Appendix B

National Survey of Children's Health
Sample Frame and Sampling Flags Creation

2021 National Survey of Children's Health sample frame[1]

John Voorheis and Maria Perez-Patron
Center for Economic Studies
US Census Bureau
April 7, 2021

This document describes using administrative records to build a sample frame for the National Survey of Children's Health (NSCH) for 2021.

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, and 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: 1) the Numident, 2) a list of Social Security Number applicants with data updated from various administrative records, and 3) the Census Household Composition Key (CHCK, formerly called 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 (MAF-ARF), a file that links person identifiers with the latest location updates from a variety of administrative data (see Figure 3). For Sample year 2021, we provide additional granularity to the information provided in Stratum 1. In addition to a flag for whether there are

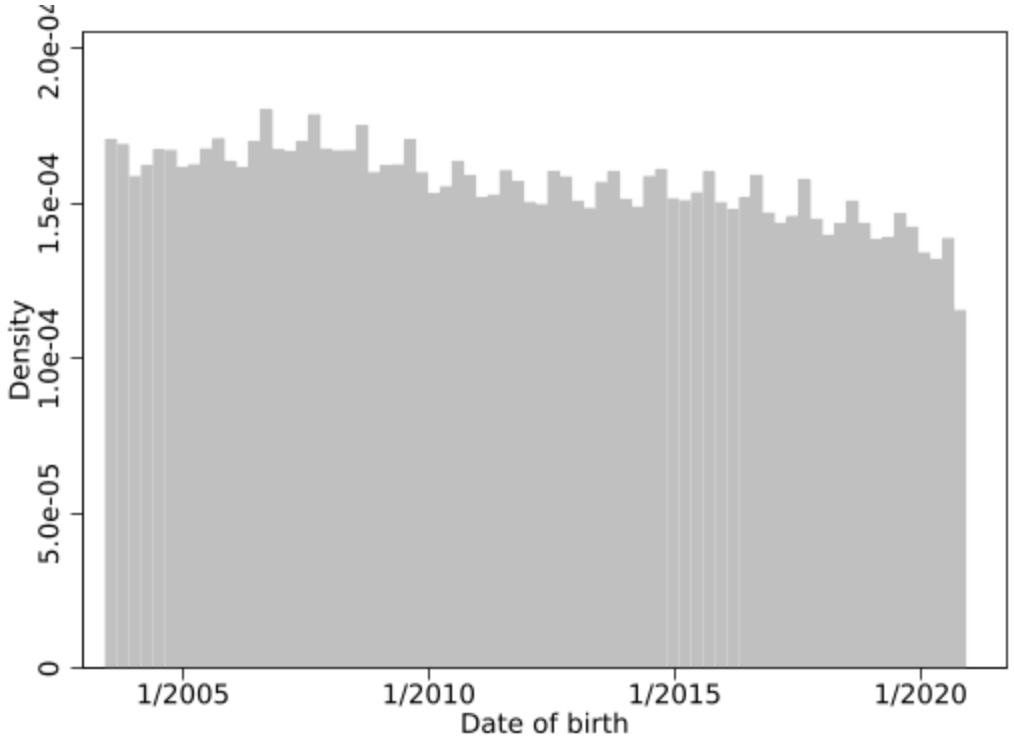
[1] All results have been reviewed to ensure that no confidential information is disclosed. The statistical summaries reported in this paper have been cleared by the Census Bureau's Disclosure Review Board, release authorization number CBDRB-FY22-CES019-001.

any children under 18 at a MAFID, we provide flags for whether there are any young children (under 5, stratum 1a) or only older children (5-17, stratum 1b), based on the date of birth information in the Numident.

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 73,520,000 children in the most recent Numident who will be aged 0–17 years on June 1, 2021 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 December 1, 2020, Numident



The CHCK file was updated in March 2020 for NSCH sample frame production.

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 the CHCK file.

Census Household Composition Key File

The CHCK uses data from Census surveys 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 CHCK are: the Census Numident, the 2010 Census Unedited File, the IRS 1040 and 1099 files, the Medicare Enrollment Database (MEDB), Indian HealthService 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 CHCK file generated in 2020, there are 64,700,000 unique records for children who will be aged 0–17 years on June 1, 2021.

Let us consider how many children from the Numident have been linked to a parent in the CHCK 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.

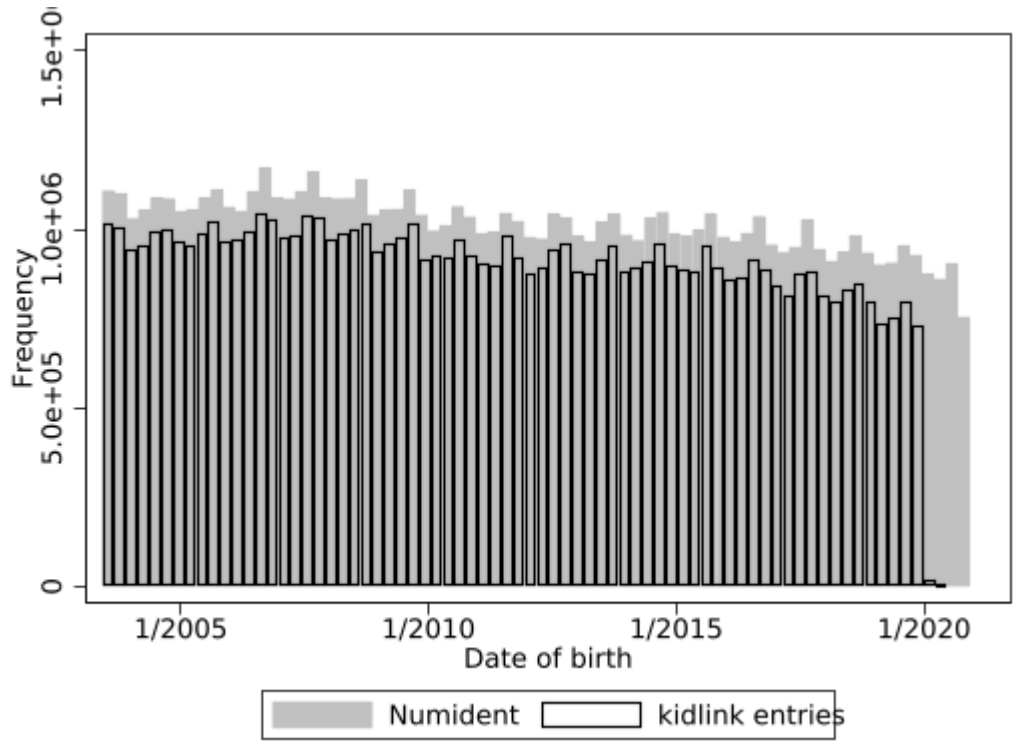
Table 1: Child-parent links in the CHCK file relative to the Numident population, aged 0–17 years as of 2021, 2020 CHCK file.

Type of link	Frequency	Percent
Mother and father	51,430,000	70%
Mother only	11,300,000	15%
Father only	1,966,000	2.7%
ACS link	68,000	0.1%
No link	8,754,000	12%
All children in Numident	73,520,000	100%[2]

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

[2] Note that numbers in this table may not add up correctly due to rounding required for disclosure avoidance.

Figure 2. Frequency distributions of date of birth, Numident vs. CHCK entries, aged 0–17 years as of June 1, 2020



The CHCK file was updated in 2020 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 84,130,000[3] children in the Numident, 68,110,000 are matched directly to a MAFID. Out of 72,300,000 CHCK-matched mothers, about 67,530,000 are matched to a MAFID. Out of 61,330,000 CHCK-matched fathers, about 57,250,000 are matched to a MAFID.

[1]All unweighted counts and estimates in this document are rounded in accordance with Census Disclosure Review Board rules.

For each child observation from the Numident, we now have three possible MAFIDs: the child-to-MAF-ARF MAFID, the child-to-CHCK-to-mother-to-MAF-ARF MAFID, the child-to-CHCK-to-father-to-MAF-ARF MAFID, and the child-to-ACS parent-to-MAF-ARF MAFID. We allocate a single MAFID to each child using that order. First, we assign the directly identified child MAFID (65,470,000 cases). If the MAFID is missing, we assign the mother MAFID (5,294,000 cases). Finally, if the MAFID is still missing, we assign the father MAFID (2,055,000 cases). That leaves 11,270,000 children from the Numident not assigned MAFIDs (a MAFID match rate of 86.6%).

There are some MAFIDs associated with a great number of children. As an example, out of 72,860,000 associated with a MAFID, 7,862,000 children are associated with a MAFID with more than 20 child-MAFID links.

The 72,860,000 children associated with a MAFID are then collapsed down to 38,280,000 unique MAFIDS. This implies 1.9 children per household for households assigned a flag.

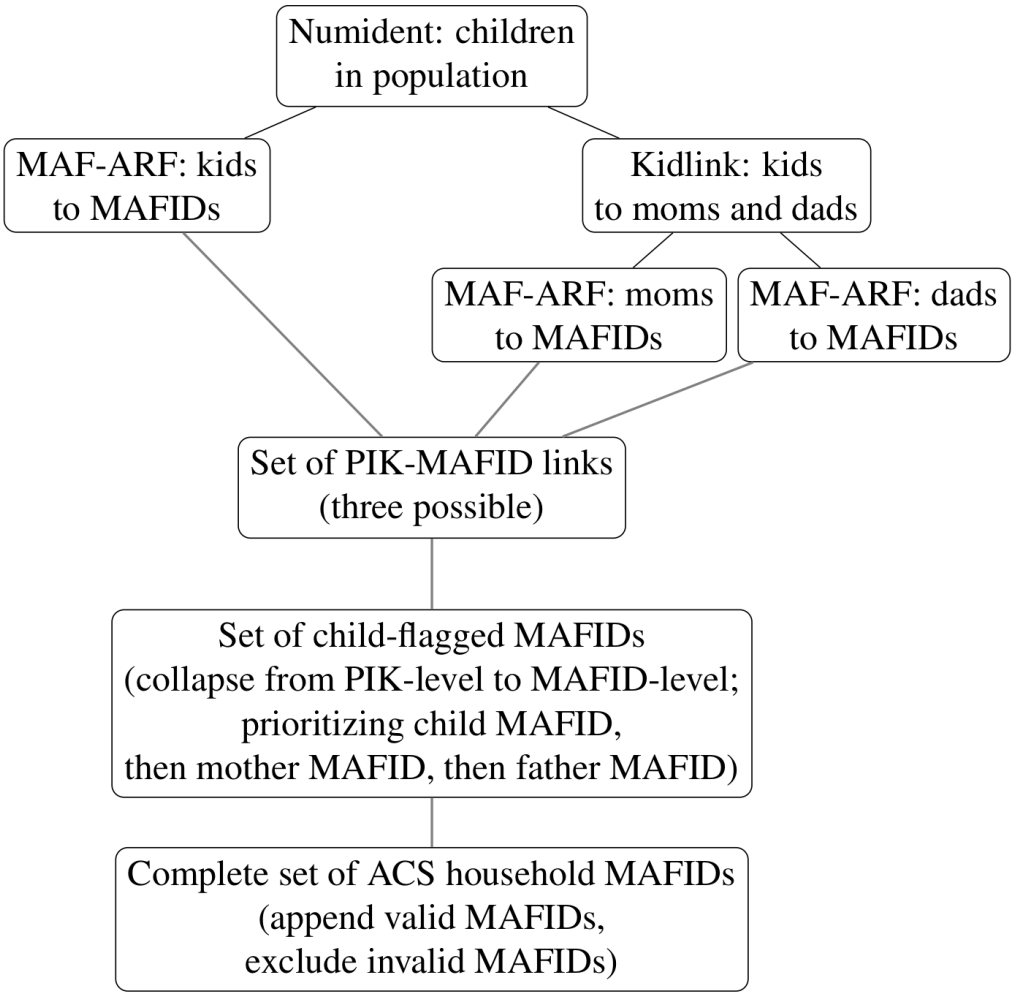
We then need to scale up the MAFID list to the universe of MAFIDs to allow sampling of unflagged households. A merge of the 38,280,000 unique child-flagged MAFIDS with the 2020 MAF-X file matches 38,280,000 MAFIDS with child flags, removes 173,600,000 MAFIDS with child flags. The sample frame file now has about 209 million valid MAFIDS of which 38,280,000 include child flags. Compare this with the 2011 ACS, in which about 37 million out of 115 million households included related children.[4]

Stratum 1 construction visualization

Figure 3 shows a visualization of the sample frame construction.

Figure 3: Stratum 1 construction

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



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

In 2016, the Stratum 1 flag performed well. That is, the surveyed sample contained approximately the same rate of children 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 2021 as in previous years, we further bifurcate Stratum 2 into those households more likely to have children and those households less likely to have children.

Households are 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 minimizing the size of Stratum 2a while also maintaining 95% coverage of children in Strata 1 and 2a.

Definitions

Population or sample concepts

- 2019 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 2019 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):

- 2019 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 2019
  - `any_kid_deduct_max`: does any tax form associated with this MAFID have any deduction related to children?[5]
  - `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 homeowner-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 `frame2021_child_present_bystate.csv`.

Sample frame objective function

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

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

Stratum 2a is defined as:

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

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

Stratum 2b is defined as:

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

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

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

State-specific Stratum 2a is defined as:

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

Stratum 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 ($\bar{C}$)[6], 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 Stratum 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 2019 ACS microdata.

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

Table 2: Optimal 2021 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	C_hat_S2
US	2,079,000[7]	0.22	0.46	0.33	0.76	0.14	0.04	0.11	0.95	31	0.00
AL	33,000	0.21	0.53	0.26	0.71	0.12	0.05	0.10	0.95	23	0.03
AK	8,000	0.15	0.53	0.32	0.71	0.13	0.17	0.13	0.89	-1	-0.48
AZ	40,000	0.21	0.45	0.34	0.76	0.15	0.04	0.11	0.95	31	0.06
AR	20,000	0.21	0.58	0.20	0.74	0.12	0.07	0.11	0.95	15	0.01
CA	191,000	0.26	0.38	0.36	0.78	0.19	0.05	0.12	0.95	37	0.11
CO	35,000	0.22	0.41	0.36	0.78	0.16	0.04	0.10	0.95	36	0.08
CT	20,000	0.21	0.40	0.39	0.78	0.15	0.04	0.10	0.95	38	0.08
DE	6,500	0.19	0.42	0.39	0.72	0.11	0.03	0.08	0.95	37	0.05
DC	4,000	0.16	0.75	0.09	0.66	0.07	0.09	0.07	0.95	7	-0.03
FL	109,000	0.19	0.41	0.39	0.68	0.13	0.03	0.08	0.95	38	0.06
GA	49,000	0.24	0.46	0.30	0.72	0.15	0.06	0.12	0.95	27	0.06
HI	8,500	0.13	0.62	0.25	0.68	0.23	0.05	0.18	0.95	21	0.10
ID	11,000	0.23	0.41	0.36	0.79	0.15	0.04	0.10	0.95	34	0.09
IL	86,000	0.22	0.44	0.34	0.77	0.14	0.05	0.10	0.95	31	0.05
IN	43,500	0.22	0.45	0.33	0.74	0.13	0.04	0.10	0.95	31	0.05
IA	31,500	0.19	0.64	0.16	0.81	0.09	0.20	0.10	0.94	-1	-0.15
KS	24,000	0.22	0.42	0.36	0.78	0.15	0.05	0.11	0.95	31	0.06
KY	30,000	0.22	0.55	0.23	0.77	0.13	0.06	0.12	0.95	20	0.03
LA	26,000	0.22	0.46	0.32	0.67	0.15	0.05	0.11	0.95	31	0.07
ME	15,500	0.14	0.45	0.41	0.80	0.09	0.03	0.06	0.95	33	0.03
MD	34,000	0.25	0.41	0.35	0.78	0.15	0.04	0.10	0.95	34	0.07
MA	38,000	0.21	0.42	0.37	0.80	0.14	0.03	0.09	0.95	37	0.07
MI	91,500	0.20	0.36	0.45	0.78	0.13	0.03	0.08	0.95	43	0.07
MN	68,000	0.21	0.39	0.40	0.82	0.12	0.04	0.09	0.95	37	0.05
MS	16,000	0.22	0.70	0.07	0.68	0.11	0.17	0.12	0.95	-1	-0.18
MO	45,000	0.21	0.45	0.34	0.77	0.13	0.04	0.10	0.95	31	0.05
MT	10,000	0.15	0.71	0.14	0.76	0.09	0.17	0.10	0.91	-1	-0.17
NE	19,000	0.22	0.60	0.18	0.81	0.09	0.13	0.10	0.95	5	-0.05
NV	17,000	0.23	0.47	0.31	0.72	0.15	0.05	0.11	0.95	30	0.06
NH	10,500	0.18	0.41	0.41	0.79	0.12	0.03	0.08	0.95	38	0.06
NJ	48,500	0.23	0.39	0.38	0.79	0.18	0.04	0.11	0.95	37	0.09
NM	14,000	0.17	0.70	0.13	0.70	0.12	0.19	0.12	0.93	-1	-0.17
NY	119,000	0.20	0.45	0.34	0.75	0.15	0.04	0.11	0.95	33	0.08
NC	62,000	0.21	0.44	0.35	0.76	0.14	0.04	0.10	0.95	33	0.07
ND	8,400	0.18	0.68	0.14	0.75	0.09	0.15	0.10	0.94	-1	-0.16
OH	80,000	0.22	0.38	0.41	0.78	0.14	0.03	0.09	0.95	40	0.07

[7] Note that the state population totals do not add up to the national population due to rounding required by Census disclosure avoidance rules

OK	40,500	0.21	0.64	0.15	0.73	0.13	0.11	0.12	0.95	7	-0.01
OR	24,500	0.20	0.46	0.34	0.77	0.13	0.04	0.10	0.95	31	0.05
PA	105,000	0.20	0.39	0.42	0.80	0.13	0.03	0.09	0.95	40	0.06
RI	5,700	0.19	0.47	0.34	0.78	0.13	0.04	0.09	0.95	34	0.06
SC	30,000	0.21	0.43	0.36	0.72	0.13	0.04	0.09	0.95	34	0.06
SD	8,700	0.20	0.69	0.11	0.80	0.10	0.19	0.11	0.94	-1	-0.19
TN	39,500	0.22	0.43	0.35	0.74	0.15	0.04	0.10	0.95	33	0.07
TX	130,000	0.26	0.46	0.28	0.75	0.18	0.07	0.14	0.95	25	0.07
UT	17,500	0.30	0.43	0.27	0.81	0.19	0.07	0.15	0.95	23	0.08
VT	7,900	0.15	0.77	0.08	0.80	0.06	0.13	0.07	0.95	-1	-0.11
VA	49,500	0.24	0.36	0.40	0.79	0.16	0.04	0.10	0.95	40	0.09
WA	43,500	0.23	0.45	0.32	0.79	0.15	0.05	0.11	0.95	31	0.07
WV	12,500	0.16	0.70	0.15	0.75	0.10	0.15	0.11	0.89	-1	-0.19
WI	70,000	0.20	0.39	0.42	0.80	0.13	0.03	0.08	0.95	40	0.06
WY	3,900	0.19	0.61	0.20	0.75	0.11	0.06	0.10	0.95	15	0.01

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 2020. 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.

State-specific performance

Table 3 shows the overlap between the MAFID and ACS distributions by state. In 2021, 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 were in Alaska, Hawaii, New Mexico, Texas, and Utah.

Table 3: [8] NSCH strata, ACS, all addresses audit

State	N	p(S1)	p(S2)	p(S3)	p(C|S1)	p(C|S2)	p(C|S3)	p(C|!S1)	p(!S3|C)
US	2,079,000	0.22	0.44	0.35	0.76	0.14	0.05	0.11	0.94
AL	33,000	0.21	0.53	0.26	0.71	0.12	0.05	0.10	0.95
AK	8,000	0.15	0.53	0.32	0.71	0.13	0.17	0.13	0.89
AZ	40,000	0.21	0.45	0.34	0.76	0.15	0.04	0.11	0.95
AR	20,000	0.21	0.58	0.20	0.74	0.12	0.07	0.11	0.95
CA	191,000	0.26	0.38	0.37	0.78	0.19	0.05	0.12	0.95
CO	35,000	0.22	0.41	0.36	0.78	0.15	0.04	0.10	0.95
CT	20,000	0.21	0.40	0.39	0.78	0.15	0.04	0.10	0.95
DE	6,500	0.19	0.42	0.39	0.72	0.11	0.03	0.08	0.95
DC	4,100	0.16	0.75	0.09	0.66	0.07	0.09	0.07	0.95
FL	109,000	0.19	0.42	0.39	0.68	0.13	0.03	0.08	0.95
GA	49,000	0.24	0.46	0.30	0.72	0.15	0.06	0.12	0.95
HI	8,700	0.13	0.62	0.25	0.68	0.23	0.05	0.18	0.95
ID	11,000	0.23	0.41	0.36	0.79	0.15	0.04	0.10	0.95
IL	86,000	0.22	0.44	0.34	0.77	0.14	0.05	0.10	0.95
IN	43,000	0.22	0.44	0.33	0.74	0.13	0.04	0.10	0.95
IA	31,000	0.19	0.64	0.16	0.81	0.09	0.20	0.10	0.94
KS	24,000	0.22	0.42	0.36	0.78	0.15	0.05	0.11	0.95
KY	30,000	0.22	0.55	0.23	0.77	0.13	0.06	0.12	0.95
LA	26,000	0.22	0.46	0.32	0.67	0.15	0.05	0.11	0.95
ME	15,500	0.14	0.46	0.41	0.80	0.09	0.03	0.06	0.95
MD	34,000	0.25	0.41	0.35	0.78	0.15	0.04	0.10	0.95
MA	38,000	0.21	0.42	0.37	0.80	0.14	0.03	0.09	0.95
MI	91,500	0.20	0.35	0.45	0.78	0.13	0.03	0.08	0.95
MN	68,000	0.21	0.39	0.40	0.82	0.12	0.04	0.09	0.95
MS	16,000	0.22	0.70	0.07	0.68	0.11	0.17	0.12	0.95
MO	45,000	0.21	0.45	0.34	0.77	0.13	0.04	0.10	0.95
MT	10,000	0.15	0.71	0.14	0.76	0.09	0.17	0.10	0.91
NE	19,000	0.22	0.60	0.18	0.81	0.09	0.13	0.10	0.95
NV	17,000	0.23	0.47	0.31	0.72	0.15	0.05	0.11	0.95
NH	10,000	0.18	0.41	0.41	0.79	0.12	0.03	0.08	0.95
NJ	48,500	0.23	0.39	0.38	0.79	0.18	0.04	0.11	0.95
NM	14,000	0.17	0.70	0.13	0.70	0.12	0.19	0.12	0.93
NY	119,000	0.20	0.45	0.34	0.75	0.15	0.04	0.11	0.95
NC	62,000	0.21	0.44	0.35	0.76	0.14	0.04	0.10	0.95
ND	8,400	0.18	0.68	0.14	0.75	0.09	0.15	0.10	0.94
OH	80,000	0.22	0.38	0.41	0.78	0.14	0.03	0.09	0.95
OK	40,500	0.21	0.64	0.15	0.73	0.13	0.11	0.12	0.95
OR	24,500	0.20	0.46	0.34	0.77	0.13	0.04	0.10	0.95
PA	105,000	0.20	0.39	0.42	0.80	0.13	0.03	0.09	0.95

[8] National Survey of Children's Health sample frame

RI	5,700	0.19	0.47	0.35	0.78	0.13	0.04	0.09	0.95
SC	30,000	0.21	0.44	0.36	0.72	0.13	0.04	0.09	0.95
SD	8,700	0.20	0.69	0.11	0.80	0.10	0.19	0.11	0.94
TN	39,500	0.22	0.43	0.34	0.74	0.15	0.04	0.10	0.95
TX	130,000	0.26	0.46	0.28	0.75	0.18	0.07	0.14	0.95
UT	17,500	0.30	0.43	0.27	0.81	0.19	0.07	0.15	0.95
VT	7,900	0.15	0.77	0.08	0.80	0.06	0.13	0.07	0.95
VA	49,500	0.24	0.36	0.40	0.79	0.16	0.04	0.10	0.95
WA	43,500	0.23	0.45	0.32	0.79	0.15	0.05	0.11	0.95
WV	12,500	0.16	0.70	0.15	0.75	0.10	0.15	0.11	0.89
WI	70,000	0.20	0.39	0.42	0.80	0.13	0.03	0.08	0.95
WY	3,900	0.19	0.61	0.20	0.75	0.11	0.06	0.10	0.95

We additionally audit the frame against an early release file of 2020 ACS microdata, as shown in table 4.

Table 4: [9] NSCH strata, ACS 2020, all addresses audit

State	N	p(S1)	p(S2)	p(S3)	p(C|S1)	p(C|S2)	p(C|S3)	p(C|!S1)	p(!S3|C)
US	1,300,000	0.23	0.42	0.35	0.84	0.11	0.04	0.08	0.94
AL	19,000	0.23	0.51	0.26	0.80	0.10	0.03	0.08	0.96
AK	4,500	0.19	0.62	0.19	0.77	0.16	0.38	0.21	0.77
AZ	24,500	0.22	0.44	0.34	0.83	0.12	0.03	0.08	0.95
AR	11,500	0.23	0.59	0.19	0.82	0.10	0.06	0.09	0.96
CA	125,000	0.27	0.35	0.38	0.84	0.15	0.03	0.09	0.96
CO	23,500	0.24	0.40	0.37	0.86	0.11	0.03	0.08	0.95
CT	13,500	0.23	0.37	0.40	0.88	0.12	0.03	0.07	0.96
DE	4,000	0.21	0.40	0.39	0.84	0.09	0.03	0.06	0.95
DC	2,700	0.17	0.75	0.08	0.79	0.06	0.05	0.06	0.98
FL	65,000	0.20	0.39	0.40	0.79	0.11	0.03	0.06	0.95
GA	28,500	0.25	0.45	0.30	0.82	0.12	0.04	0.09	0.96
HI	5,700	0.15	0.62	0.23	0.76	0.26	0.06	0.21	0.96
ID	7,100	0.24	0.39	0.38	0.85	0.15	0.05	0.10	0.93
IL	55,500	0.23	0.42	0.35	0.86	0.11	0.05	0.08	0.93
IN	27,500	0.23	0.42	0.35	0.85	0.11	0.04	0.08	0.95
IA	21,000	0.21	0.64	0.15	0.87	0.06	0.22	0.09	0.87
KS	16,000	0.23	0.39	0.39	0.86	0.11	0.06	0.08	0.92
KY	19,000	0.23	0.54	0.23	0.85	0.10	0.04	0.08	0.97
LA	14,500	0.24	0.43	0.33	0.77	0.12	0.03	0.08	0.96
ME	8,000	0.18	0.44	0.38	0.85	0.08	0.03	0.05	0.95
MD	22,500	0.25	0.39	0.36	0.86	0.11	0.04	0.07	0.96
MA	26,000	0.22	0.39	0.38	0.87	0.12	0.03	0.07	0.96

[9] National Survey of Children's Health sample frame

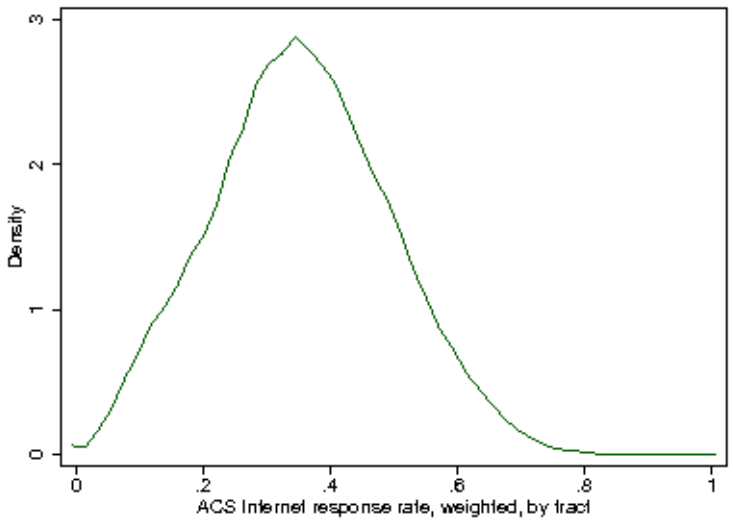
MI	53,500	0.22	0.33	0.45	0.87	0.11	0.03	0.06	0.95
MN	44,000	0.23	0.36	0.41	0.89	0.10	0.03	0.06	0.94
MS	9,500	0.25	0.72	0.04	0.76	0.09	0.27	0.10	0.96
MO	28,000	0.23	0.43	0.34	0.86	0.11	0.03	0.07	0.96
MT	6,200	0.18	0.73	0.09	0.83	0.08	0.24	0.10	0.90
NE	13,000	0.21	0.62	0.17	0.87	0.07	0.19	0.09	0.88
NV	10,500	0.23	0.44	0.33	0.80	0.11	0.03	0.08	0.96
NH	6,200	0.21	0.39	0.40	0.86	0.09	0.03	0.06	0.95
NJ	31,500	0.25	0.37	0.37	0.87	0.15	0.03	0.09	0.96
NM	7,400	0.18	0.75	0.07	0.76	0.10	0.29	0.11	0.92
NY	72,500	0.22	0.43	0.36	0.83	0.13	0.04	0.09	0.95
NC	36,500	0.23	0.41	0.36	0.83	0.11	0.03	0.07	0.96
ND	5,400	0.20	0.70	0.10	0.86	0.08	0.25	0.10	0.90
OH	51,500	0.22	0.35	0.42	0.86	0.11	0.03	0.07	0.95
OK	25,500	0.25	0.63	0.12	0.78	0.11	0.13	0.12	0.94
OR	17,000	0.22	0.44	0.35	0.86	0.10	0.03	0.07	0.95
PA	64,000	0.21	0.36	0.43	0.88	0.10	0.03	0.06	0.95
RI	3,700	0.21	0.43	0.36	0.86	0.10	0.02	0.07	0.96
SC	17,500	0.23	0.42	0.35	0.81	0.11	0.03	0.07	0.96
SD	5,900	0.21	0.71	0.08	0.87	0.08	0.24	0.09	0.93
TN	24,500	0.24	0.41	0.35	0.82	0.11	0.03	0.08	0.95
TX	77,000	0.27	0.44	0.29	0.83	0.14	0.04	0.10	0.96
UT	12,000	0.31	0.42	0.27	0.86	0.16	0.08	0.13	0.94
VT	4,500	0.18	0.77	0.05	0.86	0.07	0.16	0.07	0.96
VA	32,000	0.25	0.34	0.41	0.85	0.12	0.03	0.07	0.95
WA	31,000	0.24	0.43	0.33	0.86	0.12	0.03	0.08	0.96
WV	7,200	0.19	0.74	0.08	0.83	0.07	0.23	0.08	0.92
WI	44,500	0.21	0.36	0.43	0.86	0.11	0.02	0.06	0.96
WY	2,500	0.20	0.61	0.19	0.82	0.14	0.06	0.12	0.96

Local-area Internet-accessibility

Here we describe the construction of a tract-varying Internet-accessible household flag.

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 4 shows the kernel-smoothed distribution of tract-level Internet response for the 2013–2014 ACS survey years.

Figure 4: Kernel-smoothed probability distribution function of tract-level ACS Internet response rate, ACS paradata, 2013–2014 survey years

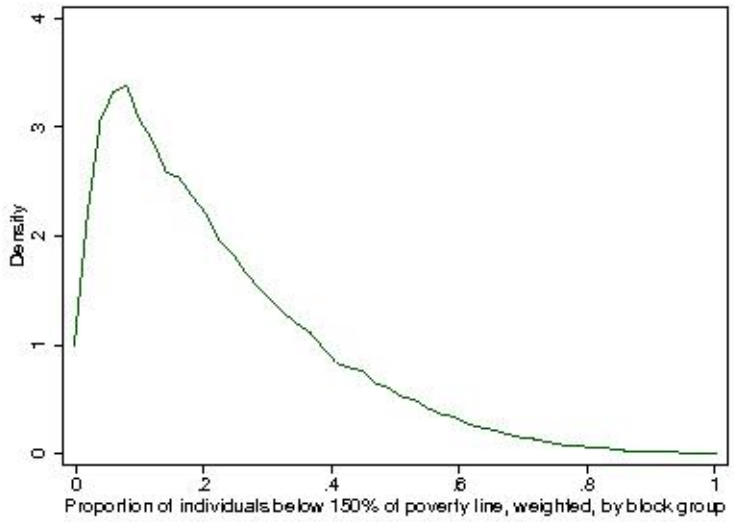


To construct an Internet-access flag, we use the first tercile for a cut-off. A block is considered to have low Internet access if the Internet accessibility index is below the first tercile of the block-level distribution. For low-population blocks, we replace missing values of the block-varying low-Internet flag with the modal value from the corresponding block group. 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.)

Local-area household income relative to the poverty rate

The frame has a set of poverty variables from the 2019 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.

Figure 5: Kernel-smoothed probability distribution function of block-group-level 150% poverty rate, ACS, 2019 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 5.

Table 5: 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
stratum1a	Stratum 1a identifier	MAFID	byte	{0, 1}	no
stratum1b	Stratum 1b 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
web_low	Low web use (lowest tritile)	Tract	byte	0, 1	No
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
mailvaldf	Valid mailing address	MAFID	byte	{0, 1}	yes

Filename: nsch_pop_file.sas7bdat
Population: all MAFIDs in 2020 MAF-X
Unit of observation: household (MAFID)