Full length article

Representativeness of the low-income population in the Health and Retirement Study

Erik Meijer [a,*], Lynn A. Karoly [b]

[a] University of Southern California and RAND Corporation, United States
[b] RAND Corporation, United States

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, we exploit the SSA administrative data matched to the HRS sample and compare the distribution of the matched SSA variables against the distribution of the same variables for the same population in the SSA databases. Overall, the HRS is representative for the population it covers. However, for some subgroups in the low-income population, there are some differences.

© 2016 Elsevier B.V. All rights reserved.

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 and Suzman, 1995; NIA, 2007). Often, researchers and policy makers are specifically interested in the low-income population that is eligible for various means-tested programs such as Supplemental Security Income (SSI) and Medicaid. The validity of such studies that rely on the HRS depends on whether it accurately represents the size and composition of the low-income population. If this 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 et al. (2009, 2010) found evidence that suggests that the HRS may not accurately represent the low-income population. Specifically, they 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 the survey data, they found considerable divergence in estimates of LIS eligibility depending on whether the estimate was based primarily on HRS or SIPP data. Analyses of the potential sources of differences between the two data sources suggested that, even after taking sampling weights into account, the HRS sample may underrepresent low-income individuals. Particularly striking is that the estimated 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 differences in sample composition. As mentioned above, our previous results are suggestive of

* Corresponding author at: University of Southern California, Center for Economic and Social Research, 635 Downey Way, Los Angeles, CA 90089-3332, United States.

E-mail address: erik.meijer@usc.edu (E. Meijer).

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. 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 et al. (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. While these studies suggest there may be issues of measurement error in income and wealth measures collected by the HRS or other surveys, 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.

HRS and SSA data sources

In our empirical analysis, 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, 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.

HRS data: sampling history, cohorts, and administrative data match

The HRS is designed 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. 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, born 1954–1959, which was added in 2010, but data for this cohort were not yet available for this study.

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 our version of the matched administrative data files was obtained through permissions granted in 2004–2008, although a significant number of matched records for the original HRS and AHEAD cohorts derive from 1992 to 1996 permissions (HRS, 2010a,b).

Respondents who give permission to match to their SSA records 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 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.

---

[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.

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