2017 National Survey of Children's Health sample frame

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

March 2, 2017

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 housing units in the US on the date of the survey.

A sample frame for all households with children

The sample frame identifies three mutually exclusive strata:

- [1] Households with *explicit links to children* in administrative data.
- [2a] Households without explicit links to children in administrative data, but predicted to be *likely to have children* conditional on administrative data.
- [2b] Households without explicit links to children in administrative data, but predicted to be *unlikely to have children* conditional on administrative data.

This document first explains the construction of the Stratum 1 flag, then documents the separation of Strata 2a and 2b.

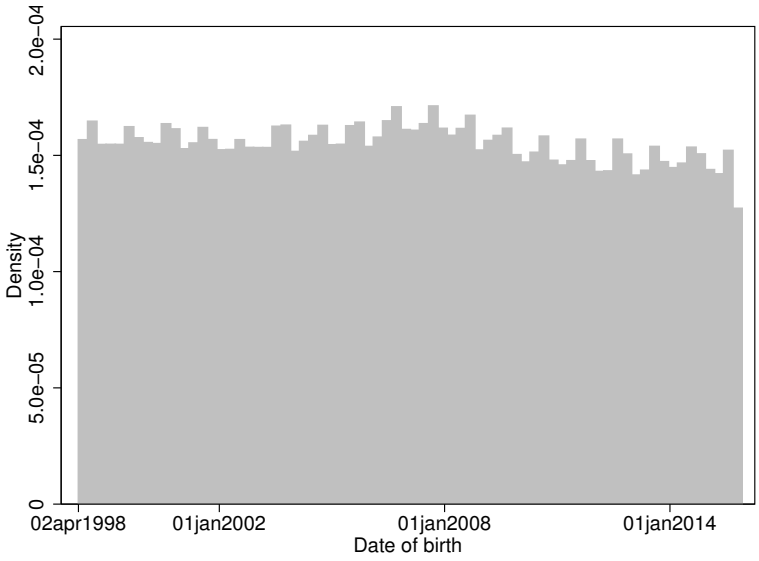
Stratum 1: identifying explicit links from children to addresses

The Stratum 1 flag for all households with explicit links to 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 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 75,156,219 children in the December 2016 Numident who will be aged 0–17 years on April 1, 2017 . Figure 1 shows the distribution of date of birth for these children.

Figure 1: Distribution of date of birth, aged 0–17 years as of April 1, 2016, December 2015 Numident



Identifying the households containing 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 the CARRA kidlink file.

CARRA kidlink

The CARRA kidlink file uses data from Census survey and federal administrative records to link children Protected Identification Keys (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.

In the CARRA kidlink file generated March 2016, there are 75,156,219 unique records for children who will be aged 0–17 years on April 1, 2017 .

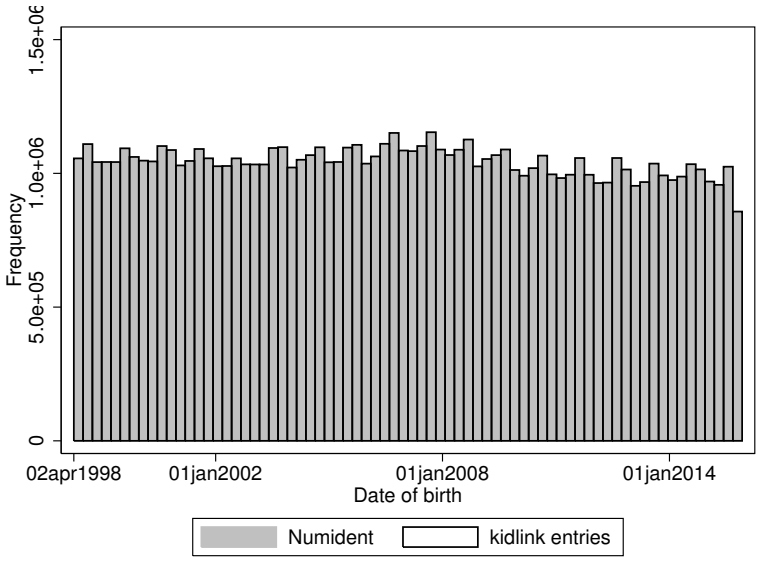
Let us consider how many children from the Numident have been linked to a parent in the CARRA kidlink file. Table 1 shows the number of children linked with both a mother and a father, linked with a mother only, linked with a father only, or not linked with any parent.

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

Type of link	Frequency	Percent
Mother and father	50,886,028	68%
Mother only	14,643,347	19%
Father only	3,050,257	4.1%
No link	6,576,587	8.8%
All children in Numident	75,156,219	100%

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

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



The CARRA kidlink file will be updated in March 2017 for NSCH 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 75,156,219 children in the Numident, 59,841,686 are matched directly to a MAFID. Out of 65,529,375 kidlink-matched mothers, 60,031,595 are matched to a MAFID. Out of 53,936,285 kidlink-matched fathers, 49,767,120 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 (1,265,823 cases). If MAFID is missing, I assign the directly identified child MAFID (58,805,360 cases). If the MAFID is still missing, I assign the mother MAFID (6,385,866 cases). Finally, if the MAFID is still missing, I assign the father MAFID (2,101,661 cases). That leaves 6,597,509 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 68,558,710 children associated with a MAFID, 296,808 children are associated with a MAFID with more than 20 child-MAFID links.

The 68,558,710 children associated with a MAFID are then collapsed down to 36,642,194 unique MAFIDS. This implies 1.87 children per household for households assigned a flag.

We then need to scale up the MAFID list to the universe of MAFIDs to allow sampling of unflagged households. A merge of the 36,642,194 unique child-flagged MAFIDS with the January 2017 ACS MAF-X file matches 36,609,700 MAFIDS with child flags, removes 32,494 MAFIDS with child flags, and adds 159,897,403 MAFIDs without child flags. The sample frame file now has 196,507,103 valid MAFIDS, of which 36,609,700 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]

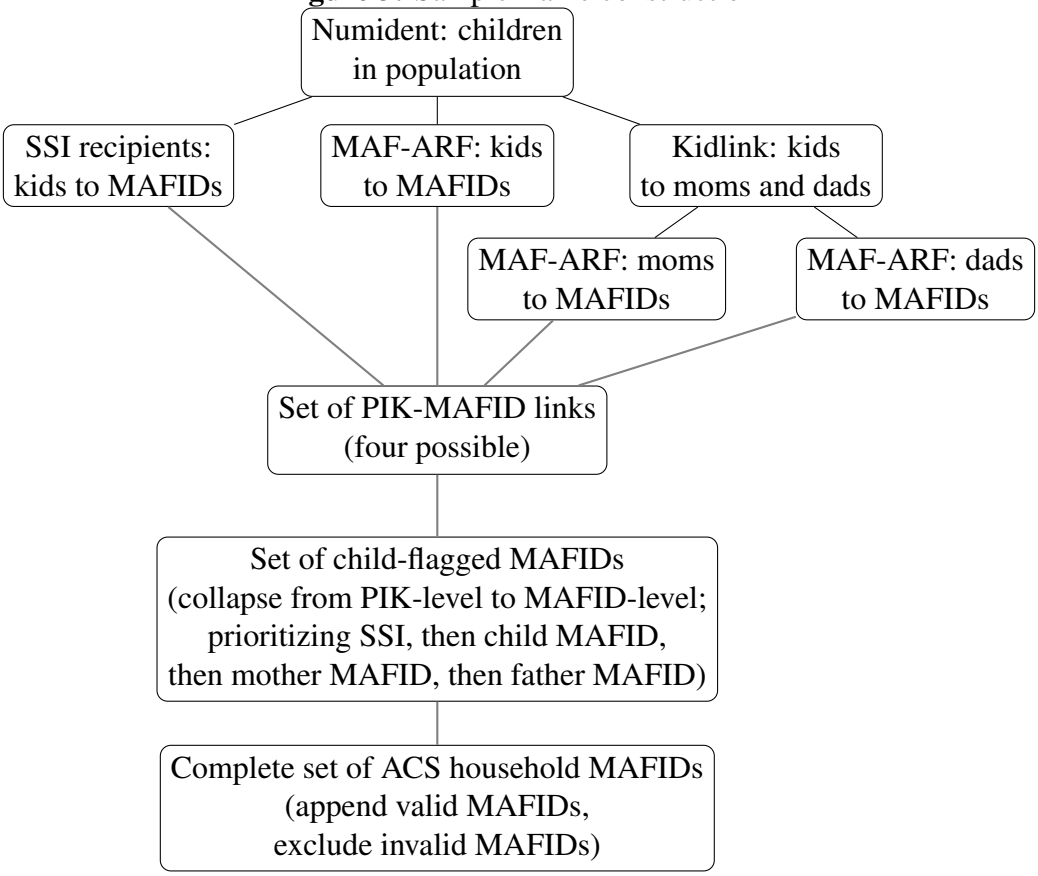
The MAF-ARF will be updated in March 2017 for NSCH sample frame production.

Stratum 1 construction visualization

Figure 3 shows a visualization of the sample frame construction.

[1] http://www.census.gov/prod/2013pubs/p20-570.pdf

Figure 3: Sample frame construction



Strata 2a and 2b: identifying probabilistic links from children to addresses

In 2016, the Stratum 1 flag performed well. That is, it contained approximately the same rate of children after sampling, as had been predicted before the survey. The survey team would like to further increase the sampling efficiency of the survey by adding more information to the second stratum. By definition, Stratum 2 does not have explicit links from children to households in the administrative data. In 2017, we will further bifurcate Stratum 2 into those households more likely to have children and those households less likely to have children.

Households will be assigned to Stratum 2a based on a model of child presence as a function of variables available in administrative data for all households in the MAF. The model is estimated with data from the most recent year of the ACS, in which child presence can be observed. Then parameter estimates from that model can be used to predict the likelihood of child presence for all households. These models are estimated separately for each state, and the threshold for bifurcation is based on an objective of maximizing the size of Stratum 2b while also maintaining 95% coverage of households with children in Strata 1 and 2a.

Definitions

Population or sample concepts

- 2015 ACS sample, edited and swapped
  - unit of observation is the household, unless noted otherwise
  - sample includes sampled vacant dwellings, unless noted otherwise
- MAF
  - population but restricted to MAFIDs marked as valid for ACS

Sample frame notation

- h indexes household
- s indexes states
- C equals 1 if a household has any children, 0 otherwise
- Strata:
  - S_1: household with children
  - S_{2a}: household likely to have children
  - S_{2b}: household unlikely to have children
- Strata sizes:
  - $p(S_1)$
  - $p(S_{2a})$
  - $p(S_{2b})$
- Strata child rates:
  - $p(C|S_1)$
  - $p(C|S_{2a})$
  - $p(C|S_{2b})$
- Coverage with unsampled S_{2b}:
  - $p(S_1 \cup S_{2a}|C)$

Model

Our goal is a scalar measure of the likelihood of a child being associated with a MAFID. This measure must be available for all ACS-valid MAFIDs in the MAF. Using a sample in which the presence of children is observable, we will estimate a model of child presence. The regressors used to make the index prediction must be observable for all MAFIDs (i.e., to predict outside of the estimation sample to the entire MAF).

The general model is:

$$C_h = f(X_h; \theta),$$

where C is equal to one if a household includes any children and zero otherwise, X is a vector of characteristics available for all households, and θ is an unknown vector of parameters.

We estimate the model using the most recent ACS 1-year sample:

$$E[C_h|X_h] = f(X_h; \hat{\beta}_{ACS}) \text{ for households } h \text{ in the ACS.}$$

With parameter estimates from the ACS, we make predictions for the entire MAF:

$$\hat{C}_h = f(X_h; \hat{\beta}_{ACS}) \text{ for households } h \text{ in the MAF.}$$

In practice, we estimate models separately for each state. We do this to account for systematic differences in administrative records coverage and MAF quality across states. The model can now be specified as:

$$E[C_{hs}|X_{hs}] = f(X_{hs}; \hat{\beta}_{s,ACS}) \text{ for households } h \text{ in state } s \text{ in the ACS,}$$

where s is the MAFID's state and the parameters $\hat{\beta}_{s,ACS}$ now vary across states. The state-specific predictions become:

$$\hat{C}_{hs} = f(X_{hs}; \hat{\beta}_{s,ACS}) \text{ for households } h \text{ in state } s \text{ in the MAF.}$$

Estimation

The model above is estimated as a linear probability model separately for each state using the edited and swapped 2015 ACS sample. The outcome is `child_present`, a flag for whether a child is present at the sampled MAFID.

The following covariates are included (with associated data sources) and are available for each MAFID (except where a missingness flag is used):

- 2015 ACS 5-year published aggregate data
  - `acs_blkgrp_childrate_lvout`: proportion of residents of block group who are children, excluding the own-observation child counts from the numerator and denominator
- MAF-ARF
  - `female2050`: flag for female between ages 20 and 50 at MAFID
  - `adult2050`: flag for adults between ages 20 and 50 at MAFID
  - `coresid_sexdiff`: flag for coresidence of men and women between ages 20 and 50 at MAFID
  - `miss_adult2050`: flag for missingness from MAF-ARF

- IRS 1040 filings, tax year 2015
  - `any_kid_deduct_max`: does any tax form associated with this MAFID have any deduction related to children?[2]
  - `itemized_max`: does any tax form associated with this MAFID use itemized deductions?
  - `miss_any_kid_deduct_max`: flag for MAFIDs without associated tax forms
- VSGI NAR commercial data
  - `vsgi_nar_homeowner_max`: does any observation associated with this MAFID record it as homeowener-occupied?
  - `miss_vsgi_nar_homeowner_max`: flag for MAFIDs without associated VSGI data
- Targus commercial data
  - `targus_homeowner_0`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_A`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_B`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_C`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_D`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_E`: various flags for homeowner-occupied MAFID
  - `targus_homeowner_F`: various flags for homeowner-occupied MAFID
  - `miss_targus_homeowner`: flag for MAFIDs without associated Targus data

Parameter estimates are stored in the file `frame2017_child_present_bystate.csv`.

Sample frame objective function

In order to choose an optimal Strata 2a, we use the following objective function:

- Minimize the size of Strata 2a while maintaining coverage of at least 95%

Strata 2a is defined as:

$$S_{2a} = \{\text{households in the MAF with } \hat{C}_h > \bar{C} \text{ but not in } S_1\}.$$

Strata 2b is defined as

$$S_{2b} = \{\text{households in the MAF but not in } S_1 \text{ or } S_{2a}\}.$$

With state-specific modeling, the objective function and coverage constraint also becomes state specific:

[2]The following IRS variable were used to make this variable: child exemptions and EITC qualifying children.

- Minimize the size of Strata 2a in each state while maintaining coverage of at least 95% in each state

State-specific Strata 2a is defined as:

$$S_{2a} = \{\text{households in the MAF with } \hat{C}_{hs} > \bar{C}_s \text{ but not in } S_1\}.$$

Strata 2b is defined as before.

Optimization algorithm

The optimization parameter is a threshold on the child-present prediction probability, such that MAFIDs with values above the threshold are assigned to Stratum 2a. Starting at a low threshold ($\bar{C}$)[3], follow this algorithm:

1. Under the current threshold $\bar{C}$, calculate the proportion of MAFIDs in Stratum 2a, $p(S_{2a})$, and the coverage of Strata 1 and 2a under no sampling of Strata 2b, ($p(S_1 \cup S_{2a}|C)$).
2. If $p(S_{2a}) > 0$ and $p(S_1 \cup S_{2a}|C) \geq 0.95$, then increase the child prediction threshold $\bar{C}$ one step (e.g., 0.01) and return to (1). If $p(S_1 \cup S_{2a}|C) < 0.95$, then the previous threshold $\bar{C}$ is the optimal cutoff for S_{2a}.

Under state-specific modeling, this algorithm is applied separately to each state.

Optimal strata

Table 2 shows the optimal strata under a 95% coverage constraint for Strata 1 and 2a. The coverage constraint assumes non-sampling of Stratum 2b. The notation is as defined above. The strata were optimized separately for each state using parameter estimates from separate state regressions of child presence in the 2015 ACS microdata.

Auditing the sample frame against the ACS

To examine the performance of the administrative records used to build the sample 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. In vacant housing units, we assign zero children. These estimates should reflect the NSCH survey production process.

[3]The most conservative starting threshold would be at $p(S_1)$, where $p(S_{2b}) = 0$.

Table 2: Optimal 2017 NSCH strata with 95% coverage constraint, state-level optimization

State	N	$p(S1)$	$p(S2)$	$p(S3)$	$p(C\|S1)$	$p(C\|S2)$	$p(C\|S3)$	$p(C\|!S1)$	$p(!S3\|C)$	q	$\hat{C}_{S2}$
US	2,305,707	0.24	0.33	0.43	0.78	0.17	0.03	0.10	0.95	43	0.10
AL	37,368	0.22	0.38	0.40	0.74	0.16	0.03	0.10	0.95	38	0.14
AK	9,394	0.19	0.62	0.19	0.71	0.19	0.05	0.15	0.95	20	0.13
AZ	44,652	0.22	0.34	0.43	0.76	0.20	0.03	0.10	0.95	43	0.14
AR	22,498	0.22	0.43	0.36	0.77	0.17	0.04	0.11	0.95	34	0.14
CA	215,206	0.28	0.31	0.41	0.82	0.21	0.04	0.11	0.95	43	0.14
CO	37,863	0.24	0.33	0.44	0.84	0.18	0.03	0.10	0.95	43	0.14
CT	23,154	0.24	0.30	0.46	0.80	0.17	0.03	0.09	0.95	43	0.14
DE	7,156	0.22	0.19	0.59	0.78	0.18	0.02	0.07	0.96	46	0.16
DC	4,697	0.20	0.37	0.43	0.71	0.12	0.02	0.07	0.95	33	0.09
FL	120,642	0.21	0.29	0.50	0.72	0.15	0.02	0.08	0.95	50	0.13
GA	56,483	0.27	0.34	0.40	0.76	0.19	0.04	0.11	0.95	40	0.15
HI	9,819	0.17	0.54	0.28	0.72	0.24	0.04	0.17	0.95	28	0.13
ID	11,439	0.24	0.32	0.44	0.77	0.18	0.03	0.09	0.95	38	0.16
IL	97,181	0.25	0.33	0.42	0.79	0.18	0.03	0.10	0.95	43	0.14
IN	47,914	0.25	0.27	0.48	0.78	0.19	0.03	0.09	0.95	47	0.17
IA	33,959	0.23	0.24	0.53	0.83	0.16	0.03	0.07	0.95	50	0.17
KS	26,796	0.25	0.27	0.48	0.79	0.17	0.03	0.08	0.95	46	0.17
KY	33,621	0.23	0.36	0.41	0.77	0.17	0.03	0.10	0.95	40	0.14
LA	30,576	0.25	0.37	0.38	0.70	0.16	0.04	0.10	0.95	37	0.15
ME	17,240	0.15	0.45	0.40	0.79	0.10	0.02	0.07	0.95	36	0.12
MD	38,955	0.27	0.28	0.45	0.80	0.18	0.03	0.09	0.95	44	0.15
MA	42,884	0.24	0.30	0.47	0.82	0.16	0.03	0.08	0.95	46	0.14
MI	100,300	0.22	0.25	0.53	0.80	0.16	0.02	0.07	0.95	54	0.16
MN	72,731	0.23	0.21	0.56	0.84	0.18	0.03	0.07	0.95	56	0.17
MS	18,393	0.25	0.42	0.33	0.71	0.17	0.04	0.12	0.95	31	0.14
MO	50,265	0.23	0.31	0.46	0.79	0.18	0.03	0.09	0.95	46	0.16
MT	11,319	0.17	0.47	0.36	0.76	0.15	0.03	0.09	0.95	33	0.13
NE	20,792	0.24	0.27	0.49	0.82	0.17	0.03	0.08	0.95	46	0.17
NV	18,322	0.24	0.33	0.43	0.74	0.17	0.03	0.10	0.95	40	0.13
NH	11,118	0.19	0.31	0.50	0.81	0.13	0.02	0.07	0.95	41	0.14
NJ	56,618	0.25	0.31	0.43	0.82	0.19	0.04	0.11	0.95	43	0.14
NM	16,057	0.18	0.47	0.35	0.73	0.18	0.03	0.11	0.95	33	0.14
NY	137,665	0.23	0.40	0.38	0.76	0.18	0.03	0.11	0.95	38	0.12
NC	68,734	0.23	0.35	0.42	0.78	0.15	0.03	0.09	0.95	41	0.14
ND	9,543	0.20	0.36	0.44	0.80	0.16	0.03	0.08	0.95	38	0.15
OH	89,658	0.24	0.28	0.48	0.80	0.16	0.03	0.08	0.95	49	0.16
OK	45,705	0.23	0.45	0.32	0.73	0.19	0.04	0.12	0.95	33	0.15
OR	26,908	0.23	0.30	0.47	0.82	0.16	0.03	0.08	0.95	44	0.14
PA	118,520	0.22	0.30	0.48	0.81	0.15	0.03	0.08	0.95	49	0.14
RI	6,630	0.22	0.34	0.44	0.79	0.15	0.03	0.09	0.95	34	0.12
SC	32,766	0.23	0.32	0.45	0.74	0.15	0.03	0.09	0.95	43	0.14
SD	9,981	0.22	0.32	0.45	0.79	0.19	0.03	0.09	0.95	40	0.17
TN	43,918	0.25	0.32	0.43	0.77	0.16	0.03	0.09	0.95	43	0.15
TX	146,469	0.28	0.34	0.38	0.78	0.21	0.04	0.13	0.95	40	0.15
UT	18,497	0.33	0.28	0.39	0.83	0.23	0.04	0.12	0.95	37	0.19
VT	8,924	0.15	0.46	0.39	0.83	0.12	0.03	0.08	0.95	33	0.13
VA	53,822	0.26	0.28	0.46	0.81	0.16	0.03	0.09	0.95	46	0.15
WA	47,760	0.25	0.30	0.45	0.82	0.19	0.03	0.10	0.95	44	0.14
WV	15,110	0.17	0.49	0.34	0.74	0.18	0.03	0.12	0.95	31	0.12
WI	75,272	0.22	0.23	0.56	0.83	0.17	0.02	0.07	0.95	56	0.17
WY	4,413	0.20	0.46	0.34	0.74	0.17	0.03	0.11	0.96	28	0.13

Table 3 shows the overlap between the MAFID and ACS distributions with respect to whether any children were present in the household.

Table 3: 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	92.2%	7.8%	100.0%
Any children	25.2%	74.8%	100.0%
Total	74.6%	25.4%	100.0%
N (ACS households)	2,322,722		

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 4 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 3.

Table 4: 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	99.0%	1.0%	100.0%
Any children 0–2	71.8%	28.2%	100.0%
Total	97.5%	2.5%	100.0%
N (ACS households)	2,322,722		

Table 5 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 6 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 7 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 8 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 5: 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	97.1%	2.9%	100.0%
Any children 3–5	38.5%	61.5%	100.0%
Total	92.9%	7.1%	100.0%
N (ACS households)	2,322,722		

Table 6: 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	96.9%	3.1%	100.0%
Any children 6–8	35.7%	64.3%	100.0%
Total	92.3%	7.7%	100.0%
N (ACS households)	2,322,722		

Table 7: 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	96.9%	3.1%	100.0%
Any children 9–11	33.4%	66.6%	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 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	97.0%	3.0%	100.0%
Any children 12–14	32.3%	67.7%	100.0%
Total	92.1%	7.9%	100.0%
N (ACS households)	2,322,722		

Table 9 shows the overlap between the MAFID and ACS distributions with respect to whether any children aged 15–17 years were present in the household.

Table 9: 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	96.9%	3.1%	100.0%
Any children 15–17	31.3%	68.7%	100.0%
Total	91.9%	8.1%	100.0%
N (ACS households)	2,322,722		

State-specific performance

Table 10 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, Mississippi, Nevada, and South Carolina. The highest rates of Type 2 error are in Alaska, Hawaii, New Mexico, Texas, and Utah.

Small-area paper-only response probability

Since 2012, the American Community Survey (ACS) respondents have been able to submit survey forms over the internet in addition to completing and mailing back a paper questionnaire. We used 2016 ACS response mode choices summarized at the block group and other block group and tract-level characteristics to model Web and paper response mode probabilities by block group. Sample households will be located within block groups and assigned a paper-only response probability. The 30% of households with the highest paper-only response probability will be flagged as 'High Paper' and will receive a paper questionnaire with the initial web invitation. For very new housing units without assigned Census blocks, we assign a value of zero for this binary variable (i.e., the default for these new households is high Internet accessibility.)

In addition to ACS response mode, we modeled block group response mode probabilities using multinomial logisitic regression on adult education (%HS, %College, %Grad), poverty (%in poverty, %between 100% and 150% of poverty threshold), adult age, race, ethnicity, foreign born, and rural/urban status. Modeled response mode probabilities are given a household weight of 10 and averaged with obserserved ACS response mode probabilities to offset sampling error in very small samples.

Local-area household income relative to the poverty rate

The frame has a set of poverty variables from the 2015 5-year American Community Survey file. These variables measure the proportion of households with household income in an interval defined by the poverty rate. Figure 5 shows the kernel-smoothed probability distribution function of the proportion of households in the block group that have household income less than 150% of the poverty rate.

Table 10: 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	71.5	28.5	24.3	7.5	92.5	75.7	37,511
Alaska	72.7	27.3	23.7	12.6	87.4	76.3	9,534
Arizona	73.1	26.9	23.6	8.5	91.5	76.4	44,646
Arkansas	70.4	29.6	25.0	8.7	91.3	75.0	22,495
California	78.3	21.7	29.7	9.3	90.7	70.3	217,111
Colorado	80.2	19.8	25.3	8.2	91.8	74.7	37,691
Connecticut	77.6	22.4	24.9	7.3	92.7	75.1	23,385
Delaware	73.0	27.0	25.1	6.4	93.6	74.9	7,367
District of Columbia	64.9	35.1	19.3	5.7	94.3	80.7	4,693
Florida	66.7	33.3	22.7	6.1	93.9	77.3	121,828
Georgia	71.6	28.4	29.9	8.9	91.1	70.1	57,019
Hawaii	71.3	28.7	18.0	16.5	83.5	82.0	9,856
Idaho	76.4	23.6	27.2	8.0	92.0	72.8	11,545
Illinois	75.0	25.0	26.5	8.1	91.9	73.5	97,583
Indiana	75.6	24.4	27.2	7.0	93.0	72.8	48,569
Iowa	79.2	20.8	26.5	5.8	94.2	73.5	34,025
Kansas	76.6	23.4	27.7	7.2	92.8	72.3	26,961
Kentucky	73.1	26.9	25.7	8.5	91.5	74.3	34,115
Louisiana	66.6	33.4	27.8	8.9	91.1	72.2	31,206
Maine	75.3	24.7	17.7	5.1	94.9	82.3	17,636
Maryland	76.3	23.7	28.5	6.8	93.2	71.5	39,331
Massachusetts	78.3	21.7	24.8	6.8	93.2	75.2	43,395
Michigan	76.9	23.1	24.2	5.4	94.6	75.8	100,990
Minnesota	81.0	19.0	25.9	5.2	94.8	74.1	72,611
Mississippi	69.3	30.7	27.8	9.3	90.7	72.2	18,761
Missouri	74.7	25.3	24.7	6.8	93.2	75.3	50,595
Montana	74.4	25.6	19.0	7.6	92.4	81.0	11,567
Nebraska	80.4	19.6	27.4	6.4	93.6	72.6	21,002
Nevada	69.3	30.7	25.6	8.5	91.5	74.4	18,288
New Hampshire	78.0	22.0	21.8	5.9	94.1	78.2	11,239
New Jersey	79.2	20.8	26.3	8.3	91.7	73.7	57,087
New Mexico	71.2	28.8	22.0	10.4	89.6	78.0	16,173
New York	72.4	27.6	23.9	9.1	90.9	76.1	138,735
North Carolina	73.7	26.3	25.7	7.4	92.6	74.3	68,857
North Dakota	76.2	23.8	25.1	6.5	93.5	74.9	9,642
Ohio	75.5	24.5	25.9	6.2	93.8	74.1	90,191
Oklahoma	70.4	29.6	26.6	9.7	90.3	73.4	46,397
Oregon	79.0	21.0	23.6	7.0	93.0	76.4	26,748
Pennsylvania	76.6	23.4	23.9	5.8	94.2	76.1	120,084
Rhode Island	75.9	24.1	22.8	7.7	92.3	77.2	6,819
South Carolina	69.3	30.7	25.1	6.8	93.2	74.9	32,989
South Dakota	76.4	23.6	23.6	7.3	92.7	76.4	9,957
Tennessee	73.2	26.8	26.6	7.7	92.3	73.4	44,043
Texas	73.8	26.2	31.6	10.7	89.3	68.4	146,897
Utah	80.1	19.9	35.6	10.5	89.5	64.4	18,761
Vermont	79.5	20.5	17.6	7.0	93.0	82.4	9,097
Virginia	76.6	23.4	28.3	7.1	92.9	71.7	54,668
Washington	76.3	23.7	26.0	7.5	92.5	74.0	47,839
West Virginia	70.5	29.5	18.3	9.1	90.9	81.7	15,434
Wisconsin	79.7	20.3	24.1	5.8	94.2	75.9	75,291
Wyoming	72.5	27.5	22.6	9.8	90.2	77.4	4,458

Figure 4: Kernel-smoothed probability distribution function of anticipated response mode probabilities

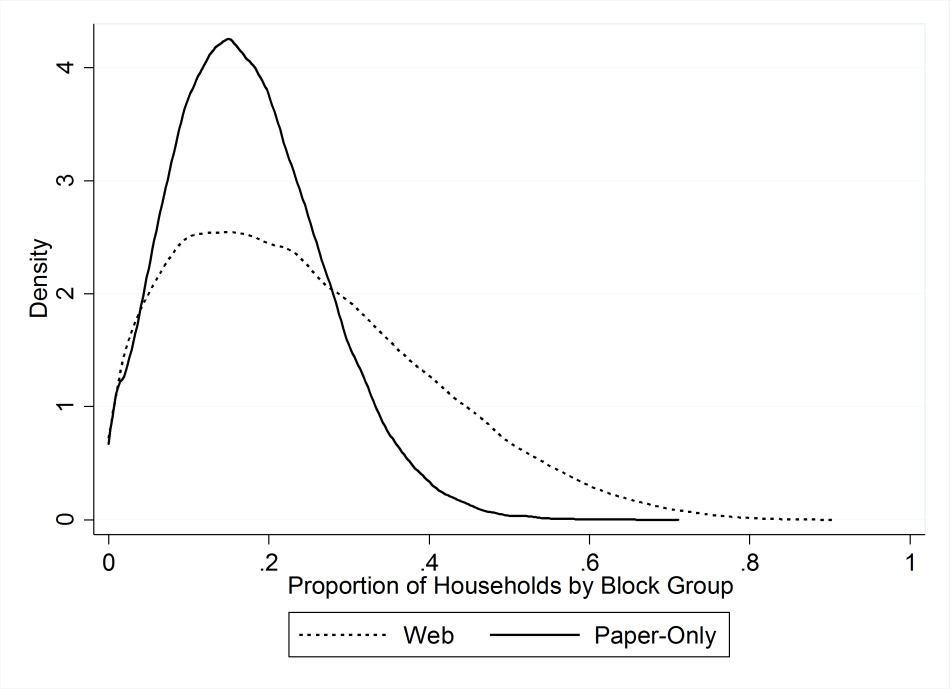
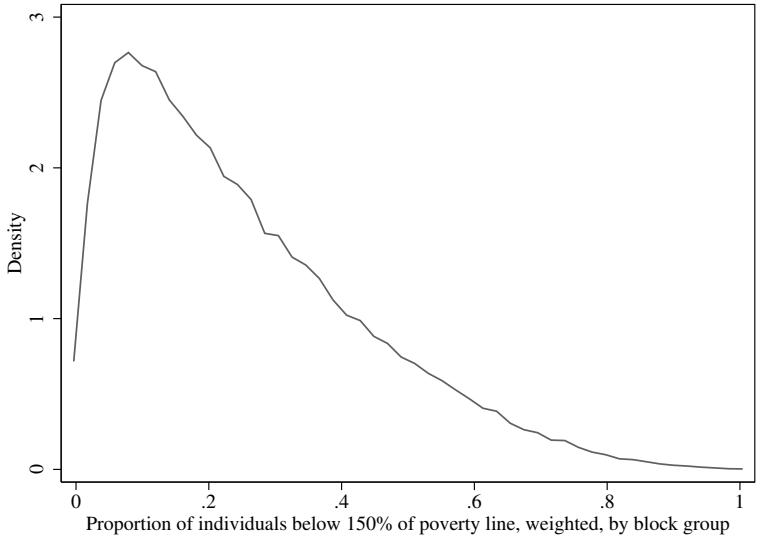


Figure 5: Kernel-smoothed probability distribution function of block-group-level 150% poverty rate, ACS, 2015 5-year file



Final sample frame data layout

The component data files are merged together based on MAFID. The data layout for this combined file is given in Table 11.

Table 11: NSCH population data file layout

Variable name	Label	Level of variation	Type	Domain	Any missing?
mafid	Master Address File ID	MAFID	long	9 digits	no
maf_curstate	State	State	str2		no
maf_curcounty	County	County	str3		no
maf_curblktract	Tract	Tract	str6		yes
maf_curblkgrp	Block group	Block group	str1		yes
maf_curblk	Block	Block	str4		yes
stratum1	Stratum 1 identifier	MAFID	byte	$\{0, 1\}$	no
stratum2a	Stratum 2a identifier	MAFID	byte	$\{0, 1\}$	no
stratum2b	Stratum 2b identifier	MAFID	byte	$\{0, 1\}$	no
acs_tract_net_response	ACS Internet response	Tract	float	$[0, 1]$	yes
blkgrp_lt_100_povrate	Pr. HH w/ inc. $<$ 100% poverty rate	Block group	float	$[0, 1]$	yes
blkgrp_100_150_povrate	Pr. HH w/ inc. 100–150% poverty rate	Block group	float	$[0, 1]$	yes
blkgrp_150_185_povrate	Pr. HH w/ inc. 150–185% poverty rate	Block group	float	$[0, 1]$	yes
blkgrp_185_200_povrate	Pr. HH w/ inc. 185–200% poverty rate	Block group	float	$[0, 1]$	yes
blkgrp_gt_200_povrate	Pr. HH w/ inc. $>$ 200% poverty rate	Block group	float	$[0, 1]$	yes
blkgrp_lt_150_povrate	Pr. HH w/ inc. $<$ 150% poverty rate	Block group	float	$[0, 1]$	yes

Filename: nsch_pop_file.sas7bdat
Population: all MAFIDs in January 2017 MAF-X
Unit of observation: household (MAFID)
Number of observations: 196,507,103
Filesize: 26GB