April 8, 2019

State Oversampling in the National Survey of Children's Health:
Feasibility, Cost, and Alternative Approaches

Some states have expressed an interest in sponsoring an oversample of the National Survey of Children's Health (NSCH) to inform state-level decision making around various priorities. This was an option in previous iterations of the NSCH and will become a renewed possibility beginning with the 2020 NSCH.

This document outlines the purpose and types of oversampling, feasibility, costs, and alternative approaches to sub-state oversampling, namely the use of synthetic or model-based local area estimates. It also includes answers to frequently asked questions.

I. PURPOSE AND TYPES OF OVERSAMPLES:

Oversamples can support more targeted assessment, program planning, and evaluation:

- State-wide oversampling increases the number of completed surveys per state which may enable reporting for smaller populations or rare outcomes (e.g. AIAN or autism) with greater precision

- Sub-state oversampling by geography enables reporting of local area estimates (e.g. city, county, or region).

II. FEASIBILITY:

These two types of oversampling involve different levels of complexity and feasibility.

(a) **State-Wide Oversampling:** An oversample to increase the number of survey participants within a state is straightforward and feasible for all states. In this case, the oversample is distributed throughout the state in roughly the same geographic distribution as the population. The larger sample would allow for state-level estimates of rare populations or outcomes, but may not facilitate sub-state (e.g., county-level) estimates.

(b) **Sub-State Oversampling:** Oversampling to produce county- or other sub-state-level estimates is more complex. The oversample is designed to supplement the NSCH sample, which is distributed proportionately by population across the state, to ensure a minimum sample size in each targeted sub-state geography. There is flexibility in the sub-state geographies that are targeted, but targeting more areas, especially areas with small populations relative to the rest of the state, will increase the oversample required and, thus, the cost of the project.

- Example: In California, targeting a minimum sample size in each of the 58 counties is probably not feasible. It is more feasible to group the smallest counties with other neighboring counties to substantially reduce the required oversample. In the case of California, we can group

counties into 40 units representing 32 stand-alone counties and 8 county groups. This grouping of counties reduces the extra sampled needed by about 40%.

Should a state wish to sponsor either type of an oversample, the additional survey responses would be included with the NSCH public use microdata files released each fall. To protect the confidentiality of respondents, county and other sub-state geographic indicators are only made available on restricted-access microdata files and, therefore, would not be listed on the public use data files. Thus, states that sponsor a sub-state oversample would need to analyze data and create their county-level estimates using the confidential files available only through the U.S. Census Bureau's Research Data Centers (RDC).[1] The process for accessing an RDC includes a researcher securing Special Sworn Status and obtaining access to a local RDC or travel to Census Headquarters in Maryland (approximately 2 months), and finally review and approval of estimates through the Census Bureau's Disclosure Review Board (DRB) prior to public release (approximately 2-4 weeks). RDC access may involve additional fees; a list of locations is available at https://www.census.gov/about/adrm/fsrdc/locations.html.

III. COSTS:

General Calculations: The cost for either a statewide or sub-state oversample is driven by four factors: 1) the cost per sampled address; 2) the estimated number of sampled addresses per completed survey (topical questionnaire); 3) the desired sample size; and 4) the number of years to achieve the desired sample size.

1) **The estimated cost per sampled address for the 2020 NSCH is $13**; this includes the cost of materials (letters, envelopes), postage, incentives, processing (data entry and cleaning), and survey planning and management. Attachment A provides a general breakdown of the cost per sampled address. The cost per sampled address will be re-evaluated annually to account for changes in incentive, material and shipping costs.

2) **On average, the ratio of sampled addresses to completed surveys is about 7:1**, but that ratio varies across states. This ratio is based on the probability a sampled address represents a household with children (not a business, vacant address, or household without children) that, in turn, completes a topical questionnaire; it can be calculated for each state in a particular year by dividing addresses by completed surveys in Attachment C of the methodology reports. Attachment B of this document lists the estimated ratio of addresses to completed topical questionnaires by state for the 2020 NSCH.

3) **Desired sample size depends on the goal of oversampling.** Achieving reasonable reliability of key estimates may be a good guide to determining a target sample size for the population of interest (e.g., AI/AN in a state-wide oversample or county in a sub-state oversample). Reliability is commonly measured with the coefficient of variation (CV) or relative standard error, which is the standard error as a percentage of the prevalence estimate, with a larger CV

[1] We recognize the burden associated with having to use an RDC and are exploring alternatives to promote data access (see FAQs).

indicating poorer relative reliability. The CV is dependent on the standard error and the prevalence of indicators, and improves with increasing sample size and increasing prevalence. If the CV exceeds 30%, estimates are commonly suppressed or flagged as unreliable. It should be emphasized that an overall sample size may not accommodate reliable analysis of subgroups, including various Title V National Performance Measures within population domains (e.g., adolescents or CSHCN). Attachment C shows reporting reliability across National Performance Measures and may be helpful in determining total target sample sizes for populations of interest. For example, an overall county-level sample size of ~150 would be necessary to yield ~30 CSHCN (23% of the unweighted national sample size) for the medical home CSHCN performance measure. A denominator of at least 30 is necessary to meet MCHB reporting standards but estimates would still have wide confidence intervals and poor precision at that minimum sample size. For full population performance measures, such as preventive dental visit and adequate insurance, a sample size of 150 would enable reporting without any reliability flags. However, a sample size of 300 would be needed to enable reporting for all performance measures and about half would still carry reliability flags.

4) **Pooling data over multiple years of the general sample reduces cost by requiring less oversample to meet a target sample size**. For example, if a state seeks a target sample size of 150 AI/AN children and has ~50 completed surveys in the general annual sample, a total of 150 surveys could be completed by purchasing 100 surveys in one year (50 base + 100 additional), 25 surveys per year over two years (100 base + 50 additional), or combining three years of data without oversampling (150 base).

Additional considerations for sub-state oversampling: As noted in the feasibility section, two additional factors influence the cost of a sub-state oversample:

1) *The number of sub-state units*. Grouping counties into county groups, for example, can substantially reduce the requirements of the sub-state oversample.
2) *The distribution of population across units*. For example, if population is concentrated in one county, the base sample will also be concentrated in that county (proportional to the population), so less of the base sample is contributing to the minimum sample requirements in the other counties.

Census can work with individual states to determine an appropriate plan to address both data and budgetary concerns.

Example Sub-State Oversampling Calculations: The scenarios presented below are modeled using the California example referenced earlier, targeting 32 counties and 8 county groups in that state. We anticipate these scenarios to represent a practical upper boundary in terms of the required oversample and cost. For this example, assuming a target number of 150 completed surveys per county, we estimate that:[2]

[2] These are example estimates provided for comparison of various scenarios. Actual costs must be determined on a case-by-case, year-by-year basis by the U.S. Census Bureau.

- A 5-year investment would cost *California approximately $64,000 per year for a total of $320,000.* This would buy:

    - 4,910 additional sampled households per year (or 24,548 total) plus the current base NSCH sample level in California, and would result in a net over the five years of 150 completed surveys per county/county group.

- A 3-year investment would cost *California approximately $123,000 per year for a total of $368,000*. This would buy:

    9,446 additional sampled households per year (or 28,337 total) plus the current base NSCH sample level in California, and would result in a net over the three years of 150 completed surveys per county/county group.

- A 1-year investment would cost *California approximately $430,000*. This would buy:

    - 32,789 additional sampled households plus the current base NSCH sample level in California, and would result in 150 completed surveys per county/county group.

The first year that this county-level oversampling could be implemented would be 2020, yielding county group-level data in 2022 after 3 years of data collection or 2024 after 5 years of data collection. Alternatively, we could combine data from the 2016 to 2020 NSCH, and conduct a one-year oversample in 2020 to supplement the 5-year base sample. The aggregate cost is slightly lower than the 5-year example above – $300,000 for 23,000 additional sampled households – because we can leverage the relatively large 2016 NSCH sample, and county-level estimates would be available as soon as the 2020 NSCH data is released in 2021.

See **Attachment D** for a relative cost by pooled years and minimum sample per county compared to a three-year investment for 150 completed surveys per county. Attachment D presents relative costs for a range of possibilities, but practically, 50 completed surveys is a minimum requirement for direct county-based estimates in most cases.

Example: A target of 50 interviews per county would incur only 24% of the cost of 150 interviews per county.

IV. ALTERNATIVES

(a) Synthetic Estimation: Synthetic or indirect estimation can also be used to produce county-level or other sub-state estimates with publicly available data. [3] This is generally accomplished by

[3]See "Local Uses of National and State Data" (http://www.childhealthdata.org/docs/nsch-docs/local-use-of-state-data-and-synthetic-estimates.pdf)

multiplying sub-state sociodemographic characteristics from the American Community Survey (ACS) with state-level prevalence estimates from the NSCH for those sociodemographic characteristics. For example, county-level obesity estimates could be indirectly estimated by applying the state-level prevalence of obesity by race/ethnicity and poverty categories to the county-level distribution or proportion of children in each race/ethnicity and poverty category combination as estimated by the ACS.

However, this approach has limitations. Imposing prevalence rates of key health measures from a state population to a county based on the sociodemographic characteristics of that county can mask true geographic differences by assuming that variation is only a function of composition.

(b) Model-Based Estimation: County-level estimates can also be derived through multilevel regression models that nest observations within counties using non-publicly available geographic information through an RDC.[4] Bayesian approaches can smooth or shrink imprecise county-level estimates toward a spatially weighted or overall state mean. Model-based estimates can be improved with sub-state oversampling, but with far lower sample requirements than a sub-state oversampling project designed for direct estimates alone. The requirements will vary by state, outcome of interest, and geographic granularity.

The Census Bureau's NSCH team is unable to provide technical support on analytic alternatives to direct estimates but can work with states to design an oversample and provide RDC access to confidential files for approved projects.

V. FREQUENTLY ASKED QUESTIONS

If my state is interesting in exploring or pursuing an oversample, what are the next steps?

After you've determined your specific interest (i.e. statewide or sub-state oversampling) and a target sample size, the Census Bureau can determine feasibility and develop a custom cost estimate. If the state wishes to proceed, a sampling plan and various agreements would need to be developed and approved. To begin discussions regarding your interest, please contact Drs. Ashley Hirai (AHirai@hrsa.gov) and Scott Albrecht (Scott.Albrecht@census.gov).

What is the deadline for sponsoring an oversample for 2020? Will this option continue in subsequent years?

For a 2020 oversample, the sampling plan must be finalized by mid-July to allow for the development and approval of various agreements with funding transfer by December of 2019. This option and timeline is expected to continue for each subsequent year.

[2] For example, see Kramer, M. R., Raskind, I. G., Van Dyke, M. E., Matthews, S. A., and Cook-Smith, J. N. (2016). Geography of Adolescent Obesity in the U.S., 2007–2011. *American Journal of Preventive Medicine,* 51(6), 898–909.

Can you tell us how many extra oversample surveys we will need to meet a certain target number for a given population group or per county over 1, 3, and 5 years?

Yes, this is something we can answer. By contrast, a state will have to determine their target, how sub-state units are delineated, and how many years they are willing to wait for multi-year estimates.

Can we target sampling for a particular demographic group rather than doing a general state-wide oversample?

We can't directly oversample households with a certain demographic characteristic since this detail is unknown for individual households. However, we may be able to target the oversample to counties or block groups with a disproportionate share of a population characteristic (e.g. AIAN) as estimated from the American Community Survey. These projects will be evaluated on a case-by-case basis and could incur additional costs.

If our state purchases an oversample, why can't we get a special file to analyze that data with county identifiers? Do we really need to access the RDC?

We are exploring multiple options to promote data access, both directly through a special file and indirectly through a resource center that could access the RDC to produce estimates. A special direct file may be an option through some modification or noise infusion so that observations could not be linked or identified in the public use file. A special state-specific file may incur additional costs.

Are local or state-specific questions an option if we purchase an oversample?

It is unclear whether or when it will be feasible to add local/state survey item options. It would involve more costs to tailor both web and print questionnaires. Thus, adding global questions to the survey is a more immediately feasible option.

For an oversample dataset, will a new "weight" be calculated by the Census Bureau's NSCH team for county-level analyses?

Any oversampled cases will be part of the Public Use File and indistinguishable from other cases (county identifiers will not be included on a Public Use File). The weights for cases in oversampled states will account for the additional complexity of the oversample and will include a higher level of detail in weighting adjustments, including county characteristics.

Attachment A – Estimated Total Cost per Sampled Address for NSCH 2020

Planning and Survey Management	$0.98
Mailed Materials, Postage, Package Assembly, & QA	$6.84
Incentives	$3.41
Sorting, check-in, and data capture of mailed returns	$0.60
Customer Assistance	$0.67
Data Processing & Editing	$0.50
Total	$13.00
Estimates based on 2018 costs; costs will be re-evaluated annually	

Attachment B - Ratio of Sampled Addresses to Completed Topical Questionnaires by State, Estimated for the 2020 NSCH

State	Addresses per Completed Topical Questionnaire	State	Addresses per Completed Topical Questionnaire
Alabama	7.8	Missouri	5.5
Alaska	8.9	Montana	7.0
Arizona	7.4	Nebraska	5.5
Arkansas	9.1	Nevada	7.8
California	5.9	New Hampshire	5.8
Colorado	5.1	New Jersey	5.4
Connecticut	5.4	New Mexico	9.5
Delaware	6.4	New York	6.9
District of Columbia	6.8	North Carolina	6.5
Florida	7.8	North Dakota	6.0
Georgia	7.6	Ohio	5.4
Hawaii	7.5	Oklahoma	8.2
Idaho	5.4	Oregon	5.3
Illinois	5.8	Pennsylvania	5.1
Indiana	5.9	Rhode Island	6.4
Iowa	5.1	South Carolina	7.4
Kansas	5.4	South Dakota	5.4
Kentucky	6.9	Tennessee	6.3
Louisiana	9.5	Texas	7.2
Maine	6.1	Utah	4.0
Maryland	5.3	Vermont	6.4
Massachusetts	4.6	Virginia	5.1
Michigan	4.9	Washington	4.9
Minnesota	3.6	West Virginia	8.7
Mississippi	9.7	Wisconsin	4.1
		Wyoming	8.5

Attachment C – Estimated Sample Size, CV, and CI Width by Title V National Performance Measure

National Performance Measure	Applicable Population	% of overall unweighted sample (2016-2017)	Approximate Prevalence (2016-2017)	Parameter	Total Subgroup Sample Size (e.g., county or AI/AN)										
					25	50	100	150	200	250	300	350	400	450	500
Developmental Screening	9-35 months	11%	31%	~ Sample Size	3	6	11	17	22	28	34	39	45	50	56
				~ CV	115%	82%	60%	48%	43%	38%	34%	32%	30%	28%	27%
				~ CI width	100%	99%	73%	59%	52%	46%	42%	39%	36%	34%	33%
Physical Activity	6-11 years	30%	28%	~ Sample Size	7	15	30	45	60	75	90	105	120	135	150
				~ CV	82%	56%	39%	32%	28%	25%	23%	21%	20%	19%	18%
				~ CI width	89%	61%	43%	35%	30%	27%	25%	23%	22%	20%	19%
	12-17 years	41%	18%	~ Sample Size	10	21	41	62	82	103	124	144	165	186	206
				~ CV	90%	62%	44%	36%	31%	28%	26%	24%	22%	21%	20%
				~ CI width	64%	44%	32%	26%	22%	20%	18%	17%	16%	15%	14%
Bullying - Perpetration	12-17 years	41%	5%	~ Sample Size	10	21	41	62	82	103	124	144	165	186	206
				~ CV	183%	126%	90%	74%	64%	57%	52%	48%	45%	42%	40%
				~ CI width	37%	25%	18%	15%	13%	11%	10%	10%	9%	8%	8%
Bullying - Victimization	12-17 years	41%	21%	~ Sample Size	10	21	41	62	82	103	124	144	165	186	206
				~ CV	82%	57%	41%	33%	29%	26%	23%	22%	20%	19%	18%
				~ CI width	68%	47%	33%	27%	24%	21%	19%	18%	17%	16%	15%
Adolescent Well-Visit	12-17 years	41%	79%	~ Sample Size	10	21	41	62	82	103	124	144	165	186	206
				~ CV	22%	15%	11%	9%	8%	7%	6%	6%	5%	5%	5%
				~ CI width	68%	47%	34%	27%	24%	21%	19%	18%	17%	16%	15%
Medical Home	CSHCN 0-17 years	23%	43%	~ Sample Size	6	11	23	34	45	57	68	79	91	102	114
				~ CV	63%	46%	32%	26%	23%	20%	19%	17%	16%	15%	14%
				~ CI width	100%	79%	54%	45%	39%	35%	32%	29%	27%	26%	24%
	Non-CSCHN 0-17 years	77%	50%	~ Sample Size	19	39	77	116	155	193	232	271	309	348	386
				~ CV	31%	22%	15%	13%	11%	10%	9%	8%	8%	7%	7%
				~ CI width	60%	42%	30%	24%	21%	19%	17%	16%	15%	14%	13%
Transition	CSHCN 12-17 years	12%	17%	~ Sample Size	3	6	12	18	23	29	35	41	47	53	58
				~ CV	173%	122%	86%	71%	62%	56%	51%	47%	44%	41%	39%
				~ CI width	100%	80%	57%	46%	41%	36%	33%	31%	29%	27%	26%
	Non-CSHCN 12-17 years	30%	14%	~ Sample Size	7	15	30	44	59	74	89	103	118	133	148
				~ CV	126%	86%	61%	50%	43%	39%	35%	33%	31%	29%	27%
				~ CI width	69%	47%	33%	27%	24%	21%	19%	18%	17%	16%	15%
Preventive Dental Visit	1-17 years	96%	80%	~ Sample Size	24	48	96	144	193	241	289	337	385	433	482
				~ CV	14%	10%	7%	6%	5%	4%	4%	4%	3%	3%	3%
				~ CI width	43%	31%	22%	18%	15%	14%	12%	12%	11%	10%	10%
Smoking - Household	0-17 years	100%	16%	~ Sample Size	25	50	100	150	200	250	300	350	400	450	500
				~ CV	63%	44%	31%	26%	22%	20%	18%	17%	16%	15%	14%
				~ CI width	38%	27%	19%	16%	13%	12%	11%	10%	10%	9%	9%
Adequate Insurance	0-17 years	100%	68%	~ Sample Size	25	50	100	150	200	250	300	350	400	450	500
				~ CV	18%	13%	9%	7%	6%	6%	5%	5%	5%	4%	4%
				~ CI width	49%	35%	24%	20%	17%	15%	14%	13%	12%	12%	11%

Abbreviations: CV, Coefficient of Variation=Standard Error/Estimate; CI, 95% Confidence Interval

Assumes an average design effect of 1.8

Suppressed: sample size<30

Unreliable: CV>30% or CI width >20% points

Reportable without flags

Attachment D – Estimated Relative Cost of Sub-State Oversampling Projects by Years Pooled and Minimum Interviews per County/County Group (3 Years, 150 Cases = 100)

Min. Interviews per County/ County Group	Years				
	1	2	3	4	5
20	10	7	6	4	3
30	18	14	11	9	8
40	25	21	17	15	13
50	33	28	24	21	19
60	41	35	31	28	25
70	49	43	38	34	31
80	57	51	45	41	38
90	65	59	53	48	44
100	73	67	60	55	51
110	81	75	68	63	58
120	89	83	76	70	66
130	97	91	84	78	73
140	106	98	92	86	80
150	114	106	**100**	94	88
160	122	114	108	102	95
170	130	122	116	109	103
180	138	130	124	117	111
190	146	138	132	125	119
200	154	146	140	133	127
210	163	154	148	141	135
220	171	162	156	149	143
230	179	170	164	157	151
240	187	179	172	165	159
250	195	187	179	173	167
260	203	195	187	181	175
270	212	203	195	189	183
280	220	211	203	197	191
290	228	219	211	205	198
300	236	227	219	213	206