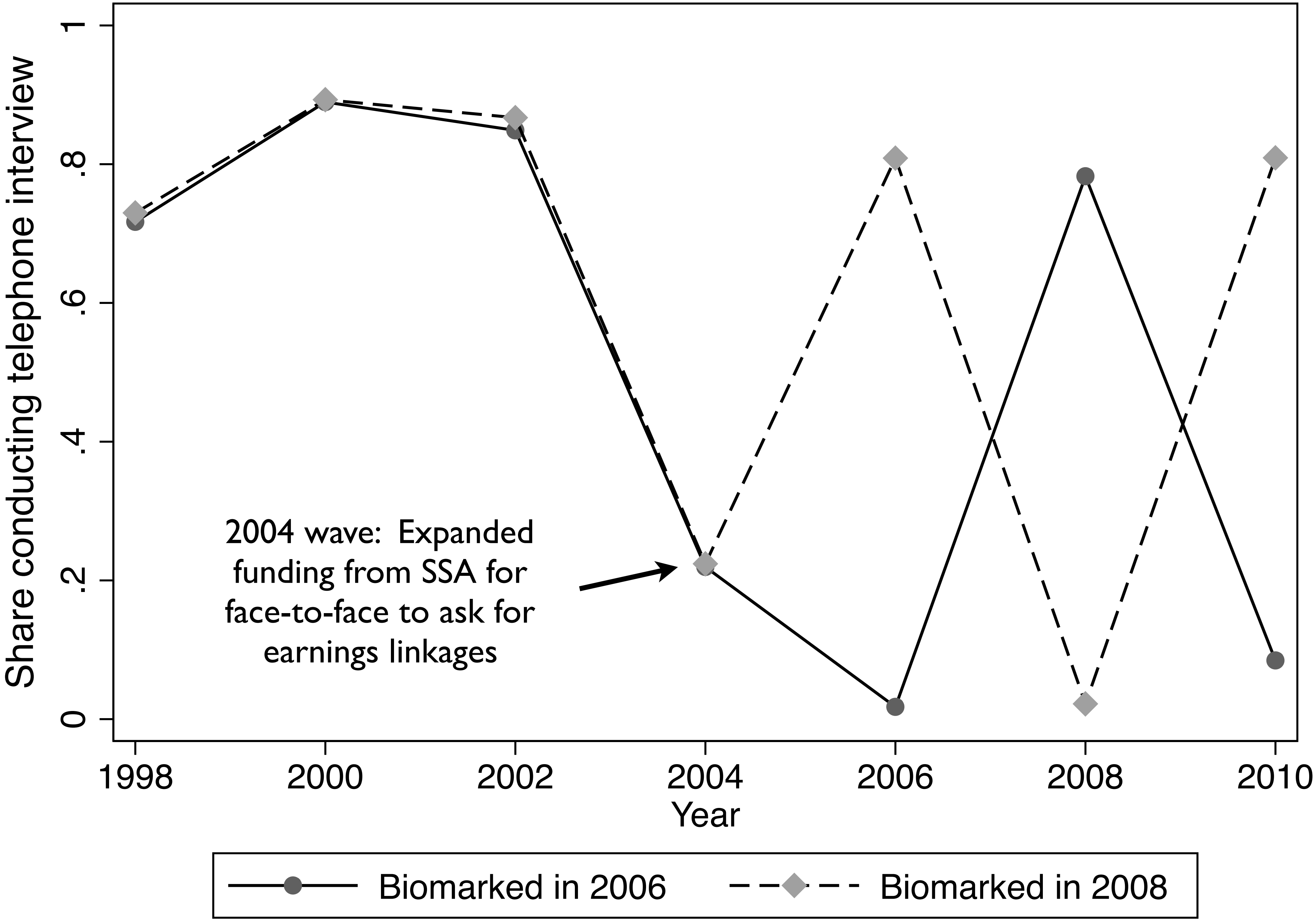
- Baseline interview with household always conducted face-to-face, computer assisted (CAPI)

- Subsequent interviews were primarily by telephone before 2004

- Core survey takes about 2 hours

- Starting in 2006, randomly rotating halves of the sample are visited for enhanced face-to-face and asked to consent to physical measures & biomarkers

Share on telephone was uniformly high before 2004, now alternates every other wave



2004 wave: Expanded funding from SSA for face-to-face to ask for earnings linkages

Variable names

- We'd ideally like a variable name to reflect what it measures
- In the underlying HRS data, this most often doesn't happen
- But in the RAND version, there's an heroic attempt to do this, all within 8 characters

Example: weight

	2014 weight	2016 weight
Homer Simpson	250	255
Marge Simpson	180	182
Kirk Van Houten	200	202
Luann Van Houten	170	172

HRS variable names aren’t mnemonic

			HRS 2014	HRS 2016
	2014 weight	**2016 weight**	**GC139**	**HC139**
Homer Simpson	250	255	250	255
Marge Simpson	180	182	180	182
Kirk Van Houten	200	202	200	202
Luann Van Houten	170	172	170	172

HRS variable names aren’t mnemonic

			HRS 2014	HRS 2016
	2014 weight	2016 weight	GC139	HC139
Homer Simpson	250	255	250	255
Marge Simpson	180	182	180	182
Kirk Van Houten	200	202	200	202
Luann Van Houten	170	172	170	172

The naming convention contains *some* information:
“O” is the 15th letter, but 2016 was the 13th wave (they skipped “I”)
The “C” is from Section C: Physical Health

RAND conventions help

	RAND HRS	RAND HRS	HRS 2014	HRS 2016
	r12weight	**r13weight**	**GC139**	**HC139**
Homer Simpson	250	255	250	255
Marge Simpson	180	182	180	182
Kirk Van Houten	200	202	200	202
Luann Van Houten	170	172	170	172

2014 was the 12th biennial wave counting from 1992

And the “**r**” prefix says the variable describes the **r**espondent

More RAND magic: spouse variables

	RAND HRS	RAND HRS	RAND HRS	RAND HRS
	r12weight	**r13weight**	**s12weight**	**s13weight**
Homer Simpson	250	255	180	182
Marge Simpson	180	182	250	255
Kirk Van Houten	200	202	.u	.u
Luann Van Houten	170	172	.u	.u

Kirk and Luann got divorced at the end of wave 3 in 1996 but have remained in the dataset
(They got remarried in 2007, but let’s pretend that didn’t happen)

More RAND magic: spouse variables

	RAND HRS	RAND HRS	RAND HRS	RAND HRS
	r12weight	**r13weight**	**r12mstat**	**r13mstat**
Homer Simpson	250	255	1.married	1.married
Marge Simpson	180	182	1.married	1.married
Kirk Van Houten	200	202	5.divorced	5.divorced
Luann Van Houten	170	172	5.divorced	5.divorced

Kirk and Luann got divorced at the end of wave 3 in 1996 but have remained in the dataset
(They got remarried in 2007, but let’s pretend that didn’t happen)

The magic of spousal variables

$$y_{it} = \alpha_i + \beta x_{it} + \beta^s x_{it}^s + \epsilon_{it}$$

- You might imagine a spouse's x^s might well affect a respondent's y
- Blatantly obvious ways that aren't worth a regression: spousal income raises household income 1-for-1
- Interesting ways: spousal characteristics like education might affect respondent's behavior, like exercise
- The RAND file allows you to quickly look at this

Pan-HRS identifier: HHID+PN = hhidpn

	RAND HRS	RAND HRS	RAND HRS		
	r12weight	**r13weight**	**hhid**	**pn**	**hhidpn**
Homer Simpson	250	255	010001	010	010001010
Marge Simpson	180	182	010001	020	010001020
Kirk Van Houten	200	202	010002	020	010002020
Luann Van Houten	170	172	010002	010	010002010

If string: hhidpn = hhid+pn

If numeric: hhidpn = hhid + 1000*pn

Moving ahead: stacking or reshape-long

hhidpn	r12weight	r13weight	r12vgactx	r13vgactx
010001010	250	255	3	4
010001020	180	182	1	1
010002020	200	202	4	4
010002010	170	172	2	2

r*vgactx = inde**x** of self-reported **vig**orous physical **act**ivity

A *lower number* means more frequent exercise

Might want to explore:

$$weight_{it} = \alpha_i + \beta\ activity_{it} + \epsilon_{it}$$

In the “wide” shape here, y could be r12weight or r13weight
But can we use them both simultaneously?

hhidpn	r12weight	r13weight	r12vgactx	r13vgactx
010001010	250	255	3	4
010001020	180	182	1	1
010002020	200	202	4	4
010002010	170	172	2	2

$$weight_{it} = \alpha_i + \beta\ activity_{it} + \epsilon_{it}$$

hhidpn	r12weight	r13weight	r12vgactx	r13vgactx
010001010	250	255	3	4
010001020	180	182	1	1
010002020	200	202	4	4
010002010	170	172	2	2

hhidpn	rweight	rvgactx	wave
010001010	250	3	12
010001010	255	4	13
010001020	180	1	12
010001020	182	1	13
010002020	200	4	12
010002020	202	4	13
010002010	170	2	12
010002010	172	2	13

Stacking observations and creating a new time variable allows us to use both y_{it} and y_{it+1} while addressing special circumstances implied by repeated observations

$$weight_{it} = \alpha_i + \beta\ activity_{it} + \epsilon_{it}$$

Especially when the x variable is a *treatment* applied to some units and not others, **panel fixed effects** is a common approach

It is like generalized difference-in-differences

Random effects is another common estimation strategy

hhidpn	rweight	rvgactx	wave
010001010	250	3	12
010001010	255	4	13
010001020	180	1	12
010001020	182	1	13
010002020	200	4	12
010002020	202	4	13
010002010	170	2	12
010002010	172	2	13

$$weight_{it} = \alpha_i + \beta \; activity_{it} + \epsilon_{it}$$

Suggested workflow to see whether HRS can help you

- Can HRS tell me about X?
- Think about ways that you might phrase X in a questionnaire and identify some keywords
- Check the RAND HRS file documentation PDF
  - Either keyword-search or look at *View: Table of Contents: Data Codebook: Contents*
- Check out the raw master codebook text file in a recent wave, a file like `h16core/h16cb/h2016.txt`
- For X's that might be obscure, check the online Question Concordance

Contents of the RAND HRS Data File

- In version L there were 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
F	Pensions	7
G	Health insurance	9
H	Family structure	5
I	Retirement plans, expectations	21
J	Employment history	20

Some special content areas beyond the core data

- Mortality (see earlier)
- Exit interviews and bequests
- Family structure and the RAND family file
- Consumption and Activities Mail Survey (CAMS)
- Childhood health retrospectives
- Restricted access files: (a) Social Security, (b) Medicare, (c) geocodes
- Biomarkers and Genetic Data

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

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

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

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; RAND CAMS spending data 2015 (v.2) also includes ’11, ’13, ‘15
- v.B 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

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

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

FIGURE A-3 **HRS supplemental off-year surveys**

Consumption and Activities Mail Survey (CAMS) biennial from 2001-2015+

CAMS is administered by mail to a random subsample of about 4,000 HRS core respondents; the survey collects extensive information about individuals' time use and household patterns of spending

Aging, Demographics and Memory Study 2001, 2002, 2006, 2008

ADAMS is an in-home neuropsychological assessment designed to provide a diagnostic determination of dementia or cognitive impairment without dementia; the study aims to estimate the prevalence of dementia as well as risk factors and outcomes

Prescription Drug Study 2005, 2007, 2009

The Prescription Drug Study (called the Health and Well-Being Study in 2009) is designed to track changes in prescription drug use and coverage as Medicare Part D—the federal prescription drug benefit—was implemented; administered by mail to 3,500- 5,000 HRS respondents; the 2009 wave added new content on experienced well-being

Internet Surveys 2003, 2006-2007, 2009, 2011, 2013

Web-based surveys developed in conjunction with the RAND Corporation; covered topics include internet use/ social media, health literacy, childhood health, cognition, well-being, residence history, income, assets, expectations, consumption, retirement preferences, prescription drug use, health behavior, annuities, and sibling transfers

HRS Mail Survey 1999

First mail survey designed to evaluate the impact on response rate of questionnaire length and impact of participation in the mail survey on core response rates; topics include health and health care use, psychosocial and attitudinal factors, housing and employment, spending preferences

Human Capital Mail Survey 2001

Surveyed a random subsample of about 4,000 HRS 2000 respondents by mail about parental economic investments in education, children's educational attainment, and the costs associated with attending college

Diabetes Study 2003

Conducted by mail with a subsample of about 2,000 HRS respondents who reported having diabetes in the 2000 or 2002 core interview; the focus was diabetes care, self-management, and health care utilization

Disability Vignette Survey 2007

Interviewed about 4,000 HRS respondents about their own health and disabilities; then told respondents vignettes that provide descriptions of people in different states of health and asked respondents to rate the level of disability of the hypothetical person

Health Care Mail Survey 2011

Conducted by mail with a subsample of about 7,000 HRS respondents on topics in health care, including access, utilization, policy, and veterans' health services

Health Care and Nutrition Survey 2013

Conducted by mail with a subsample of over 8,000 HRS respondents on topics in health care access and satisfaction, food security, food expenditures, and nutritional intake including vitamins and other supplements

Veterans Mail Survey 2013

Conducted by mail with a subsample of over 1,800 HRS respondents who had ever served in the active military. Contains questions about health care of veterans and their experiences in the military

Life History Mail Survey 2015

Contains questions about residential history from birth to age 50 and about educational history and experiences of HRS participants

Obtaining access to HRS restricted-use files

- Some of these files — SSA earnings, e.g., — had required 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
- Involvement of the Human Subjects Institutional Review Board (IRB) is required and usually time-consuming

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!

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, may also be available — there were lab issues early on that produced some challenges

Genetic data

- Many thanks to Amal Harrati, who wrote her Berkeley dissertation with these and is now at Stanford

- 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 with variation in humans

- 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

Access

- HRS has been amazing with methods of putting the data securely into the hands of researchers

- You can usually get the data onto a standalone machine if you want to, with precautions and destruction requirements

- Another nice option for many data is to *access them remotely and securely* via secure Windows Remote Desktop

  - Data lives on their servers. You can see it but can't copy it

  - You can upload and download stuff through clearance

  - I can't speak highly enough of how smoothly this works

Harmonized HRS

- Product of USC Program on Global Aging, Health, and Policy — Jinkook Lee and Eileen Crimmins — and Gateway to Global Aging Data
- Version A released in February 2018
- Designed to be similar to RAND HRS   (and it is very similar!)
- Pairs with:

Harmonized ELSA (England)	Harmonized LASI (India)
Harmonized SHARE (Europe + Israel)	Harmonized MHAS (Mexico)
Harmonized KLoSA (South Korea)	Harmonized TILDA (Ireland)
Harmonized JSTAR (Japan)	Harmonized CRELES (Costa Rica)
Harmonized CHARLS (China)	

Statistical languages

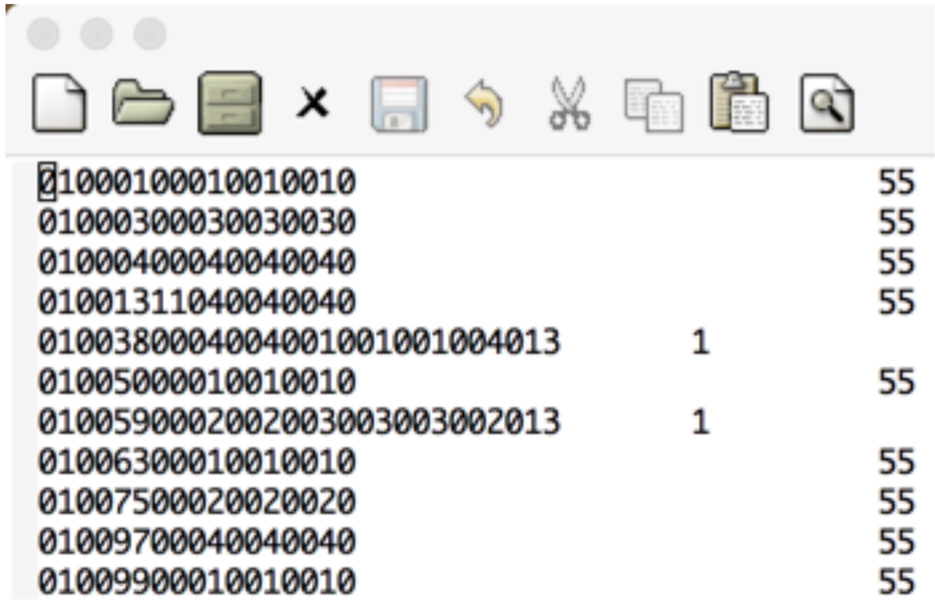
- You can use whatever you like
- But complexity implies you should probably use something that keeps the complexity
- SAS and STATA do a nice job of storing rich descriptors
- R is free and awesome (so I’m told)

Nature of the prefab HRS data files

- RAND files, Harmonized files, etc.
- Three “versions” to fit all: **`SAS`**, **`SPSS`**, and **`STATA`**
- No underlying text-file version of the data

Nature of the underlying HRS data files

- Three “versions” to fit all: `SAS`, `SPSS`, and `STATA`
    - Underlying data are in “fixed width” text files `*HyyX_Z.da` where `yy` = year, `X` = section, `Z` = respondent or other
    - Program files instruct `SAS`, `SPSS`, and `STATA` to extract variables and label them, their values, etc.



```
01000100010010010                            55
01000300030030030                            55
01000400040040040                            55
01001311040040040                            55
0100380004004001001001004013       1
01005000010010010                            55
0100590002002003003003002013       1
01006300010010010                            55
01007500020020020                            55
01009700040040040                            55
01009900010010010                            55
```

Using R

- Loading the data into R — and understanding what's what — are probably the main challenges

- Some thoughts. Seems like translating data files > starting from scratch

  - Package 'foreign' allows you to read SAS, STATA, SPSS files

    - `read.dta()` places variables from a v12 Stata dataset into data frames

    - "missing values are correctly handled"

    - "The data label, variable labels, timestamp, and variable/dataset characteristics are stored as attributes of the data frame"

  - Package 'rio' also imports, with extra functions for keeping labels etc.

Using Python

- Ironically:
  - Python's `pandas` package was in name inspired by `pan`el `da`ta
  - One way one could go was how pandas did it, with 3D data frames called `pandas.Panel()`
  - This seems not to have taken off
- `pandas` is a good Swiss Army Knife, with good treatment of missing values and many other things
- If you use Python, I suggest `pandas` but you probably already knew that. Sorry

Merging mechanics in Stata with an original data file

First step: Read in original data with `infile`

```
. clear

. set more off

. infile using H08B_R.dct

. save H08B_R.dta, replace

. set more on


. gen hhidpn = HHID + PN

. destring hhidpn, replace


. merge 1:1 hhidpn using rndhrs_l.dta
```

- If you want to join data to individuals, the linking variable is **hhidpn**
- In most HRS Stata .dct files, HHID and PN are read as strings
- Concatenate strings with “+” then destring because the RAND file uses numeric (long) hhidpn
- Beware of precision issues with hhidpn, double check your work

Alternative merging mechanics in Stata

```
. destring HHID, replace

. destring PN, replace

. gen long hhidpn = HHID*1000 + PN

. merge 1:1 hhidpn using rndhrs_l.dta
```

- Much of the RAND documentation implies they use this top formulation
- Beware precision issues
- I prefer the string concatenation shown on the previous slide

```
. merge 1:1 HHID PN using H08B_R.dta
```

- You can also merge by both HHID and PN if you want
- But the RAND file only has hhidpn, so that requires an extra step

Key variables and naming conventions in the RAND file

`hhidpn`	Individual identifier
`hacohort`	HRS birth cohort assignment (e.g.,AHEAD, HRS, CODA)
`rabyear`	Respondent’s birth year
`ragender`	Respondent’s sex
`raedyrs`	Respondent’s education in years
`rKiwstat`	Respondent’s interview status in wave K (mortality and more)
`inwK`	Whether respondent is in wave K
`sKgender`	Sex of spouse of R in wave K
`sKedyrs`	Education in years for spouse of R in wave K

Reshaping mechanics in Stata

	hhidpn	r1shlt	r2shlt	r3shlt
1	1010	2	3	4
2	2010	3	3	2

```
. clear

. use rndhrs_l.dta

. keep hhidpn r*shlt

. reshape long r@shlt, i(hhidpn)
  j(wave)
```

(no line break in the reshape command)

- RAND HRS file comes in a “wide” format (at left)
- Each individual gets a single row
- For longitudinal analysis, often want “long” format: each row a person-wave

	hhidpn	rshlt	wave
1	1010	2	1
2	1010	3	2
3	1010	4	3
4	2010	3	1
5	2010	3	2
6	2010	2	3

Application: Health status and smoking (1/4)

- Does smoking worsen health? Only a social scientist would question whether it in fact does

- (Aside: Yes, I believe smoking is bad for you. Kids, don't smoke.)

- Let's use this question to explore the RAND HRS file. We want to examine:

  $$y_i = \alpha + \beta x_i + X_i B + \varepsilon_i \qquad (1)$$

  where x is smoking, y is self-reported health, and X contains controls

- In HRS and many other surveys, self-reported health is "inverted," with 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor

- Smoking is bad for health if $\partial y/\partial x > 0$

- With the HRS, we can estimate equation (1) by (i) ordinary least squares, by (ii) instrumental variables if we link in cigarette prices with geocodes, and by (iii) panel estimators

Application: Health status and smoking (2/4)

$$y_i = \alpha + \beta x_i + X_iB + \varepsilon_i \qquad (1)$$

- Ordinary least squares estimation is likely to be biased by omitted variables when we don't have the right X's
  - There are probably unmeasurables that prompt people to have bad health AND to smoke
  - (And maybe health influences the decision to smoke)
- A good instrumental variable like cigarette prices, which are external to the individual and health, could reveal a causal pathway of smoking
- Panel fixed-effects can also shed light on equation (1) but do not solve all problems of causality; it merely reveals differences in x's associated with differences in y's

Application: Health status and smoking (3/4)

`hhidpn`	Individual identifier	`rKagey_m`	Respondent's age in years at middle of interview in wave K
`ragender`	Respondent's sex (2 categories)	`rKshlt`	Respondent's self-reported health status (12,3,4,5, 1 = excellent, 5 = poor)
`raracem`	Respondent's race (3 categories: white, black, other)	`rKsmoken`	Respondent self-reports smoking cigarettes now
`rahispan`	Respondent's Hispanic status	`raedyrs`	Respondent's years of education

Application: Health status and smoking (4/4)

```
. use "/Users/redwards/Data/HRS/RAND/randLstataSE/
  rndhrs_l.dta"

. keep hhidpn r*shlt r*smoken ragender raracem rahispan
  r*agey_m raedyrs

. reg r10shlt r10smoken i.ragender i.raracem i.rahispan
  raedyrs r10agey_m, vce(robust)

. drop raedyrs

. reshape long r@shlt r@smoken r@agey_m, i(hhidpn ragender
  raracem rahispan) j(wave)

. reg rshlt rsmoken i.ragender i.raracem i.rahispan ragey_m
  i.wave, vce(robust)

. tsset hhidpn wave

. xtreg rshlt rsmoken i.ragender i.raracem i.rahispan ragey_m
  i.wave, fe vce(robust)
```