A sample frame built from administrative records for the National Survey of Children's Health

Keith Finlay
Center for Administrative Records
Research and Applications
US Census Bureau
keith.ferguson.finlay@census.gov
301-763-6056

February 3, 2016

This document describes using administrative records to build a sample frame for the National Survey of Children's Health (NSCH).

Population of interest

The population of interest is all children residing in the US on the date of the survey. The survey should provide oversamples of children with chronic disabilities. The sample frame should also provide some information about household access to the Internet.

A sampling frame for all households with children

The sampling frame for all households with children comes from three data sources: the Numident, a list of Social Security Number applicants with data updated from various administrative records; and the CARRA kidlink file, a prototype linkage between children and parents based on Census and administrative records. Household addresses are updated with the Master Address Auxiliary Reference File, a file that links person identifiers with the latest location updates from a variety of administrative data.

Using the Numident to identify children

The Numident is based on off the all individuals who have been assigned Social Security Numbers. Demographic data from the Numident is updated from federal tax data and various administrative records. There are 71,873,129 children in the 2015 Numident who will be aged 0–17 years on April 1, 2016 . Figure 1 shows the distribution of date of birth for these children.

The Numident is updated monthly.

Identifying the households containing the children in the Numident

To sample households with children, we must connect the children in the Numident to the households in which they live. We do this with two files: the 2010 Census Unedited File and the CARRA kidlink file.

CARRA kidlink

The CARRA kidlink file uses data from Census survey and federal administrative records to link children PIKs to parent PIKs. We can use this file to identify the parents of children in the Numident. The source data for the CARRA kidlink file are: the Census Numident, the 2010 Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian Health Service database (IHS), Selective Service System (SSS), and Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS) data from the Department of Housing and Urban Development. Of these, the IRS 1040 provides the most significant information.

There are 68,519,439 unique records for children in the 2014 kidlink who will be aged 0–17 years on April 1, 2016 .

Let us consider how many children from the Numident have been linked to a parent in the CARRA kidlink file. Table 1 shows the number of children linked with both a mother and a father, linked with a mother only, linked with a father only, or not linked with any parent.

Figure 2 compares the distributions of date of birth for these children against the distribution shown in Figure 1.

The current vintage of the CARRA kidlink file is 2014. This explains the missing kidlink entries for the youngest children in the Numident (the very right side of the distribution in Figure 2). The CARRA kidlink file will be updated by April 1, 2016, with the newest versions of the input files for final sample frame production.

Updating household location using the MAF-ARF

In order to update household location, we use a Census dataset called the Master Address Auxiliary Reference File (MAF-ARF). The MAF-ARF links person identifiers to address identifiers using Census survey data and federal administrative data. The source data for the MAF-ARF file are: the Census Numident, the 2010 Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian Health Service database (IHS), Selective Service System (SSS), and Public and Indian Housing (PIC) and Tenant Rental Assistance Certification System (TRACS) data from the Department of Housing and Urban Development, and National Change of Address data from the US Postal Service. Of these, the IRS 1040 provides the most significant information.

Out of 71,873,129 children in the Numident, 55,763,902 are matched directly to a MAFID. Out of 59,762,607 kidlink-matched mothers, 54,630,103 are matched to a MAFID. Out of 48,628,252 kidlink-matched fathers, 44,866,739 are matched to a MAFID.

For each child observation from the Numident, we now have four possible MAFIDs: the SSI MAFID, the kid to MAF-ARF MAFID, the child-to-kidlink-to-mother-to-MAF-ARF MAFID, and the child-to-kidlink-to-father-to-MAF-ARF MAFID. I allocate the single MAFID using that order. First, I assign the SSI MAFID (1,713,591 cases). If MAFID is missing, I assign the directly identified child MAFID (54,341,170 cases). If the MAFID is still missing, I assign the mother MAFID (4,932,199 cases). Finally, if the MAFID is still missing, I assign the father MAFID (1,662,678 cases). That leaves 9,223,491 children from the Numident not assigned MAFIDs (a MAFID match rate of 87.2%).

There are some MAFIDs associated with a great number of children. As an example, out of 62,649,638 children associated with a MAFID, 295,328 children are associated with a MAFID with more than 20 child-MAFID links.

The 62,649,638 children associated with a MAFID are then collapsed down to 34,353,877 unique MAFIDS. This implies 1.82 children per household for households assigned a flag.

We then need to scale up the MAFID list to the universe of valid MAFIDs to allow sampling of unflagged households. A merge of the 34,353,877 unique child-flagged MAFIDS with the ACS MAF-X file matches 30,717,480 MAFIDS with child flags, removes 3,636,397 MAFIDS with child flags, and adds 100,608,270 MAFIDs without child flags. The sample frame file now has 131,316,961 valid MAFIDS, of which 30,717,480 MAFIDS include child flags . Compare this with the 2011 ACS, in which 37,147,503 out of 114,991,725 households included related children.[1]

The MAF-ARF will be updated by April 1, 2016, with the newest versions of the input files for final sample frame production.

A sampling frame for disabled children

A subpopulation of children with disabilities comes from the Social Security Agency's Supplemental Security Income (SSI) program. Children with certain disabilities from households with low-enough income are eligible for an SSI subsidy.[2]

There are 1,849,126 unique records for children in the 2014 SSR who will be aged 0–17 years on April 1, 2016 . Figure 3 compares the distributions of date of birth for these children against the distribution shown in Figure 1.

There are a number of sampling concerns with using SSI recipients for the disability oversample:

- Using this subpopulation to create a disability sample would likely oversample children with severe disabilities. The list of "compassionate allowances" can be found at the Social Security site.[3]
- Conditioning on SSI may introduce nonrandom selection on household income for three reasons.

[1] http://www.census.gov/prod/2013pubs/p20-570.pdf
[2] http://www.socialsecurity.gov/pubs/EN-05-10026.pdf
[3] http://ssa.gov/compassionateallowances/conditions.htm

  - Households with lower income are more likely to be eligible[4], but eligibility may raise household income through the subsidy.
  - These children are also more likely to get Medicaid support for healthcare expenses.
  - Parental labor supply may be affected by the severity of the disability.

- Children on SSI are certainly more likely to be in the Numident.
- It may take some months for disabled children to show up in SSI. In the age distribution figure, it's clear that SSI recipients are older than children from the Numident.
- Our Supplemental Security Record (SSR) data file includes the variable `PSTAT-CUR`, which is the payment status code for the current month. We do not have retrospective information on the child's payment status.

Addresses for SSI recipients

Of the children on SSR, 0.878 have been matched to a MAFID. (This compares with a MAF match rate of 0.882 for the entire SSI data file.) We believe that the children on SSI have relatively updated addresses, so those children can be linked directly to Master Address File IDs. The SSI file is updated annually. The current SSR file is from 2014. We will use a 2015 vintage for the NSCH.

Sample frame construction visualization

Figure 4 shows a visualization of the sample frame construction.

Auditing the sample frame against the ACS

To examine the performance of the administrative records used to build the sampling frame, we merge the list of MAFIDs constructed above with the American Community Survey housing-unit sample from 2014. Currently, this audit uses unedited ACS data (i.e., item nonresponse are left as missing and are not imputed including children's age). If item nonresponse is random with respect to the presence of children in the household, this should not cause any systematic bias in the audit.

All estimates are weighted with the housing-unit-level weights, which include weight for vacant units (214,137 vacant housing units in the 2014 ACS). In vacant housing units, we assign zero children. These estimates should reflect the NSCH survey production process.

Table 2 shows the overlap between the MAFID and ACS distributions with respect to whether any children were present in the household.

[4] http://ssa.gov/ssi/text-child-ussi.htm

Child flag performance by age group

We are particularly interested in the coverage of young children. In this section, we show how the child flags perform for specific age groups. These are stricter tests since any deviation in age beyond the age interval will cause either a Type 1 or Type 2 error.

Table 3 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 0–2 years were present in the household. Given that the input administrative records used to construction the child flags are 1–2 years old and that the ACS data are from 2014, it is not surprising that the overlap for children aged 0–2 years is much lower than the overall rate shown in Table 2.

Table 4 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 3–5 years were present in the household. By ages 3–5, overlap between the child flag and the ACS data is above 60%.

Table 5 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 6–8 years were present in the household.

Table 6 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 9–11 years were present in the household.

Table 7 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 12–14 years were present in the household.

Table 8 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 15–17 years were present in the household.

State-specific performance

Table 9 shows the overlap between the MAFID and ACS distributions by state. The smallest oversample strata are in Hawaii, Maine, Vermont, and West Virginia. The largest oversample strata are in California, Texas, and Utah. The highest rates of Type 1 error are in DC, Florida, Louisiana, and Mississippi. The highest rates of Type 2 error are in Alaska, Hawaii, Texas, and Utah.

An Internet-accessible household flag

Here I describe the construction of tract- and block-varying Internet-accessible household flags. The data come from American Community Survey paradata and IRS 1040 filing mode data.

Since 2012, ACS respondents have been able to submit survey forms over the Internet. ACS paradata record whether a respondent chose the online option. The ACS paradata has been summarized at the tract level. Our Internet-accessible household measure is equal to a weighted proportion of the respondents that chose to submit the ACS survey over the Internet if given the option to do so. Figure 5 shows the kernel-smoothed distribution of tract-level Internet response for the 2013–2014 ACS survey years.

We also get a measure of local Internet accessibility from IRS 1040 filing data. Filers have a choice of filing electronically. We identify filers who file electronically but without a paid preparer, and infer that these individuals file at home using the Internet. We then calculate an electronic self file rate by Census block. Figure 6 shows the distribution of Census-block-level electronic self file for tax year 2014.

To synthesize the information from these two measures into a single index, we use principal components analysis (PCA). PCA is a data reduction technique that finds the linear combination of the two input variables that maximizes the variance of a single index. PCA uses standardized forms of the variables, so the predicted index is also in standardized form. Figure 7 shows the distribution of the predicted Census-block-level score from the PCA. We can consider this variable a Census-block-level Internet accessibility index.

A 150% poverty rate flag

Here I describe the construction of a block-group-varying flag for the proportion of households below 150% of the poverty line. The data come from 2014 5-year American Community Survey file. Figure 8 shows the distribution of block-group-level 150% poverty rate flag.

Final sample frame data layout

The three component data files are merged together based on MAFID. The data layout for this combined file is given in Table 10.

List of Figures

Figure 1: Distribution of date of birth, Numident, aged 0–17 years as of April 1, 2016

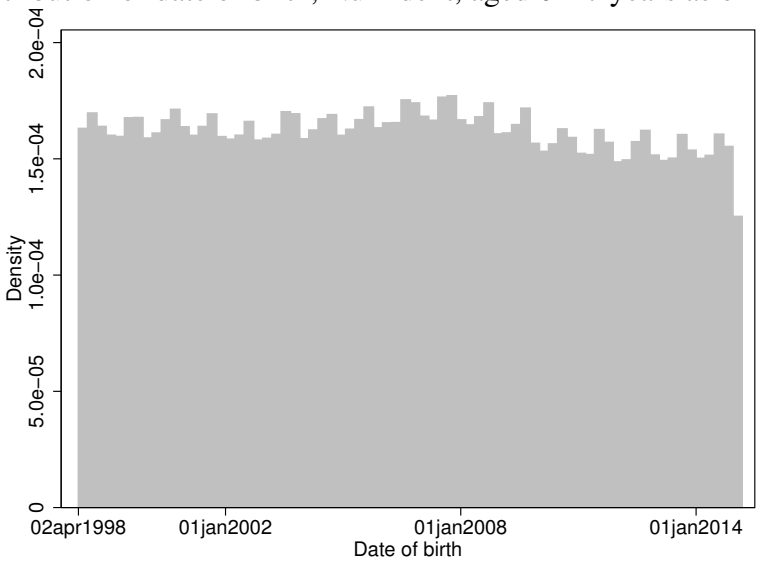


Figure 2: Frequency distributions of date of birth, Numident vs. kidlink entries, aged 0–17 years as of April 1, 2016

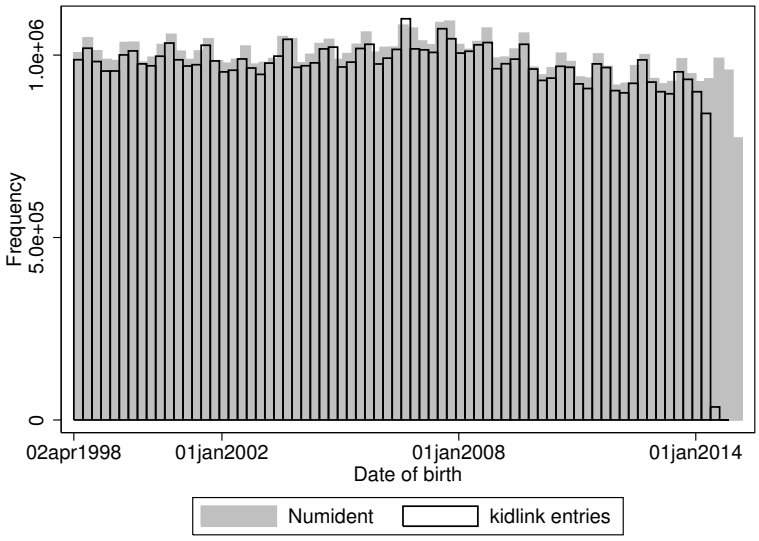


Figure 3: Distributions of date of birth, Numident vs. SSI recipients, aged 0–17 years as of April 1, 2016

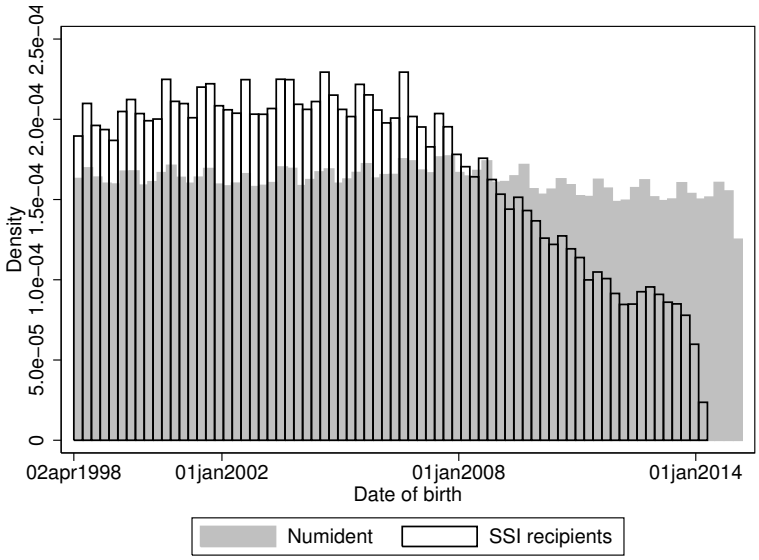


Figure 4: Sample frame construction

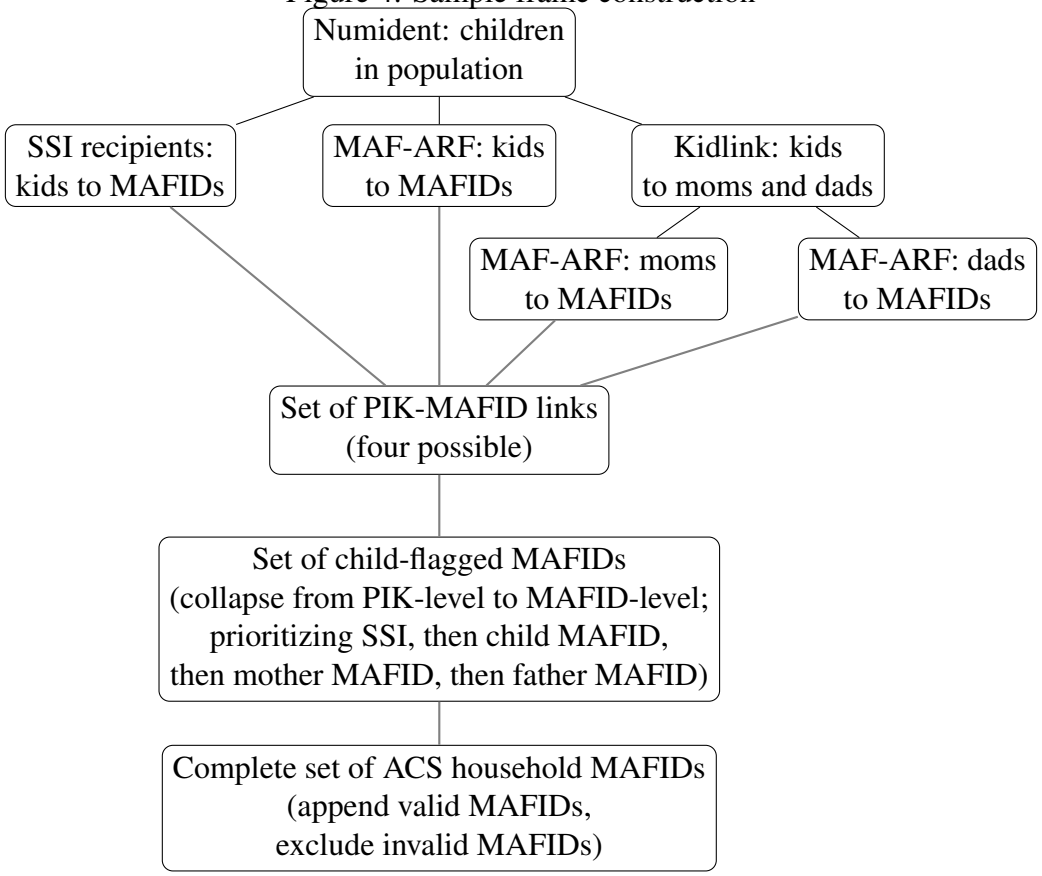


Figure 5: Kernel-smoothed probability distribution function of tract-level ACS Internet response rate, ACS paradata, 2013–2014 survey years

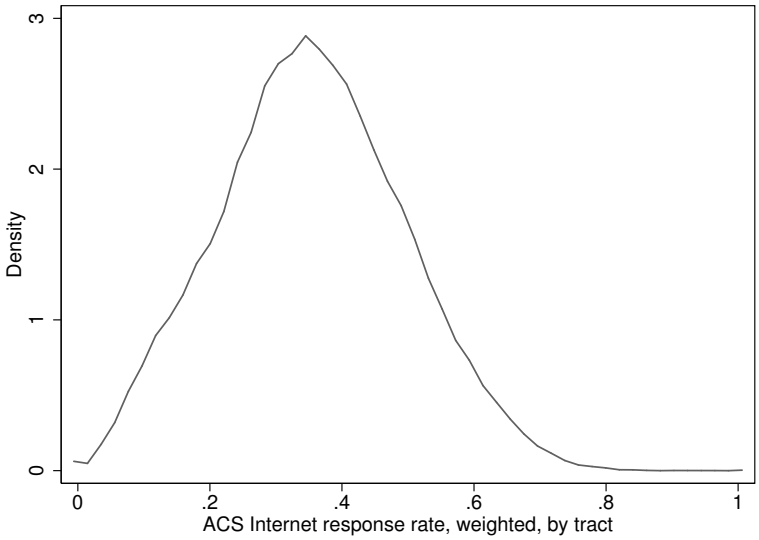


Figure 6: Kernel-smoothed probability distribution function of the Census-block-level rate of IRS electronic self 1040 filing, IRS 1040 data, tax year 2014

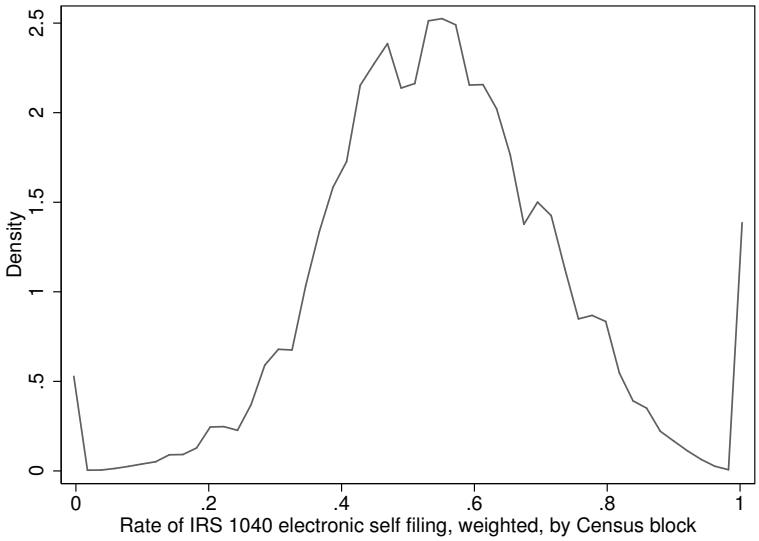


Figure 7: Kernel-smoothed probability distribution function of the Internet accessibility index, ACS paradata (2013–2014 survey years) and IRS 1040 data (tax year 2014)

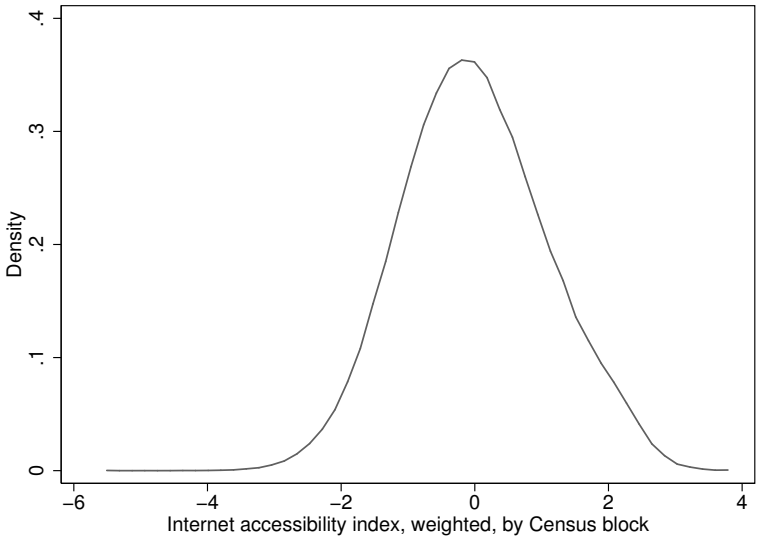
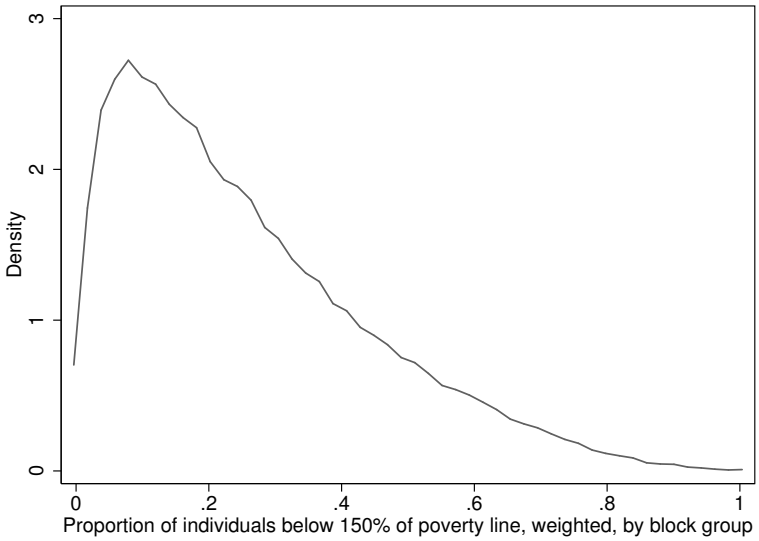


Figure 8: Kernel-smoothed probability distribution function of block-group-level 150% poverty rate, ACS, 2014 5-year file



List of Tables

Table 1: Child-parent links in the CARRA kidlink file relative to the Numident population, aged 0–17 years as of April 1, 2016

Type of link	Frequency	Percent
Mother and father	45,918,616	64%
Mother only	13,843,991	19%
Father only	2,709,636	3.8%
No link	9,400,886	13%
All children in Numident	71,873,129	100%

Table 2: Comparison of NSCH child flags and ACS data, any children in household, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children	Any children	Total
No children	90.3%	9.7%	100.0%
Any children	22.7%	77.3%	100.0%
Total	74.6%	25.4%	100.0%
N (ACS households)	2,322,722		

Table 3: Comparison of NSCH child flags and ACS data, any children in household aged 0–2 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 0–2	Any children 0–2	Total
No children 0–2	98.3%	1.7%	100.0%
Any children 0–2	62.0%	38.0%	100.0%
Total	97.5%	2.5%	100.0%
N (ACS households)	2,322,722		

Table 4: Comparison of NSCH child flags and ACS data, any children in household aged 3–5 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 3–5	Any children 3–5	Total
No children 3–5	96.7%	3.3%	100.0%
Any children 3–5	38.2%	61.8%	100.0%
Total	92.9%	7.1%	100.0%
N (ACS households)	2,322,722		

Table 5: Comparison of NSCH child flags and ACS data, any children in household aged 6–8 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 6–8	Any children 6–8	Total
No children 6–8	96.6%	3.4%	100.0%
Any children 6–8	35.4%	64.6%	100.0%
Total	92.3%	7.7%	100.0%
N (ACS households)	2,322,722		

Table 6: Comparison of NSCH child flags and ACS data, any children in household aged 9–11 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 9–11	Any children 9–11	Total
No children 9–11	96.6%	3.4%	100.0%
Any children 9–11	33.0%	67.0%	100.0%
Total	92.1%	7.9%	100.0%
N (ACS households)	2,322,722		

Table 7: Comparison of NSCH child flags and ACS data, any children in household aged 12–14 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 12–14	Any children 12–14	Total
No children 12–14	96.7%	3.3%	100.0%
Any children 12–14	31.7%	68.3%	100.0%
Total	92.1%	7.9%	100.0%
N (ACS households)	2,322,722		

Table 8: Comparison of NSCH child flags and ACS data, any children in household aged 15–17 years, 2014 ACS, housing unit weights including vacants

	Observed ACS households		
NSCH child flags	No children 15–17	Any children 15–17	Total
No children 15–17	96.6%	3.4%	100.0%
Any children 15–17	31.1%	68.9%	100.0%
Total	91.9%	8.1%	100.0%
N (ACS households)	2,322,722		

Table 9: Comparison of NSCH child flags and ACS data, any children in household, 2014 ACS, housing unit weights including vacants, by state

NSCH frame	Any children (a)			No children (b)			N (c)
ACS obs. children	Any (d)	None (e)	Total	Any (f)	None (g)	Total	
	(d)/(a)	(e)/(a)	(a)/(c)	(f)/(b)	(g)/(b)	(b)/(c)	(c)
State	×100	×100	×100	×100	×100	×100	
Alabama	71.9	28.1	21.2	9.9	90.1	78.8	37,511
Alaska	72.4	27.6	19.8	15.6	84.4	80.2	9,534
Arizona	74.5	25.5	21.3	10.0	90.0	78.7	44,646
Arkansas	71.6	28.4	21.6	11.0	89.0	78.4	22,495
California	80.1	19.9	27.3	10.9	89.1	72.7	217,111
Colorado	82.7	17.3	22.8	9.9	90.1	77.2	37,691
Connecticut	79.6	20.4	23.2	8.2	91.8	76.8	23,385
Delaware	76.1	23.9	22.3	7.9	92.1	77.7	7,367
District of Columbia	68.6	31.4	16.6	6.9	93.1	83.4	4,693
Florida	69.0	31.0	19.7	7.8	92.2	80.3	121,828
Georgia	74.4	25.6	25.5	11.7	88.3	74.5	57,019
Hawaii	72.4	27.6	14.8	18.4	81.6	85.2	9,856
Idaho	80.2	19.8	23.0	10.6	89.4	77.0	11,545
Illinois	78.2	21.8	24.6	8.7	91.3	75.4	97,583
Indiana	78.4	21.6	24.5	8.5	91.5	75.5	48,569
Iowa	83.3	16.7	23.9	7.0	93.0	76.1	34,025
Kansas	79.3	20.7	25.6	8.3	91.7	74.4	26,961
Kentucky	75.6	24.4	22.3	10.6	89.4	77.7	34,115
Louisiana	68.0	32.0	24.2	11.1	88.9	75.8	31,206
Maine	77.3	22.7	15.7	6.3	93.7	84.3	17,636
Maryland	79.2	20.8	25.0	9.0	91.0	75.0	39,331
Massachusetts	82.3	17.7	22.7	7.5	92.5	77.3	43,395
Michigan	80.0	20.0	22.4	6.2	93.8	77.6	100,990
Minnesota	83.9	16.1	23.5	6.7	93.3	76.5	72,611
Mississippi	70.1	29.9	24.1	11.9	88.1	75.9	18,761
Missouri	76.4	23.6	22.4	8.3	91.7	77.6	50,595
Montana	77.5	22.5	17.2	8.5	91.5	82.8	11,567
Nebraska	84.1	15.9	24.1	8.4	91.6	75.9	21,002
Nevada	72.6	27.4	20.7	11.4	88.6	79.3	18,288
New Hampshire	82.4	17.6	19.5	6.9	93.1	80.5	11,239
New Jersey	81.0	19.0	24.4	9.5	90.5	75.6	57,087
New Mexico	71.7	28.3	18.9	12.6	87.4	81.1	16,173
New York	75.0	25.0	19.9	11.6	88.4	80.1	138,735
North Carolina	76.0	24.0	22.2	9.7	90.3	77.8	68,857
North Dakota	81.1	18.9	20.0	9.8	90.2	80.0	9,642
Ohio	78.1	21.9	24.0	7.2	92.8	76.0	90,191
Oklahoma	71.9	28.1	22.6	12.3	87.7	77.4	46,397
Oregon	82.3	17.7	21.6	7.9	92.1	78.4	26,748
Pennsylvania	79.8	20.2	21.6	7.0	93.0	78.4	120,084
Rhode Island	79.4	20.6	21.4	8.0	92.0	78.6	6,819
South Carolina	72.4	27.6	21.6	8.8	91.2	78.4	32,989
South Dakota	76.3	23.7	20.6	10.0	90.0	79.4	9,957
Tennessee	75.0	25.0	23.3	10.0	90.0	76.7	44,043
Texas	76.1	23.9	26.5	14.3	85.7	73.5	146,897
Utah	82.5	17.5	31.2	13.8	86.2	68.8	18,761
Vermont	83.1	16.9	15.9	7.8	92.2	84.1	9,097
Virginia	79.8	20.2	24.7	9.4	90.6	75.3	54,668
Washington	80.6	19.4	22.9	9.0	91.0	77.1	47,839
West Virginia	74.3	25.7	14.4	11.2	88.8	85.6	15,434
Wisconsin	82.5	17.5	22.2	6.9	93.1	77.8	75,291
Wyoming	76.7	23.3	19.0	11.7	88.3	81.0	4,458

Table 10: NSCH population data file layout

Variable name	Label	Format	Domain
mafid	Master Address File ID	long integer	9 digits
maf_curstate	State	str2	
maf_curcounty	County	str3	
maf_curblktract	Tract	str6	
maf_curblkgrp	Block group	str1	
kids_00_02	Number of children aged 0–2 years	integer	≥ 0
kids_03_05	Number of children aged 3–5 years	integer	≥ 0
kids_06_08	Number of children aged 6–8 years	integer	≥ 0
kids_09_11	Number of children aged 9–11 years	integer	≥ 0
kids_12_14	Number of children aged 12–14 years	integer	≥ 0
kids_15_17	Number of children aged 15–17 years	integer	≥ 0
any_ssi	Any children in household on SSI?	byte	$\{0, 1\}$
acs_tract_net_response	Tract-level ACS Internet response	float	$[0, 1]$
block_elf_self_rate	Block-level 1040 electornic self file rate	float	$[0, 1]$
block_net_access_index	Block-level Internet accessibility index	float	$(-\infty, \infty)$
blockgroup_150povrate	Block group-level 150% poverty rate	float	$[0, 1]$
Filename: nsch_pop_file.sas7bdat			
Population: all MAFIDs with valid housing unit types			
Unit of observation: household (MAFID)			
Number of observations: 131,316,961			
Filesize: 9GB			