Using Health and Retirement Study Data

A Guide for New Users

July 2025

Amanda Sonnega, PhD

Statistical code provided by Ryan McCammon, Chichun Fang, Christopher Greene, and Sergio Martinez

HRS | HEALTH AND RETIREMENT STUDY
A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

hrs.isr.umich.edu

Contact Information

Health and Retirement Study
Institute for Social Research
University of Michigan
426 Thompson Street
Ann Arbor, MI 48104
hrsquestions@umich.edu

HRS Co-Investigator Team

David R. Weir, Ph.D., Co-Director, University of Michigan
Kenneth M. Langa, M.D., Ph.D., Co-Director, University of Michigan
Jessica Faul, Ph.D., Associate Director, University of Michigan
Jennifer Ailshire, Ph.D., University of Southern California
Michael D. Hurd, Ph.D., RAND
Theodore Iwashyna, M.D., Ph.D., University of Michigan
Jodi Schimmel Hyde, Ph.D., Mathematica
Richard N. Jones, Ph.D., Brown University
Sharon L.R. Kardia, Ph.D., University of Michigan
Lindsay Kobayashi, Ph.D., University of Michigan
Helen G. Levy, Ph.D., University of Michigan
Jennifer Manly, Ph.D., Columbia University
Kathleen M. McGarry, Ph.D., University of California, Los Angeles
Jacqui Smith, Ph.D., University of Michigan
Bharat Thyagarajan, Ph.D., University of Minnesota
Brady West, Ph.D., University of Michigan

NIA Program Officials

Lis Nielsen Ph.D., Director, Division of Behavioral and Social Research
John Phillips, Ph.D., HRS Program Officer, Division of Behavioral and Social Research
Jonathan King, HRS Project Scientist, Division of Behavioral and Social Research

Acknowledgements

HRS is grateful to the National Institute on Aging of the National Institutes of Health for primary support of this project (U01 AG009740) and to the Social Security Administration for substantial co-funding.

Suggested Citation

Sonnega, A. (2025). Using Health and Retirement Study data: A guide for new users. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Overview

The University of Michigan Health and Retirement Study (HRS) is a longitudinal survey of a representative sample of Americans over the age of 50. Supported by the National Institute on Aging (NIA) and the Social Security Administration (SSA), HRS is designed to provide reliable data on the decisions, choices, and behaviors of people as they age and respond to changes in public policy, the economy, and health. Since 1992, the biennial survey has interviewed more than 46,000 individuals. The core survey focuses on income and wealth, health and use of health services, employment, and family connection. In 2006, data collection expanded to biomarkers, genetics, and a range of information on psychological states and social contexts. The data are linked at the individual level to administrative records from the Social Security Administration, the Centers for Medicare and Medicaid Services, Veteran's Affairs, and the 1940 Decennial Census as well as contextual data on state policies, environmental conditions, and food access. At the employer level, they are linked to employer-provided pension plan information. There are three levels of data access: public, sensitive, and restricted. Public data are available to all registered users. Sensitive and restricted data require special permission. The HRS website is your portal to find out more about the study and to access the data. This HRS Guide is intended to help new users get started using the data.

Visit the Website: hrs.isr.umich.edu

Email the HRS helpdesk at hrsquestions@umich.edu

Email the RAND HRS helpdesk at randhrshelp@rand.org

Table of Contents

Figures

CHAPTER 1

Introduction

Our Goal with this Guide

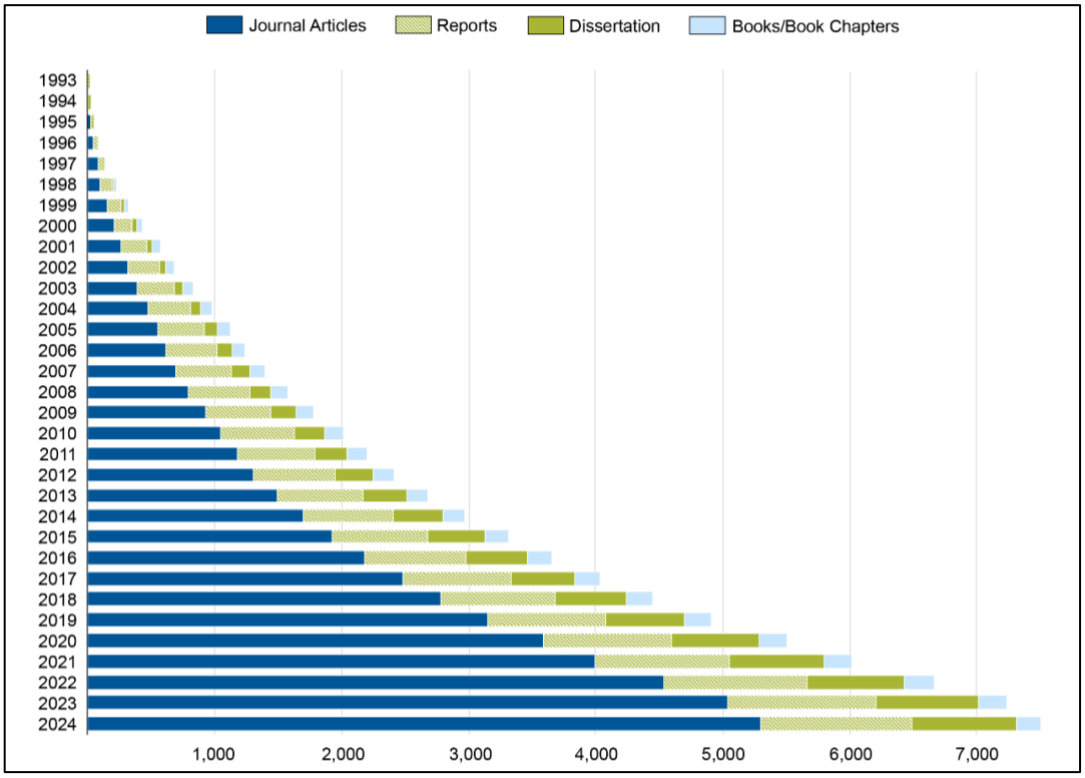
The Health and Retirement Study (HRS) is a key source for researching significant trends in retirement, health, and other changes in life as we age in the United States (U.S.). HRS is a comprehensive resource that covers various aging-related topics. It is used to explore a wide array of questions about aging.

The HRS is a large and complex study, but the basic design is a survey that:

- Is nationally representative of the U.S. population over age 50
- Follows individuals and their spouses or partners from the time of their entry into the survey until death
- Maintains a steady state design by introducing a new 6-year birth cohort of respondents every 6 years
- Is highly multidisciplinary

HRS has two main goals: to gather important data about aging and to make it available to the public. The use of HRS data is expanding, leading to a surge in related publications, as seen in Figure 1. Our goal with this guide is to provide those who are new to the data with the tools they need to get started using the data.

Figure 1. Publications Using HRS Data by Type



The HRS website may be challenging for newcomers to use. This guide is meant to be used alongside the website. It will direct you to specific parts of the site, each time starting from one of the main links on the HRS homepage:

- About
- Documentation
- Data Products
- Media & Publications

This guide highlights sections that are helpful for beginners and offers additional support as you become more experienced with the data.

 *Visit the Website: hrs.isr.umich.edu*

HRS Original Data and RAND HRS Data Products

It is crucial for those who are new to the HRS to understand the difference between the HRS original data and the RAND HRS data. The HRS original data come from interviews as well as mail and internet surveys conducted by staff at the University of Michigan (UM). They are processed by UM staff and made available on the HRS website. HRS original data can be difficult to use, however, especially for those who are new to the data. Therefore, researchers at the RAND Corporation take some of these data and further process them into user-friendly products, detailed in Chapter 4 and summarized in Figure 5. It is also important to learn how to use the HRS original data since not all HRS original data are included in the RAND HRS data products. The RAND HRS Longitudinal File is a RAND HRS data product that combines information from multiple survey sections and for all waves of HRS core data. It is often a good place to get started.

 *VISIT THE WEBSITE: Data Products>RAND HRS Products*

The All-Important Cross-Wave Tracker File

The HRS Cross-wave Tracker File is also crucial for new users. It helps track individual respondents and their interview details across different survey waves, providing basic demographics, interview statuses, and cross-sectional weights for each eligible respondent. It is a wide data file with one row for every individual who has ever been a respondent in the HRS. The columns are variables across all waves. The Tracker file is updated with data from new waves and new respondents about once a year. We advise reading the Cross-wave Tracker file data description, which gives more detail about its structure and variables.

 *VISIT THE WEBSITE: Documentation>Data Descriptions*

As you start using the HRS, the Cohort Profile is a useful overview of the study, accessible through our online bibliography.

Sonnega A. and others. Cohort Profile: the Health and Retirement Study (HRS).

VISIT THE WEBSITE: Media & Publications>Online Bibliography

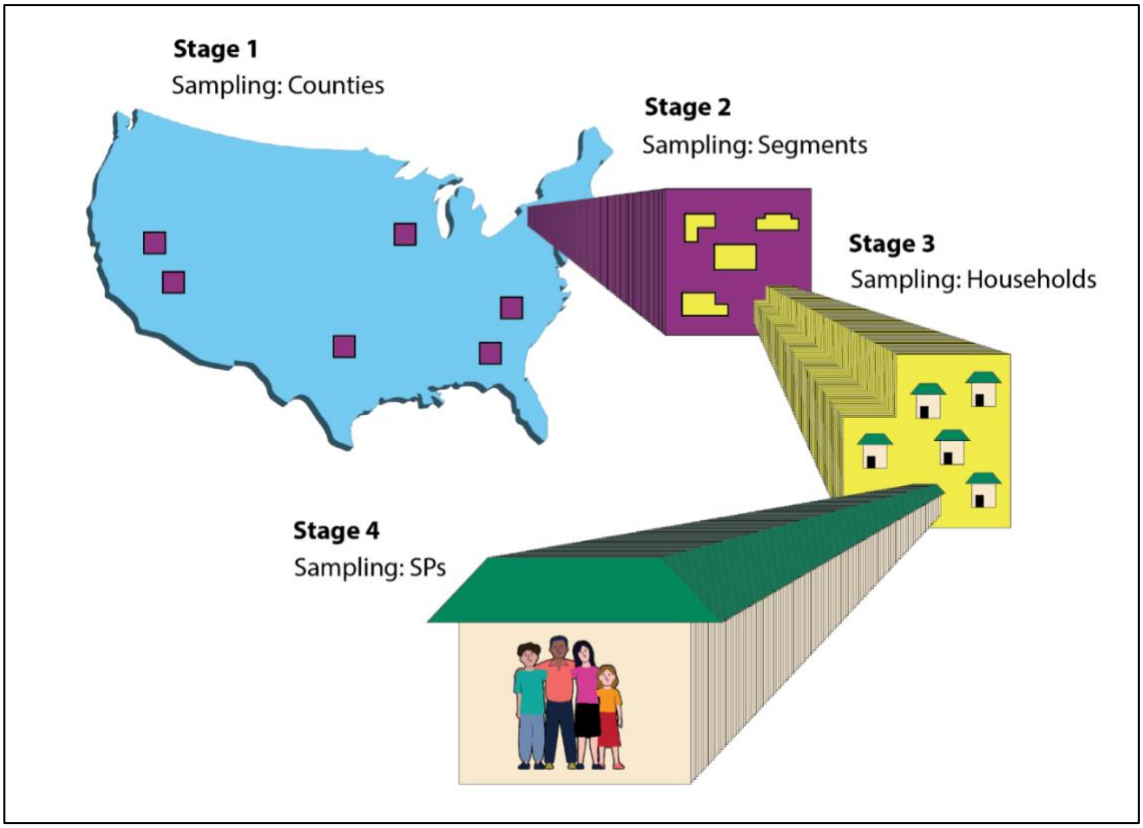
CHAPTER 2

Survey Design

HRS Multistage Complex Sample Design

To draw a nationally representative sample of the U.S. population over age 50, HRS employs a complex sample design with two main design features: stratification and clustering. Figure 2 depicts a basic multistage design.

Figure 2. Generic multistage complex sample design [reprinted with permission]



Heeringa, S.G., West, B.T., and Berglund, P.A. (2017). Applied Survey Data Analysis, Second Edition. Chapman Hall / CRC Press: Boca Raton, FL.

In a national sample, the process involves multiple steps. In stage 1, large areas called primary sampling units (PSUs), typically counties, are selected. Then in stage 2, within these PSUs, smaller areas like cities or towns called secondary sampling units (SSUs) are chosen (strata). In stage 3, specific sections of homes (clusters) are picked, and finally, in stage 4 individuals within those homes are selected for interviews.

The HRS sample is based on a multi-stage, area-clustered, stratified sample design. It consistently separates large PSUs, like major cities, which are always included, from smaller

PSUs, chosen based on their population size. This ensures that denser areas are more likely to be selected.

The HRS sample design also includes oversampling of Hispanic and African American households. This increases their numbers in the HRS to allow for more accurate analysis and comparisons within these racial/ethnic groups.

The HRS focuses on the U.S. population over 50 who do not live in institutions. People in the study who move into nursing homes after joining are still interviewed every two years, making the sample representative of older adults both at home and in nursing facilities post-enrollment.

The HRS sample design results in varied selection chances for respondents, so using sampling weights is crucial to avoid bias when making generalizations about the U.S. population. Standard methods for estimating variance are incorrect for use with a complex sample design, and corrections using design variables are necessary. Chapter 8 gives instructions on how to properly use these weights and design variables in analysis.

A video tutorial is available that reviews these concepts in detail: Sonnega A. HRS sample Design, Weighting, and Complex Variance Estimation.


VISIT THE WEBSITE: Documentation>Video Tutorials

More information about the sampling design and the development of weights for 2016 and beyond is also available on the website.

- Lee S. and others. HRS 2016 Sampling Weights.
- Heeringa S.G. and Connor J. Technical Description of the Health and Retirement Study Sample Design.
- Heeringa S.G. Technical Description of the Asset and Health Dynamics Among the Oldest Old (AHEAD) Study Sample Design.


VISIT THE WEBSITE: Documentation>Survey Design and Methodology

Creating a Longitudinal Cohort

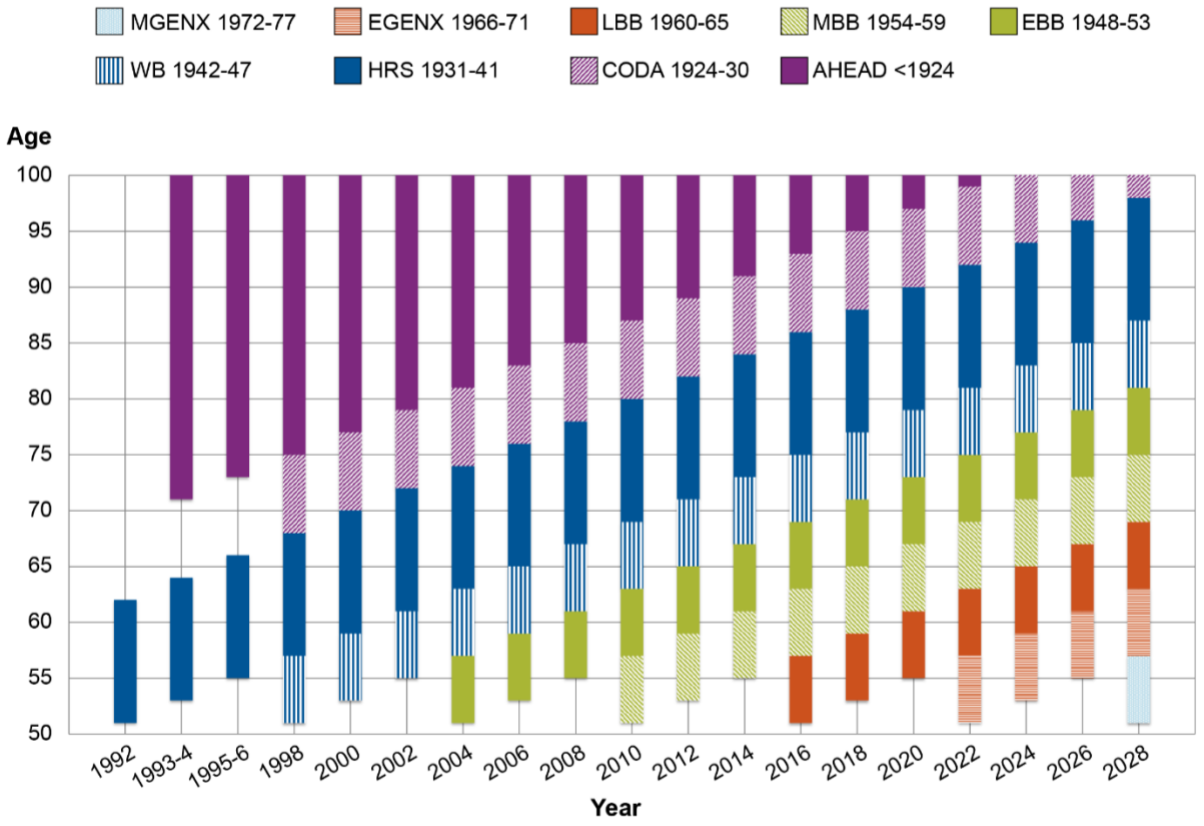
As shown in Figure 3, since it began in 1992, the HRS has gradually added respondents over time. The first cohort, the HRS cohort, included people born between 1931 and 1941 (and their spouses or partners who can be any age), who have been interviewed every two years since 1992. In 1993, the study expanded to include the Assets and Health Dynamics of the Oldest Old (AHEAD) cohort, for individuals born in 1924 or earlier, who were at least 70 at that time.

Over time, the HRS expanded by merging existing cohorts and adding new ones to fill age gaps for those over 51 in the U.S. In 1998, it merged the initial cohorts and added the Children of the Depression (CODA, born 1924-1930) and War Babies (born 1942-1947). HRS now regularly adds

younger cohorts: Early baby boomers (born 1948-1953) in 2004, Mid baby boomers (born 1954-1959) in 2010, Late baby boomers (born 1960-1965) in 2016, and the first part of Generation X (eGenx, born 1966-1971) in 2022.

For all cohorts, both members of a couple (both married and unmarried but living together) are included in the sample shown in Figure 3 and are considered respondents. And remember, while the sampled respondent is in the eligible age range, the spouse or partner can be of any age. Age-ineligible spouse/partners will have a zero weight in any wave that their birth cohort is not represented, but note that many younger spouses or partners will age into study age-eligibility in a later wave and will then be assigned a positive weight.

Figure 3. HRS longitudinal cohort sample design



Who is Interviewed?

As previously described, the HRS samples households. Initial recruitment into the study happens at the doorstep of a home. The interviewer checks if there is anyone living in the household in the cohort-eligible age range to participate in the study. The questions first confirm that the person answering is a household member, meaning they live, eat, and sleep there. Then, they are asked to count how many people in the household are in three different age groups: 18-44, 45-59, and 60 or older.

When a household has at least one person aged 45 or older, the interviewer collects more information about everyone aged 18 and above living there. This includes their name, birth year or age, ethnicity, race, and whether they are in a couple. The survey then chooses up to two qualified individuals for an in-depth interview. If a couple is living in the household and at least one partner is cohort-eligible, ***both partners are selected for the interview***. During these interviews, the chosen respondents may talk about their parents, children, and others who assist them, but only the selected individuals, not their relatives or helpers, are interviewed. Both members of a couple are considered respondents in the HRS.

Of course, household composition can change over time as people are, for example, widowed, divorced, and remarried. HRS tracks all of these changes and seeks to enroll any new partners. Chapter 7. Data Structure and Management provides the information you need to track changes in households over time if that is of interest for your research.

Family Respondent and Financial Respondent

In the HRS, respondents answer questions about themselves as well as about their households. In coupled households, questions related to family details or financial matters like income and wealth, are directed to only one member of a couple. The person who answers family-related questions during a survey wave is labeled the Family Respondent (or Family R), and the one who answers financial questions is called the Financial Respondent (or Financial R). These roles are identified by special wave-specific markers xFAMR and xFINR found in the Tracker file. If the selected person living in the household is single, that individual is both the Family R and the Financial R.

Interview Modes and Design Features

The biennial waves of data collection shown in Figure 3 are referred to in HRS as the core interview. Since the first wave in 1992, baseline core interviews have been conducted face-to-face in respondents' homes. Through 2002, panel (follow-up) interviews were typically conducted by telephone unless the respondent is over age 80, in which case panel interviews could also be face-to-face.

Enhanced Face-to-Face Interview

From 2006 onwards, HRS has used a strategy for the core biennial interview that combines both in-person and telephone interviews. Respondents are randomly divided into two groups: one group receives a detailed in-person interview, which includes physical health assessments, biomarkers, and a paper-and-pencil psychosocial questionnaire. This in-person interview with the extra measurement is called the enhanced face-to-face or (EFTF) interview. The other group is interviewed over the phone and answers just the standard set of questions asked every two years. Each respondent's type of interview—whether in-person or by phone—switches with every new survey wave

Figure 4 shows how the EFTF interview plan works. There are two groups, "A" and "B". Group "A" had their EFTF interview in 2006, while Group "B" did not. In 2008, Group "B" had their turn for the EFTF interview, and Group "A" did not. After that, each group takes turns for the enhanced interview with every new survey wave: Group "A" in 2010, Group "B" in 2012, and so on, alternating each time.

This means that while core survey data are available every wave on the full sample, the physical, biomarker, and psychosocial measures are available every wave on only half of the full core sample – either A or B – and every four years for all A and B respondents. The enhanced interview can take place at the baseline interview or at follow-up.

Figure 4. Design of the enhanced face-to-face interview

2006	2008	2010	2012	2014	2016	2018	2020
A	B	A	B	A	B	A	B

A = First random half sample; B = Second random half sample

The variable EFTFASSIGN can be found in the Tracker file. It is a flag for which sample a given respondent is in, A or B.

Web Mode

Starting in 2018, HRS gives some households the option to complete their main survey online. These households are chosen based on whether all the members said they use the internet regularly during the previous survey. However, households scheduled for the in-depth face-to-face interviews during a particular survey cycle are not given the web option. To understand how offering surveys online affects participation rates, data quality, and the answers provided, about 40% of the households that could do the survey online continue to do it in the usual way (by phone or in-person) as a comparison group. In the questionnaire, any instructions that are only for the web interviews are highlighted in teal.

The Tracker file has two flag variables that help identify web cases and controls: IWMODE and WEBCONTROL. In addition, starting in 2018, there are mode indicators in each of the HRS core section data files.

Proxy Respondents

Another important design feature of the HRS is the use of proxy respondents. Respondents who are unwilling or unable to do an interview themselves are offered the opportunity to use a proxy, who is usually a spouse or other family member. A research paper available in the bibliography provides more information on this topic: Weir D.R. and others. Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: a comparison of HRS and ELSA.

The process of proxy selection is documented on the study website. Scroll down to the section on Survey Design: HRS Staff. Proxy Selection in the HRS

VISIT THE WEBSITE: Documentation>Survey Design and Methodology >Survey Design

Tracker file provides a flag variable called PROXY with detail on the circumstances of the proxy interview.

Follow-up to Nursing Homes

Although the baseline core interviews are conducted with community-dwelling persons only, respondents who move to nursing homes and other senior living facilities after the baseline wave are retained and interviewed there in subsequent waves. Thus, although the study does not sample nursing home residents, over time the sample has come to represent the population residing in nursing homes in the US.

An indicator variable called xNURSHM (where x is a wave indicator) is available in Tracker that provides detail on the circumstances of the nursing home interview. A nursing home weight is also available beginning in 2000.

CHAPTER 3

Survey Content

Designing a Multidisciplinary Survey

The HRS is complex and covers many topics because it combines expertise from several areas. The study is run as a partnership between the University of Michigan and the National Institute on Aging (NIA). A team of experts from various fields leads the study, deciding its goals and what questions to ask in surveys. The NIA also has a Data Monitoring Committee (DMC) made up of experts in aging research who give advice on the success of the HRS and suggest improvements to its methods and topics. To support this, the DMC requests reports every six years on new and important topics. These reports are written by experts who review how well HRS is doing in specific areas and give suggestions for the future. Reports from the reviews conducted in 2010, 2016, and 2021 are available on the website.

VISIT THE WEBSITE: Documentation>DMC Review Papers

The Biennial Core Interview

The main part of the HRS is the core interview that happens every two years. This interview usually lasts about two hours and is where most of the study's information comes from. It includes questions about various issues that affect the lives and health of older adults:

- Health: physical/psychological self-report, conditions, disabilities; biomarkers and genetics; cognitive testing; health behaviors (smoking, drinking, exercise)
- Health Services: utilization including preventive care, expenditure, insurance, out-of-pocket spending
- Labor Force: employment status/history, retirement, earnings, disability, retirement, type of work
- Economic Status: income by source, wealth by asset type, capital gains/debt, consumption; linkage to pensions, Social Security earnings/benefit histories
- Family Structure: extended family, proximity, transfers to/from of money, time, housing
- Expectations: decision making, subjective probabilities

For much more detail on many of the major content areas, HRS has developed a series of content video tutorials that are available on the website under the Documentation link.

- Brown C. Employment, Retirement and Pensions in the HRS; 2022.
- Fang C. Pensions and Social Security Benefits Wealth in the HRS; 2022.
- Langa K. Health and Physical Functioning in the HRS; 2022.
- Levy H. Health Insurance and Medical Care in the HRS; 2022.
- Main R. RAND HRS Family Data; 2022.
- Ryan L. Cognition Data in the HRS; 2022.

VISIT THE WEBSITE: Documentation>Video Tutorials

Experimental Modules

Each wave, HRS includes short survey sections called experimental modules. These modules are added onto the end of the core interview, last about three minutes, and touch on a variety of topics. Some topics are new, while others add more detail to areas already covered in the main survey. Generally, these modules are given to a random sample of about 10 percent of the individuals participating in the core interview. Each respondent receives only one experimental module per wave. In the Codebooks, they are listed as Section M (1992-2000) or Section V (2002 and later). You can read more about the specific content of the experimental modules under Documentation.

VISIT THE WEBSITE: Documentation>Experimental Modules

Interviewer Observations

In large national surveys like the HRS that include in-person data collection, the interviewers are asked to take notes (thumb nails) on what the household and surrounding neighborhood look like. These interviewers are taught to pay attention to different signs to make sure their notes are correct. The HRS writes down these notes and shares them with researchers in files that anyone can use. These data are in section TN for thumb nail.

While these observations are useful to survey methodologists, researchers using HRS data have found these data to be quite useful in substantive applications. The following publication illustrates the use of these data and is available in the HRS online bibliography: Freedman V.A. and others. Neighborhoods and disability in later life.

VISIT THE WEBSITE: Media & Publications>Online bibliography

Exit and Post-exit Interviews

If a respondent passes away, the study tries to do a follow-up interview called an "exit interview" with someone who knew them well, like their spouse, child, or another person close to them. This interview aims to gather information about health care costs, family matters, and what was happening in the person's life near the end, as well as how their belongings and money were handled after they died. If there are still questions left after the exit interview, especially about the settling of the person's estate, a "post-exit interview" might be done in the next survey wave. The post-exit interview is triggered when the exit proxy respondent reports that a deceased respondent's estate, will, or trust has not been settled.

Question text for exit interview questions appears in gray or purple in the questionnaire text. Exit and post exit interview data are distributed as separate data files in the public data download area for each core wave.

A Biosocial Survey

As described already, in 2004, the HRS introduced a new enhanced in-person interview, the Enhanced Face-to-Face (EFTF), which collects a wide range of measures of physical function, finger stick blood spots for blood-based biomarkers (through 2016), DNA samples, venous blood collection, and a psychosocial questionnaire. The Venous Blood Study adds even greater biomarker measurement.

Physical Functioning Measures

Interviewers administer a series of physical measures and tests at the EFTF including:

- grip strength
- timed walk
- lung function
- balance
- measured height and weight
- waist circumference
- blood pressure

Variables for measures of physical functioning are public data available in section I in each core wave.

Biomarkers

Using a finger prick, dried blood spots were collected 2006 through 2016 for clinical biomarkers including:

- cholesterol
- hbA1c
- C-reactive protein
- Cystatin C

In 2016, the HRS began collecting samples of venous blood, which enhances the existing set of biomarkers by adding biomarkers related to immune system functioning, information about age-related changes at the molecular and cellular levels, and epigenetic markers.

Data for both dried blood spot biomarkers (2006-2016) and the 2016 Venous Blood Study (VBS) are available as Sensitive Health Data. A video tutorial describes these data resources: Faul J. Biomarkers and Physical Measures in the HRS; 2022.

VISIT THE WEBSITE: Documentation>Video Tutorials

A range of User Guides related to biomarker data are available on the website.

- Crimmins E.M. and others. Documentation of Physical Measures, Anthropometrics and Blood Pressure in the Health and Retirement Study.
- Crimmins E.M. and others. Documentation of Biomarkers in the 2006 and 2008 Health and Retirement Study.

- Crimmins E.M. and others Documentation of Biomarkers in the 2010 and 2012 Health and Retirement Study.
- Crimmins E.M. and others. Documentation of Blood-Based Biomarkers in the 2014 Health and Retirement Study.
- Crimmins E.M. and others. Venous Blood Collection and Assay Protocol in the 2016 Health and Retirement Study.

 *VISIT THE WEBSITE: Documentation>User Guides and Documentation Reports*

Genetics

Salivary DNA is collected during the EFTF to obtain DNA samples. These samples are genotyped and available through an external data repository. Several derived data products, however, are available on the HRS website including:

- APOE and Serotonin Transporter Alleles (2021)
- Candidate genes and SNP files
- Exome Data (2006-2010)
- Epigenetic Clocks
- Polygenic Score Data
- Telomere Data

Full GWAS data are available through dbGaP and NIAGADS. APOE and Serotonin Transporter Alleles and Telomere Data are available as Sensitive Health Data. Candidate Gene and SNP File data and Exome Data are available as Restricted Data. Epigenetic clocks and Polygenic Score Data are available as Public Data.

 *VISIT THE WEBSITE: Data Products>Genetic Data*

A video tutorial provides a high-level overview on the genetic data resources, and several user guides are available.

Faul J. Genetics in the HRS.

 *VISIT THE WEBSITE: Documentation>Video Tutorials*

Faul J. and Smith J.A. Health and Retirement Study: Genetic Data Consortia Collaboration.

Ware E.B. and others. HRS Polygenic Scores: 2006-2010 Genetic Data.

 *VISIT THE WEBSITE: Documentation>User Guides*

Psychosocial Functioning

At the end of the EFTF interview, respondents are given a self-administered questionnaire on psychosocial functioning to be completed and mailed back to the study office that asks about respondents' well-being, various sources of stress, social relationships, personality and beliefs, and experiences at work. It is section LB, listed toward the bottom of the questionnaire page. This stands for Leave Behind, since it is left behind at the end of the EFTF interview. Appendix Figure 1 summarizes the content of the psychosocial questionnaire.

Documentation includes a video tutorial and a user guide.

Smith J. Psychosocial and Well-being Data in the HRS.

VISIT THE WEBSITE: Documentation>Video Tutorials

Smith J. and others. Psychosocial and Lifestyle Questionnaire 2006-2022.

VISIT THE WEBSITE: Documentation>User Guides

Off-Year Studies

HRS also conducts supplemental off-year surveys with subsamples of 3,000 to 7,000 respondents from the core sample. These off-year studies are conducted in the years between core interview waves and can be linked to the core data. The off-year studies are summarized in Appendix Figure 2.

Some of the Off-Year studies contain sensitive health information and are available as Sensitive Health Data. Others are available as Public Data (scroll down the page to Off-Years Studies). See Chapter 4 for more detail on the difference between public, sensitive, and restricted data. A video tutorial describing the off-years and internet studies is available: Helppie-McFall B. Off-year and Internet Studies in the HRS.

VISIT THE WEBSITE: Documentation>Video Tutorials

Longitudinal and Cross-Wave Data Products

HRS creates several data products that combine data from longitudinal information into aggregated data files within specific topic areas. These include:

- Cross-Wave Child Proximity
- Cross-Wave Childhood Health and Family Aggregated Data
- Cross-Wave Imputation of Cognitive Functioning Measures

- Cross-Wave Census Region/Division and Mobility File
- Cross-Wave Marital History Aggregated Data
- Employer-Sponsored Pension Wealth from Current Jobs in 2016
- Cross-Wave Prospective Social Security Wealth Measures of Pre-Retirees
- Polygenic Score Data (PGS)

External Linkages

HRS data are also linked to external sources of information that are available as restricted data products. See Chapter 5 for more on accessing Restricted Data.

Social Security Administration Earnings and Benefits

For respondents who consent, HRS obtains administrative records from the Social Security Administration on the Social Security earnings and benefits of HRS respondents. In addition, the Disability Analysis File (DAF) contains historical, longitudinal, and one-time data on HRS respondents who participated in the Supplemental Security Income (SSI) or Social Security Disability Insurance (SSDI) programs.

Centers for Medicare and Medicaid Services Claims Information

Research files created by the Medicare & Medicaid Resource Information Center (MedRIC) link HRS data to detailed information about health care utilization from records maintained by the Centers for Medicare & Medicaid Services (CMS). The current CMS data linkage includes HRS respondents interviewed through the 2018 wave who have consented to the Medicare data linkage.

Veterans Administration

This dataset contains health care records for HRS respondents who utilized services in the Veterans Affairs (VA) healthcare system between the years of 1999 and 2013. For this data linkage project, the HRS identified veterans (both living and deceased) based on their self-report of military service in a prior wave of the HRS. A finder file that included HRS respondents who were deceased veterans and living veterans who gave informed consent was sent to the VA Information Resource Center (VIReC), where a probabilistic match of VA and HRS identifiers was performed.

The VA data included in this data set include: 1) VA enrollment and eligibility information; 2) Utilization of hospital, outpatient, and long-term care services; 3) Costs; 4) Pharmacy (prescription and non-prescription medications); and 5) Laboratory and radiology services.

Contextual Data Resource (HRS-CDR)

Information on the socioenvironmental context of HRS respondents such as health care and food access, air quality, and crime rates is available as a collection of user-friendly datasets, the

HRS-CDR, that enable researchers to study the impact of place on health and well-being among HRS respondents.

Employer-provided Pension Plans

Periodically, the HRS obtains pension Summary Plan Descriptions (SPDs) from the employers of study respondents. These SPDs are then analyzed and coded, and the plan description data, along with specific data from the respondents, are prepared for analysis through the Pension Estimation Program (PEP). This program is designed to estimate the pension entitlements held by respondents of the HRS, based on the plan formulas and benefit provisions obtained from the linked sample of pension providers.

HRS-O*NET Linkage

The 2022 Health and Retirement Study Linkage to Occupational Information Network (O*NET) Data provides a linkage between measures and variables provided in the O*NET and detailed 2010 Census occupational codes found in 2010-2020 restricted access HRS data.

Cognition Data Resources

Measurement of cognition has been part of the core HRS survey since the beginning in 1992. Two additional HRS studies provide in-depth cognitive assessment. The Aging, Demographics, and Memory Study (ADAMS) is a supplemental study fielded from 2001 to 2009 in the HRS that conducted in-person clinical assessments to gather information on respondents' cognitive status. A diagnosis of dementia, cognitive impairment but not demented (CIND), or non-case was assigned on the basis of this assessment.

The Harmonized Cognitive Assessment Protocol (HCAP) is a substudy within the HRS that is part of an ongoing international research collaboration to measure and understand dementia risk within ongoing longitudinal studies of aging around the world. It includes a carefully selected set of established cognitive and neuropsychological assessments and informant reports to better characterize cognitive function in older adults. A dedicated cognition page aggregates all the resources for cognition data in the HRS.

VISIT THE WEBSITE: Data Products>Cognition Data

Life History Data Resources

The collection of retrospective information about the early life of HRS respondents began in the core biennial HRS interviews in 1992 and has evolved over time. For example, the initial 1931-1941 HRS cohort provided information about the education of their parents in 1992, childhood geographical location in 1996, and additional background about family context in 1998. From 2008 onward, questions about childhood family context, together with multiple items about

childhood health conditions and diagnoses were added. Marital, fertility, and employment histories have also been collected in the core interviews at study entry and repeated over time.

Thus, the HRS core interviews provide rich life history data, but additional data collection was necessary to fill in some gaps and enhance information about childhood and early-to-mid adulthood. The Life History Mail Survey (LHMS) adds residential, educational, employment, partnership, and caregiving information prior to age 50. A dedicated life history page aggregates all the resources for life history data in the HRS.

VISIT THE WEBSITE: Data Products>HRS Life History Data Resources

COVID Data Resources

HRS added COVID-19-related questions to the 2020 core interview (Section COV) and to the psychosocial self-administered questionnaire (pages 36-43). In addition, a new Contextual Data Resource on state-level COVID-19 policies has been added as part of the HRS restricted data products. A supplementary 2021 COVID-19 Mailout Survey provides a wide range of information related to respondents' experiences of the pandemic. A dedicated COVID data page aggregates all the resources for COVID data in the HRS.

VISIT THE WEBSITE: Data Products>COVID-19 Research Initiatives in the HRS International Network

Imputations

HRS respondents always have the option to skip questions that they either do not want to or are unable to answer. Some sections of the questionnaire tend to have more "missing." One of the ways to address this issue is through survey design. One area that HRS researchers recognized early on in the study as being likely to have more missing data: any question that queried a respondent on a specific dollar amount, such as their income, wealth, or out-of-pocket medical expenses. To help reduce the amount of "missingness" associated with these questions, HRS researchers developed the method of "unfolding brackets." Bracketed data are also used to help impute missing data. To learn more about unfolding brackets, scroll down to the Survey Design section on the Survey Design & Methodology page under Documentation to access: Heeringa S.G.. and others. Unfolding brackets for reducing item nonresponse in economic surveys.

VISIT THE WEBSITE: Documentation>Survey Design and Methodology

HRS also employs statistical imputation to fill in missing data for the sections with high levels of missingness. RAND has conducted the imputations of financial information that are part of the

RAND HRS Longitudinal File and the RAND HRS Detailed Imputations File. Imputations are also conducted on cognition information that are part of the RAND HRS Longitudinal File.

To learn more about the cognition imputations, scroll down to the Cognition section on the User Guides page under Documentation to access: McCammon R.J. and others. Health and Retirement Study Imputation of Cognitive Functioning Measures: 1992-2018.

VISIT THE WEBSITE: Documentation>User Guides

To learn more about the RAND imputations, see the RAND HRS Detailed Imputations File.

VISIT THE WEBSITE: Data Products>RAND HRS Products>RAND HRS Detailed Imputations File 2020

Conducting Cross-national Comparisons

As population aging has progressed in every region of the world, the success of the HRS has generated substantial interest in collecting similar data in a similar manner. The International Family of Studies represent a growing network of longitudinal studies of aging around the world for which HRS is the model. Harmonization of design and content is facilitated by Health and Retirement Studies Around the World, a network for the harmonization of aging studies

The HRS family of studies share some common characteristics including being population representative of older adults, longitudinal, multidisciplinary, having coordinated survey instruments, enhanced economic data, and integrated biomarkers. All HRS family studies distribute their data to the research community. For those interested in conducting comparative research using the HRS family of studies, the Gateway to Global Aging Data is another useful resource created and maintained by researchers at the University of Southern California. The Gateway to Global Aging site offers a digital library of survey questions, a search engine for finding comparable questions across surveys, and identically defined variables for cross-country analysis.

VISIT THE WEBSITE: About>Partnerships and Collaborations>International Family of Studies

CHAPTER 4

RAND HRS Data Products

Researchers at the RAND Corporation have created user-friendly versions of much of the HRS public data, the RAND HRS Data Products. The RAND Center for the Study of Aging creates the RAND HRS data products under subcontract from HRS.

While you can access the RAND HRS data products on the HRS website, you will also want to visit the website for the RAND Center for the Study of Aging, where the RAND HRS data products are produced. There are many useful resources including a PowerPoint presentation that provides more detail on the topics covered in this Guide and a Frequently Asked Questions (FAQ) link that contains a wealth of practical guidance on using the data.

All RAND HRS data products are available to download from the HRS website.

VISIT THE WEBSITE: Data Products>RAND HRS Products

The RAND HRS data products do some important things that make using the data easier:

- They have renamed variables in a user-friendly, consistent way.
- They have constructed many variables, such as household income and wealth, which would be extremely time-consuming for you to create on your own.

The first two RAND HRS data products we describe here are the best starting place for a new user of the HRS data. Other RAND HRS data products are likely to be of greater interest after you are already up and running.

Questions regarding RAND HRS data products can be directed to randhrshelp@rand.org.

RAND HRS Fat Files

The HRS original data from the core biennial interview are distributed in a set with one data file for each of the many core interview sections. The RAND HRS Fat Files merge all section files together into one large file. The RAND HRS Fat Files contain all the original variables collected at the respondent or household level, except those from "Other Person" sections (i.e., data on children, siblings, household members, helpers, and transfers). The files are easily merged with one another and with the RAND HRS Longitudinal File, which can expedite the preparation of analytic files from HRS.

RAND HRS Longitudinal File

As noted, the RAND HRS Longitudinal File contains public information (i.e., no sensitive or restricted data) from the HRS Core and Exit interview, with derived variables covering a large

range of topics. It uses information from the current Tracker, Region and Mobility, and Master ID files. All variables are named and derived consistently across survey years, and any cross-wave differences are documented. Every individual who has ever completed at least one HRS Core Interview has a record in this file. This includes individuals who were age-eligible (born in eligible years) at the time of their first interview, spouses that were not age-eligible at baseline, and spouses who married an age-eligible respondent between survey waves.

The RAND HRS Longitudinal File is a relatively easy entry point to the HRS data. Although the file contains a rich subset of information from the Core and Exit Interviews, it is not exhaustive. In cases where researchers cannot find all variables they need, the file can easily be merged with other RAND HRS data products as well as HRS original data files.

RAND HRS Detailed Imputations File

The RAND HRS Detailed Imputations File contains the detailed income and wealth components that are used in the construction of the aggregated income, wealth, and medical expenditures summary measures included in the RAND HRS Longitudinal File.

RAND HRS Family Data

The RAND HRS Family Data files are user-friendly versions of the HRS family data. They contain a cleaned and processed collection of variables related to the respondent's family. The files include a subset of available characteristics of all children of HRS respondents and spouses, data on children-in-law, and data on grandchildren of the respondent. They are longitudinal files that link HRS child families within waves, and link HRS children across waves.

RAND HRS CAMS Data

The RAND HRS CAMS Data File is a user-friendly version of Part B of the Consumption and Activities Mail Survey (CAMS). It contains annualized, cleaned, and aggregated spending and consumption variables. Specifically, total household spending and household consumption are calculated across all categories and for these subsets of spending: nondurables, durables, housing, and transportation.

RAND HRS Tax Calculations

The RAND HRS Tax Calculations data contain information about federal, state, and FICA taxes for HRS survey respondents. The NBER Internet TAXSIM calculator is used to perform these calculations.

RAND HRS Exit/Post-Exit Interview and Finder Files

The RAND HRS Exit/Post-Exit Interview data files merge all the original variables from each respondent-level questionnaire section to create a single respondent-level dataset for each wave. The files include all respondents from both the exit and post-exit Interviews.

Citing HRS Original and RAND HRS Data Products

Citation instructions for the HRS original data:

In text:

> "The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan."

In references:

> "Health and Retirement Study, ([insert Product Name]) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (year)."

For early release data products, please use the following:

> "This analysis uses Early Release data from the Health and Retirement Study, ([insert Product Name]), sponsored by the National Institute on Aging (grant number NIA U01AG009740) and conducted by the University of Michigan. These data have not been cleaned and may contain errors that will be corrected in the Final Public Release version of the dataset."

A good general citation for the design and content of the HRS is:

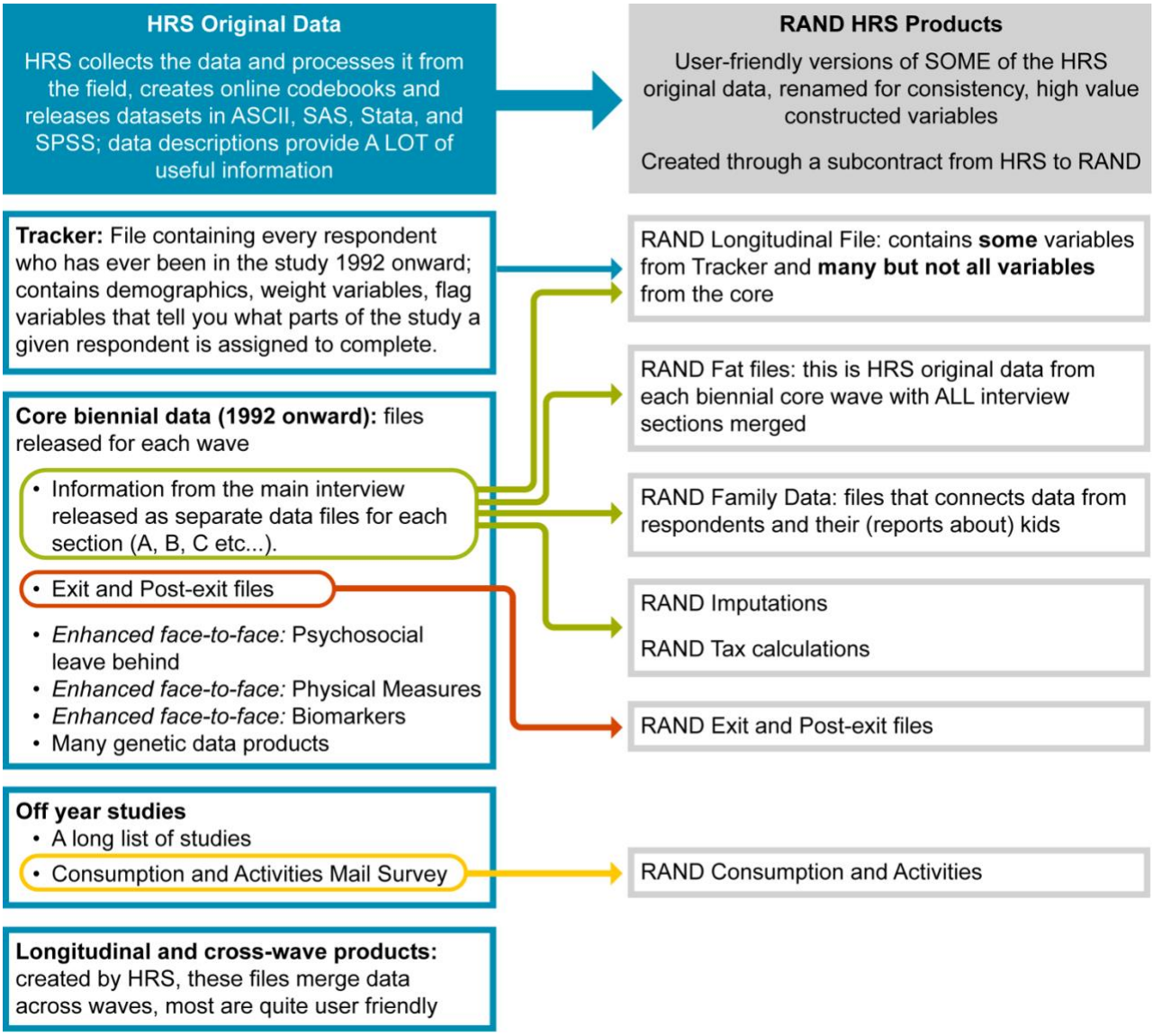
> Sonnega A. and others. Cohort Profile: the Health and Retirement Study.

Citation instructions for RAND HRS Data products are available on the RAND Center for the Study of Aging website under Resources for Users, Citing Instructions.

The Relationship between HRS Original Data and RAND HRS Data Products

Figure 5 shows the relationship between HRS original data and RAND HRS data products.

Figure 5. The relationship between HRS original data and RAND HRS data products



CHAPTER 5

Three Levels of Data Access

There are three levels of data security: public, sensitive, and restricted.

Public Data Access

The vast majority of the data are public data available to all registered users. You simply create an account and login to begin accessing public data files from the website as described in the Data Products section of Chapter 6 and the Register for HRS Original Data section of Chapter 7.

Sensitive Data Access

Sensitive health data, like biomarkers and information on prescription drugs, require an extra step in registration. Access to any of the Biomarker and Sensitive Health data products requires that you complete a data user agreement and a data order form online. At the end of the paragraph at the top of the Sensitive Health Data page, there is a link to the Sensitive Health Data Order Form.

These products contain sensitive health information:

- ADAMS
- Biomarkers
- Telomere
- APOE and Serotonin Transporter Allelles
- HCAP
- 2019 Health Survey (including detailed medication data)
- 2009 Health and Well-being Study (including prescription drug data)
- 2018 Marijuana attitudes experimental module
- 2005 and 2007 Prescription Drug Study
- 2016 Venous Blood Study

VISIT THE WEBSITE: Data Products>Sensitive Health Data

Restricted Data Access

All of the external linkages in the HRS are restricted because they contain confidential information. HRS also makes some confidential survey information available as restricted data, for example, detailed occupational and geographical data. There are two ways to access restricted data: the virtual desktop infrastructure (VDI) system and traditional licensing.

Access with Virtual Desktop Infrastructure (VDI)

Nearly all restricted data can be accessed through the VDI system that allows users to remotely connect through a secure connection from their own desktop, to a secure data enclave maintained by the Michigan Center for the Demography of Aging (MiCDA) that houses the restricted data. HRS provides access to the VDI free of charge.

Access with Traditional License

In some circumstances, users may need to apply to obtain and use HRS restricted data on their own secure systems under the traditional license. Reasons may include:

- Custom software not supported in the VDI
- Custom hardware configurations that cannot be mimicked in the VDI
- Large third-party data sets
- Contracts with third parties that prohibit the hosting of their data in the VDI

Note that CMS Data are only available with Traditional Licensing.

For both VDI and traditional license, use the same login credentials you obtained for the public data. This will allow you to begin your application.

VISIT THE WEBSITE: Data Products>Restricted Data

CHAPTER 6

Using the Website

The HRS website is your portal to find out more about this complex study and to access and get started using the data. This section highlights particularly useful sections.

Documentation

Questionnaires

A good place to start learning about the content of the survey is to look through the Questionnaires. These are printed representations of the interview sections. The top half of the questionnaire page shows the biennial core interview content by section across waves. You click on the section letter to view the questionnaire. Note that some questions are asked every second, third, or fourth waves, so-called Alternate Wave questions. Details are available in the Alternate Wave Questions Master Chart, which is linked at the top of the Questionnaire page.

There are two main sections on the questionnaire page. The top half links to questionnaires associated with the biennial core interview by content section. Note that in some cases, the letter indicating each section has changed over the years. The bottom half of the page displays the off-year and health studies.

Within the questionnaire, you can find information that is preloaded into the questionnaire at the beginning. For example, if you click on Section C (Health) for 2020, in a panel interview, information about chronic conditions the respondent reported previously is preloaded into the interview.

It is important to note that the questionnaire contains information describing the flow through the survey, including branch points. Examining the branch points will help you understand why some questions have a lower sample size than you might have expected. For example, only respondents who report having diabetes will be asked the follow-up question about the stage of their disease. You will also see at the top of the questionnaire that, as noted, exit interview questions appear in fuchsia or gray.

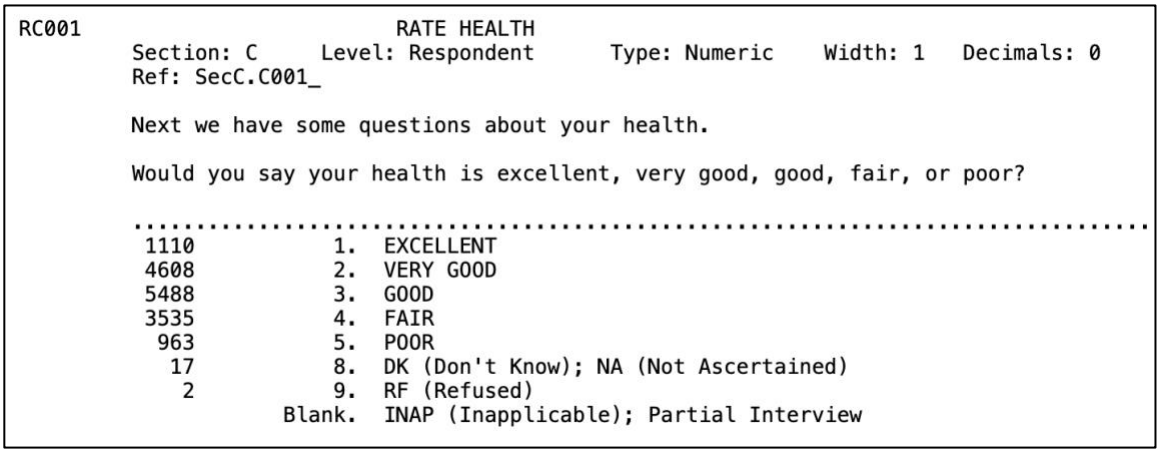
VISIT THE WEBSITE: Documentation>Questionnaires

Codebooks

Every HRS data file has a codebook that provides frequency distributions on all variables in that data file. The codebook is organized by questionnaire section. Clicking on a particular section opens to another window. On the left is a listing of all of the variables in this section. On the right are the frequency distributions. The variable name on the left is a hyperlink to the frequency distribution for that variable. For example, clicking on Section C. Physical Health (respondent) for the 2020 Core, if we further scroll down to RC001, which is the variable name

for self-rated health, the frequency distribution for that variable appears on the right (see Figure 6). This is also a good way to see the code frame for a particular question.

Figure 6. 2020 core release codebook for Section C



```
RC001                          RATE HEALTH
         Section: C     Level: Respondent      Type: Numeric    Width: 1   Decimals: 0
         Ref: SecC.C001_

         Next we have some questions about your health.

         Would you say your health is excellent, very good, good, fair, or poor?

         ..........................................................................
          1110           1.  EXCELLENT
          4608           2.  VERY GOOD
          5488           3.  GOOD
          3535           4.  FAIR
           963           5.  POOR
            17           8.  DK (Don't Know); NA (Not Ascertained)
             2           9.  RF (Refused)
                      Blank.  INAP (Inapplicable); Partial Interview
```

Note that if you find a variable in the codebook that has a much smaller sample than you might have expected, it is probably because of the skip patterns as noted above. It may also be an alternate wave question, and you can consult the chart to find out if it is.

VISIT THE WEBSITE: Documentation>Codebooks

Data Descriptions

Each data product also has a Data Description associated with it. **This is the most important documentation you can read for any HRS data product you want to use.** The Data Description is your best source of information for understanding and using the associated data product.

VISIT THE WEBSITE: Documentation>Data Descriptions

User Guides and Documentation Reports

User Guides and Documentation Reports offer detailed information about the topics covered in the survey. They explain how the survey questions were created and where they came from. These guides are useful for researchers analyzing the survey data because they go beyond basic codebooks and questionnaires. They give detailed explanations of the survey questions, how variables have changed over time, statistical summaries of the responses, analyses of the quality of the data, details about how each survey is organized, and guidance on how to handle the data in special cases.

VISIT THE WEBSITE: Documentation>User Guides and Documentation Reports

Survey Design and Methodology

The Data Collection Table on the Survey Design and Methodology link (see Figure 7) shows the history of HRS data collection efforts, with links to additional information about each data product. For cross-year products, the table shows which years are included in the release. All acronyms in the table are spelled out at the top. Then for each year below, you can see the biennial, off-year, sensitive health, restricted data, and derived data products. Each of these buttons takes you to an overview information page about that data product.

Figure 7. Data Collection Table

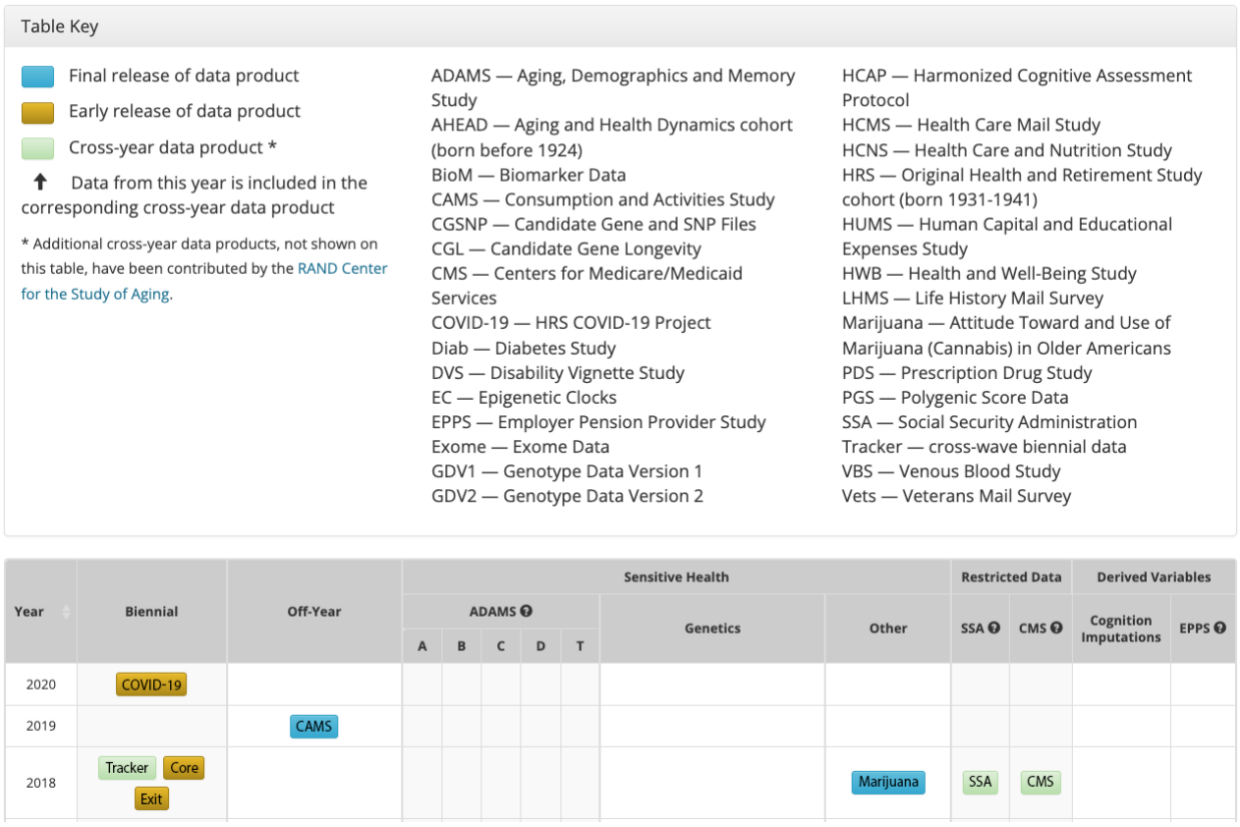


Table Key

Final release of data product
Early release of data product
Cross-year data product *
↑ Data from this year is included in the corresponding cross-year data product

* Additional cross-year data products, not shown on this table, have been contributed by the RAND Center for the Study of Aging.

ADAMS — Aging, Demographics and Memory Study
AHEAD — Aging and Health Dynamics cohort (born before 1924)
BioM — Biomarker Data
CAMS — Consumption and Activities Study
CGSNP — Candidate Gene and SNP Files
CGL — Candidate Gene Longevity
CMS — Centers for Medicare/Medicaid Services
COVID-19 — HRS COVID-19 Project
Diab — Diabetes Study
DVS — Disability Vignette Study
EC — Epigenetic Clocks
EPPS — Employer Pension Provider Study
Exome — Exome Data
GDV1 — Genotype Data Version 1
GDV2 — Genotype Data Version 2
HCAP — Harmonized Cognitive Assessment Protocol
HCMS — Health Care Mail Study
HCNS — Health Care and Nutrition Study
HRS — Original Health and Retirement Study cohort (born 1931-1941)
HUMS — Human Capital and Educational Expenses Study
HWB — Health and Well-Being Study
LHMS — Life History Mail Survey
Marijuana — Attitude Toward and Use of Marijuana (Cannabis) in Older Americans
PDS — Prescription Drug Study
PGS — Polygenic Score Data
SSA — Social Security Administration
Tracker — cross-wave biennial data
VBS — Venous Blood Study
Vets — Veterans Mail Survey

Year	Biennial	Off-Year	Sensitive Health							Restricted Data		Derived Variables	
			ADAMS					Genetics	Other	SSA	CMS	Cognition Imputations	EPPS
			A	B	C	D	T						
2020	COVID-19												
2019		CAMS											
2018	Tracker Core Exit								Marijuana	SSA	CMS		

The Survey Design and Methodology page also provides documentation related to incentive token amounts, sample sizes and response rates, weight information, survey design, and imputations.

VISIT THE WEBSITE: Documentation>Survey Design and Methodology

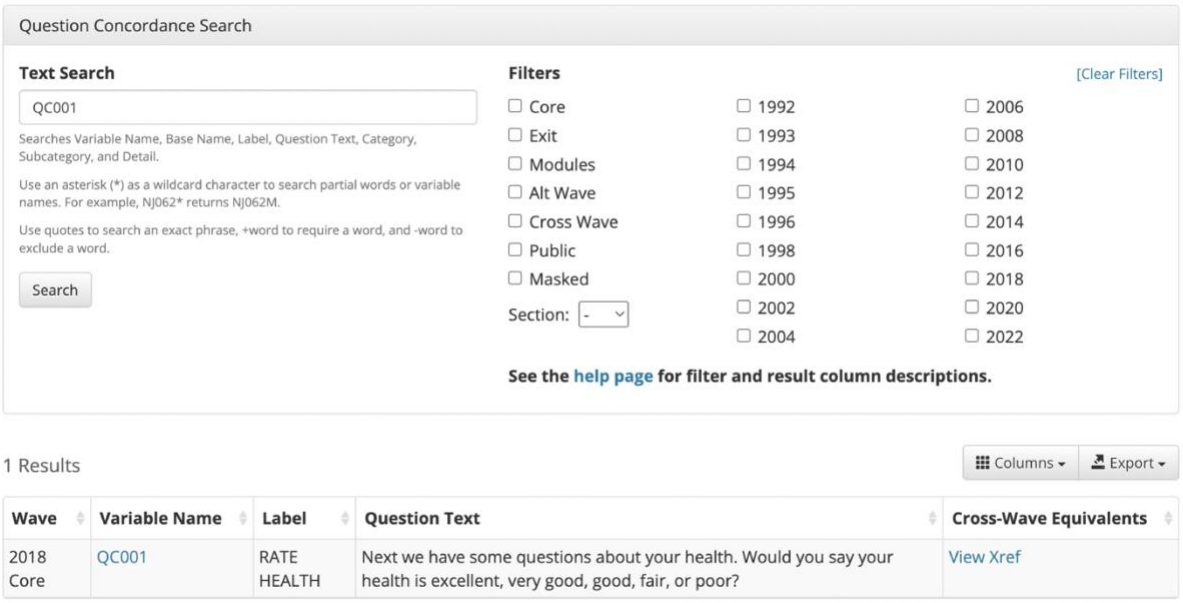
Concordance

The Concordance is a question search tool that helps you quickly see which questions are in which waves. Type what you're looking for in the search box. Use quotation marks to search this exact phrase. There are filters for the type of variable and for the wave. For example, in

Figure 8, we have typed C001, which is the variable name for self-rated health. The search returns a table that displays the wave, variable name, label, question text, and a cross-wave equivalents (xref). Clicking on the variable name will take you to the codebook entry for that item. You could also search for a phrase, like "self-rated health."

The cross-wave equivalents or cross reference (xref) is a feature that helps users find out if a specific variable from the survey data is available in other survey waves and under what names. This can be very useful, but it's not entirely accurate, particularly for earlier waves of data. The way variables were named was a bit inconsistent in the past but became more standardized from the year 2000 onwards. So, when looking at variable information from earlier surveys, like those from 1992 and 1994, be cautious. For instance, although self-rated health information exists in those years, they might not show up in the xref feature.

Figure 8. Concordance tool



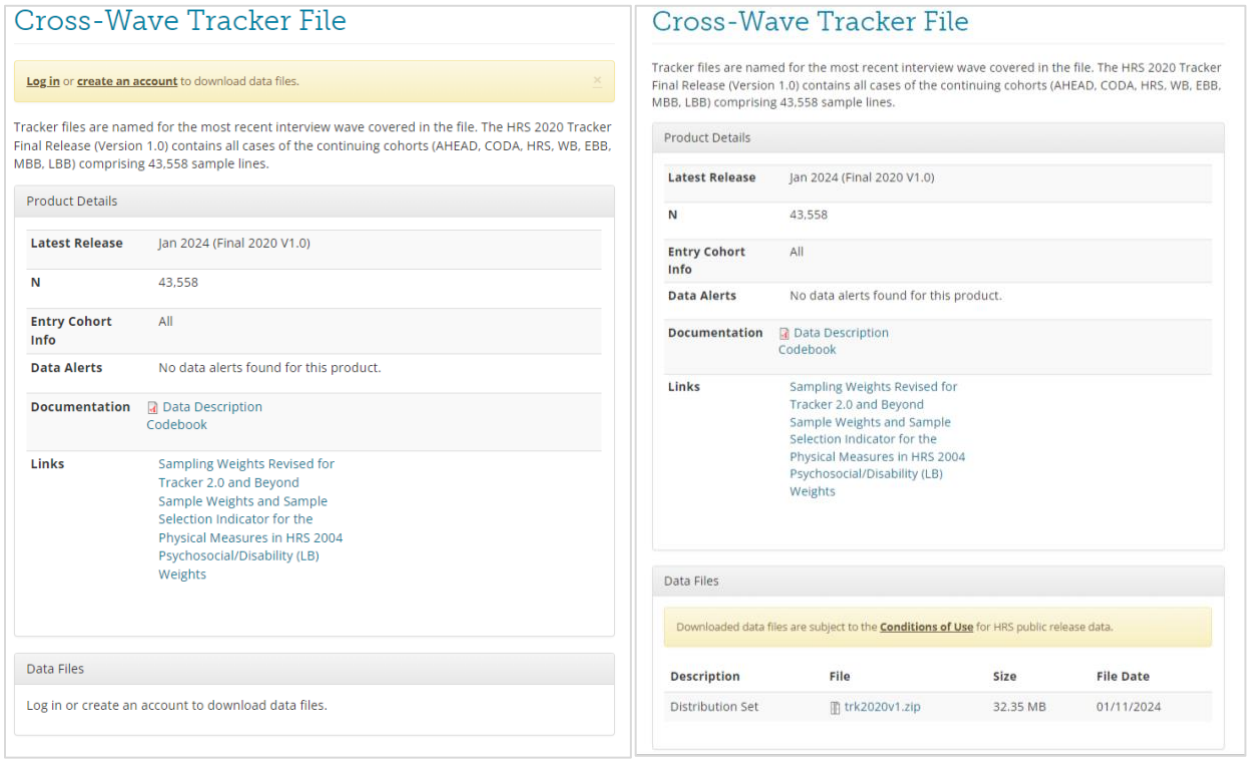
VISIT THE WEBSITE: Documentation>Question Concordance

Data Products

On the Data Products page, you'll find links sorted into three groups based on the type of data: public, sensitive health, and restricted. For the public and sensitive health categories, clicking a link for a specific data file will bring you to a page with more details about that file. At the top of this page, there is an overview that includes the release date and version of the file, the number of respondents and households included, which groups of respondents are in the file, when the data was gathered, any important notices about the data, and links to related documents like the data description, questionnaire, and codebook. If you haven't logged in,

you'll only be able to see this general information. However, once you log in, you'll also see the actual data files available for download below the documentation section (see Figure 9).

Figure 9. Showing what's available before and after logging in



Data Alerts

Data alerts are notices of errors, corrections, or problems in HRS early and final public data releases and associated documentation.

VISIT THE WEBSITE: Data Products>Data Alerts

Media & Publications

Online Bibliography

HRS keeps track of research publications that use its data. Researchers agree to inform HRS about their publications as part of the conditions for using the data. The HRS team also proactively looks through publication databases to add to their bibliography of research.

The Bibliography page shows the total number of related publications and offers options to download the list in different formats, like RTF for Word. You can sort the bibliography by the

first author's last name, title, type of publication, and year. There's a search box for more specific queries, and options to filter the search by publication type, year, and keywords.

As you search, the number of entries that match your criteria will be updated, and you can download these results as an RTF file. Each listed publication includes a link to PubMed or Google Scholar, which often gives you access to the full-text paper. This resource is beneficial for seeing how others have used HRS data and for generating your own research ideas.

VISIT THE WEBSITE: Media & Publications>Online Bibliography

CHAPTER 7

Data Structure and Management

Although we recommend beginning with the RAND HRS Longitudinal File, we are going to begin this chapter by describing the HRS original data structure and various merges that new users are likely to need to be able to do. We then describe the RAND HRS data files, which implement some of these merges. For both the HRS original data and the RAND HRS data, we provide important information that will help you begin using the data and merging data to create data extracts for your specific data analyses. All HRS original and RAND HRS data are released in "wide" format. We end this chapter with guidance on converting the files to "long" format, which may be useful in certain analyses.

Why do I need both RAND HRS data products and HRS original data files?

It is important to remember that the RAND HRS Longitudinal file does not include all of the biennial core data, nor does it include information from HRS off-years studies, sensitive data, restricted data, or even all of the tracking information in the Tracker file. The RAND HRS Longitudinal File has many advantages especially for new users but for seasoned researchers as well. For example, RAND constructs many variables that would take a lot of time for individual researchers to create such as total household wealth and income. Thus, in practice, most researchers make use of RAND HRS data products merged with HRS original files. This chapter will give you the tools you need to work with all of these data files.

Data Management with HRS Original Data

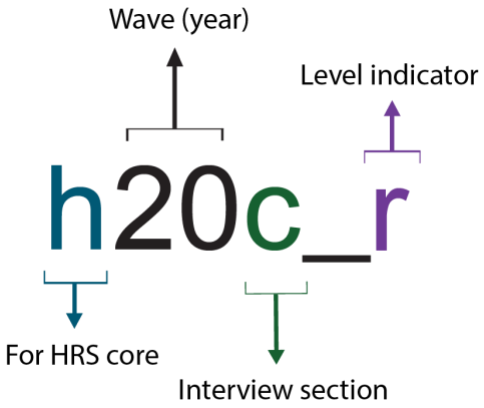
Several elements will help you manage the data. This section begins with information to help you understand the data file naming conventions, levels of data files, and the naming conventions of HRS original variables. All data files can be merged using identification variables.

Data File Naming Conventions

As noted above, the HRS original data files for the core interview are provided as a set for each wave, with individual data files for each interview section. Figure 10 provides an example of a data file name from the 2020 wave for Section C (health), which is a respondent level file. The "h" at the beginning of the file name represents the HRS core interview. Exit files from the core begin with an "x." The second two places are a number indicating the wave (1992-2022). The third part of the file name is a letter indicating the interview section. Then, there is an underscore followed by a letter that indicates the level of the file. Appendix Figure 3 provides a table with all 2020 public data files names listed by section. Note that this table is included in the data description for each core wave.

Appendix Figure 4 provides a summary of biennial core section content for 2022 and the letter that corresponds to each section in the file name. Appendix Figure 5 provides a key to the last part of the file name, the level indicator, which we describe next.

Figure 10. Data file naming convention, example for section C: health status



Levels of Core and Exit Data

In the core survey, most questions are asked of all respondents and apply specifically to the respondents themselves, such as self-rated health or current employment status. Some questions are asked about the household, such as the number of people living in the household. For these questions, in two-respondent households, household level questions are asked of one respondent who is designated as the financial respondent, family respondent, or cover screen respondent (the first respondent interviewed) on behalf of the entire household.

The core survey contains a range of questions about the respondents' children, siblings, and helpers. To facilitate use of these data, HRS Core data releases for each wave contain files at seven other levels: household member-and-child, sibling, helper, transfer-to-child, transfer-from-child, jobs, and pension. These are described in more detail in Appendix A. The two most commonly used levels of HRS original core data files are:

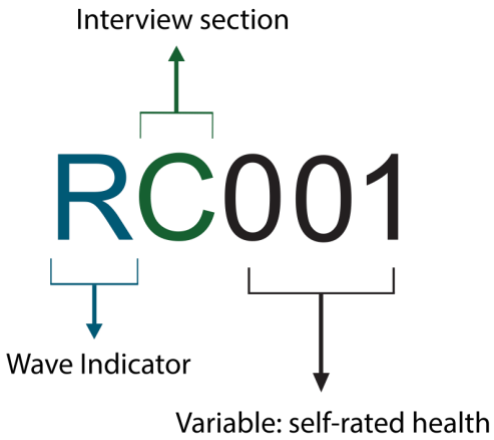
Respondent-level files contain questions that were asked of all respondents about themselves (or asked of a proxy about the respondent if the respondent was not able to give an interview). The files contain one record for each respondent or proxy who gave an interview. These files are available for core and exit files for all years.

Household-level files contain questions asked about the household. A coversheet respondent answered family questions on behalf of the entire household. Similarly, a family respondent answered family questions on behalf of the entire household, and a financial respondent answered household-level financial questions on behalf of the entire household. The household-level files contain one record for each household in which at least one interview was obtained. These files are available for core files for all years.

Variable Naming Conventions

Variable naming conventions in the HRS original data are not completely consistent, especially in the earlier waves. Beginning in 2000 (and especially in 2002), the variable naming convention depicted in Figure 11 was adopted and is reasonably consistent since then. The first letter in HRS original variable names is a letter that indicates the wave. Appendix Figure 6 provides the key to the wave indicators (column 2, HRS prefix). The second letter indicates the content section. Appendix Figure 4 provides a key to biennial core sections for the 2022 wave. Then, the numeric portion is the variable number. In the example, the first letter of the variable name is R, which indicates that the variable comes from the 2022 wave of core data. The second letter indicates that the variable comes from section C (health), and the variable number is 001, which is for self-rated health.

Figure 11. Variable naming convention 2000 onward, example for 2020 Section C: self-rated health status



Identification Variables

Two primary identification variables (HHID and PN) uniquely identify a particular respondent record in a data file. Other identification variables, such as sub-household ID, allow merging with files from other levels (e.g., household level or helper level).

Household Identification or HHID

As described in the first section of this guide, HRS seeks to interview individuals within households. Both members in coupled households are interviewed. Over time, HRS tracks complex and changing household structures including divorce, widowhood, and remarriage. Therefore, we need one variable that identifies the household and another that identifies a person within that household.

HHID is a 6-digit character variable assigned to a household at baseline. It is stable across all waves of data. It uniquely identifies the original household and any households derived from that household in subsequent waves.

Person Number or PN

Person number is a three-digit, character variable that is unique within an original household (HHID) that does not change across waves. It has values of 010 and 020 that are assigned to original respondents in a household. 010 is one member of a couple and 020 is the other member. Anything other than 010 or 020 (e.g., 011, 012, 021, etc…) are new spouse/partners of respondents who get re-married/re-partnered.

Sub-household ID or SUBHH

As noted, household composition can change over time as people experience widowhood, divorce/split and remarriage. Sub-household ID is a one digit, character variable. In combination with HHID, xSUBHH (where x is the letter wave indicator) uniquely identifies a household at a given wave. Sub-household IDs can be different at each wave. In the baseline interview, all households for the wave are considered 'original' households and have a SUBHH of zero. Respondents can have a sub-household identifier at a particular wave even if they did not provide an interview. The codes for SUBHH include:

0. Original household
1. Sub-household, split off from original
2. Sub-household, split off from original
3. Deceased respondent household
4. Deceased respondent household
5. Sub-household, split off a household that already split into a '1' and '2'
6. Sub-household, split off a household that already split into a '1' and '2'
7. Used when two respondents split and then recombine with each other
8. Sub-household, split off a household that already split into a '1' and '2'
9. Not in the sample this wave

Figures 12a-12d show examples of how the SUBHH changes with these potential household changes wave to wave.

Figure 12a. Married/partnered couple stays together

2006	2008
Married	Married
HHID=123456, PN=010, KSUBHH=0 HHID=123456, PN=020, KSUBHH=0	HHID=123456, PN=010, LSUBHH=0 HHID=123456, PN=020, LSUBHH=0

Figure 12b. One person in a married/partnered couple dies

2006	2008
Married	010 dies between waves
HHID=123456, PN=010, KSUBHH=0 HHID=123456, PN=020, KSUBHH=0	HHID=123456, PN=010, LSUBHH=3 HHID=123456, PN=020, LSUBHH=0

Figure 12c. Married/partnered couple divorces

2006	2008
Married	Divorced
HHID=123456, PN=010, KSUBHH=0 HHID=123456, PN=020, KSUBHH=0	HHID=123456, PN=010, LSUBHH=1 HHID=123456, PN=020, LSUBHH=2

Figure 12d. Married/partnered couple divorces, and one remarries

2006	2008
Married	Divorced, 010 remarries between waves
HHID=123456, PN=010, KSUBHH=0 HHID=123456, PN=020, KSUBHH=0	HHID=123456, PN=010, LSUBHH=1 HHID=123456, PN=011, LSUBHH=1 HHID=123456, PN=020, LSUBHH=2

Partner Person Number or PPN

xPPN is a three digit variable that identifies the person number of the spouse or partner of a sample member if the sample member is part of a couple. For all the sample members still alive in a given wave, a concatenation of their own person numbers (PN) with xPPN in ascending order would generate xCOUPID. For deceased sample members, xPPN gives the person numbers of their spouses or partners in the last wave during which they were alive. More detail about the PPN is available in the data description for Tracker.

VISIT THE WEBSITE: Documentation>Data Descriptions>Tracker

Other Person Number or OPN

OPN is a 3-digit, character variable. In combination with HHID and xSUBHH, OPN uniquely identifies another person that respondents provide information about, such as:

- Household members
- Children

Or in combination with HHID and PN, OPN identifies a

- Helper
- Sibling

Specifically, OPN is used in sections E, F, and G to keep track of siblings, children, and helpers, parents, and other household members, in data collection loops. Whereas, other identification variables are available in the Tracker file, the OPN is available in the core data file for each wave. More detail about the OPN is available in the data description for each core data product.

VISIT THE WEBSITE: Documentation>Data Descriptions>2020 HRS Core

Overlap Cases

Overlaps refer to cases that have multiple primary IDs and require special handling in constructing longitudinal files and in merging Tracker data to wave-specific files. The variables HHID and PN reflect the current status of the case, while overlap cases also have a former HHID and PN values from a previous wave. These former values are provided in the identification variables OVHHD and OVPN. More detail is provided in Appendix B.

Getting Ready to Use the HRS Original Data

As you get ready to use the data, there are several things you should be thinking about.

- Begin by identifying the variables you need for your analysis and the data files that contain them. It is a good idea to select only the variables you need from each file, since smaller data files are more manageable. Remember that the Concordance tool can help you search for variables in the HRS original data.

- If you are using longitudinal information for the same variable, it is a good idea to understand any potential differences in the question wording or changes in the code frame that may have occurred over time. Don't forget to consult the Alternate Wave Chart, for variables that are only asked every other wave.

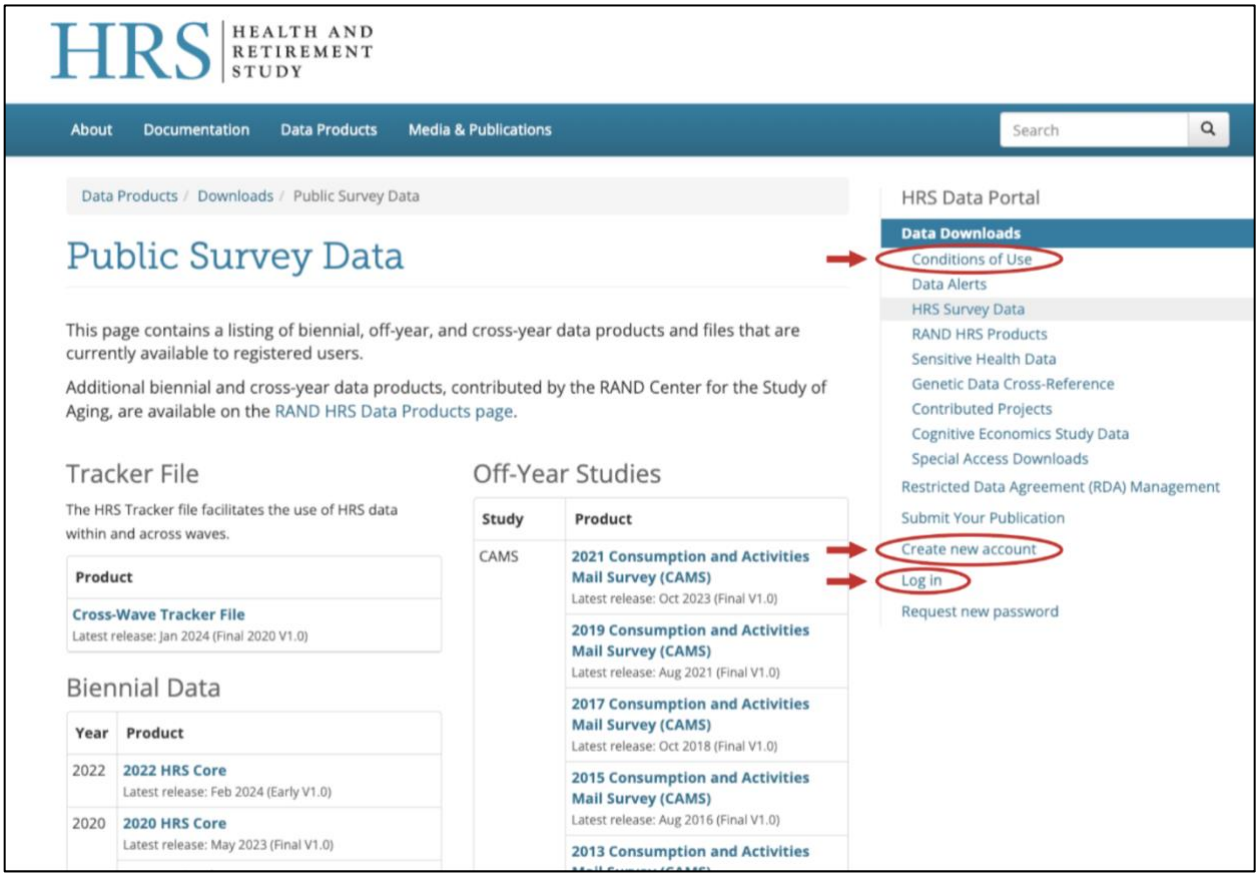
- Think about what the unit of analysis is: the individual, the household, a couple, respondent-caregiver, or other? Knowing what level of analysis you want will help you know what merges you need to do. If you are merging files, will you have one record per respondent, one record per household or something else? What identification variables will be required to merge the various files that contain variables needed for your analysis?

- If you are merging variables across data files, what type of merge will be required? Will the merge be a one-to-one matching of records (e.g., respondent-to-respondent) or a one-to-many (e.g., household-to-respondent) matching?

Register for the HRS Original Data

Begin by registering to access the data. To register for the public data, go to the Data Products link from the homepage. On the right-hand side of the page, you will see the set of links under HRS Data Portal shown in Figure 13.

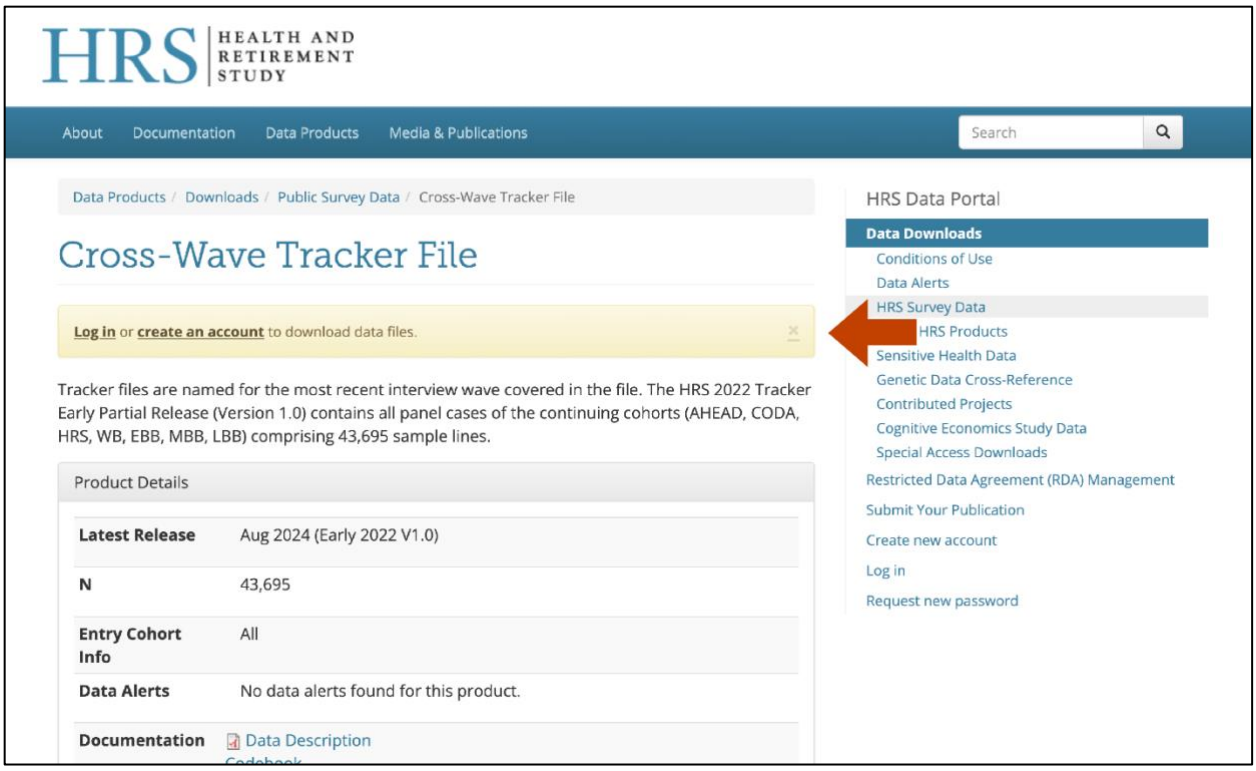
Figure 13. HRS Data Portal



After that, simply complete the registration information on the Create New Account link. You will receive login credentials that you will use each time you login to the website.

Come back to this page to use your credentials to login. The other way you can create an account or login after creating an account is a yellow bar across the top of the page for **every** HRS data file. See Figure 14 for an example to download the Cross-wave Tracker File.

Figure 14. Create account and login on the data products page for the Tracker File



Do read the Conditions of Use! This contains important information about citing the data in your manuscripts, about emailing hrspublications@umich.edu when you have research products to share (dissertation, thesis, journal article, book, book chapter, report, etc.), and about letting us know if your registration information changes.

Of key importance is the provision for *very limited* third party distribution of the data. You will also find here instructions for submitting replication datasets that may be required by journals where you are submitting your HRS research.

Stay informed: During registration, you will have the opportunity to sign up to receive quarterly user newsletters and data announcements. We recommend that you sign up for both.

Download and Get Ready to Use the HRS Original Data

Now you are ready to begin downloading data from the HRS website. Go to the data file you are interested in using and login. After you are logged in you will see the data files for download. We describe here the distribution files, focusing particularly on core data and our recommendation for a file structure to download to your PC or laptop. Note that the distribution set includes program statements to "build" the data files. From 2018 onward, the distribution set includes the data files already built. However, it also includes the program statements.

Distribution Files

The files are packaged for download from the HRS web site in two different ways – as one large .zip file that contains six smaller .zip files, one .pdf file, and one .txt file, or the eight smaller files available individually for separate download. For example, for the 2016 wave of core data, the combined file is H16core.zip.

The distribution set includes the data description, codebook, and questionnaire for the file. The data are available in ASCII, SPSS, SAS, and Stata file formats.

Note that when you open the zipped file for the core interview for any given biennial wave, you will see separate data files for each separate section (e.g., section C: Health etc.).

Directory Structure and Contents

While a particular setup is not required for using HRS files, we suggest a directory structure. By using this directory structure, you will not have to change the path name in your program statement files. If you use a different structure, just change the directory references in the program statement files.

Again, using the 2016 Core Wave as an example. The individual .zip files for separate download are: Data file H16da.zip contains data files. Program statement files H16sas.zip contains SAS program statements. H16sps.zip contains SPSS program statements. H16sta.zip contains STATA program statements. Documentation files H16cb.zip contains the codebook. H16qn.zip contains the questionnaire.

Directory	**Contents**
c:\hrs2016	Files downloaded from Web site
c:\hrs2016\codebook	Unzipped files from H16cb.zip
c:\hrs2016\data	Unzipped files from H16da.zip
c:\hrs2016\qnaire	Unzipped files from H16qn.zip
c:\hrs2016\sas	Unzipped files from H16sas.zip
c:\hrs2016\spss	Unzipped files from H16sps.zip
c:\hrs2016\stata	Unzipped files from H16sta.zip

Decompress the selected .zip files into the appropriate subdirectories.

Program Statements

All core data files for waves 1992 through 2016 need to be “built.” Each data file comes with associated SPSS, SAS, or STATA program statements to read the data. Files containing SPSS statements are named with .SPS extension, those with SAS statements with a .SAS extension, and those with STATA statements with .DO and .DCT extensions. The statement files are named beginning with the same prefix as the corresponding data file. For example, SAS statements in the file H16A_R.SAS go with the H16A_R.DA data file. Note that beginning with the 2018 wave,

HRS provides data files that are already "built," so you do not need to apply the program statements to get them ready to use.

Merging HRS Original Data Files

For many analyses, you may need to merge various HRS data files. We provide detail for the most common merges. Recall that HRS original files are provided at different levels previously described. Merging between and across files levels requires the use of the identification variables already described. The statistical code to accomplish all of the merges is available on the HRS website in SAS, STATA, SPSS, and R.

VISIT THE WEBSITE: Documentation>New User Guide

Merging Two or More Respondent-level Files

The most common type of merge with HRS original data is a respondent level merge of two or more respondent-level data files. Figure 15 illustrates this merge with an example dataset from two respondent level data files: the Tracker file (TRK2022TR_R) and H20C_R, the file containing data from Section C. Health for 2020. When we combine these files using the identifiers HHID and PN, we get one new merged file (TRK_H20_R).

Figure 15. Merging two or more respondent level files

Merging Two Household-level Files

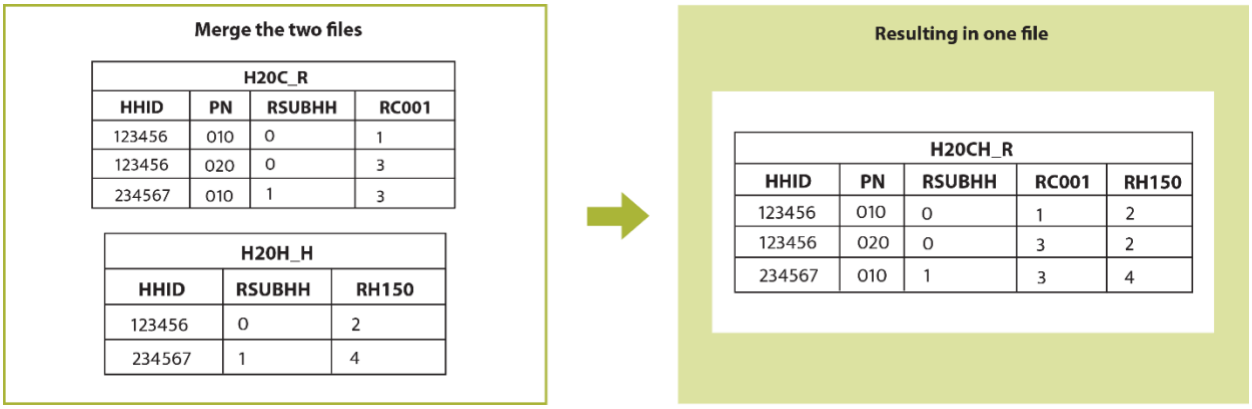
To create a household-level file with data from two or more household-level files, merge the household-level files using HHID and xSUBHH where xSUBHH is the current-wave SUBHH (x represents the HRS wave indicator, see Appendix Figure 6. Wave indicators and cohorts represented). Figure 16 illustrates this merge with an example dataset from two household level data files for wave R (2020): the file containing data from Section E. Family Structure (H20E_H) for the 2020 wave and the file containing data from Section H. Housing (H20H_H). When we combine these files using the identifiers HHID and RSUBHH we get one new merged file (H20EH_H).

Figure 16. Merging two or more household level files

Merging Household-level Data onto a Respondent-level File

Sometimes we need to merge data from a household level file with respondent-level data. Here we are putting the household level variable onto each respondent-level record, maintaining a respondent-level file. To do this, we merge the respondent-level file(s) and the household-level with an example dataset from one respondent level data file and one household level file: H20C_R and H20H_H. When we combine these files using the identifiers HHID and RSUBHH we get one file (H20CH_R).

Figure 17. Merging household level data onto a respondent level file

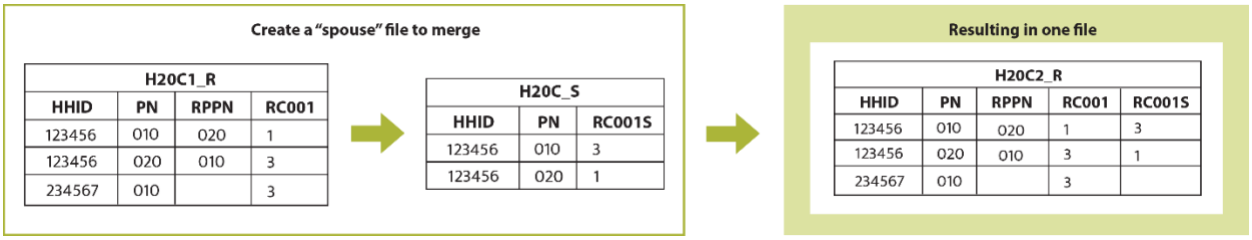


Combining Data from both Household Members

Some analyses are facilitated by combining data from two respondents from the same household. For example, if you wanted to study the impact of a partner's self-reported health status on the respondent's own health status, this would require data to be by structured so that respondent level data in wide format (one row for each respondent) has corollary partner data on the same line. Specifically, here we are performing a respondent-level merge to add information from one partner's record to the other. Figure 18 shows that we start with a respondent-level file, H20C1_R. Then we create a "spouse" file (H20C_S) by assigning the value of RPPN (a wave-specific partner person number for 2020) to PN (PN = RPPN), and renaming the "spouse" variable by adding S to the end (RC001S). Keep only HHID, the new PN (formerly RPPN), and the new spouse variables of interest, which yields the new dataset H20C_S.

These two files are then merged by HHID and PN to create H20C2_R, a file that contains self-reported health for each R and their partner on the same record (RC001 and RC001S).

Figure 18. Combining data from both household members in a respondent-level file



Merging Respondent-level Files across Two or More Waves

Figure 19 illustrates merging respondent level files longitudinally using two respondent level data files: H20C_R and H18C_R, the files containing data from Section C. Health for the 2018 and 2020 waves. When we combine these files using the identifiers HHID and PN we get one new merged file (H20_3R).

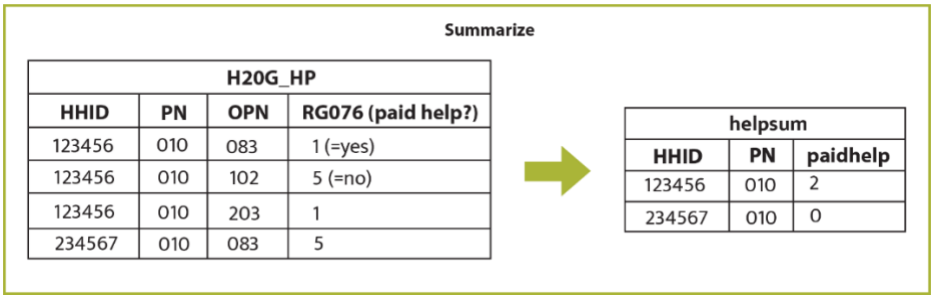
Figure 19. Merging respondent-level files across two or more waves

This graphic illustrates the **union** of the two waves where you end up with a merged file that includes respondents who participated in **either** wave. If you wanted to create a dataset that reflected the **intersection** of two waves, you would end up with a merged file that contained only respondents who had participated in **both** waves.

Merging a Helper-level File with a Respondent-level File

As shown in Figure 20, Data files like H20G_HP contain one record for each helper identified by the respondent in Section G. Functional Limitations and Helpers. The records in the helper file are uniquely identified by HHID, PN, and OPN (other person number). In the example below, respondent 123456/010 has three helpers, while respondent 234567/010 has one helper. Suppose we want to know the number of paid helpers the respondent has. To do this, we would count the number of records for each respondent (HHID/PN) with a value of '1 ' for RG076 in the helper file. We summarize this information by creating a dataset with a single record per respondent (helpSum). This file can then be merged to a respondent-level file by HHID/PN (see the previous example "Merging two or more respondent-level files").

Figure 20. Merging a helper-level file with a respondent-level file

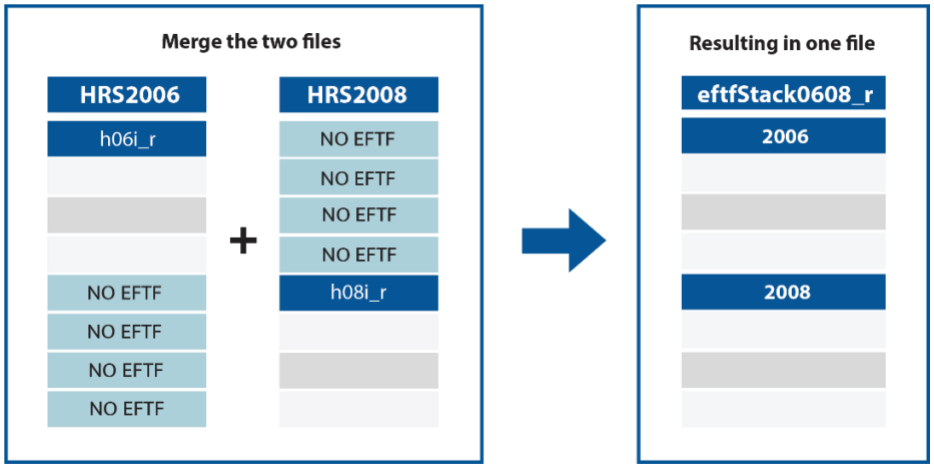


Stacking Enhanced Face-to-Face Interview Data from Adjacent Waves

As shown in Figure 4, data from the EFTF are on a random half of the core sample each wave. Some researchers choose to combine the data from the two half samples, usually combining 2006 and 2008, 2010 and 2012, 2014 and 2016, and 2018 and 2020. You can think of this as creating four longitudinal waves of EFTF data. In order to do this, we need to address the fact that the variable names change over time. In the HRS original data, from 2002 onward, most variable names are the same across waves, except for the alphabetic wave prefix. Note that some of the variables in the LB (psychosocial questionnaire) change over time, so you will need to rename these individually.

Figure 21 shows this for 2006 and 2008 for data from the enhanced face-to-face interview. Begin with the data files that include the variables you want from the 2006 EFTF Physical Measures section (h06i_r) and 2008 EFTF Physical Measures section (h08i_r). Next, rename the variables of interest so that the names are the same in the two datasets. This will allow the variables from adjacent wave EFTF half-samples to be easily combined. Finally, simply concatenate (or stack) the two datasets, creating a new dataset, eftfStack0608_r, with cases from both of the half-samples.

Figure 21. Stacking Enhanced Face-to-Face interview data from adjacent waves



Data Management with RAND HRS Data Products

As with the HRS original data, data management with the RAND HRS Data Products requires an understanding of some of the key features of the data including the variable naming conventions, the identification variables, and the structure of the data. As we will show, part of the processing that goes into creating RAND HRS Products involves some of the merges we have just discussed. The two most commonly used RAND HRS data products are the RAND HRS Longitudinal File and the RAND HRS Fat Files.

RAND HRS Longitudinal File Structure

To create the RAND HRS Longitudinal File, RAND researchers merge a large selection of variables from across the core biennial sections, including the exit interviews, for all available waves. They also merge some household level information, such as household income.

An important feature of the RAND HRS Longitudinal File is that it also merges some spouse/partner data (for coupled households) onto to their partner's record (as discussed in Combining data from both household members). To distinguish one respondent's information from the other, RAND uses the prefix designation of "S" in the variable name, as shown in the section of this guide describing the RAND HRS Longitudinal File Variable Naming Conventions. This makes it convenient for individual-level analyses that consider spouse or partner characteristics as potentially influencing an individual. Because there are separate records (lines) for each member of the couple, the data are also structured to allow you to conduct couples-levels "dyadic" analyses.

Consider the case of Shawn and Casey, who are a couple both providing interviews in the HRS (see Figure 22). We also have respondents who are not partnered, such as Ethel, and there is no need for an S variable for that individual. In the RAND HRS Longitudinal File, they would appear as:

Figure 22. Structure of the RAND HRS Longitudinal File

HHID	PN	R15SHLT	S15HLTH
123456	010 (Shawn)	1	3
123456	020 (Casey)	3	1
234567	010 (Ethel)	3	.U

.U=unmarried/partnered

As noted, HRS tracks changes in relationship status over time. The RAND Longitudinal File (like the Tracker File) contains records for all members of an original household as well as subsequent re-partnerships stemming from it. For example, if Shawn and Casey split, their records will remain in the data, but their S variables will now be missing (.U) from the wave that they were no longer partnered until (if) either of them re-partners. New partners are enrolled as respondents.

RAND creates a household wave-specific identification variable called HwHHID, which is HHID plus the wave-specific SUBHH and the PN. To provide additional assistance in tracking re-partnerings, they also include a variable called SwHHIDPN. xPPN and xPN_SP are used in the creation of SwHHIDPN, which assigns the Respondent's spouse each wave. If respondents change spouses between wave 1 and wave 2, S1HHIDPN will be different than S2HHIDPN. Otherwise, it will be the same or 0/missing (for no spouse from divorce/widow/etc).

Finally, a very important feature of the RAND HRS Longitudinal File is that is includes a wide range of constructed variables. For example, the HRS interview collects large amounts of information on sources of household income and household wealth. The Longitudinal File contains wave-specific summary variables for each of these important variables (HwITOT for total household income and HwATOTA for total household wealth). Another example of a constructed variable in the Longitudinal File that draws on a large number of HRS original variables is tenure in longest held job (RwJLTEN).

VISIT THE WEBSITE: Data Products>RAND HRS Products>RAND HRS Longitudinal File 2020

RAND HRS Longitudinal File Codebook

RAND creates very detailed codebooks that you will consult often as you use the data. Note that the codebook for the RAND HRS Longitudinal File codebook is very large, but is a searchable pdf and listed under the documentation.

It is very worthwhile to familiarize yourself with the structure of the codebook and be sure to read through the entry for any variables you use so that you are aware of what assumptions, if any, were used in the process of creating it. There is a lot of important introductory information at the beginning of the codebook, such as "What's New" in each data release, how to download the data and set up your computing environment, some simple programming examples (such as performing a merge), and so forth.

For each variable, the codebook includes:

- Variable list: for each wave, contains name, label and variable type (continuous or categorical)
- Descriptive statistics: N, mean, standard deviation, minimum, maximum
- Categorical variable codes: all variable values and frequencies by survey wave
- How constructed: defines what information the variable captures and describes missing values
- Cross wave differences in the HRS original data: describes any changes in the survey that may impact longitudinal analysis
- HRS original data variables used: name and label of all original HRS variables used to construct the variable

Note that the codebook provides frequency distributions for the "S" variables, but these do not represent separate individuals represented by the "R" variables. For any given wave, the count for the "S" variable is included in the count for the "R" variable. And if you subtracted the count

for any “R” variable in a given wave from the count for the “S” variable (R5CESD minus S5CESD, for example), you would have roughly the count in that wave of single households.

VISIT THE WEBSITE: Data Products>RAND HRS Products>RAND HRS Longitudinal File 2020

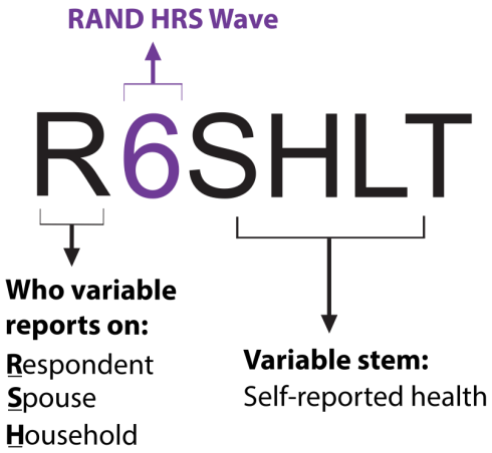
RAND HRS Fat Files

The RAND HRS Fat files are individual respondent-level datasets for each wave of the HRS that include all the variables from each respondent-level and household-level section merged into a single (very large) respondent-level file. These files can easily be merged with other RAND HRS data products (by HHIDPN), or with the any of the HRS original data files. The RAND HRS Fat Files are not processed and contain only HRS original variables.

RAND HRS Longitudinal File Variable Naming Conventions

One of the things that can be confusing using HRS original data is that the variable names, especially in the earlier waves of core data, were inconsistently named. All variables in the RAND HRS Longitudinal File are renamed to be longitudinally consistent across survey waves and follow a standardized naming convention, depicted in Figure 23.

Figure 23. RAND HRS Data Products variable naming conventions



The first letter indicates “who” the variable reports on: “R” indicates the respondent, “S” indicates a spouse or partner, and “H” indicates the entire household. Note that the “S” variables, which provide information from the respondent’s spouse, are provided for convenience only. Respondents’ spouses are still considered respondents in their own right and will have their own observation (row) with their own respondent-level variables, just as they do in the HRS original core data.

The second portion of the variable name is a number that indicates the RAND HRS wave number that corresponds to a survey year. Appendix Figure 6 shows the correspondence between the HRS original data wave indicator and the RAND HRS wave indicator. The last part of the variable name is the variable stem which, in this case, refers to self-reported health.

Some variables are not wave-specific and therefore do not have a wave number. These are mostly identifiers and demographic variables that do not change by wave. These variables have an "A" instead of a wave number (EX: RABYEAR is the Respondent's birth year). Also, the Exit Interview variables have an "E" instead of the wave number.

Identification Variables

All RAND HRS Data Products can easily be merged with other RAND HRS data products and with any of the HRS original data files.

The primary identification variables HHID and PN are included in the RAND HRS Longitudinal File (and all RAND HRS data products). RAND also creates a version of the original household identification (HHID) and person number (PN) that is simply a combination of the HHID and PN variables provided in Tracker. To merge any RAND product, you can use HHIDPN, but to merge to an HRS original data product, you have a couple of options.

You can merge on HHID and PN or you can create RAHHIDPN (a character version of HHID PN). Below we provide the code to create RAHHIDPN with the HRS original data.

SAS: RAHHIDPN = HHID || PN;

Stata: genstr9 rahhidpn=hhid+ pn;

SPSS: String RAHHIDPN (A9).

Compute RAHHIDPN = concat(HHID,PN).

Creating an Analysis File

The RAND HRS Longitudinal File contains a large part of the HRS original core data, and many users complete their research projects using only this data file. As you have seen, however, there is a vast amount of data collected in other parts of the study that you are likely to want to combine with the resources of the RAND HRS Longitudinal File. Even if you are drawing more heavily on data from other parts of the study, you will likely want to use some constructed variables, such as total household wealth and total household income. The RAND HRS Longitudinal File is a respondent-level file, so just follow the code for the appropriate merge given your needs.

Transforming Data Files File from "Wide" to "Long"

The HRS original data and RAND HRS data products are all in wide format, including those that contain cross-wave longitudinal data, such as the Tracker file and the RAND HRS Longitudinal File. That means that there is one record (row) for each respondent (HHID PN), and variables are in the columns for the longitudinal waves. It is often useful to restructure to a "long" format with one record per respondent per wave (HHID PD STUDYYR). Figure 24 illustrates this

restructuring graphically for the several waves of the Tracker file using the example of the variable interview type (IWTYPE).

Figure 24. Transforming the data from wide to long

TRK2016									
HHID	PN	GENDER	AIWTYPE	BIWTYPE	CIWTYPE	DIWTYPE	EIWTYPE	FIWTYPE	GIWTYPE
123456	010	1	1	99	1	99	11	99	99

longTrk					
HHID	PN	GENDER	STUDYYR	HRSPREFIX	IWTYPE
123456	010	1	1992	a	1
123456	010	1	1993	b	99
123456	010	1	1994	c	1
123456	010	1	1995	d	99
123456	010	1	1996	e	11
123456	010	1	1998	f	99
123456	010	1	2000	g	99

The code to accomplish this transformation is available on the HRS website in SAS, STATA, SPSS, and R.

VISIT THE WEBSITE: Documentation>New User Guide

CHAPTER 8

Applying Weights in Analyses

Once you have the analytic dataset ready, you are ready to begin data analysis. In this guide, we provide guidance on how to apply the HRS weight and design variables in analyses. We also provide the code for statistical analysis for some simple analyses, demonstrating how to apply the weight and complex sample design variables.

Weighting and Complex Variance Estimation

The major implication of the HRS complex sample design is that sample units, both individuals and households, have different probabilities of sample selection. Inferences to the U.S. population will be biased unless sampling weights are used in calculating descriptive statistics such as means and totals. Sample weights account for differential selection. As noted in Chapter 2, the sample is geographically stratified and clustered so variance estimates using standard formulae are not appropriate and need to be corrected in analyses using weights and strata and cluster indicators.

HRS provides the needed sampling weights. The baseline sampling weight starts from a base weight that accounts for differential probability of selection. It is the inverse of the probability of selection. It is fixed and, by design, greater than zero. Then we account for non-response by adjusting the weight for differential probabilities of survey participation among sampled units. The final step in creating the weight is a post-stratification that projects the sample to the target population size within groups defined by characteristics including birth cohort, marital status, race, and gender. The current reference data source is the American Community Survey.

The design variables used in analysis to adjust the variance for stratification and clustering are named STRATUM and SECU, respectively. The Tracker file contains the wave-specific weight variables and the design variables that you will need when using the weights. The data description for Tracker contains very useful information regarding these variables.

STRATUM is a variable that defines the sampling strata of the HRS data, and reflects the stratification at each stage of sample selection. Note that the number of values of STRATUM increases starting with the 2016 wave because of changes to the sampling strategy. The sampling error computation unit (SECU) variable is the stratum random half-sample code for the calculation of variance estimates using the balanced repeated replication method or approximate “two-per-stratum” Taylor series method. These common approaches to variance estimation under a complex sample design require two clusters per stratum. SECU is used as the cluster indicator for purposes of variance estimation assuming “ultimate cluster selection” of households from PSUs. You always need to two primary sampling units (PSU)s per stratum for the analysis to work.

Cross-sectional Weighting

In cross-sectional analysis, analysts should use either the respondent (xWGTR) or household (xWGTHH) weight for the year of data collection of interest and for the level of measurement of the measure of interest (respondent or household). The wave-specific weights (xWGTR where x is the wave indicator) are cross-sectional weights and allow for the data from represented birth cohorts in that wave to be used for population inference in the year of data collection. Appendix Figure 7 displays the variable names of the respondent and household weights available in Tracker (HRS Original) and the RAND HRS Longitudinal File (RAND HRS).

Longitudinal Weighting

A common question given the longitudinal nature of the data, is how to apply weights in longitudinal analyses. HRS does not construct longitudinal weights, although it is possible for users to do so. In longitudinal analyses the user needs to make choices about whether to weight and, if so, whether to use the base-year weight or the terminal-year weight. The terminal year weight is appropriate for retrospective analysis where the goal is to model or describe the histories of individuals or households who were in the population as of the terminal year. When it is possible for new respondents to enter the sample at different waves, the user needs to consider the implications of differential left censoring. Terminal-year weighting does not correct for differential left censoring.

The base-year weight is more appropriate for prospective analyses where the goal is to model or describe the future experiences of the base-year population. Most event history models are of this nature. In these models, we take a sample of individuals or households “at risk” and follow them through time until either the event of interest occurs or observation ceases. In these analyses the impact of sample attrition needs to be considered. Base-year weighting does not correct for attrition, and this attrition may result in serious biases if the attrition propensity is correlated with the propensity to experience the event of interest. An excellent resources is the 2nd edition of Applied Survey Data Analysis by Heeringa, West, and Berglund. Chapter 11 of this book focuses entirely on the Analysis of Longitudinal Complex Sample Survey Data.

Preserving the Sample Structure in Weighted Analyses

When you are interested in analyses within subsamples and you are using the weights and sample design indicators (STRATUM and SECU), it is best to preserve the sample structure by not subsetting the sample. The sample design indicators are designed to apply to the full sample for each wave. You would not get valid standard errors if, for example, you selected a subsample of only women in 2020 and then applied RWGTR (the respondent-level weight for the 2020 wave), SECU and STRATUM in your analyses. Rather, if you were interested in only women, you would use gender as a domain variable in analyses. You would obtain results for men and women separately, but with the full sample structure preserved. In the analysis examples, we note the placement of the domain variable.

Each program (SAS, Stata, SPSS, and R) handles the set up you may need to "tell" the program you are using data that employ a complex sample design and are using weights and design variables in the analysis.

The code to create two-way tables and to conduct logistic regression using weights, design variables, and the subpopulation (domain) command is available on the HRS website in SAS, STATA, SPSS, and R.

VISIT THE WEBSITE: Documentation>New User Guide

Correcting Error When At least One Stratum Contains Only One Cluster

To account for the complex sample design using typical variance estimation methods, it is necessary to have more than one cluster (HRS variable: SECU) per stratum (HRS variable: STRATUM). Using methods for subpopulations or domains in contemporary statistical analysis software, it is uncommon for the situation of a stratum with just one cluster to occur. One instance when this may happen using HRS is when analyzing data from one of the experimental modules found in Section V of the questionnaire. These modules are typically asked of a random 10% of the overall sample for a given wave, and there may be strata with only one cluster with non-missing data. In this circumstance, it is appropriate to collapse adjacent stratum as needed to maintain two clusters per stratum.

The code to combine strata is available on the HRS website in SAS, STATA, SPSS, and R.

VISIT THE WEBSITE: Documentation>New User Guide

Appendix Figures

Appendix Figure 1. Overview of the Content of the Psychosocial and Lifestyle SAQ: 2006–2022

	2006	2008	2010	2012	2014	2016	2018	2020	2022
50% EFTF Subsample	A	B	A	B	A	B	A	B	A
Social Participation – Activity Engagement		x	x	x	x	x	x	x	x
Life Satisfaction (Diener)	x	x	x	x	x	x	x	x	x
Social Network Composition (Kin/Friends)	x	x	x	x	x	x	x	x	x
Spouse – support, closeness – time together	x	x	x	x	x	x	x	x	x
Child – support, contact frequency	x	x	x	x	x	x	x	x	x
Family - support, contact frequency, type	x	x	x	x	x	x	x	x	x
Friends - support, contact frequency, type	x	x	x	x	x	x	x	x	x
Cynical Hostility	x	x	x	x			x	x	x
Optimism / Pessimism	x	x	x	x	x	x	x	x	x
Hopelessness	x	x	x	x	x	x	x	x	x
Loneliness	x	x	x	x	x	x	x	x	x
Neighborhood Disorder / Social Cohesion	x	x	x	x	x	x	x	x	x
Perceived Personal Control (Agency)	x	x	x	x	x	x	x	x	x
Domain Specific Control (Efficacy)	x	x	x	x	x	x	x	x	x
Positive and Negative Affect	x	x	x	x	x	x	x	x	x
Religiosity / Spirituality	x	x	x	x	x	x	x	x	x
Subjective Age / Self Perceptions of Aging		x	x	x	x	x	x	x	x
Perceived Everyday Discrimination	x	x	x	x	x	x	x	x	x
Attributions of Everyday Discrimination	x	x	x	x	x	x	x	x	x
The "Big 5" Personality Traits	x	x	x	x	x	x	x	x	x
Risk Attitudes (Risk Preferences)					x	x	x	x	x
Need for Cognition			x	x	x	x	x	x	
Concerns (Fears) about Aging									x
Purpose in Life (Psychological Well-being)	x	x	x	x	x	x	x	x	x
Domain-Specific Satisfaction		x	x	x	x	x	x	x	x
Self Esteem (Rosenberg)							x	x	x
End-of-Month Financial Strain	x	x	x	x	x	x	x	x	x
Ongoing Chronic Stressors	x		x	x	x	x	x	x	x
Perceived Stress Scale (PSS)							x	x	x
Anxiety (last week - Beck)	x	x	x	x			x	x	x
Subjective Social Status (Ladder)	x	x	x	x	x	x	x	x	x
Stressful Life Event – Last 5 Years	x	x	x	x				x	x
Access to & Activities with Modern Devices								x	x
Barriers to Device Use									x
Day Reconstruction / Experienced WB				x	x	x	x	x	x
Day Reconstruction – Activities / Time Use				x	x	x	x	x	x
Financial Well-being								x	x
Currently Working	x	x	x	x	x	x	x	x	x
Job Lock		x	x	x	x	x	x	x	x
Perceived Ability to Work		x	x	x	x	x	x	x	x
Work/Non-work Interference & Enhancement	x	x	x	x	x	x	x	x	x
Job Satisfaction	x	x	x	x	x	x	x	x	x

2020-2022 COVID-19 Pandemic Module

	2006	2008	2010	2012	2014	2016	2018	2020	2022
Specific COVID-related Worries								x	x
Changes in Family / Friend Connections: - Changes in Activities with Family & Friends - Stress due to Changes - Changes in Support Given / Received - Relationship Quality Changes								x x x x x	
Experienced Loneliness								x	
Experienced Discrimination								x	
Pandemic Social Distance Behaviors								x	
Activity Frequency Changes								x	
Learn New Device or Application								x	x
New Activities Using Devices									x
Personal / Household Work Impact								x	
Emotional Stress								x	
Resilience - Coping Strategies								x	x
Comments about Positive Experiences								x	x

Constructs Deleted Over Time

	2006	2008	2010	2012	2014	2016	2018	2020	2022
Well-being									
Positive and Negative Affect (MIDUS)	x								
Psychological WB: Growth and Self-Acceptance	x								
Experienced Well-being Yesterday				x	x	x	x		
Lifestyle									
Social Participation	x								
Community Meeting Attendance	x								
Frequency of Prayer		x							
Retrospective Social Participation		x	x	x					
Social Connections									
Partner Division of Labor					x	x			
Social Effort / Reward Balance	x	x	x	x					
Self-related Beliefs									
Compassionate and Self-Image Goals						x			
Personality									
Self-control / Impulsiveness			x	x					
Conscientiousness and Impulsiveness		x	x						
Anger (Spielberger scale)	x	x	x	x					
Work									
Work / Family Priorities	x								
Chronic Work Discrimination	x	x	x	x					
Job Satisfaction and Job Stressors	x	x	x	x					
Work Environment		x	x	x					
Coworker Support		x	x	x					
Supervisor Support		x	x	x					

Moved to Life History Mail Survey: 2015–2019

	2006	2008	2010	2012	2014	2016	2018	2020	2022
Quality of Relationships with Parents Early in Life		x	x	x					
Unusual Living Circumstances (homelessness, incarceration before age 50)				x	x	x			
Experiences of Lifetime Discrimination	x	x	x	x					
Lifetime Traumas	x	x	x	x					
Quality of Relationship with Mother Early in Life		x	x	x					
Lifetime Traumas before the Age of 18	x	x	x	x					

Appendix Figure 2. Summary of Off-year Studies

Off Year Studies that are Public Data

Consumption and Activities Mails Survey (CAMS) Biennial Mail Survey (2001-2021)

Cross sectional Mail Surveys

- 2021 HRS Perspectives on the Pandemic
- 2015/2017/2019 Life History Mail Survey (LHMS)
- 2013 Health Care and Nutrition Mail Survey (HCNS) (non-sensitive survey information)
- 2013 Veteran's Mail Survey
- 2011 Health Care Mail Survey (HCMS) (non-sensitive survey information)
- 2007 Disability Vignette Survey (DVS)
- 2001 Human Capital and Educational Expenses Survey (HUMS)
- 1999 HRS Mailout Survey

Internet surveys conducted in 2003, 2006/2007, 2009, 2011, 2013

Topics include: Internet use/social media; Cognition expectations; Health literacy; Psychosocial topics; Well-being; Childhood health; Residence history; Sibling transfers; Household spending; Income and assets; Prescription drug use; Health and disability; Health behaviors; Retirement preferences; Annuities.

Off Year Studies that are Sensitive Data

Aging Demographic and Memory Study (ADAMS) (2007, 2009, 2011)-in home clinical interview to diagnose dementia

Cross sectional Mail Surveys

- 2019 Health Survey
- 2013 Health Care and Nutrition Mail Survey (HCNS) (portion of the survey with sensitive health data)
- 2011 Health Care Mail Survey (HCMS) (portion of the survey with sensitive health data)
- 2009 Health and Well-being Study (HWB)
- 2003 Diabetes Study

Appendix Figure 3. HRS 2022 Data File Names by Section

Dataset	Section Description
H22a_r	Coverscreen (Respondent)
H22a_h	Coverscreen (Household)
H22b_r	Demographics
H22c_r	Physical Health
H22d_r	Cognition
H22e_fc	Family Structure (Children) and Transfers (From Child)
H22e_h	Family Structure (Children) (Household)
H22e_mc	Family Structure (Children) (HH Member Child)
H22e_tc	Family Structure (Children) and Transfers (To Child)
H22f_r	Parents, Siblings and Transfers (Respondent)
H22f_sb	Parents, Siblings and Transfers (Siblings)
H22g_hp	Functional Limitations and Helpers (Helper)
H22g_r	Functional Limitations and Helpers (Respondent)
H22h_h	Housing
H22i_r	Physical Measures
H22io_h	Interviewer Observations-Housing (Household)
H22io_r	Interviewer Observations (Respondent)
H22j_r	Employment
H22j2_p	Pension
H22j3_r	Retirement
H22lb_r	Leave-Behind Questionnaires
H22m_r	Disability
H22n_r	Health Services & Insurance
H22p_r	Expectations
H22pr_h	Preload (Household)
H22pr_jb	Preload (Jobs)
H22pr_mc	Preload (HH Member Child)
H22pr_p	Preload (Pension)
H22pr_r	Preload (Respondent)
H22pr_sb	Preload (Siblings)
H22q_h	Assets and Income
H22r_h	Asset Change
H22s_r	Widowhood and Divorce
H22t_r	Wills and Life Insurance
H22tn_r	Thumbnails
H22v_r	Modules
H22w_r	Event History, Internet Use and Social Security
H22y_r	Time Calculations

Appendix Figure 4. Summary of Biennial Core Section Content for 2022

Section Indicator	Section Content
PR	Preload
A	Coverscreen
B	Demographics
C	Physical Health
D	Cognition
E	Family Structure
F	Parents, Siblings and Transfers
G	Functional Limitations and Helpers
H	Housing
I	Physical Measures and Biomarkers
J	Employment
J2	Pensions
J3	Retirement and Social Security
M	Disability
N	Health Services and Insurance
P	Expectations
Q	Assets and Income
R	Asset Change
S	Widowhood and Divorce
T	Wills and Life Insurance
V	Modules
W	Event History, Internet Use and Social Security
Y	Time Calculations
IO	Interviewer Observations
LB	Leave-Behind Questionnaires
TN	Thumbnails

Appendix Figure 5. Indicators for Levels of Core Data

Indicator	What level (who is it about)	Data Structure
R	Respondent Level	One record per respondent
HP	Helper Level	One record per helper per respondent
SB	Sibling Level	One record per sibling per respondent
H	Household Level	One record per household
MC	Household Member & Child Level	One record per child/household member per respondent
TC	Transfer to Child Level	One record per monetary transfer a respondent gave to a child
FC	Transfer from Child Level	One record per monetary transfer a child gave to a respondent

Appendix Figure 6. Wave Indicators and Cohorts Represented

Year	HRS Wave Indicator	RAND Wave	Representative Birth Cohorts	Representative Ages	HRS Cohorts
1992	A (V)	1	1931-1941	51-61	HRS
1993	B (V)	2	<= 1923	70+	AHEAD
1994	C (W)	2	1931-1941	53-63	HRS
1995	D	3	<= 1923	72+	AHEAD
1996	E	3	1931-1941	55-65	HRS
1998	F	4	<= 1947	51+	HRS, AHEAD, CODA, WB
2000	G	5	<= 1947	53+	HRS, AHEAD, CODA, WB
2002	H	6	<= 1947	55+	HRS, AHEAD, CODA, WB
2004	J	7	<= 1953	51+	HRS, AHEAD, CODA, WB, EBB
2006	K	8	<= 1953	53+	HRS, AHEAD, CODA, WB, EBB
2008	L	9	<= 1953	55+	HRS, AHEAD, CODA, WB, EBB
2010	M	10	<= 1959	51+	HRS, AHEAD, CODA, WB, EBB, MBB
2012	N	11	<= 1959	53+	HRS, AHEAD, CODA, WB, EBB, MBB
2014	O	12	<= 1959	55+	HRS, AHEAD, CODA, WB, EBB, MBB
2016	P	13	<= 1965	51+	HRS, AHEAD, CODA, WB, EBB, MBB, LBB
2018	Q	14	<=1965	53+	HRS, AHEAD, CODA, WB, EBB, MBB, LBB
2020	R	15	<=1965	55+	HRS, AHEAD, CODA, WB, EBB, MBB, LBB
2022	S	16	<=1971	51+	HRS, AHEAD, CODA, WB, EBB, MBB, LBB, EGENX

Appendix Figure 7. Wave-specific Respondent and Household Weights

	HRS Original		RAND HRS	
Core Wave	Respondent	Household	Respondent	Household
1992	AWGTR	AWGTHH	R1WTRESP	R1WTHH
1993	BWGTR	BWGTHH	R2WTRESP	R2WTHH
1994	CWGTR	CWGTHH	R2WTRESP	R2WTHH
1995	DWGTR	DWGTHH	R3WTRESP	R3WTHH
1996	EWGTR	EWGTHH	R3WTRESP	R3WTHH
1998	FWGTR	FWGTHH	R4WTRESP	R4WTHH
2000	GWGTR	GWGTHH	R5WTRESP	R5WTHH
2002	HWGTR	HWGTHH	R6WTRESP	R6WTHH
2004	IWGTR	IWGTHH	R7WTRESP	R7WTHH
2006	JWGTR	JWGTHH	R8WTRESP	R8WTHH
2008	LWGTR	LWGTHH	R9WTRESP	R9WTHH
2010	MWGTR	MWGTHH	R10WTRESP	R10WTHH
2012	NWGTR	NWGTHH	R11WTRESP	R11WTHH
2014	OWGTR	OWGTHH	R12WTRESP	R12WTHH
2016	PWGTR	PWGTHH	R13WTRESP	R13WTHH
2018	QWGTR	QWGTHH	R14WTRESP	R14WTHH
2020	RWGTR	RWGTHH	R15WTRESP	R15WTHH
2022	SWGTR	SWGTHH	R16WTRESP	R16WTHH

Appendices

Appendix A. Levels of Core Data Files

As noted in the main text, the respondent-level and the household-level files are the most commonly used. Other level files are described in detail here.

Sibling Level Files (SB) consists of characteristics of the respondent's siblings. If a respondent had at least one living parent, he/she was asked a variety of questions about his/her siblings (Section F). The sibling file contains one record for each sibling of a respondent. Each respondent reports on his/her own parents and siblings. Sibling data are also stored in the preload section, H20PR_SB.

Household Member and Child Level Files (MC) contain information provided by the family respondent or financial respondent about each household member or child of the respondent or of the respondent's spouse or partner. The files contain one record for each household member (other than the respondent or the respondent's spouse or partner) or child.

This information can come from Section A. (coverscreen), the family respondent in Section E. Family Structure, and/or the preload section.

Helper Level Files (HP) contains information provided by each respondent about helpers other than the respondent's spouse or partner. A helper is a person or organization reported by the respondent as providing help with ADLs or IADLs. If the helper was a child, the record contains information about the help provided by the child and the child's spouse or partner, if any. The file contains one record for each helper (or, if the helper is a married or partnered child, the helping couple) for each respondent reporting help. If a child helped both mother and father, the file will contain two records – one of mother's report of the child's helping her and one for father's report of the child's helping him.

Transfer-to-Child-Level Files (TC) contains information provided by the family respondent about transfers of money to a child or grandchild. The file contains one record for each transfer to a child or grandchild.

Transfer-from-Child-Level (FC) contains information provided by the family respondent about transfers of money from a child or grandchild. The file contains one record for each transfer from a child or grandchild.

Job level file (JB) contain information on jobs the respondent has reported over time. Current jobs are represented, as well as past employment with unresolved pensions from past waves. The files have one record for each job represented.

Pension Level Files (P) contain information about pensions that the respondent has reported over the years during their involvement with the HRS. Each file contains one record for each pension that has been reported and followed up on in the current data collection.

Appendix B. Overlap Cases

Overlaps refer to cases that have multiple primary IDs and require special handling in constructing longitudinal files and in merging Tracker data to wave-specific files. The variables HHID and PN reflect the current status of the case, while overlap cases also have a former HHID and PN values from a previous wave. These former values are provided in the variables OVHHD and OVPN. There are two types of overlap cases in the HRS data. First, there were a number of original HRS 1992 (Wave 1) households eligible to be either an HRS or AHEAD household. An interview was attempted for all of them in HRS 1992. Afterwards, a random subsampling was performed, resulting in 60% of the cases remaining in HRS and the remainder assigned to AHEAD. We refer to the 134 cases transferred to AHEAD as “HRS inter-study overlap” cases. Among those 134 cases assigned to AHEAD, 108 were actually interviewed in AHEAD, and 26 were not. Second, there are a number of cases belonging to what we call “household merge overlap” resulting from intermarriage among respondents who entered the study in separate households. The OVHHID and OVPN columns (and corresponding variables in tracker) hold the individual’s original HHID and PN values, the HHID and PN columns (and corresponding variables in tracker) hold their current ID values and the final column lists the first year in which their current HHID and PN values were used. These cases are documented in the Tracker file data description. There is also a flag variable called OVRESULT that provides additional information about the overlap cases and OVYEAR that provides the year that the intermarriages occurred.