Appendix B

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

This document was approved by the Disclosure Review Board on Jan 14, 2021 (DRB Approval No. CBDRB-FY21-POP001-0052).

2020 National Survey of Children's Health sample frame

Garret Christensen and John Voorheis
Center for Economic Studies
US Census Bureau

April 2, 2020

This document describes using administrative records to build a sample frame for the National Survey of Children's Health (NSCH) for 2020.

Click Here to Enter DRB Delegated Authority Approval Number

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.

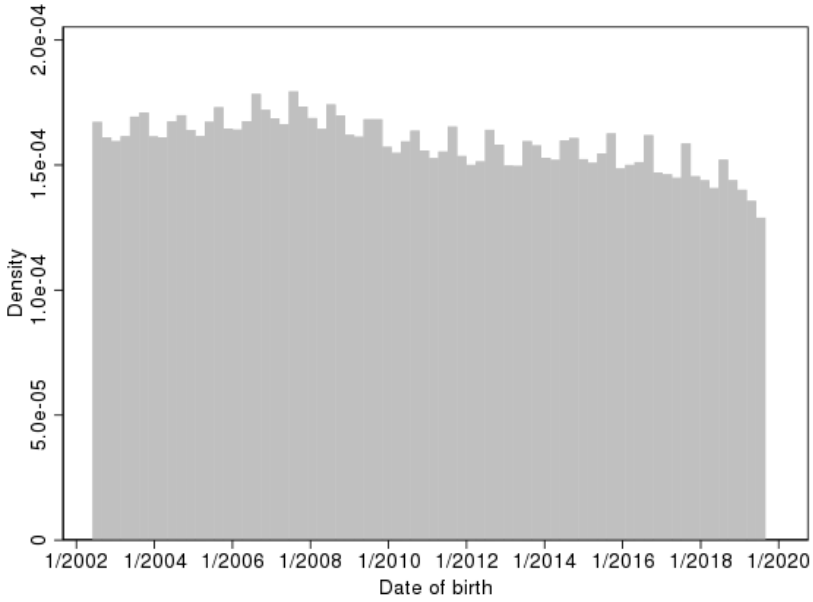
Stratum 1: identifying explicit links from children to addresses

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, 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 83,650,000 children in the 2019 Numident who will be aged 0–17 years on June 1, 2020. 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 June 1, 2020 (2019 Numident)



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 March 2019, there are 64,440,000 unique records for children who will be aged 0–17 years on June 1, 2020.

In addition to the links between parents and children available in the CHCK, we will also utilize the links between household members which can be measured in the American Community Survey, which is not an underlying data source for the CHCK. For each child in the Numident aged 0-17 on June 1, 2020, we harvest relationships with the head of household and the spouse of the head of household. We then use these links to supplement the links in the CHCK.

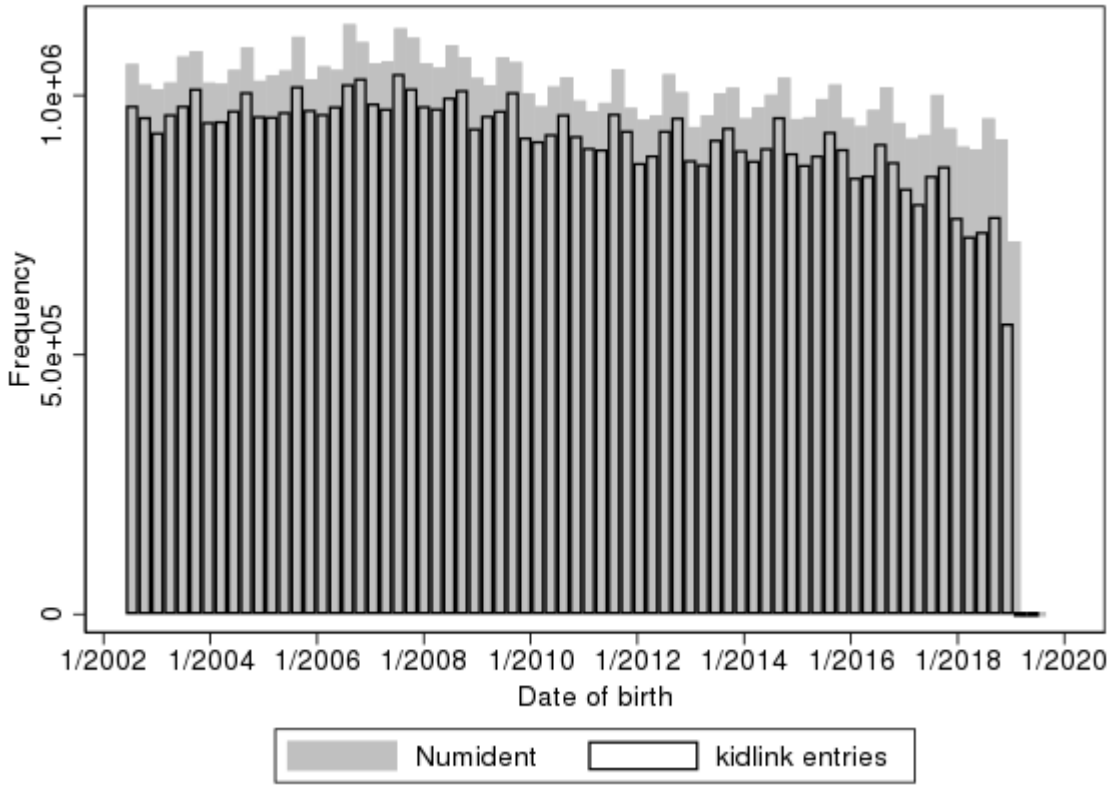
Let us consider how many children from the Numident have been linked to a parent in the CHCK file or to a parent in the ACS. Table 1 shows the number of children linked with both a mother and a father, linked with a mother only, linked with a father only, linked with a parent in the ACS or not linked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

Table 1: Child-parent links in the CHCK file relative to the Numident population, aged 0–17 years as of 2020, March 2019 CHCK file and ACS

Type of link	Frequency	Percent
Mother and father	58,140,000	70%
Mother only	13,800,000	17%
Father only	2,423,000	2.9%
ACS link	97,600	0.1%
No link	9,185,000	11%
All children in Numident	83,650,000	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. CHCK entries, aged 0–17 years as of June 1, 2019



The CHCK file was updated in March 2019 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 83,650,000[1] children in the Numident, 68,390,000 are matched directly to a MAFID. Out of 71,940,000 CHCK-matched mothers, about 66,400,000 are matched to a MAFID. Out of

[1] All unweighted counts and estimates in this document are rounded to no more than four significant figures in accordance with Census Disclosure Review Board rules on rounding.

60,560,000 CHCK-matched fathers, about 56,010,000 are matched to a MAFID. Additionally, out of 9,433,000 ACS-matched parents, 8,790,000 are matched to a MAFID.

For each child observation from the Numident, we now have multiple 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 (65,750,000 cases). If the MAFID is missing, we assign the mother MAFID (4,893,000 cases). Then, if the MAFID is still missing, we assign the father MAFID (2,032,000 cases). Finally, if the child, CHCK mother and CHCK father MAFIDs are missing, we assign the ACS parent MAFID (43,000 cases). That leaves 10,930,000 children from the Numident not assigned MAFIDs (a MAFID match rate of 86.9%).

There are some MAFIDs associated with a great number of children. As an example, out of 72,720,000 associated with a MAFID, 7,440,000 children are associated with a MAFID with more than 20 child-MAFID links.

The 77,600,000 children associated with a MAFID are then collapsed down to 38,160,000 unique MAFIDS. This implies 1.91 children per household for households assigned a flag.

For 2020, we apply one additional step in the construction of stratum 1. We use administrative HUD PIC and TRACS data, which contain flags for the number of children present at the household level for all public housing and voucher households, to enhance the existing stratum 1 process. We merge all MAFIDs not assigned a stratum 1 flag using the above CHCK-MAF-ARF process with the most recent data on all public housing and voucher households in the PIC-TRACS data. We will then assign a stratum 1 flag to all households which have a child present flag in the HUD data. This adds 185,000 households to stratum 1.

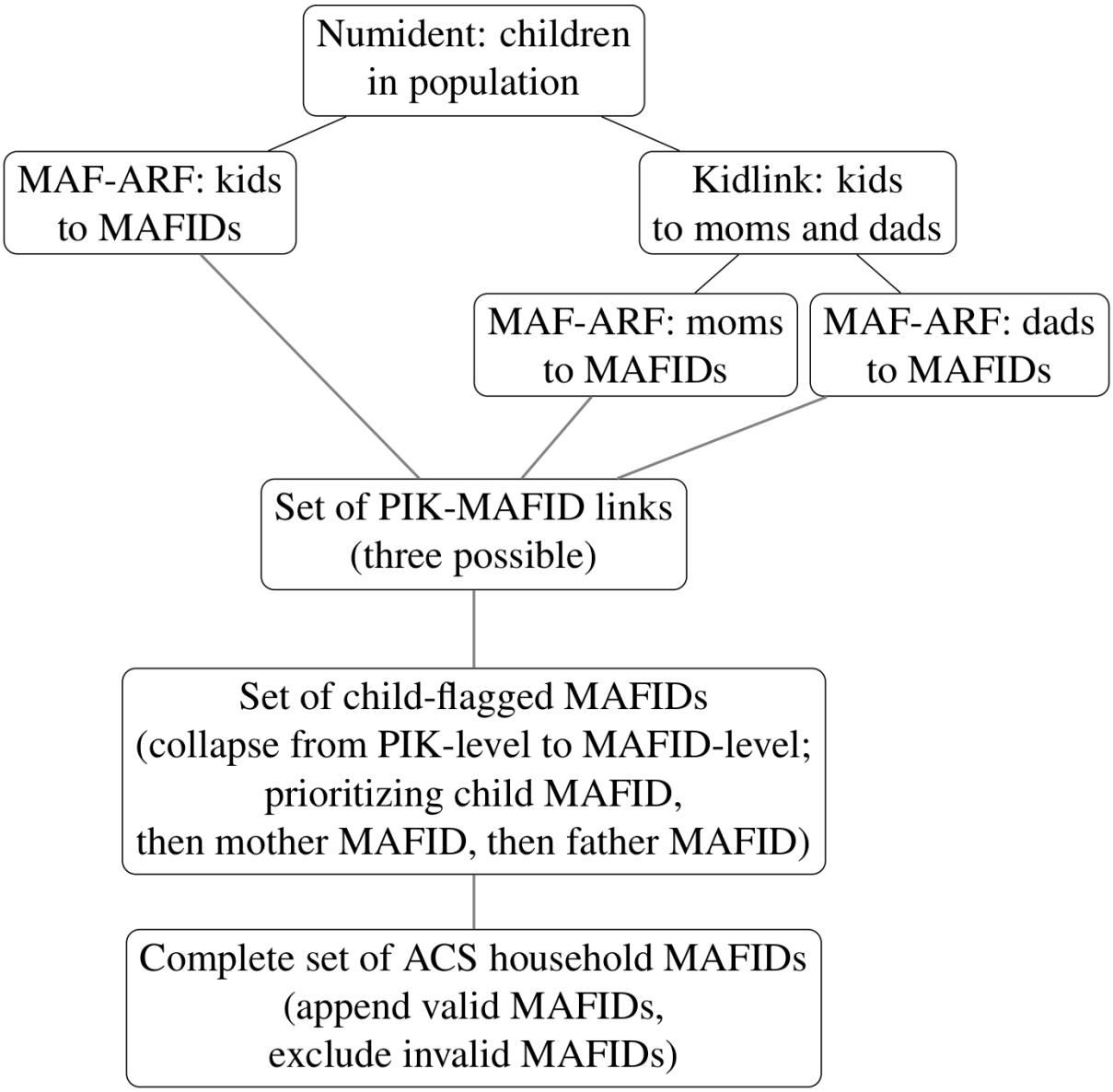
We then need to scale up the MAFID list to the universe of MAFIDs to allow sampling of unflagged households. A merge of the 38,160,000 unique child-flagged MAFIDS with the January 2019 ACS MAF-X file matches 38,160,000 MAFIDS with child flags, removes 171,000,000 MAFIDS with child flags, and adds 400 MAFIDs without child flags. The sample frame file now has about 209 million valid MAFIDS. Compare this with the 2011 ACS, in which about 37 million out of 115 million households included related children.[2]

Stratum 1 construction visualization

Figure 3 shows a visualization of the sample frame construction.

Figure 3: Stratum 1 construction

[2] 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 2020 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

- 2018 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 2018 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):

- 2018 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 2018
  - any_kid_deduct_max: does any tax form associated with this MAFID have any deduction related to children?[3]
  - 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
- VSGI CRD commercial data
  - homeowner_crd: is this address owner occupied in the VSGI CRD data
  - any_child_crd: are there any children at this address
  - num_children_crd: number of children at this address
  - child_age_*: flags for child age ranges
  - miss_*_crd: flags for missingness in the CRD 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

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

Parameter estimates are stored in the file frame2018_child_present_bystate.csv. In general these models provide a reasonable fit, explaining between 40 and 50 percent of the in-sample variation for most states.

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:

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

Stratum 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:

- 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:

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

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 ($\bar{C}$)[4], 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, ($p(S_1 \cup S_{2a}|C)$).

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

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 2018 ACS microdata.

Table 2: [5] NSCH optimal Strata

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	2143000	0.22	0.446	0.334	0.759	0.149	0.043	0.108	0.953	31	0.032
AL	35000	0.208	0.535	0.257	0.697	0.133	0.053	0.11	0.95	21	0.034
AK	8700	0.141	0.54	0.319	0.719	0.144	0.197	0.155	0.88	-1	-0.199
AZ	41500	0.205	0.473	0.322	0.75	0.16	0.046	0.119	0.951	28	0.06
AR	20500	0.215	0.54	0.246	0.723	0.137	0.06	0.115	0.95	20	0.035
CA	201000	0.268	0.367	0.364	0.765	0.189	0.044	0.122	0.952	37	0.11
CO	35500	0.227	0.41	0.363	0.789	0.16	0.039	0.106	0.95	36	0.08
CT	21500	0.228	0.374	0.398	0.79	0.157	0.035	0.099	0.952	40	0.083
DE	6800	0.19	0.361	0.448	0.745	0.139	0.029	0.085	0.952	43	0.066
DC	4300	0.174	0.594	0.232	0.655	0.076	0.038	0.066	0.951	21	0.014
FL	113000	0.196	0.401	0.404	0.681	0.142	0.028	0.088	0.952	40	0.069
GA	51500	0.243	0.453	0.304	0.732	0.168	0.052	0.126	0.953	28	0.071
HI	9200	0.139	0.613	0.248	0.696	0.241	0.053	0.189	0.951	21	0.101
ID	11000	0.215	0.455	0.33	0.78	0.161	0.044	0.113	0.951	30	0.084
IL	89000	0.226	0.407	0.367	0.768	0.16	0.041	0.109	0.952	34	0.072
IN	44000	0.228	0.422	0.35	0.76	0.152	0.043	0.107	0.951	33	0.062
IA	32000	0.197	0.645	0.158	0.796	0.086	0.206	0.098	0.942	-1	-0.164
KS	24500	0.221	0.394	0.385	0.776	0.157	0.04	0.106	0.953	34	0.068
KY	31000	0.222	0.569	0.209	0.767	0.135	0.072	0.119	0.95	17	0.012
LA	27500	0.224	0.454	0.322	0.682	0.149	0.045	0.11	0.951	31	0.078
ME	16000	0.141	0.483	0.376	0.77	0.093	0.03	0.07	0.95	28	0.03
MD	35500	0.245	0.38	0.374	0.785	0.162	0.041	0.106	0.95	37	0.081
MA	39500	0.213	0.414	0.373	0.803	0.143	0.036	0.095	0.952	37	0.075

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

MI	94500	0.2	0.35	0.45	0.787	0.141	0.028	0.083	0.953	43	0.075
MN	69000	0.209	0.363	0.428	0.828	0.14	0.033	0.087	0.952	40	0.071
MS	17000	0.222	0.55	0.228	0.702	0.136	0.065	0.118	0.951	18	0.024
MO	46500	0.21	0.446	0.344	0.761	0.134	0.039	0.096	0.952	31	0.052
MT	10500	0.15	0.698	0.151	0.77	0.098	0.151	0.105	0.926	-1	-0.188
NE	19500	0.206	0.585	0.209	0.8	0.106	0.1	0.104	0.95	9	-0.034
NV	18000	0.227	0.451	0.322	0.72	0.153	0.046	0.113	0.951	31	0.066
NH	10500	0.174	0.491	0.334	0.807	0.1	0.035	0.076	0.952	31	0.046
NJ	50000	0.233	0.373	0.394	0.793	0.183	0.038	0.114	0.952	38	0.101
NM	15000	0.166	0.689	0.146	0.677	0.119	0.176	0.126	0.934	-1	-0.153
NY	124500	0.206	0.476	0.318	0.753	0.15	0.043	0.11	0.951	30	0.068
NC	64000	0.215	0.438	0.348	0.746	0.155	0.041	0.109	0.951	33	0.073
ND	8800	0.177	0.682	0.141	0.776	0.081	0.169	0.091	0.935	-1	-0.152
OH	83000	0.219	0.406	0.376	0.772	0.148	0.034	0.097	0.954	37	0.07
OK	43500	0.215	0.658	0.128	0.735	0.131	0.139	0.132	0.95	3	-0.046
OR	25500	0.209	0.448	0.342	0.782	0.134	0.041	0.097	0.951	31	0.055
PA	110000	0.2	0.378	0.422	0.795	0.137	0.032	0.087	0.952	40	0.067
RI	5900	0.2	0.371	0.429	0.758	0.152	0.031	0.092	0.95	43	0.086
SC	30500	0.212	0.441	0.347	0.719	0.133	0.039	0.096	0.95	33	0.06
SD	9000	0.186	0.69	0.124	0.782	0.102	0.201	0.113	0.934	-1	-0.191
TN	40500	0.23	0.415	0.355	0.747	0.152	0.039	0.104	0.952	34	0.076
TX	137000	0.256	0.46	0.284	0.752	0.18	0.064	0.141	0.951	25	0.073
UT	18000	0.306	0.4	0.294	0.825	0.189	0.066	0.14	0.95	25	0.097
VT	8100	0.149	0.618	0.233	0.797	0.082	0.044	0.072	0.952	17	0.004
VA	50500	0.24	0.394	0.365	0.783	0.158	0.041	0.106	0.95	36	0.082
WA	45500	0.23	0.419	0.351	0.796	0.16	0.042	0.109	0.95	34	0.078
WV	13500	0.151	0.684	0.165	0.73	0.104	0.171	0.117	0.872	-1	-0.185
WI	71500	0.199	0.364	0.437	0.807	0.145	0.031	0.088	0.951	41	0.07
WY	4100	0.172	0.754	0.074	0.777	0.106	0.146	0.11	0.951	1	-0.106

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 2018. This is an in-sample audit, and as such it will by design meet the 95% threshold.

All estimates are weighted with the housing-unit-level weights, which include weight for vacant units (about 200,000 vacant housing units in the 2018 ACS). In vacant housing units, we assign zero children. These estimates should reflect the NSCH survey production process.

State-specific performance

In 2020, the smallest oversample strata were 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: [6] 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	2143000	0.22	0.437	0.343	0.759	0.152	0.043	0.108	0.95
AL	35000	0.208	0.535	0.257	0.697	0.133	0.053	0.11	0.95
AK	8700	0.141	0.54	0.319	0.719	0.144	0.197	0.155	0.88
AZ	41500	0.205	0.471	0.324	0.75	0.16	0.046	0.119	0.951
AR	20500	0.215	0.54	0.246	0.723	0.137	0.06	0.115	0.95
CA	201000	0.268	0.367	0.365	0.765	0.189	0.045	0.122	0.951
CO	35500	0.227	0.411	0.362	0.789	0.16	0.039	0.106	0.95
CT	21500	0.228	0.374	0.398	0.79	0.157	0.035	0.099	0.952
DE	6800	0.19	0.361	0.448	0.745	0.139	0.029	0.085	0.952
DC	4300	0.174	0.595	0.231	0.655	0.076	0.038	0.066	0.951
FL	113000	0.196	0.401	0.403	0.681	0.142	0.028	0.088	0.952
GA	51500	0.243	0.452	0.305	0.732	0.168	0.052	0.126	0.953
HI	9200	0.139	0.613	0.248	0.696	0.241	0.053	0.189	0.951
ID	11000	0.215	0.454	0.331	0.78	0.161	0.044	0.113	0.951
IL	89000	0.226	0.407	0.368	0.768	0.16	0.041	0.109	0.952
IN	44000	0.228	0.422	0.35	0.76	0.152	0.043	0.107	0.951
IA	32000	0.197	0.645	0.158	0.796	0.086	0.206	0.098	0.942
KS	24500	0.221	0.393	0.386	0.776	0.158	0.04	0.106	0.953
KY	31000	0.222	0.569	0.209	0.767	0.135	0.072	0.119	0.95
LA	27500	0.224	0.454	0.322	0.682	0.149	0.045	0.11	0.951
ME	16000	0.141	0.486	0.373	0.77	0.093	0.03	0.07	0.95
MD	35500	0.245	0.38	0.375	0.785	0.162	0.042	0.106	0.95
MA	39500	0.213	0.414	0.373	0.803	0.143	0.036	0.095	0.952
MI	94500	0.2	0.351	0.449	0.787	0.141	0.028	0.083	0.953
MN	69000	0.209	0.363	0.428	0.828	0.14	0.033	0.087	0.952
MS	17000	0.222	0.549	0.229	0.702	0.136	0.065	0.118	0.951
MO	46500	0.21	0.446	0.344	0.761	0.134	0.039	0.096	0.952
MT	10500	0.15	0.698	0.151	0.77	0.098	0.151	0.105	0.926
NE	19500	0.206	0.586	0.209	0.8	0.105	0.1	0.104	0.95
NV	18000	0.227	0.452	0.321	0.72	0.153	0.046	0.113	0.951
NH	10500	0.174	0.492	0.333	0.807	0.101	0.034	0.076	0.954
NJ	50000	0.233	0.372	0.394	0.793	0.183	0.038	0.114	0.952
NM	15000	0.166	0.689	0.146	0.677	0.119	0.176	0.126	0.934
NY	124500	0.206	0.476	0.318	0.753	0.15	0.043	0.11	0.951
NC	64000	0.215	0.438	0.348	0.746	0.155	0.041	0.109	0.95

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

ND	8800	0.177	0.682	0.141	0.776	0.081	0.169	0.091	0.935
OH	83000	0.219	0.405	0.376	0.772	0.148	0.034	0.097	0.954
OK	43500	0.215	0.658	0.128	0.735	0.131	0.139	0.132	0.95
OR	25500	0.209	0.448	0.343	0.782	0.135	0.041	0.097	0.951
PA	110000	0.2	0.379	0.421	0.795	0.137	0.032	0.087	0.952
RI	5900	0.2	0.371	0.429	0.758	0.152	0.031	0.092	0.95
SC	30500	0.212	0.442	0.346	0.719	0.133	0.039	0.096	0.95
SD	9000	0.186	0.69	0.124	0.782	0.102	0.201	0.113	0.934
TN	40500	0.23	0.415	0.355	0.747	0.152	0.039	0.104	0.952
TX	137000	0.256	0.459	0.284	0.752	0.18	0.064	0.141	0.951
UT	18000	0.306	0.4	0.294	0.825	0.19	0.066	0.14	0.95
VT	8100	0.149	0.62	0.231	0.797	0.082	0.044	0.072	0.952
VA	50500	0.24	0.394	0.366	0.783	0.157	0.042	0.106	0.95
WA	45500	0.23	0.419	0.35	0.796	0.16	0.042	0.109	0.95
WV	13500	0.151	0.684	0.165	0.73	0.104	0.171	0.117	0.872
WI	71500	0.199	0.364	0.437	0.807	0.145	0.031	0.088	0.951
WY	4100	0.172	0.753	0.075	0.777	0.106	0.147	0.11	0.95

We additionally audit the frame out of sample against an early release file of 2019 ACS microdata, as shown in table 4. 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. Performance is slightly lower than the in-sample audit in table 3, but is in line with the results in previous sample years.

Table 4: [7] NSCH strata, ACS2019, 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	1889000	0.233	0.416	0.351	0.834	0.127	0.046	0.09	0.938
AL	29000	0.232	0.516	0.252	0.785	0.116	0.045	0.092	0.955
AK	5800	0.183	0.598	0.219	0.748	0.176	0.435	0.246	0.718
AZ	36000	0.221	0.457	0.322	0.815	0.131	0.054	0.099	0.933
AR	17500	0.231	0.531	0.239	0.804	0.125	0.062	0.106	0.945
CA	181000	0.275	0.35	0.374	0.826	0.16	0.038	0.097	0.953
CO	32500	0.236	0.399	0.364	0.858	0.126	0.031	0.081	0.957
CT	19000	0.223	0.367	0.411	0.87	0.129	0.025	0.074	0.96
DE	5700	0.209	0.363	0.428	0.815	0.106	0.028	0.064	0.946
DC	3800	0.169	0.572	0.259	0.749	0.067	0.032	0.056	0.952
FL	97000	0.209	0.39	0.402	0.778	0.117	0.028	0.072	0.949
GA	45000	0.263	0.432	0.306	0.801	0.133	0.047	0.097	0.949

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

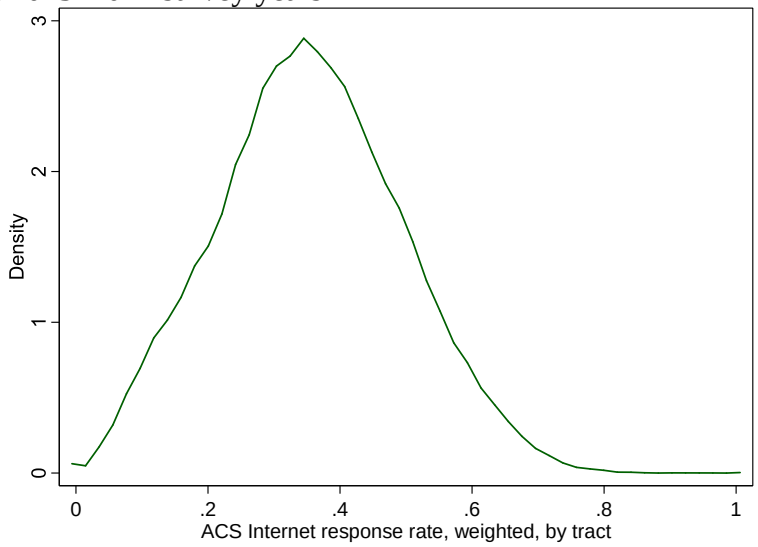
HI	7700	0.146	0.606	0.248	0.744	0.252	0.062	0.197	0.944
ID	9700	0.257	0.411	0.333	0.849	0.137	0.042	0.095	0.951
IL	80500	0.229	0.386	0.385	0.841	0.13	0.049	0.09	0.927
IN	40000	0.234	0.404	0.362	0.842	0.13	0.042	0.088	0.942
IA	29500	0.204	0.643	0.153	0.876	0.063	0.236	0.096	0.859
KS	22000	0.233	0.375	0.392	0.862	0.137	0.058	0.097	0.917
KY	27500	0.231	0.566	0.203	0.834	0.108	0.057	0.095	0.956
LA	22500	0.243	0.427	0.33	0.754	0.135	0.043	0.095	0.944
ME	12000	0.177	0.497	0.326	0.838	0.087	0.033	0.066	0.947
MD	31500	0.263	0.359	0.378	0.852	0.126	0.033	0.078	0.956
MA	35500	0.222	0.397	0.381	0.869	0.12	0.027	0.075	0.959
MI	80500	0.219	0.333	0.448	0.857	0.119	0.03	0.068	0.944
MN	61500	0.228	0.332	0.44	0.886	0.12	0.034	0.071	0.943
MS	14000	0.242	0.532	0.226	0.756	0.134	0.078	0.117	0.935
MO	41000	0.225	0.431	0.345	0.841	0.12	0.043	0.086	0.942
MT	8600	0.172	0.736	0.093	0.816	0.096	0.253	0.113	0.9
NE	17500	0.225	0.582	0.193	0.872	0.078	0.158	0.098	0.888
NV	16000	0.239	0.446	0.315	0.799	0.128	0.034	0.089	0.959
NH	8900	0.212	0.442	0.346	0.867	0.095	0.028	0.066	0.959
NJ	45000	0.25	0.355	0.395	0.859	0.155	0.032	0.09	0.955
NM	11500	0.193	0.718	0.088	0.75	0.111	0.289	0.131	0.898
NY	108000	0.222	0.453	0.325	0.82	0.138	0.036	0.095	0.955
NC	55000	0.231	0.413	0.356	0.827	0.129	0.035	0.086	0.951
ND	7300	0.205	0.694	0.102	0.846	0.081	0.245	0.102	0.902
OH	75000	0.227	0.382	0.391	0.855	0.119	0.031	0.075	0.952
OK	34000	0.236	0.664	0.1	0.779	0.126	0.186	0.133	0.935
OR	23000	0.215	0.444	0.341	0.847	0.109	0.041	0.079	0.943
PA	95500	0.211	0.359	0.429	0.873	0.113	0.029	0.067	0.948
RI	5200	0.208	0.352	0.44	0.834	0.125	0.036	0.076	0.931
SC	26500	0.228	0.421	0.35	0.796	0.112	0.035	0.077	0.949
SD	7700	0.217	0.698	0.085	0.863	0.098	0.266	0.116	0.919
TN	36500	0.239	0.401	0.36	0.818	0.129	0.037	0.085	0.949
TX	119000	0.273	0.439	0.288	0.816	0.152	0.052	0.113	0.95
UT	16000	0.331	0.393	0.277	0.872	0.173	0.068	0.129	0.95
VT	6400	0.182	0.587	0.231	0.862	0.088	0.039	0.074	0.958
VA	46000	0.253	0.377	0.37	0.863	0.123	0.027	0.076	0.963
WA	41000	0.242	0.404	0.354	0.853	0.134	0.034	0.087	0.956
WV	11000	0.18	0.717	0.103	0.822	0.084	0.226	0.102	0.9
WI	62500	0.214	0.342	0.444	0.868	0.125	0.029	0.071	0.947
WY	3400	0.208	0.735	0.056	0.802	0.108	0.238	0.117	0.949

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

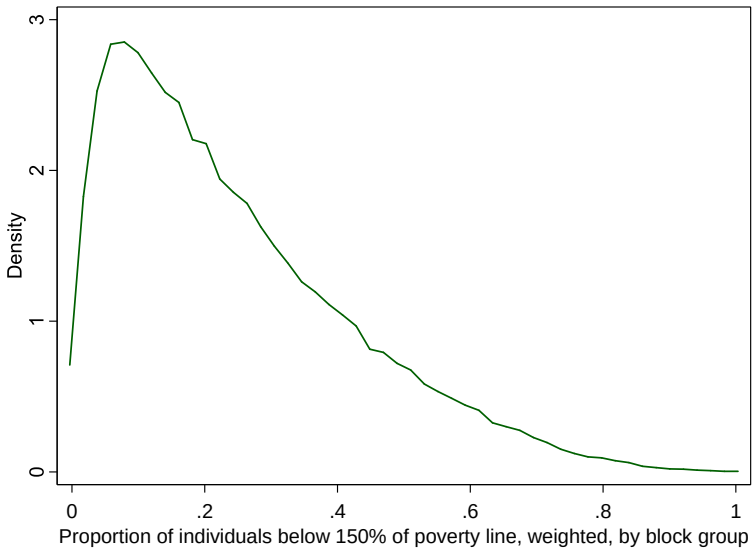


To construct an Internet-access flag, we use the first tritile for a cut-off. A block is considered to have low Internet access if the Internet accessibility index is below the first tritile of the block-level distribution. For low-population blocks, we replace missing values of the block-varying low-Internet 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 2018 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, 2018 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 2.

Table 2: NSCH population data file layout

Variable name	Label	Level of variation	Type	Any missing?
mafid	Master Address File ID	MAFID	long	no
maf_curstate	State	State	str2	no
maf_curcounty	County	County	str3	no
maf_curblktract	Tract	Tract	str6	yes
maf_curblkgrp	Block group	Blocck group	str1	yes
maf_curblk	Block	Block	str4	yes
stratum1	Stratum 1 identifier	MAFID	byte	no
stratum2a	Stratum 2a identifier	MAFID	byte	no
stratum2b	Stratum 2b identifier	MAFID	byte	No
kids_00_02	Number of children aged 0–2 years	MAFID	integer	no
kids_03_05	Number of children aged 3–5 years	MAFID	integer	no
kids_06_08	Number of children aged 6–8 years	MAFID	integer	no
kids_09_11	Number of children aged 9–11 years	MAFID	integer	no

kids_12_14	Number of children aged 12–14 years	MAFID	integer	no
kids_15_17	Number of children aged 15–17 years	MAFID	integer	no
blkgrp_185_200_povrate	Pr. HH w/ inc. 185–200% poverty rate	Block group	float	yes
blkgrp_gt_200_povrate	Pr. HH w/ inc. > 200% poverty rate	Block group	float	yes
blkgrp_lt_150_povrate	Pr. HH w/ inc. < 150% poverty rate	Block group	float	yes
mailvaldf	Valid mailing address	MAFID	byte	yes

Filename: nsch_pop_file.sas7bdat

Population: all MAFIDs in 2019 MAF-X

Unit of observation: household (MAFID)