University of Michigan Retirement Research Center

Working Paper
WP 2013-273

Representativeness of the Low-Income Population in the Health and Retirement Study

Erik Meijer and Lynn A. Karoly

M R
R C

Project #: UM11-16

Representativeness of the Low-Income Population in the Health and Retirement Study

Erik Meijer
University of Southern California and RAND Corporation

Lynn A. Karoly
RAND Corporation

July 2013

Michigan Retirement Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48104
www.mrrc.isr.umich.edu
(734) 615-0422

Acknowledgements

This work was supported by a grant from the Social Security Administration through the Michigan Retirement Research Center (Grant # 5 RRC08098401-03-00). The findings and conclusions expressed are solely those of the author and do not represent the views of the Social Security Administration, any agency of the Federal government, or the Michigan Retirement Research Center.

Representativeness of the Low-Income Population in the Health and Retirement Study

Abstract

We study to what extent the Health and Retirement Study (HRS) is representative of all income groups, but with a particular emphasis on low-income groups. To focus on the HRS sample composition and abstract from potential measurement issues associated with measures of income and program participation, we exploit the SSA administrative data matched to the HRS sample and compare their distribution against the distribution of the same variables for the same population in the SSA databases. We find that overall, for cohorts and years that can be most reliably compared, the distributions are very similar and conclude that the HRS is representative for the population it covers. However, for some subgroups in the low-income population (e.g., recipients of Supplemental Security Income, Medicaid beneficiaries), there are some differences and thus we caution against estimating population totals for such small subpopulations. The HRS samples for which restricted matched administrative data are available are often not representative of a broad population of interest, because not all HRS respondents were asked permission to match in any given year. Therefore, the restricted HRS data files are generally not suitable for estimating population distributions, although they are still very useful for modeling purposes.

Citation

Meijer, Erik, and Lynn A. Karoly (2013). "Representativeness of the Low-Income Population in the Health and Retirement Study." Michigan Retirement Research Center (MRRC) Working Paper, WP 2013-273. http://www.mrrc.isr.umich.edu/publications/papers/pdf/wp273.pdf

Authors' Acknowledgements

This research was supported by the Social Security Administration through the Michigan Retirement Research Center (MRRC), project UM11-16. We thank Paul Davies at SSA for helpful discussions and Thuy Ho at SSA for assistance with the data. Roald Euller at RAND provided valuable programming support. We also thank Lynn Fisher at SSA, Michael Hurd at RAND, and participants at the April 2011 MRRC Researcher Workshop for comments at various stages of the project.

1. Introduction

The Health and Retirement Study (HRS) is a key data source used to analyze the health and economic status of the middle-aged and older population in the United States (Juster & Suzman, 1995; NIA, 2007). Many analyses are specifically interested in the low-income population that is eligible for various means-tested programs such as Supplemental Security Income (SSI) and Medicaid. However, the validity of studies that rely on the HRS depends on whether it accurately represents the size and composition of the low-income population. If that population is over- or under-represented, estimates of the share of the population as well as the number eligible for specific programs or with characteristics that capture dimensions of well-being (e.g., poor health, income below poverty), may be affected. Thus, it is vital to understand the extent of bias, if any, in the representativeness of the HRS sample of the low-income population.

Meijer, Karoly, and Michaud (2009, 2010) find evidence to suggest that the HRS may not accurately represent the low-income population. Specifically, the study used survey data from the HRS and the Survey of Income and Program Participation (SIPP), matched to Social Security Administration (SSA) administrative records, to estimate the size of the population eligible for the Medicare Part D Low-Income Subsidy (LIS). After carefully accounting for methodological issues such as panel data attrition, selective matching of survey and administrative data, and measurement error in survey data, the study found considerable divergence in estimates of LIS eligibility depending on whether the estimate was based primarily on HRS or SIPP data. Additional analyses of the potential sources of differences between the two data sources suggests that, even after taking sampling weights into account, the HRS sample may underrepresent low-income individuals. Particularly striking is that the number of noninstitutionalized Medicare beneficiaries aged 65 and older who are also enrolled in Medicaid, a Medicare Savings program, or SSI, is almost 50% higher in the SIPP-based estimates for 2006 than in the HRS-based estimates for the same year.

Further investigation of the differences in the two data sources is required, however, to reach a more definitive conclusion regarding the representativeness of the low-income population in the HRS. Hence, this paper undertakes a rigorous assessment of the representativeness of the low-income population in the HRS by using matched SSA administrative data on earnings and beneficiary payments and comparing the resulting distributions to marginal distributions that are directly taken from SSA records. The SSA records cover the entire population, so they provide a benchmark distribution, for any given income component, with which we can compare the distribution of the same administrative data measure in the administrative data sets matched to the HRS survey sample. Because the variables originate in the same SSA records, discrepancies between the distributions found in the direct SSA records and the distributions found in the administrative data sets matched to the HRS must be due to the HRS sample composition. As mentioned above, our previous results are suggestive of such

differences, but these analyses studied specific populations and were aimed at answering specific research questions that do not allow us to make more general statements about the sample composition of the HRS. Any deviation from representativeness in the HRS sample composition will not only affect the administrative data matched to the HRS sample, but many, if not most, HRS survey variables as well and thus problems with representativeness have wide-ranging implications.

Several studies to date have examined the quality of the income and wealth data in the HRS and other surveys such as the SIPP, for example, Scholz and Seshadri (2008), Sierminska, Michaud, and Rohwedder (2008), and Czajka and Denmead (2008). These studies compare income or wealth distributions across different data sets, and distributional differences are typically attributed to the quality of the variables involved, that is, quality of measurement at the individual level. For Medicaid beneficiary status, survey measurement quality, in particular underreporting, has been studied directly by matching subsamples of respondents with administrative sources (Card, Hildreth, and Shore-Sheppard, 2004; Davern, Klerman, and Ziegenfussi, 2007). While these studies suggest there may be issues of measurement error in income and wealth measures collected by the HRS or other surveys, as well as measures of program participation, we are not aware of any analyses to date that have considered the potential for the HRS to over- or under-represent the low-income population, a vital issue for the accuracy of research and policy analyses that focus on the low-income population using the HRS.

Our paper proceeds as follows. In the next section, we provide relevant information on the HRS data, the matched HRS-SSA administrative data, and the SSA administrative data sources that we rely on for our analysis. The third section describes the approach we take to investigating the representativeness of the HRS sample and the analytic challenges that arise given a number of data limitations. The fourth section presents the results, while a final section discusses the findings and their implications.

2. HRS and SSA Data Sources

Our empirical analysis focuses on the HRS which has become a widely used source of data for studying the population age 50 and above. We exploit the fact that the HRS data have been matched to SSA administrative records, although complexities are introduced based on when HRS respondents were asked for permission to link to such data and the extent to which permission was obtained. Thus, in the next section, we first detail the structure of the HRS samples and cohorts and the availability of the matched data. We then describe the SSA administrative data sources that we also rely on.

2.1 HRS Data: Sampling History, Cohorts, and Administrative Data Match

Initially, the HRS was a sample of individuals born in 1931-1941 and their spouses, and thus aimed to be representative of this population. Since 1998, the HRS is intended to be representative of the population age 50 years and older. Because the SSA administrative

data cover a broader population, we need to select our sample from the SSA administrative data to reflect the population the HRS is intended to represent. Therefore, we first describe the HRS sampling history and how it affects the target population.[1] Table 1 provides a summary of the HRS cohort structure.

The first wave of the HRS was conducted in 1992.[2] It sampled individuals born 1931-1941 and their spouses of any age.[3] This is called the "original HRS cohort" or simply the HRS cohort. In 1993, the AHEAD study (Assets and Health Dynamics Among the Oldest Old) conducted its first wave. At the time, it was a separate study, although closely related to the HRS (Soldo et al., 1997). It sampled individuals born in 1923 or earlier and their spouses of any age (including some couples who had been interviewed as part of the HRS the year before, the "overlap" cases). This is called the AHEAD cohort. HRS waves 2 and 3 were conducted in 1994 and 1996, respectively, and AHEAD wave 2 was conducted in 1995.

In 1998, the HRS and AHEAD studies were combined, and the combined study was also called HRS. Thus, the 1998 wave is wave 4. In this wave, the sample was expanded to include two additional cohorts. The CODA (Children of the Depression Age) cohort consists of individuals born 1924-1930 and the WB (War Babies) cohort consists of individuals born 1942-1947. Again, their spouses of any age were also included in the study. However, in selecting respondents, individuals from these birth years whose spouses were born in 1923 or earlier or 1931-1941 were not part of the sampling frame, because such couples were already represented in the AHEAD and HRS cohorts. Thus, after combination and expansion, the HRS was intended to be a representative sample of individuals born in 1947 or earlier and their spouses of any age.

From 1998 onward, HRS waves are conducted biennially and every six years, a new cohort is introduced, which covers the next six birth years and spouses that were born in those same years or later. The first such refreshment sample was the EBB (Early Baby Boomer) cohort, born 1948-1953 (and their spouses born 1948 or later), added in 2004. The second addition was the MBB (Mid Baby Boomer) cohort, which was added in 2010, but data for this cohort were not yet available for this study.

[1] For more detail on the HRS sampling structure, see National Institute on Aging (2007).

[2] In most waves, the field period that started in the year mentioned concluded in the next year, so not all respondents were interviewed in the same calendar year. We will refer the year the field period started as the year of the wave, as is common in descriptions of the HRS.

[3] For the purpose of sample selection, the HRS treats cohabitation the same as marriage. We refer to both married and unmarried partners as "spouses," as in most of the HRS documentation.

Table 1: HRS cohorts, years of sampling, and years they were asked permission to match.

Cohort	Birth years	Year of sampling	Ages at sampling	Ages in 2003	Years when asked permission to match[a]
AHEAD	-1923	1993	70+	80+	1993
CODA	1924-1930	1998	68-74	73-79	1998
HRS	1931-1941	1992	51-61	62-72	1992, 2004
WB	1942-1947	1998	51-56	56-61	1998
EBB	1948-1953	2004	51-56	50-55	2004

[a] In other years, some members were asked permission, but not the whole cohort.

The sampling frame for each new sample consists of noninstitutionalized individuals, which includes individuals in retirement homes, but not in nursing homes (and other institutions like prisons and mental hospitals). However, once in the sample, individuals are followed, even if they enter a nursing home. Furthermore, after household splits (divorces), both members are followed, even if one of them was not age-eligible for their sampling cohort. Also, any new spouses of respondents are added to the sample, regardless of their age.

As shown in Table 1, respondents are asked for permission to match their survey responses to SSA administrative records in the wave they are first interviewed, and in subsequent waves if permission was not obtained earlier.[4] Additionally, the HRS cohort was asked permission again in 2004. In 2006 and 2008, respondents in "enhanced face-to-face" interviews (as opposed to ordinary face-to-face interviews and interviews by telephone) of all cohorts except AHEAD who had not given permission in 2004 or later were again asked permission. Documentation for the HRS shows that most of the information available in the latest version of the matched administrative data files was obtained through permissions granted in 2004 to 2008, although a significant number of records for the original HRS and AHEAD cohorts derive from 1992 to 1996 permissions (HRS, 2010a, 2010b).

If a respondent gives permission to match to their SSA records, they are asked to provide their Social Security number (SSN). HRS sends the list of SSNs to SSA, along with other details, such as names and birth years to validate the match, SSA then extracts the records of these individuals from their databases and performs some postprocessing (e.g., selecting a smaller set of variables) and sends the data to HRS, which does some further postprocessing. Prior to 2006, permissions were only given retrospectively, and thus the data obtained through the 1992 permissions spanned the years from the earliest

[4] See HRS (2010b) for more detailed information about who was asked permission in which year.

available SSA records up to 1991 and data obtained through the 2004 permissions included data up to 2003. From 2006 onward, permissions are given prospectively (up to 2030), and data are updated biennially.

Subject to confidentiality agreements, strict safeguarding protocols, and restrictions on usage, these matched administrative data are available to researchers. Our study relies on the HRS Version 3.0 (2010) release of these data, the most recent available.

2.2 SSA Administrative Data

The SSA administrative data contain data that are collected for administering SSA programs, especially individuals' earnings, which are used for computing Social Security benefit entitlements, and the benefits themselves. We discuss each of the relevant administrative data sources.

Earnings data are maintained in the Master Earnings File (MEF) at SSA. Earnings from employment are obtained from W2 forms, as collected by the IRS. There are three alternative measures of earnings: total taxable earnings (Box 1), Social Security wages (Box 3), and Medicare wages and tips (Box 5). The latter earnings measure is most complete and therefore, we use this one. In particular, Social Security wages are censored at the Social Security taxable maximum.[5] More importantly for our study, certain (government) jobs are not covered by Social Security, which thus gives many zeros for individuals who otherwise have nonzero earnings. In contrast, the Medicare taxable maximum was $125,000 in 1991 and there was no maximum after 1993. Further, almost all earnings are Medicare-covered. Total taxable earnings are generally lower than Medicare earnings because of certain tax deductions, and therefore we prefer to use Medicare earnings as a measure of income. The data also include self-employment earnings, which are collected by IRS through Schedule SE. Again, there are Medicare and FICA (Social Security) versions and we use the Medicare version. Individuals may have multiple sources of earnings or self-employment earnings in a year, typically through multiple jobs. We aggregate them to an annual total.

Social Security benefits entitlement data are maintained in the Master Beneficiary Record (MBR), and actual payments data are maintained in the Payment History Update System (PHUS). Again, there are several alternative measures. We prefer to use the amount actually received by the individual in a given month, before deduction of the Medicare Part B premium. This amount is available from the PHUS files. However, it turns out that this variable was not included in all years data were matched to the HRS, so for analyses that rely on permissions given in years in which the PHUS data were not matched, we use the Monthly Benefit Amount (MBA) measure, which is the amount that

[5] For the years of interest to this study, the Social Security taxable maximum was $53,400 in 1991, $65,400 in 1997, and $87,000 in 2003 (see http://www.ssa.gov/OACT/COLA/cbb.html).

the individual was entitled to in the given month. This may differ from the paid amount because entitlements can be retroactively corrected and because entitlements are paid in the month following the month they refer to.

The MBR data also contain information about Medicare beneficiary status. This is recorded as a start date and a termination date, but only for the latest episode of coverage. It is conceivable that an individual is a Medicare beneficiary, then terminates, and later becomes a beneficiary again. For example, this can happen if Medicare eligibility is obtained through receipt of Social Security Disability Income (SSDI) and later coverage is lost after the SSDI benefits end when earnings exceed the threshold for "substantial gainful activity." If this happens, which episode of Medicare coverage we observe may depend on the date on which the data were extracted from the SSA database. However, it is rare to lose Medicare beneficiary status, so this is not likely to substantially influence our results.

In addition to Medicare status, the MBR data indicate whether the Medicare Part B premium was paid by a third party. This also takes the form of a start and stop date. Importantly, there is a code for the third party being the state Medicaid agency. If this flag is set, it implies that the individual is either a Medicaid beneficiary or enrolled in a Medicare Savings Program (MSP). The latter applies to individuals whose incomes are above the Medicaid threshold, but still quite low. This variable has been used previously by GAO (2004) and Meijer, Karoly, and Michaud (2009) to analyze Medicaid/MSP enrollment. Because Medicaid/MSP status is an indicator of low income, we include this measure in our analysis as well.

The universe of all records for these various administrative data sources is maintained by SSA. For this study, we rely on the Continuous Work History Sample (CWHS), a data set constructed by SSA for a random 1% sample of all individuals in the SSA administrative databases (see Panis et al., 2000, chapter 10). Upon our request, SSA in 2011 extracted data from their administrative databases and constructed variables for the CWHS sample exactly as they are constructed for the HRS-matched data (but without the HRS postprocessing). Because the CWHS is constructed using the same administrative data measures as those used in the HRS match, the measures in the two data sources should be highly comparable. We have analyzed the data for the CWHS sample by submitting our SAS programs, which were then executed by an SSA programmer.

3. Approach

As mentioned above, differences in the distribution of an income component across surveys or in the incidence of a measure of program participation may be due to differences in measurement quality across those surveys. To abstract from this potential for measurement error in survey data, we exploit the fact that the SSA's administrative

records have been matched to the HRS. Thus, we have access to measures of income and program participation for the HRS sample that derive from SSA administrative data sources. These same administrative data bases, housed at SSA, contain the universe of these administrative records for all individuals who have a Social Security number (SSN). If the distribution of these administrative variables in the HRS sample differs from the distribution of these same variables in the SSA database for the same population, this discrepancy cannot be attributed to better or worse measurement, because the same variables are used, as the variables for the HRS sample are just an extract of the SSA database. Hence, any difference in these distributions must be attributable to differences in sample composition. This is the basis for our empirical investigation.

In this section, we describe our analytic approach in more detail and the income and program participation measures we rely on. In principle, we could study any income source or program participation measure available in the SSA administrative records and compare its distribution in the HRS with the SSA administrative data source. Because this study originated with a concern about the representativeness of low-income individuals in the HRS, and in particular SSI and Medicaid beneficiaries, we focus on the income variables—with an emphasis on the lower tail of the income distribution—and SSI and Medicaid beneficiary status. For interpretability purposes, we construct measures that are economically meaningful, to the extent the data allow us to.

3.1 Construction of the Analysis Sample and the Comparison Variables

As shown in Table 1, the primary years in which respondents were asked permission to match with SSA records are 1992, 1993, 1998, and 2004. We focus on the 2004 sample, because it is the largest one and the most complete one in terms of both population covered and variables included in the administrative data. However, we will augment our main analyses with comparisons for 1992 and 1998 in a future version of this paper. The 1993 permissions are less useful for our purposes, for a number of reasons, and therefore we will not use the 1993 data. First, the 1993 permissions involved only the AHEAD cohort, which was substantially older at the time of sampling than most other cohorts, and therefore had a different sampling strategy (part of the sample was drawn from the Medicare enrollment database; Soldo et al., 1997). Second, as Adams et al. (2003) documented, the AHEAD sample was not representative of the whole birth year cohort it represents, because nursing home residents were excluded from the initial sample. They compared mortality in the sample with the lifetable and found that after the second AHEAD wave (1995), the sample was in line with the population.

Cohort membership in the HRS is partly a result of marital status and age of the spouse. For example, an individual born in 1949 and married to an individual born in 1940 is part of the HRS cohort (birth years 1931-1941), rather than the EBB cohort (1948-1953), because at the time of sampling the EBB cohort, the HRS cohort was already in the sample and the sampling strategy includes spouses of any age. In the administrative data, however, we have no information about spouses, let alone domestic

partners, which are treated the same for the purpose of sampling in the HRS. Hence, in our comparisons, we define cohort based only on birth year of the individual.

The HRS draws its samples from the noninstitutionalized population. Hence, the sample is intended to be representative of the noninstitutionalized population at the time of sampling. In the administrative data, we are not able to select only from the noninstitutionalized population, because institutionalization status is not known. Thus, the population from which the HRS sample is drawn differs from the population that we can study in the administrative records. However, most cohorts were sampled when their age-eligible members were in their fifties and very few individuals in that age group reside in nursing homes. Therefore, this will not bias our results considerably, except potentially for the 1998 sample from the CODA cohort. Moreover, the HRS follows respondents when they enter nursing homes and thus the sample should become representative of the whole population after a few waves. Unfortunately, however, sampling weights for nursing home residents are available in the HRS only for the 2000 and 2002 waves, which are not the waves we focus on. Thus, we can study unweighted analyses, which are potentially biased because of HRS's differential sampling probabilities (e.g., oversamples of minorities and Floridians) or weighted analyses, which exclude the nursing home residents. We do both types of analyses and generally find few noticeable differences between the two. In cases where we do find a difference, we study to what extent this is driven by the differences in the samples or by the weights.

Because we are specifically interested in any differences in the income distributions, we include all income components that are recorded in the administrative data in our analyses. As discussed in section 2, this includes earnings (wages and tips) from employment, self-employment earnings, Social Security benefits, and SSI benefits. For some of these there are multiple potential measures, with slightly different definitions. To get closest to a measure of the resources available to the individual, we use the most comprehensive ones available, with priority for measures that reflect what the individual actually received in a given month or year, as opposed to entitlements, which may be retroactively corrected. Thus, for earnings and self-employment earnings, we use the ones that are the basis for the Medicare tax, which are more inclusive than the other measures, for Social Security benefits, we use the actual income received as recorded in the PHUS files (adding any Medicare Part B premiums withheld), and for SSI we use the federal benefits received plus any state supplements received. In the benefits records matched to the HRS, there are two sub-records, reflecting primary and secondary benefits. Individuals may be eligible for Social Security benefits on account of their own past earnings, but also based on past earnings of a (former) spouse. If an individual has both types of entitlements and the spousal entitlements are higher, the individual is entitled to the higher of the two, but this is recorded as the full amount based on their own earnings in one sub-record and the difference between the two attributed to spousal benefits in the other sub-record. In the data for the CWHS sample, the variables for the second sub-record were included, but they were all blank. For maximum comparison,

therefore, we exclude the second sub-record from the analyses of the HRS sample as well.

In waves up to 2004, permission to match only included retrospective records, up to the calendar year before the date of the interview. Hence, data obtained through permissions given in 1992, 1998, and 2004 include records up to 1991, 1997, and 2003, respectively. Therefore, when studying representativeness of the sample for a target year, we study the income distribution for the prior calendar year. Specifically, we study the income distributions in 1991, 1997, and 2003. (The current version of this paper only includes comparisons for 2003.)

We will mostly focus on the sum of the included income components, which we will denote as "total observed income", "total SSA observable income", or simply "total income". This total income measure includes a large fraction of individual income, especially for lower income individuals, but it does not include income of the spouse (because we do not have information about the spouse in the administrative data), it does not include pensions and annuities (except for Social Security), asset income, and certain other income sources such as worker's compensation, unemployment insurance benefits, and government transfers other than the ones mentioned. This does not invalidate our comparisons, because we compare the same variables for the two samples, but it is important to keep this in mind when interpreting the income amounts in the tables.

In addition to total income, we study the separate components, but because not everybody receives each component, the sample sizes for these distributions are smaller, often much smaller, than for total income. Therefore, we mostly focus on the percentage of individuals who have nonzero income in each of the components.

As discussed in section 2, we also study dual eligibles (Medicare-Medicaid or Medicare Savings), which is an indicator for low household income and wealth. We classify an individual as a Medicare beneficiary if the start and termination dates of either Medicare Part A, or Part B, or both, enclose July 1 of the target year. A missing termination date means that the individual is still a beneficiary at the time of data extract, so we take this into account as well. For dual eligibility, we additionally require the start and stop dates of third party payments to enclose July 1 of the target year, and the type of third party payments to be "state buy-in".

3.2 Analytic Challenges

Our aim is to study the representativeness of the HRS survey sample, so that our conclusions will be relevant for users of the public use data. In particular, we intend to focus on the representativeness of the initial sampling. However, distributional differences may result from other sources as well, which poses some analytical challenges for the current study. As stated above, the HRS draws its samples from the noninstitutionalized population, and thus the initial sample represents a subpopulation that differs from the population in the administrative records. If a difference in the

distribution of interest is due to a difference between these two populations, it is not due to a sampling error, but to a deliberate (and common) choice in the sampling design. Fortunately, the fraction of individuals who reside in nursing homes in the birth year cohorts we study in their years of entering the sample is often very small and does not bias the results appreciably. Exceptions are the AHEAD cohort and the CODA cohort, which entered the sample at older ages. After a few years, even for these cohorts the discrepancy between the sampled population and the whole population becomes negligible. We study the AHEAD cohort only in later waves, but we intend to include comparisons for the baseline wave of the CODA cohort in a later version of this paper.

Because we will study later waves of most cohorts, discrepancies could be due to nonrandom panel attrition. If this is the case, the conclusions would apply to the later-wave survey sample, but not necessarily to the initial sample, and thus the conclusions would have more limited scope. Attrition in the HRS was studied by Michaud et al. (2011) and, to a lesser extent, Meijer, Karoly, & Michaud (2009), and both studies found little evidence for biases due to selective attrition. Thus, we can safely ignore this issue.

Like most survey data sources, the HRS includes sampling weights to make the sample more representative. These survey weights correct for deliberate differential sampling probabilities—which in the case of the HRS includes oversampling of minorities and Floridians—and differential nonresponse, including attrition. The weights are based on characteristics of the sampling design (design weights), which are then corrected to match distributions of certain demographics (such as age, sex, race, and marital status) in the population, as found in Census Bureau statistics (Census Bureau population estimates, American Community Survey, or Current Population Survey), resulting in poststratified weights. Users are generally advised to use the sampling weights if they wish to estimate population characteristics, like means and percentages, and thus our study of representativeness should aim to reflect this usage. Hence, we present weighted analyses. However, this leads to a potential selectivity, because nursing home residents do not have weights in the HRS (except in 2000 and 2002). Therefore, we also present unweighted results, which does include the nursing home residents but does not take the sampling weights into account. It turns out that in most cases, the differences between these approaches are not large.

The next potential source of bias is nonrandom selection of the sample of individuals who are asked permission to match with SSA records. As discussed above, the HRS did not ask all respondents permission in all waves. There are a few waves in which one or more complete sampling cohorts were asked permission. These are the most inclusive and thus least likely to suffer from selectivity. In the other waves, respondents who were asked permission are new spouses, respondents who were not present in an earlier wave in which they might have been asked permission, and respondents who previously refused permission. These are small and selective subsamples, and therefore we make no attempt at studying these separately. Since 2006, all individuals in "enhanced face-to-face" interviews are asked permission. We do not know whether these are random

subsamples, and we do not study 2006 or later. Note, however, that because after permission is obtained, historical data are matched, we do include the resulting data from these permissions in studying representativeness of earlier waves. That is, data for an individual who gave permission to match in 2002 or 2006 include information about 1997 income, and we do use this for studying representativity in the 1998 wave. However, the 2002 permissions do not give us data for 2003, and thus a respondent who most recently gave permission in 2002 is excluded from the analysis sample for the 2004 wave. Even for the waves in which entire cohorts were asked permission, the samples can be selective because the sample is based on the sampling cohort, whereas our comparisons study birth year cohorts, and the two are not identical. For example, in 2004, all members of the EBB sampling cohort were asked permission, but the EBB sampling cohort excludes any individuals who are in the EBB birth year cohort but have a spouse who is in one of the older cohorts. In section 4, we will show an example that illustrates the biasing effect of this difference.

After selecting the sample of individuals who were asked permission, the set of individuals who give permission, and the set for whom restricted data are obtained, may be selective. This is analogous to the attrition problem. It has been studied by Olson (1999), Haider and Solon (2000), and to a lesser extent Meijer, Karoly, and Michaud (2009), and similar to the attrition results, these studies did not find an appreciable biasing effect of nonrandom matching. Thus, we can safely ignore this issue.

However, there is a related issue that does affect our analyses, which has to do with the information provided by the HRS. We are not always able to pinpoint exactly who was asked permission to match, and we generally do not have a reliable indicator of who gave permission. An exception is 2004, for which we received a small data set from the HRS that includes a flag for whether a signed consent form was obtained from the respondent and whether the individual's record was sent to the SSA. The restricted files matched to the HRS only include records for respondents who had a record in the respective database. An individual who does not receive benefits yet does not have a record in the MBR database and thus will be missing from the restricted benefits file. Absence of a record then can be due to either not being matched at all (due to not having obtained permission or a failure of the match, e.g., a typo in the SSN) or to not having a record in the SSA database, and for the distributions it matters whether we can distinguish the true zeros from the mismatches. The CWHS sample was drawn from the master list of all SSNs (the Numident database) and thus includes true zeros, but not mismatches. For the HRS sample, we classify an observation as a match if the individual has a record in any of the restricted files we study, and as a nonmatch otherwise, but this is not a perfect measure. Therefore, in the results section, we will focus on income distributions conditional on having nonzero income, which guarantees a match and thus is insensitive to the issue, and only briefly discuss the zeros.

Conversely, there is also an issue with whether the CWHS sample represents the target population. In particular, the CWHS data contain records for individuals who do

not live in the U.S. and, more importantly, records of deceased individuals whose deaths were never recorded by SSA. The latter do not receive benefits and thus this issue also mainly affects the zeros. Appendix B studies this issue in more detail.

The HRS-matched restricted data have been postprocessed by the HRS team. For confidentiality reasons, earnings amounts have been rounded (generally to the nearest $100) and topcoded. This has not been done in the CWHS sample. Because we study percentages in broader income categories, this does not affect our comparisons. Another potential discrepancy in the data is that in the SSA databases, variables are often overwritten if later corrections are necessary. Thus, if benefit amounts are later corrected, the original amounts are overwritten. Because the data for the HRS sample were extracted at different points in time (1992-2010), and earlier than the data for the CWHS sample (2011), this introduces the possibility of discrepancies if these corrections are systematic, for example if they have a tendency to be upward corrections. We speculate that this will not be an important problem, but whenever there is a choice, we use variables that are never overwritten; in particular we use the PHUS variables whenever available. Our Medicare and dual eligibility variables are also potentially overwritten. SSA only keeps the latest episode of coverage of these in the MBR database. For Medicare, this is not a serious problem, because it is rare to lose Medicare beneficiary status. For example, in 1% or less of the observations in our study do we classify someone as a non-beneficiary because of an earlier termination date. However, for dual eligibility, this is more serious, because Medicaid beneficiary status may be updated monthly, and a pattern of temporary jobs and lay-offs may induce many short episodes of Medicaid eligibility. Given that the CWHS sample was extracted most recently, it is more likely to find a recent Medicaid episode of coverage that replaces one in, say 2003, in the CWHS sample than in the HRS sample.

In cases where there is a potential for discrepancies that are due to factors other than representativity of the HRS sample, we will interpret the results carefully in the light of alternative explanations.

4. Results

We now present the results of our analysis, focusing on results for 2003, when data are available for all five birth cohorts sampled in the HRS (see Table 1). We begin with the results for our measure of total observed income (defined above to include earnings plus Social Security and SSI benefits) and then consider income components. We then discuss results for Medicare and Medicaid beneficiary status. In all cases, we compare results for the CWHS, the SSA administrative data source, with the match-administrative data measures in the HRS sample. Results for the HRS are presented both unweighted and using the HRS sampling weights.

We note that the results presented here are preliminary. We inadvertently used primary[6] date of birth and death instead of beneficiary date of birth and death and our sample selection inadvertently removed individuals with no records in the MBR file (for example, individuals who have earnings but do not receive benefits yet). Neither of these invalidates the comparisons in this paper as long as the same procedures are applied to both data sources, but they hinder interpretation of the numbers. We will correct this in the next version of this paper.

4.1 Total Observed Income and Income Components

As noted in section 3, the HRS-matched data do not always allow a sharp distinction between income records that are missing because the match failed and income records that are missing because the individual did not have any earnings or benefits. For example, the fraction of individuals with nonzero total observed income is 90.2% in the CWHS in 2003 for individuals in the HRS birth cohort. That share is 71.9% in the corresponding unweighted HRS sample and 72.5% in the corresponding weighted HRS sample. Hence, the HRS sample contains too many zeros. However, we do not attach much importance to this discrepancy, as we cannot determine how much of this difference arises from mismatches versus true differences in the estimated share of the population with zero observed income.

In order to obtain further meaningful comparisons, we condition our remaining statistics on total observed income being nonzero, for which the match must have been successful. Figure 1 show the resulting income distribution in 2003 for the HRS birth cohort for individuals with nonzero income, both weighted and unweighted, along with the distribution based on the CWHS. We conclude that the HRS fits the population distribution quite well: there is no substantial difference between CWHS distribution and the unweighted and weighted HRS distributions.

Table 2 breaks down this distribution in observed income for the HRS cohort by sex. Females in this age group tend to have lower incomes than males, but again the distribution for the HRS sample is very similar to the one for the CWHS.

[6] The primary is the person whose earnings the benefits are based on; the beneficiary is the person receiving the benefits. We are interested in the latter. Often they are the same person, but especially for women in the studied age groups it is not uncommon that they receive benefits based on their (current, former, or deceased) spouses' earnings.

Figure 1: Distribution of total observed income in 2003, conditional on being nonzero, by sample; HRS birth year cohort.

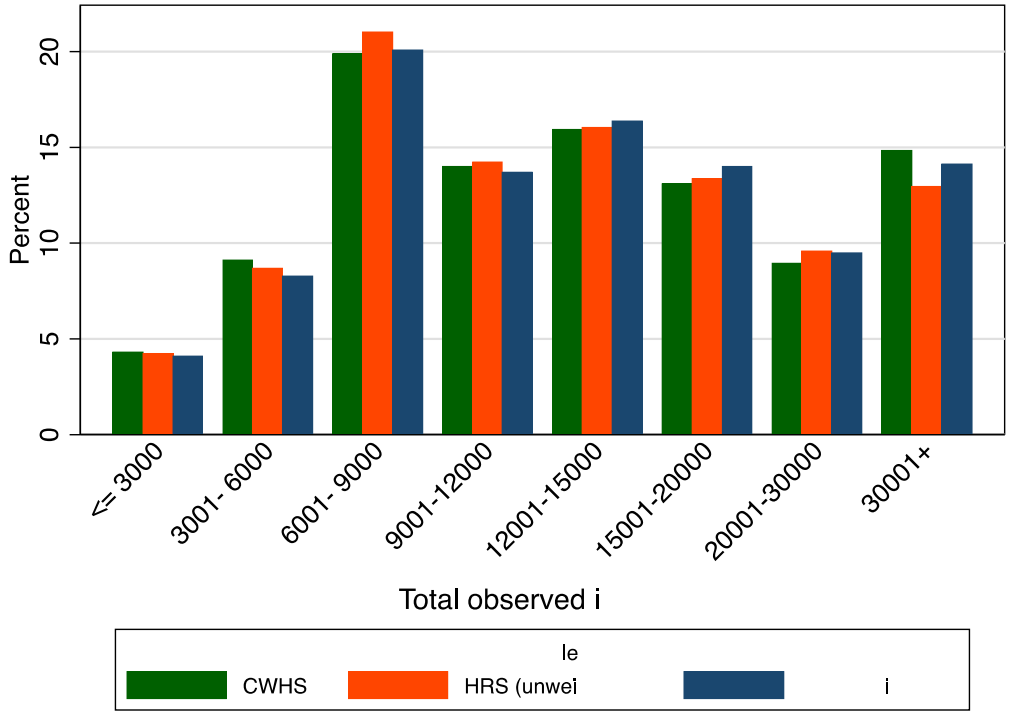


Table 2: Distribution of total income in 2003, by sample; HRS birth year cohort.

		Income distribution conditional on being nonzero ($)							
Sample	Percent nonzero	<= 3000	3001-6000	6001-9000	9001-12000	12001-15000	15001-20000	20001-30000	30000+
All									
CWHS	90.2	4.3	9.1	19.9	14.0	15.9	13.1	8.9	14.8
HRS (unw)	71.9	4.2	8.7	21.0	14.2	16.0	13.3	9.6	13.0
HRS (wgt)	72.5	4.1	8.3	20.1	13.7	16.4	14.0	9.5	14.1
Males									
CWHS	94.7	3.0	4.9	10.4	13.2	21.4	17.6	10.1	19.4
HRS (unwd)	74.4	3.2	4.9	11.6	12.2	20.9	19.1	10.0	18.0
HRS (wgt)	74.7	2.7	4.7	10.9	11.9	20.8	19.9	9.7	19.4
Females									
CWHS	85.7	5.7	13.6	30.1	14.8	10.0	8.2	7.7	9.9
HRS (unw)	69.7	5.2	12.2	29.9	16.1	11.4	7.9	9.1	8.2
HRS (wgt)	70.5	5.4	11.7	28.9	15.4	12.0	8.3	9.2	9.0

Table 3: Percentage of individuals with nonzero income components in 2003, conditional on total observed income being nonzero, by sample; HRS birth year cohort.

Sample	Wages and tips	Self-employment earnings	SS benefits	SSI benefits
CWHS	36.4	8.0	83.1	5.4
HRS (unweighted)	36.2	7.4	84.7	5.2
HRS (weighted)	36.8	7.8	84.7	3.9

Table 3 shows the percentages of respondents in the HRS birth cohort who have income from the four constituent sources that make up our measure of total observed income, conditional on having nonzero total income. Here, the fraction receiving SSI is lower in the HRS sample than in the CWHS sample, especially when applying the weights. The discrepancy we observe in SSI receipt for the HRS weighted sample, as opposed to the unweighted sample, could be due to these samples being different or to the weights changing the distribution. The former would be the case if a higher fraction of individuals who do not have a weight (namely those who are in nursing homes) are SSI beneficiaries, which does not appear implausible a priori. We computed the SSI recipiency rate using the same sample as for the weighted analyses (i.e., for those with nonzero weights), but not using the weights. These unweighted results are very similar to the unweighted results in the table: in particular, the fraction with nonzero SSI income is also 5.2%. Thus, the difference between the unweighted and weighted results is due to downweighting of respondents who received SSI benefits and not due to differences in sample composition.

We also looked at the SSI amount received, conditional on having nonzero SSI income. This analysis shows that there is also a discrepancy between the CWHS and HRS in the distribution of this income component for the HRS cohort. Specifically, the fraction of SSI recipients receiving more than $3000 is 56.5% in the CWHS sample and 42.2% in the unweighted and weighted HRS samples. Note, however, that these HRS samples are small, with about 230 individuals with nonzero SSI income in both the weighted and unweighted HRS samples.

Table 4 shows the distribution of total income for the other four birth cohorts in 2003. This analysis shows some noticeable differences, especially for the top income category for the oldest and youngest (AHEAD and EBB) respondents. The differences in Table 4 are most likely due to the selection of respondents who were asked permission in 2004 to match their SSA records. These were mostly respondents from the HRS and EBB sampling cohorts. Notably, the EBB sampling cohort that was asked for permission in 2004 does not include age-eligible individuals whose spouses were born before 1948 because those individuals would be sampled in one of the earlier cohorts, a feature that would potentially affect the representativeness of those in the cohort with matched administrative data.

Table 4: Distribution of total income in 2003, by birth year cohort and sample; other birth year cohorts.

Cohort		Income distribution conditional on being nonzero ($)							
Sample	Percent nonzero	<= 3000	3001-6000	6001-9000	9001-12000	12001-15000	15001-20000	20001-30000	30000+
AHEAD									
CWHS	86.2	2.2	11.8	20.2	28.5	20.7	12.3	3.1	1.2
HRS (unw)	89.7	2.2	12.4	17.2	30.1	20.4	10.0	3.4	4.3
HRS (wgt)	88.1	1.9	11.2	16.6	30.7	20.0	10.5	3.5	5.5
CODA									
CWHS	91.6	2.9	12.3	22.7	20.3	22.0	12.2	4.3	3.4
HRS (unw)	72.4	2.5	11.3	21.4	20.6	23.6	13.0	4.7	2.9
HRS (wgt)	72.4	2.5	10.5	20.3	21.7	23.8	12.9	5.1	3.2
WB									
CWHS	77.5	6.8	6.0	9.8	7.7	6.6	10.1	14.3	38.6
HRS (unw)	58.5	7.9	5.5	8.9	6.9	6.9	8.4	16.1	39.4
HRS (wgt)	59.1	7.1	4.6	7.5	5.8	6.9	8.3	15.9	43.9
EBB									
CWHS	80.1	7.4	7.0	15.1	11.1	8.2	11.3	13.4	26.4
HRS (unw)	75.0	4.5	5.3	7.9	7.7	6.3	7.8	17.0	43.5
HRS (wgt)	75.8	3.8	4.8	6.8	6.9	6.1	7.4	17.6	46.7

Figure 2 illustrates the relevance of this issue. Here, we have taken the HRS survey income components that correspond to the income components in the administrative data and constructed a survey version of the SSA administrative-data measure of observable income. The distribution of this survey-based income measure is shown for individuals born between 1948 and 1953 (the EBB birth year cohort), depending on whether they were sampled in the EBB cohort or HRS cohort (the cohorts that were asked permission in 2004) or another sampling cohort. This comparison shows the larger share of higher incomes among those in the EBB or HRS sampling cohorts.[7]

[7] Note that the survey income distribution has an even higher percentage of individuals in the top category. This is (partly) explained by the omission of secondary benefits from our administrative data. In Meijer, Karoly, & Michaud (2009), analyses using survey income variables gave similar results as analyses using the administrative variables, so we do not have evidence that survey income is systematically higher than administrative income.

Figure 2: Difference in income distributions between the EBB and HRS sampling cohorts and the other sampling cohorts for respondents from the EBB birth year cohort, 2003, HRS survey data.

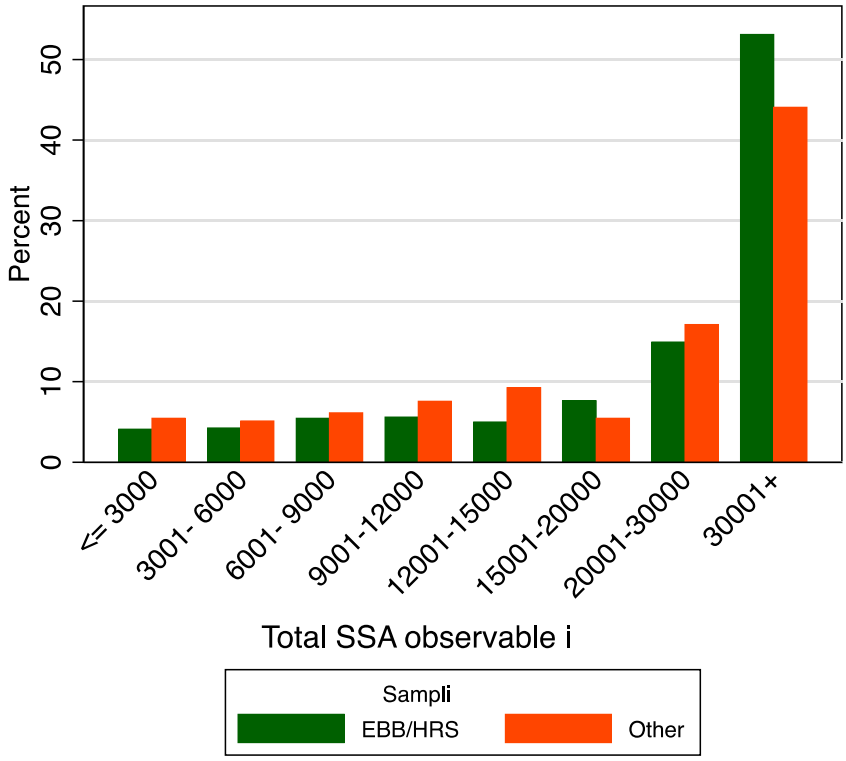


Table 5 shows the percentage with nonzero income components, conditional on a nonzero total observed income, for the other birth cohorts. Compared with the CWHS, SSI receipt is considerably less common in the HRS samples and having nonzero wages is more common, especially in the EBB cohort. This corroborates the results from Table 4 and again points at the lack of representativeness of these samples due to the selectivity in who was asked permission and is therefore included in the matched administrative data.

Table 5: Percentage of individuals with nonzero income components in 2003, conditional on total observed income being nonzero, by birth year cohort and sample; other birth year cohorts.

Cohort Sample	Wages and tips	Self-employment earnings	SS benefits	SSI benefits
AHEAD				
CWHS	4.2	1.7	96.2	6.4
HRS (unweighted)	9.9	2.5	93.4	0.9
HRS (weighted)	10.6	2.2	92.4	0.4
CODA				
CWHS	12.0	3.9	95.8	6.7
HRS (unweighted)	14.1	5.0	98.0	2.5
HRS (weighted)	14.5	5.0	97.7	2.3
WB				
CWHS	76.7	12.7	13.7	5.0
HRS (unweighted)	79.9	12.6	13.1	2.8
HRS (weighted)	81.0	13.4	12.0	2.3
EBB				
CWHS	68.3	7.8	24.7	10.0
HRS (unweighted)	84.9	11.1	8.2	3.7
HRS (weighted)	85.9	11.4	7.9	2.9

4.2 Medicare and Medicaid Beneficiary Status

In Meijer, Karoly, & Michaud (2009), we found indications that the number of dual eligibles—Medicare beneficiaries who are also Medicaid beneficiary or enrolled in an MSP—may be too low in the HRS. Medicaid or MSP beneficiary status indicates having low income. Table 6 presents the percentage of Medicare beneficiaries and the number of dual eligibles as a percentage of the number of Medicare beneficiaries. The percentage of Medicare beneficiaries may not be accurate because of a similar problem of distinguishing between true zeros and mismatches, but the fraction of dual eligibles as presented here does not suffer from this ambiguity. The table shows quite a few noticeable differences. For the WB and EBB cohorts, the two youngest age groups, the share of Medicare beneficiaries is substantially higher in the HRS sample than in the CWHS sample, which suggests that the HRS sample has too few zeros, the opposite of what we found for the income variables.

Table 6: Percentage Medicare beneficiaries and percentage dual eligibles among those in 2003, by birth year cohort and sample.

Cohort Sample	Medicare (%)	Dual eligibles (as % of Medicare)
AHEAD		
CWHS	92.1	12.2
HRS (unw)	90.1	11.6
HRS (wgt)	89.1	8.4
CODA		
CWHS	92.4	9.5
HRS (unw)	95.4	7.6
HRS (wgt)	94.9	7.4
HRS		
CWHS	58.1	8.8
HRS (unw)	61.5	11.3
HRS (wgt)	61.4	9.0
WB		
CWHS	3.4	53.3
HRS (unw)	11.2	23.4
HRS (wgt)	9.7	21.7
EBB		
CWHS	17.3	23.7
HRS (unw)	23.0	29.7
HRS (wgt)	25.8	28.6

Regarding the Medicaid/MSP measure, for the HRS and EBB cohorts, which are much more completely represented in the matched administrative data, the fraction of dual eligibles is higher in the HRS sample than in the CWHS sample, contrary to our earlier findings. For the WB cohort, it is dramatically lower. This is a small and highly selective sample, however.

Table 7 gives a breakdown of the Medicare beneficiary percentage by narrow age bands in the HRS birth cohort. As expected, there is a sharp break at age 65 in Medicare receipt. For the 62-64 age group, we see a large discrepancy in the reason why individuals were classified as not being a beneficiary. For the CWHS sample, this is predominantly because they have a later start date (i.e., the date on which Medicare beneficiary status begins), whereas for the HRS sample there is a large fraction for which

the start date is missing. This reflects the fact that the data for the CWHS sample were extracted in 2011, whereas the data for much of the HRS sample were extracted in 2004. The 62-64 age group has become eligible for Medicare in between these years, which is reflected in the updated records in the CWHS sample.

Table 7: Percentage Medicare beneficiaries and reasons for not being Medicare beneficiary, by birth year; HRS birth year cohort, 2003.

Birth years Sample	Medicare (%)	Later start date (%)	Start date missing (%)
1931-1933 (age 70-72)			
CWHS	90.7	1.6	7.7
HRS (unw)	93.3	0.6	6.1
HRS (wgt)	93.3	0.7	6.1
1934-1938 (age 65-69)			
CWHS	81.6	11.4	6.9
HRS (unw)	84.3	9.2	6.5
HRS (wgt)	84.4	8.9	6.6
1939-1941 (age 62-64)			
CWHS	0.8	86.0	13.1
HRS (unw)	3.2	44.4	52.4
HRS (wgt)	2.8	44.5	52.7

As pointed out earlier, the HRS sample of the WB cohort is relatively small and highly selective for 2003, and thus we expect important differences between the HRS sample and the CWHS sample for this birth cohort. Table 8 illustrates these differences. However, it is interesting to note that the number of dual eligibles among the female Medicare beneficiaries is almost twice as high as among the males in the CWHS sample, whereas in the HRS sample these percentages are of similar order of magnitude.

Table 8: Percentage Medicare beneficiaries and percentage dual eligibles among those, by sex; WB cohort, 2003.

Sample	Medicare (%)	Dual eligibles (as % of Medicare)
Males		
CWHS	3.9	39.2
HRS (unw)	10.2	26.8
HRS (wgt)	9.5	16.4
Females		
CWHS	2.9	72.2
HRS (unw)	11.8	21.5
HRS (wgt)	9.9	26.4

5. Discussion

The HRS is a data source that plays a key role in the analysis of the lives and well-being of the elderly and near-elderly in the United States. As such, it is one of the main sources of information for issues related to Social Security and Medicare. Indications from earlier studies led to concerns that the HRS may underrepresent low-income households. We study this issue by exploiting the SSA restricted administrative data that have been matched to the HRS. By comparing income distributions and benefit receipt in these data with the comparable measures in a 1% sample of SSA's own records, we are examining two samples from the same main administrative database with the same measures, and thus any difference in the distribution must be due to a difference in sample composition.

While this analytic strategy is valid in principle, there are some issues in the application of this method that need to be resolved before we can conclude that any discrepancy implies that the HRS is not representative of the population of interest. In practice, the distribution in the HRS sample used in the comparisons may differ from the distribution in the SSA sample for a number of potential reasons: (1) a nonrepresentative initial sample; (2) nonrandom attrition between the initial sample and the year for which we study representativeness; (3) a nonrandom subset of the respondents is asked permission to match their records with the SSA records; (4) selectivity in giving permission; (5) selectivity in successfully matching the administrative records to the survey data; and (6) an inability to limit the SSA sample to the same target population as in the HRS. Our aim is to study the representativeness of the initial sample, the first explanation. Earlier studies have shown that there is little evidence for the second, fourth, or fifth explanations: nonrandom attrition in the HRS, selectivity in giving permission, or selectivity in the match rate.

With regards to the third explanation, the subset of respondents who is asked permission is often a highly selective one. Therefore, we focus mostly on the "original HRS" cohort (birth years 1931-1941). All members of this birth year cohort were asked permission in 1992 and 2004, and thus this eliminates the selectivity in this step. Another advantage of the original HRS cohort is that we are better able to identify this population in the SSA administrative records, the sixth concern. For other HRS birth year cohorts, the members may be included in a different sampling cohort because their spouse had a different age. Information about an individual's spouse is not available in the SSA data, and thus we would not be able to clearly identify the same population as the one that was asked permission.

Another potential threat to the validity of the study is that the HRS initially samples individuals from the noninstitutionalized population, which cannot be identified in the SSA data. However, the HRS follows individuals when they enter a nursing home. We use wave 7 (2004), which is 12 years after initial sampling, and the literature has found that after this many years, the sample should represent the whole population. However, nursing home residents do not have sampling weights in the wave 7 data, which leaves us with the choice between doing unweighted analyses, which may give biased results due to differential sampling probabilities, or weighted analyses, which may give biased results because it ignores nursing home residents. We present results using both approaches and often they lead to similar inferences. Also note that the nursing home population is small.

From our main analysis, using the original HRS cohort in wave 7 (income refers to 2003), we conclude that the distribution of total observed income—earnings, Social Security benefits, and SSI benefits—for both samples are very similar, and there is not much difference between weighted and unweighted analysis. We conclude that the HRS is generally representative of the population of interest. A more detailed analysis, however, shows that the weighted HRS sample underrepresents the low-income population of SSI recipients (3.9% vs. 5.4% in the SSA sample). This is not the case for the unweighted sample (5.2%).

For the other birth year cohorts, we see more differences in the distribution of total observed income and the incidence of income components, namely SSI.. In most cases, we have been able to argue that these differences are likely due to the selectivity in which sample members are asked for permission to match. Thus, for these cohorts, the HRS sample with matched administrative data is less representative, but from this we cannot conclude that the survey sample is less representative.

In addition to income distributions and income components, we have also studied Medicaid beneficiary status. Medicaid is means tested and Medicaid beneficiary status indicates low household income and wealth, which are not otherwise observed in our data. In the data, we observe whether an individual's Medicare Part B premium is paid by the state Medicaid agency. This is done whenever the individual is a Medicaid

beneficiary or enrolled in a Medicare Savings Program, which is a program for slightly less poor, but still low-income individuals. Thus, we observe this indicator only for Medicare beneficiaries. The unweighted sample for the original HRS cohort seems to overrepresent such dual eligibles slightly, which is contrary to the indications from the literature, but there are some data limitations that do not allow us to draw firm conclusions. For other cohorts, dual eligibility is underrepresented, but as mentioned, these were selective samples because of who was asked permission to match.

In summary, we conclude from our analyses that the HRS sample is representative for broad analyses. For estimating population totals for small subpopulations, such as SSI recipients or Medicaid beneficiaries, some caution should be exercised. Researchers who intend to use the restricted data for their analyses should take note of which respondents were asked permission to match in which wave. In general, these data are often only representative of specific subpopulations.

In the course of this study, we came across numerous limitations of the matched administrative data, which have complicated our study. Some of these limitations are unavoidable, but we believe that the usefulness of the data can be improved in a number of ways that should be feasible to implement. We provide a list of potential improvements and enhancements in Appendix A.

Appendix A: Potential improvements of and useful enhancements to the administrative and public use data

In the paper, we have mentioned a number of the challenges we faced in working with the matched administrative data in the restricted-access HRS files: difficulty distinguishing true zeros from mismatches, that certain variables are only available in certain years, and other limitations of the restricted (and public use) data. This appendix provides a list of potential improvements and enhancements that we believe are useful for researchers:

1. The following variables can be added to each wave of survey data: (a) a flag indicating whether the respondent was asked permission to match with SSA records; (b) the preload variables that were used to determine whether permission was asked; (c) a flag indicating whether the respondent ultimately gave permission, i.e., a signed consent form was obtained; (d) a flag indicating whether the respondent's information was sent to SSA.

2. The following variable can be added to the restricted data for each survey respondent: (e) a flag whether the HRS received a record from SSA for this respondent.

3. It would be helpful to add a Numident extract to the restricted data files. The Numident file has very little information (although it has some demographics), but the main purpose of including it would be to indicate whether matching was successful, because everyone with a Social Security Number has a record in it. This would allow researchers to clearly distinguish zeros from mismatches.

4. The following variables can be added to the restricted data set: (f) flags indicating for each source database (Numident, MBR, MEF, SSR) whether a record was found for the respondent (and thus included in the data).

5. It would be helpful to obtain and include the variables that were not obtained in the extracts based on the 1992-2002 permissions (e.g., the PHUS variables and SSI). Although we have not studied the consent form or the legal matters related to it, we suspect that the permissions given by the respondents in those years would not restrict the variables one could match. For some variables, it may be difficult to trace back the records as they were at the end of the year before permission was given, but the PHUS records are not overwritten, so for these variables it should be straightforward.

6. For research properties, it would be very useful to keep records that are currently overwritten. For example, for Medicare and Medicaid, we only observe the start date and termination date (if any) of the latest episode of coverage at the time of record extract, and thus we do not observe earlier

episodes. Ideally, this would be implemented in the master SSA databases, for example, by replacing these start and stop dates by a set of monthly variables indicating whether the respondent was a beneficiary in that month. Alternatively, the HRS would be able to keep old versions of the data (especially now the data are updated every two years), so that information about earlier Medicare and Medicaid episodes (and some other variables that are overwritten, like disability applications) are still available for later analysis.

7. For many analyses (not necessarily related to the restricted data), it would be very helpful if the HRS survey data contained weights for nursing home residents.

Appendix B: Representativity issues with the CWHS sample and their consequences for our study

The CWHS sample is a one percent sample of all Social Security Numbers (SSNs). Because it was drawn directly from the universe of all SSNs, there is no scope for biases due to the sampling design. As with any sample, it is possible that the sample turns out to be unrepresentative due to chance, but given the size of the sample, it is extremely unlikely that this will lead to noticeable lack of representativity. However, all of this pertains to the population sampled, which is "Social Security Numbers". Each SSN is associated with an individual, but not every individual with an SSN is in the population of interest and not every individual in the population of interest has an SSN.

To start with the latter, the HRS aims to be representative of all individuals who live in the U.S., subject to certain age restrictions. This includes individuals who live in the U.S. but do not have an SSN, in particular undocumented immigrants. These individuals are not covered by the CWHS sample. Because the HRS-matched administrative data also originate from the SSA records, these are subject to the same undercoverage. Hence, this does not invalidate our analyses, but it means that the analyses are limited to the population of individuals who have an SSN. Our analysis cannot address whether the HRS is representative of the subpopulation of individuals without an SSN. Because this subpopulation is relatively small (approximately 5% of the population; Sohn & Oh, 2010), our results should also reflect representativity of the whole population, but not representativity of small subpopulations of which individuals without an SSN are a sizable part.

The converse problem, that not every SSN in the SSA database represents an individual in the target population, may happen for at least two reasons. The first is that individuals may leave the U.S. For example, a temporary foreign worker or student may return to their home country. Such individuals are assigned an SSN upon their arrival in the U.S., but this SSN is not removed from the databases after they leave. The SSA may not even be aware of their departure. Natural born and naturalized U.S. citizens may also decide to emigrate temporarily or permanently without their SSNs being removed from the databases. Likewise, individuals living in the "outlying areas" (Puerto Rico being the most prominent one) are not in the HRS target population but are included in the SSA's records.

The second reason why an SSN in the SSA's database may not represent an individual in the target population is that the individual may have died and SSA may not be aware of it. This issue has received some attention in the literature, primarily in the context of the SSA's Death Master File (DMF), which is derived from the SSA's Numident database (Hill & Rosenwaike, 2001; Schisterman & Whitcomb, 2004; Buchanich et al., 2005; Sohn & Oh, 2010; Wojcik et al., 2010). The focus of this literature is the mortality rate in a given year, and often for a specific sample or subpopulation (e.g., the participants in a clinical study). Some of these studies reach

relatively optimistic conclusions, because about 95 percent of the deaths in a year are recorded in the DMF, among individuals at older ages. The correspondence is much lower for younger individuals, because they do not receive Social Security Benefits yet, and therefore SSA does not track their whereabouts closely. The 95 percent is deemed satisfactory by some of the mentioned authors, but others view the information as not accurate enough, but still useful as a screening device.

We do not directly use the DMF. Rather, we combine information from the MBR and Numident, giving priority to the MBR information if available, because this is more accurate because SSA tracks individuals closely if they are current beneficiaries. Therefore, we believe our usage of the SSA information leads to the highest possible accuracy. However, the 95 percent in any given year (and lower percentages for younger years), combined with the emigrants, leads to an accumulation of records of older (or dead) individuals who are not in the target population. For example, we found that the CWHS sample contains a substantial number of individuals born before 1900, which we removed for our study.

Figures A1-A3 illustrate the issue for our target years. Because the CWHS is a 1 percent sample, the unweighted sample size multiplied by 100 should be an estimate of the population size. We thus computed estimates of the male and female population for each single year of age in 1991, 1997, and 2003, and compared these with corresponding estimates of the U.S. resident population obtained from the U.S. Census Bureau. Figures A1-A3 show that up to age 40, the CWHS overestimates resident population by about 20 percent in 1991, but much smaller percentages in 1997 and 2003, which is likely caused by the improvements in SSA record keeping mentioned by Hill and Rosenwaike (2001). However, in all three years, we see a steep increase in the overestimate at higher ages. Hence, even though at higher ages deaths occurring in the U.S. in a given year are better recorded, the cumulative effect of missed deaths in prior years, emigration, and the diminishing size of the target population due to mortality lead to large increases in the overestimation of population size in the CWHS sample. The difference between males and females can be attributed to the higher mortality of men, but we have made no attempts to investigate whether this accounts for the complete difference. Figures A4-A6 show the same phenomenon in absolute sizes, in five-year age bands, which shows that at older ages the absolute differences do not appear to increase noticeably, and that the rapid increase of the relative differences is due to the shrinking size of the population at those ages.

Because SSA verifies eligibility for benefits, individuals who receive Social Security benefits are generally alive. Consequently, the overrepresentation of deceased individuals in the CWHS sample leads to an overestimate of the number of individuals who do not receive any benefits and, for that matter, are not Medicare or Medicaid beneficiaries. Our comparisons mostly condition on benefits being nonzero, and thus these should not be affected by the missed deaths. The effect of emigration is more complex. Under some circumstances, individuals who do not reside in the U.S. may be

receive Social Security benefits. For the outlying areas, this is true in general. We studied the Annual Statistical Supplements to the Social Security Bulletin for the years 1992, 1998, and 2004, which have statistics for December of the prior year. From Table 5.J2 in these publications, we find that about 2.4 percent of the beneficiaries are not in the target population. Although this is not negligible, this does not account for a large part of the discrepancies in Figures A1-A6. However, as the Statistical Supplements also show (Tables 5.J6-5.J9), the benefits received in foreign countries and outlying areas are typically quite low. This may cause a distortion in the distributions of benefits of about one percentage point. This margin must be taken into account when interpreting the comparisons between the HRS and CWHS distributions.

Figure A1. Percent difference between CWHS population estimate and Census Bureau residential population estimate, by sex and age, 1991.

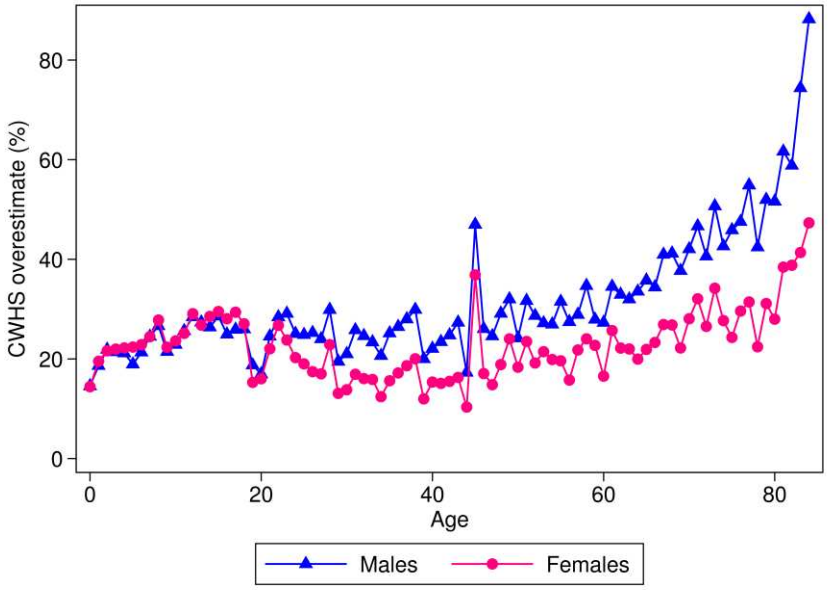


Figure A2. Percent difference between CWHS population estimate and Census Bureau residential population estimate, by sex and age, 1997.

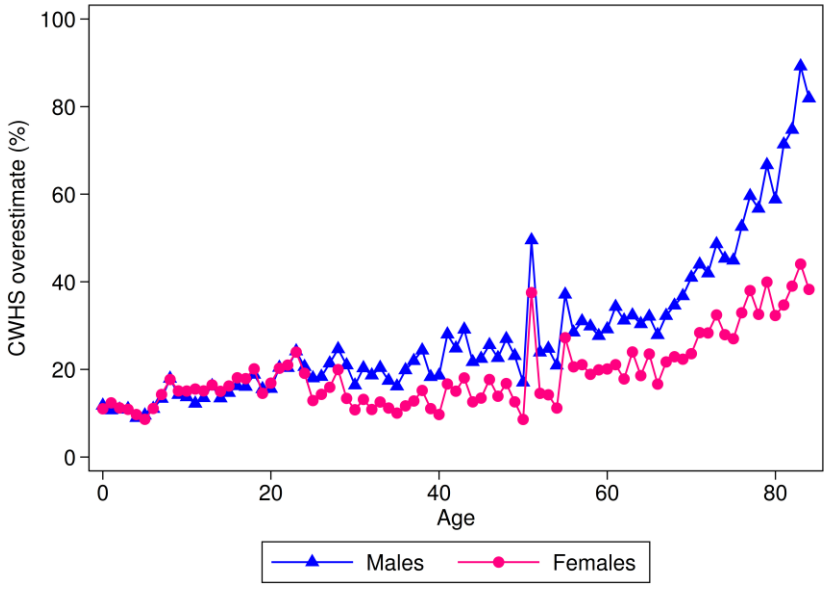


Figure A3. Percent difference between CWHS population estimate and Census Bureau residential population estimate, by sex and age, 2003.

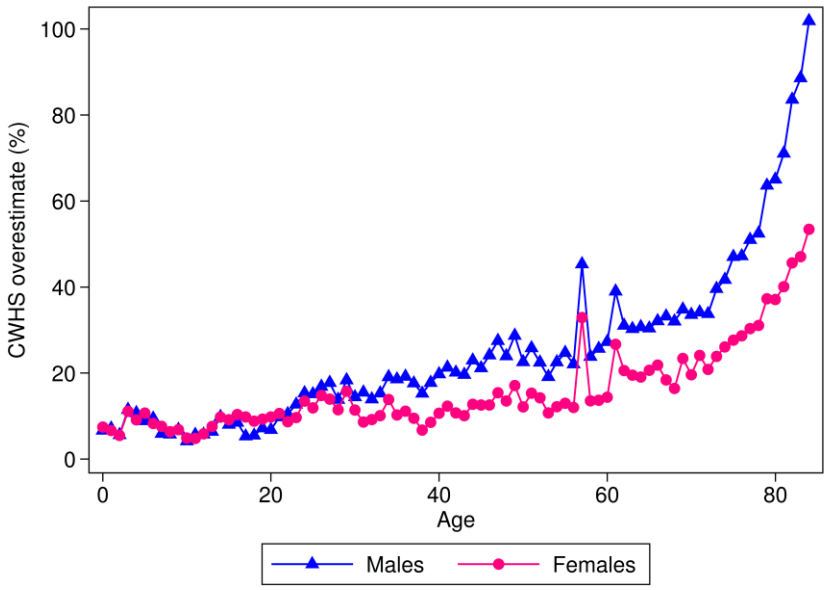


Figure A4. Population in millions, by sex and age, 1991, according to Census Bureau estimates (U.S. residential population; dark bars) and CWHS sample (light bars).

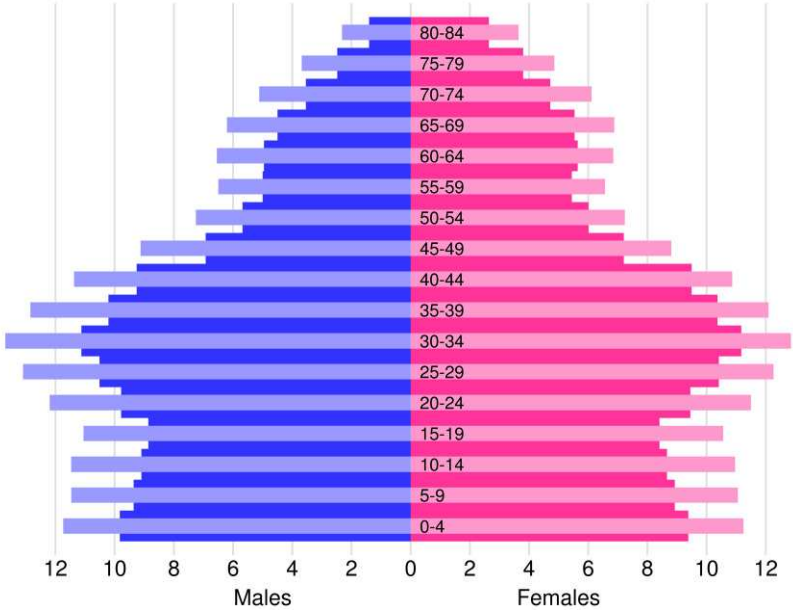


Figure A5. Population in millions, by sex and age, 1997, according to Census Bureau estimates (U.S. residential population; dark bars) and CWHS sample (light bars).

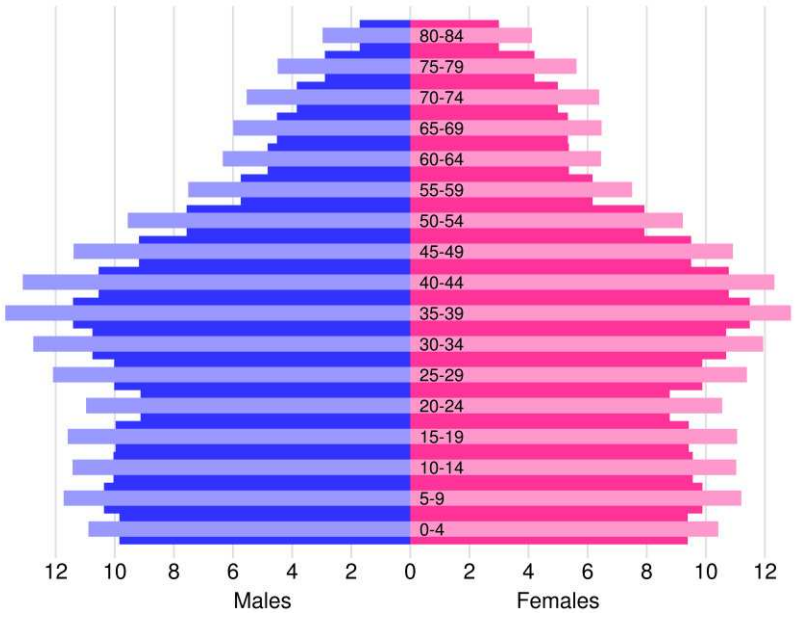
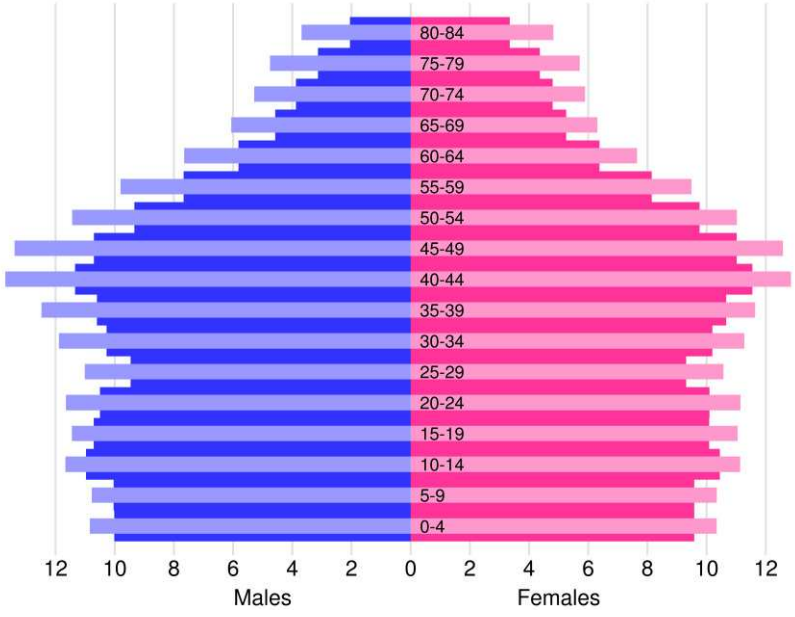


Figure A6. Population in millions, by sex and age, 2003, according to Census Bureau estimates (U.S. residential population; dark bars) and CWHS sample (light bars).



References

Adams, Peter, Michael D. Hurd, Daniel McFadden, Angela Merrill, and Tiago Ribeiro, "Healthy, Wealthy, and Wise? Tests for Direct Causal Paths Between Health and Socioeconomic Status," Journal of Econometrics, Vol. 112, No. 1, January 2003, pp. 3–56.

Buchanich, Jeanine M., David G. Dolan, Gary M. Marsh, and Jaime Madrigano, "Underascertainment of Deaths using Social Security Records: A Recommended Solution to a Little-Known Problem." American Journal of Epidemiology, Vol. 162, 2005, pp. 193–194.

Card, David, Andrew K. G. Hildreth, and Lara D. Shore-Sheppard, "The Measurement of Medicaid Coverage in the SIPP: Evidence from a Comparison of Matched Records," Journal of Business and Economic Statistics, Vol. 22, No. 4, October 2004, pp. 410–420.

Czajka, John L., and Gabrielle Denmead, Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys, final report, Washington, D.C.: Mathematica Policy Research, December 2008.

Davern, Michael, Jacob Alex Klerman, and Jeanette Ziegenfussi, Medicaid Under-Reporting in the Current Population Survey and One Approach for a Partial Correction, Santa Monica, Calif.: RAND Corporation, WR-532, 2007. As of September 28, 2012: http://www.rand.org/pubs/working_papers/WR532/

GAO, *Medicare Savings Programs: Results of Social Security Administration's 2002* outreach to low-income beneficiaries. Washington, DC: U.S. General Accounting Office, GAO-04-363, 2004. As of September 26, 2012: http://purl.access.gpo.gov/GPO/LPS48446

Haider, Steven, and Gary Solon, Non Random Selection in the HRS Social Security Earnings Sample, Santa Monica, Calif.: RAND Corporation, DRU-2254-NIA, 2000. As of September 28, 2012: http://www.rand.org/pubs/drafts/DRU2254/

Health and Retirement Study, Respondent cross-year benefits: Data description and usage, Version 3.0. Ann Arbor, MI: University of Michigan, Institute for Social Research, 2010a.

Health and Retirement Study, Respondent cross-year detail earnings: Data description and usage, Version 3.0. Ann Arbor, MI: University of Michigan, Institute for Social Research, 2010b.

Hill, Mark E., and Ira Rosenwaike, "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages." Social Security Bulletin, Vol. 64, No. 1, 2001, pp. 45–51.

Juster, F. Thomas, and Richard Suzman (1995). An overview of the Health and Retirement Study. Journal of Human Resources, 30, S7-S56.

Meijer, Erik, Lynn A. Karoly, and Pierre-Carl Michaud, Estimates of Potential Eligibility for Low-Income Subsidies Under Medicare Part D, TR-686, Santa Monica, Calif.: RAND Corporation, 2009. As of September 28, 2012: http://www.rand.org/pubs/technical_reports/TR686/

Meijer, Erik, Lynn A. Karoly, and Pierre-Carl Michaud, "Using Matched Survey and Administrative Data to Estimate Eligibility for the Medicare Part D Low Income Subsidy Program," Social Security Bulletin, Vol. 70, No. 2, May 2010, pp. 63-82.

Michaud, Pierre-Carl, Arie Kapteyn, James P. Smith, and Arthur Van Soest, "Temporary and permanent unit non-response in follow-up interviews of the Health and Retirement Study," Longitudinal and Life Course Studies, Vol. 2, 2011, pp. 145-169.

National Institute on Aging, Growing older in America: The Health and Retirement Study. Bethesda, MD: National Institute on Aging, 07-5757, 2007.

Olson, Janice A., "Linkages with Data from Social Security Administrative Records in the Health and Retirement Study," Social Security Bulletin, Vol. 62, No. 2, 1999, pp. 73–85. As of September 28, 2012: http://www.ssa.gov/policy/docs/ssb/v62n2/v62n2p73.pdf

Panis, Constantijn, Roald Euller, Cynthia Grant, Melissa Bradley, Christine E. Peterson, Randall Hirscher, and Paul Steinberg, SSA Program Data User's Manual. Santa Monica, Calif: RAND Corporation, PM-973-SSA, 2000.

Schisterman, Enrique F., and Brian W. Whitcomb, "Use of the Social Security Administration Death Master File for ascertainment of mortality status." Population Health Metrics, Vol. 2, Article 2, 2004. As of July 25, 2013: http://www.pophealthmetrics.com/content/2/1/2

Scholz, John Karl, and Ananth Seshadri, The Assets and Liabilities Held by Low-Income Families, working paper, Madison, Wisc.: University of Wisconsin, September 9, 2008. As of September 28, 2012: http://www.ssc.wisc.edu/~scholz/Research/Assets_Poverty.pdf

Sierminska Eva, Pierre-Carl Michaud, and Susann Rohwedder, "Measuring Wealth Holdings of Older Households in the U.S.: A Comparison Using the HRS, PSID and SCF," paper presented at Pensions, Private Accounts, and Retirement Savings Over the Life Course workshop, Ann Arbor, Mich., November 20, 2008. As of September 28, 2012: http://psidonline.isr.umich.edu/Publications/Workshops/2008/LC/MRS_WealthCompsv10.pdf

Sohn, Min-Woong, and Elissa Oh, "Completeness and Accuracy of Death Dates and the Implications for Studying Disease Burdens: Focus on Alternative Data Sources." In Victor R. Preedy and Ronald R. Watson, Handbook of Disease Burdens and Quality of Life Measures. New York: Springer, 2010, pp. 345–358.

Soldo, Beth J., Michael D. Hurd, Willard L. Rodgers, and Robert B. Wallace, "Asset and Health Dynamics Among the Oldest Old: An overview of the AHEAD study," Journal of Gerontology, Vol. 52B, Special Issue, 1997, pp. 1–20.

Wojcik, Nancy C., Wendy W. Huebner, and Gail Jorgensen, "Strategies for Using the National Death Index and the Social Security Administration for Death Ascertainment in Large Occupational Cohort Mortality Studies." American Journal of Epidemiology, Vol. 172, 2010, pp. 469–477.