Methods for Sampling Households to Identify 100,000 Births

Jill M. Montaquila[1], J. Michael Brick[1], Lester R. Curtin[2]
Jill M. Montaquila, Westat, 1650 Research Blvd., Rockville, Maryland 20850[1]
National Center for Health Statistics, Centers for Disease Control[2]

Abstract

The National Children's Study is a national household probability sample designed to identify 100,000 children at birth and follow the sampled children for 21 years. Data from the study will support examining numerous hypotheses concerning genetic and environmental effects on the health and development of children. The goals of the study present substantial challenges. For example, the need for preconception, prenatal, and postnatal data require identifying women in the early stages of pregnancy, the collection of many types of data, and the retention of the children over time. In this paper, we give an overview of the sample design for this study, and highlight the approaches used to address these challenges. We will also describe the rationale for the sampling choices made at each stage, the unique organizational structure of the NCS, and issues we expect to face during implementation.

Keywords: National Children's Study, longitudinal study

1. Overview of the Study

Authorized by the Children's Health Act of 2000 (PL 106-310), the National Children's Study (NCS) is a national household probability sample designed to identify 100,000 children at birth and follow the sampled children for 21 years. Data from the study will support the examination of numerous hypotheses concerning genetic and environmental effects on the health and development of children. The goals of the study present substantial challenges. For example, the need for preconception, prenatal and postnatal data require identifying women prior to pregnancy or in the early stages of pregnancy, the collection of many types of data, and the retention of the children in the sample over time. Data collection for the NCS is scheduled to begin with a pilot in 2008, with main study recruitment beginning in 2009.

A number of options for study and sample designs were considered for the NCS. While it is recognized that there are advantages and disadvantages to each of the candidate approaches, a national probability sample of all U.S. births was chosen as the design that best fulfills the goals of the study. It was decided that the sample should be designed with the primary goal of facilitating estimation and analysis at the national level, although it is likely that some estimates at the county level will be produced. Consideration was given to a non-probability design (a "center model") in which women would be recruited only at a small number of predesignated large health care centers); however, to allow for inference to the population as a whole, the decision was to use a probability design based on a household sampling model.

2. Challenges

The analytic needs of the NCS pose a number of challenges for the sample design. One challenge results from the need for a number of preconception and prenatal measures, which requires that women be enrolled into the study prior to conception. In the household design, this will be facilitated by enumeration of each eligible household and subsequent (immediate, if possible) enrollment of each age-eligible woman (each woman between the ages of 18 and 44) in the household. When a woman is enrolled in the study, her "pregnancy risk" will be estimated based on characteristics that are associated with the probability of pregnancy. A schedule of follow-up visits for data collection will be established, based on the woman's pregnancy risk classification. At each visit, her pregnancy risk will be reassessed, to determine whether the schedule for subsequent visits should be altered. The need for prenatal data collected at specific points in the pregnancy cycle means that once an enrolled woman becomes pregnant, it will be necessary for the study to be aware of her pregnancy as soon as possible; the contact protocol and the schedule of follow-ups is designed with this goal in mind.

A second challenge results from the need for postnatal and childhood measures. Since eligibility is determined at birth, the longitudinal nature of this study requires that children who move after birth be followed.

A third challenge is the need to collect environmental and ecological measures, or link them from other sources. A consequence of this requirement is that segments should be contiguous and as compact as possible, and should be constructed with the census block as the elemental unit.

A fourth challenge involves the competing needs of facilitating the production of national estimates of relationships as well as multilevel modeling. For the purpose of producing national estimates of prevalence rates, and linear or logistic regression coefficients, the goals would be to distribute the sample among a large number of segments and to form segments that are as internally heterogeneous as possible in order to minimize the effect of segment intraclass correlation on precision. For multilevel modeling, in contrast, the goals would be to have the sample clustered within a small number of segments that are internally homogeneous, in order to allow for estimation of neighborhood effects.

3. Sample Design

The decision to use the household recruitment model implies the sample design for the NCS to be a multistage probability sample of births in the United States, where the births are identified from a sample of households. The design includes two or three stages of sampling. The first stage of sampling is the selection of primary sampling units (PSUs), which are single counties or groups of contiguous counties. The second stage is the selection of smaller geographic areas (segments) from within the PSU. In general, the sampled segments comprise census blocks or combinations of blocks and are defined to roughly correspond to neighborhoods. In PSUs with large population, this sampling may be done in two stages to reduce the sampling workload. The third stage, which applies only to the very densely populated segments, involves the selection of groups of households from within the sampled segments. Each of these stages is described in further detail below.

3.1 PSU Selection

The selection of participants is based on a multistage probability design using an area frame. The sample size requirement of 100,000 births was determined by examining the sample size needs to meet specific outcomes generated from hypotheses the survey was to address. Based on operational and budgetary considerations, each area selected is targeted for 1,000 participants, thus requiring about 100 PSUs. Based upon preliminary research on between- and within-PSU variation, the target of 100 PSUs was deemed acceptable for geographic coverage for environmental exposures. A major unknown factor in planning the study is the participation rate (response rate) and the planning has involved considerable discussion on how to set an expected response rate that is both meaningful and attainable. To protect the study in its efforts to enroll 100,000 births, a total of 110 PSUs was selected.

A PSU is defined as a single county or a small number (but limited to no more than 4) of contiguous counties. Any county with over 120,000 expected births in the 4 year period of data collection was considered to be self-representing (SR), i.e., was brought into the sample with probability 1. A total of 12 counties met the criteria; to geographically balance the sample a 13th county was classified as SR. For most PSUs, the expected participant sample size will be set at 1,000 births over a 4-year period. Based on number of births, and to maintain a nearly self-weighting design, Los Angeles is targeted to have 4,000 enrolled births (or 4 PSUs), and Cook County, IL, (Chicago) and Harris County, TX, (Houston) are targeted for 2,000 enrolled births (or 2 PSUs) each. Thus, the sample can be considered to have 18 SR PSUs.

A minimum measure of size (MOS) of 2,000 births per PSU was established in an effort to obtain a target sample of 1,000 enrolled births. Many counties in the U.S. (primarily the non-metropolitan counties) have very small numbers of births (in many cases, fewer than 10 per year) and it was not always possible to form a PSU that met the two criteria of 2,000 births and no more than four counties combined. As a result, five of the selected PSUs do not meet the minimum MOS criterion and several others are close to the minimum MOS.

The PSUs were stratified according to a set of general design variables. Specifically, to ensure reasonable geographic coverage and to ensure urban/rural representation, 18 major strata were formed by crossing the 9 Census Divisions with the two-way classification of metropolitan and non-metropolitan as delineated on the Vital Statistics data files using 1990 Census based definitions. Within each major stratum, minor strata were formed to have roughly equal numbers of expected births (on the average, each minor stratum had about 160,000 expected births over 4 years). Minor strata were formed by considering size (births) of the PSU, percent minority births (American Indian, Asian, Hispanic, and Black), and/or percent low-birth-weight births. This stratification by percent minority births was done to ensure proportionate representation of different types of areas and of subpopulations rather than to disproportionately sample any subpopulations.

After all PSUs had been stratified, 1 PSU was selected from each minor stratum with probability proportional to

size; the measure of size used was the number of resident births from 1999-2002 (the latest 4 years available at the time of selection). The 110 PSU design includes 18 SR strata, 66 Metropolitan PSUs, and 26 non-metropolitan PSUs. The distribution of the 110 PSUs is given in Table 1.

Table 1. Number of sampled PSUs in each of the major strata

Census Div.	SR (Metro)	NSR: Metro	NSR: Non-Metro	Total
1	0	4	1	5
2	2	10	1	13
3	4	8	3	15
4	0	6	3	9
5	1	13	5	19
6	0	5	3	8
7	3	8	4	15
8	1	5	3	9
9	7	8	2	17

3.2 Segment Selection

The second stage of selection is sampling segments within the sampled PSUs. Segments were formed by combining contiguous census blocks in an effort to create units with measures of size (expected births) as close as possible to the target MOS. Within each PSU, geographic stratification of segments is useful because many of the characteristics that differentiate subpopulations (such as income, race/ethnicity, educational attainment and environmental measures), as well as environmental factors, tend to be geographically clustered. As with the stratification of PSUs, the stratification of segments is used in an effort to ensure proportionate representation of geographic, demographic and socioeconomic subpopulations. The strata used for segment selection may vary in their MOS, provided that the target MOS for segments within the strata vary proportionate to the stratum MOS.

The MOS used to form segments is the expected number of births in the segment over the 4-year enrollment period, accounting for anticipated changes in the population such as new construction or population declines due to migration. The MOS is computed at the census block level and aggregated to the segment level. Initially, the plan was to obtain the block MOS by applying estimated birth rates to block-level population projections. However, when it was determined that overall block-level birth counts could be obtained for this purpose, these block-level birth counts were used, and then adjusted for births that could not be geocoded to any block and for expected changes (growth/decline) in order to arrive at the block MOS. A challenge of this approach is obtaining data on resident births occurring outside the jurisdiction (e.g., births to Queens residents that occur in New Jersey); when such data cannot be obtained, adjustment factors (based on aggregate data, e.g., ZIP code-level rates) will be applied to the birth estimates to account for out-of-area births.

Within each stratum in a PSU (stratification is used in most PSUs), exactly one segment will be selected with equal probability (as discussed below). The numbers of strata used for segment selection vary from PSU to PSU, with a general guideline of between 10 and 20 (although for some very rural PSUs, slightly fewer than 10 strata may be used). Within each PSU, the exact number of segment strata and the stratification variables to be used are arrived at with input from the Study Center (SC); the SC provides input on factors such as operational concerns, important sub-PSU regions and other potential stratifiers.

Section 2 addressed several challenges that imposed constraints on segment formation and selection. The requirements of an equal probability sample of births and an equal number of births in each PSU have implications for the sampling of segments. Operational and analytic needs dictated that within sampled segments, every birth should be eligible. Thus, the segment was intended to be the final stage of selection, and the segments were to be constructed to yield the target number of births in the PSU (1,000 over four years) and were to be sampled to yield an equal probability sample of segments.

Initially, the segment selection was designed to be a single-stage selection process. However, following the creation of the segment frame, the SC is asked to review each segment in the frame to determine whether any changes should be made (provided such changes are feasible) so that the segments adhere as closely as possible to “neighborhood” boundaries. In urban PSUs with hundreds of segments in the sampling frame, such a comprehensive review would be very labor-intensive and time-consuming. Thus, in order to use resources more efficiently, the sampling protocol for large PSUs (typically those expected to have over 500 segments in total) includes an intermediate step. In these large PSUs, geographic units (e.g., block groups or contiguous blocks that are considerably larger than an individual segment) will be formed within strata, and one geographic unit sampled with probability of selection proportionate to the size (expected number of births) of the geographic unit. Segment sampling then proceeds as for the PSUs, but all of the work is done only within the sampled geographic units. Segments within the sampled

geographic unit will be approximately equal in size and one segment will be randomly selected within each geographic unit.

Let B denote the total number of births in the PSU, and within the PSU, let B^h denote the number of births in stratum h, $h = 1, 2,\ldots, H$, so that

$$B = \sum_{h=1}^{H} B_h .$$

Suppose that within stratum h, geographic units are formed, and the number of births in GU hi is given by B_{hi}.

In each stratum, exactly one GU is sampled with probability proportionate to the number of births. That is, the probability of selection of GU hi (conditional on the PSU sample) is

$$\pi_{hi} = \frac{B_{hi}}{B_h}.$$

Let N_{hi} denote the number of segments to be formed within GU hi. Within sampled GU hi, segments are formed to be as equal in size (number of births) as possible, and exactly one segment is sampled with equal probability; i.e., conditional on the selection of GU hi, the probability of selection of segment hij is

$$\pi_{hij|hi} = \frac{1}{N_{hi}}.$$

Therefore, the overall probability of selection of segment hij within the PSU (conditional on the PSU having been selected) is

$$\pi_{hij} = \pi_{hi}\pi_{hij|hi} = \frac{B_{hi}}{B_h}\frac{1}{N_{hi}}$$

and this is a constant if and only if

$$N_{hi} = k\frac{B_{hi}}{B_h}, \qquad (1)$$

where k is a constant.

Expression (1) holds if the segments are formed so that the MOS is approximately proportionate to the number of births in the stratum, or in other words, $B_{hij} \approx \frac{B_{hi}}{N_{hi}}$ for each segment hij in GU hi.

With a target of B^* births over 5 years, letting $hij \in S$ denote the sampled segments, then

$$\begin{aligned} B^* &= \sum_{hij\in S} B_{hij} \\ &\approx \sum_{hij\in S} \frac{B_{hi}}{N_{hi}} \\ &= \sum_{hij\in S} \frac{B_h}{k} \\ &= \frac{B}{k} \end{aligned}$$

So $k = \frac{B}{B^*}$, the reciprocal of the sampling fraction within the PSU.

The above shows that (1) the condition that $N_{hi} = k\frac{B_{hi}}{B_h}$ is sufficient for obtaining an epsem sample of segments, and (2) forming the segments to be as equal in size as possible, with size $B_{hij} \approx \frac{B_{hi}}{N_{hi}}$ is sufficient for obtaining the target number of births. (Note that if the B_{hij} deviate much from $\frac{B_{hi}}{N_{hi}}$, then the target B^* might not be met, or might be exceeded.)

Although the condition $B_{hij} \approx \frac{B_{hi}}{N_{hi}}$ is sufficient for obtaining the target number of births, it has not been shown to be a necessary condition. Even if the segments vary in size around $\frac{B_{hi}}{N_{hi}}$, as long as their average size is $\frac{B_{hi}}{N_{hi}}$, then in expected value, $E\left\{\sum_{hij\in S} B_{hij}\right\}$ will be B^*. However, a fixed sample size of 1,000 births in each sampled PSU is required to the extent possible and having only an expected value equal to this value is not enough. Essentially, the variance of the total number of births in the sampled segments, $V\left\{\sum_{hij\in S} B_{hij}\right\}$, might be larger than desired.

It should be noted that the above discussion assumes that the number of births in a segment B_{hij} is fixed and known. This does not take into account the effects of changes in the segment such as new construction, year-

to-year variations in the numbers of births (even in a stable area), and within-PSU variations in response rates.

3.2.1 Large blocks (chunking)

In most selected segments, household screening is attempted in all households or dwelling units (DUs) in the segment. The Study Center is responsible for conducting the screening, which will include the simultaneous listing and screening of all DUs in the segment. The Coordinating Center will provide training materials for this listing and screening activity, and the Study Center is responsible for conducting the activity within this framework.

An exception to the complete listing and screening of all DUs in a segment is for a very large segment, which cannot be subdivided during segment formation (or is found to be much larger than expected at the time of sampling). In such segments, DUs are subsampled. If one of these large segments is selected, the segment is divided into "chunks" and then a chunk is randomly sampled for listing and enrollment. For example, suppose a segment is known at the time of sampling to be twice as large as the target segment size. That segment will be assigned twice the probability of selection as other segments in the stratum. Suppose it is selected and the Study Center determines that it consists of two very large apartment buildings that are approximately equal in terms of numbers of DUs. In that case, the Study Center, in consultation with the Coordinating Center, will assign each apartment building to be a chunk, and one of the two will be randomly selected to be retained in the sample. Other approaches for chunking (depending on the situation) include using floors of apartment buildings or block faces as chunks.

3.2.2 Listing and enumeration

For the National Children's Study, once the sample of segments has been selected, lists of all residential addresses in each sampled segment must be obtained. Traditionally, in area probability samples, this has been done through a process known as "listing." Trained "listers" are sent to canvass the segment, identify segment boundaries and compile a hard-copy list of residential addresses as they move in a systematic manner through the segment. Typically, the addresses are sampled in the main office, the sampled addresses are keyed to create an electronic database, and the sampled addresses are sent back for field work.

In recent years, there has been a growing interest among survey methodologists in using U.S. Postal Service delivery files in place of listing. (See O'Muircheartaigh, Eckman, and Weiss 2002; Iannacchione, Staab, and Redden 2003; Dohrmann, Han, and Mohadjer 2006.) Evaluations of the Postal Service residential delivery files have examined two aspects of the quality of the lists for sampling purposes: coverage and geocoding errors. To examine coverage, the general approach is to compare counts of the numbers of residential units obtained from the residential delivery files (hereafter referred to as "address lists") to counts of housing units obtained from listers through the traditional listing approach and/or from the decennial census. Although this is not a perfect comparison (due to the possibilities of duplicates on the address lists and geocoding errors, as well as changes in the segment due to new construction and demolition, for example), it is useful for providing a general idea of the coverage of the address lists. Such comparisons have revealed that the address lists generally provide good coverage in urban areas (where generally, the address lists contain above 90% of the expected units based on listing or the decennial census) but substantially poorer coverage in rural areas.

A second issue with using the address lists to compile a list of all DUs in an area is geocoding error. Errors and differences in resolution in the GIS systems used to define segment boundaries and to geocode the addresses may result in improper shifts in segment boundaries, geocoding of addresses to the wrong side of a street, and the need to interpolate between known addresses. Additionally, depending on how updated the GIS database is, a proportion of addresses will not be able to be geocoded due to new streets or changes in street addresses. Some of the problems associated with the geocoding errors were expected, but the magnitude of the problem is relatively large at this low level of geographic detail. The geocoding errors are problematic even in urban areas where the coverage of the address lists is high (O'Muircheartaigh et al. 2006). For the NCS the concern associated with these geocoding problems are even more serious because a goal of the NCS is to collect data within neighborhoods for analytic reasons. This cannot be accomplished if the delivery file listings are not physically adjacent to each other.

An alternative approach that is being developed for the NCS is a "list and screen" approach. Under this approach, listing is done by the field staff. As the field staff initially moves through the segment, she lists each residential unit and attempts to complete a Household Enumeration for that unit. If no one is at home in the residential unit, the interviewer lists the unit and moves on to the next unit. This list and screen approach eliminates the need for a separate, costly listing operation. Additionally, this approach gives Study Centers the opportunity to get to know their segments.

The list and screen operation is automated, so that cases can be created and Household Screeners completed as dwelling units are identified. The address lists are used as a basic framework the interviewers can use to move through the segment and identify dwelling units.

4. Discussion

As a large-scale national longitudinal study, there are a number of challenges in the design, implementation and analysis of the NCS. This paper deals with one key design challenge—that of designing and selecting a nationally representative sample of 100,000 births in such a way that it will have analytic utility for both anticipated and unanticipated analyses.

Seven of the PSUs were chosen to be Vanguard sites—sites in which a pilot study will be conducted beginning one year prior to the start of the main study. As a result, nearly all procedures are being implemented in these Vanguard sites first. These seven PSUs were the first in which within-PSU selection was done, and the procedures for within-PSU selection have evolved as a result of the experiences and lessons learned in these seven Vanguard sites.

A few important developments occurred in the approach for forming segments. Initially, segments were formed using a manual process (i.e., manually combining blocks until the target measure of size was reached), and the idea was that the Study Center would examine each segment. However, while undertaking this effort for the Vanguard sites, it was determined that this was, at best, inefficient and, at worst, not feasible, in larger PSUs. Two important changes to the segment formation and review process resulted. First, an algorithm was developed to automate the segment formation process. Second, a two-stage selection approach was implemented in the larger PSUs to reduce the amount of review required.

Another development was a refinement of the measure of size used for segment formation. Initial plans were to estimate the number of births in each census block by applying estimated birth rates to population projections. However, when experience with this approach revealed substantial inaccuracies in these estimates and it was determined that block-level birth counts could be obtained for use in forming segments, the approach was changed to using the block-level birth counts as the basis for the measure of size; the modeled birth estimates are still computed for comparative purposes, to assess the quality of the birth data.

A third development was in the approach used for listing. The development and use of an automated address list-assisted listing application represents advancement over the traditional hard copy approach to listing. Because all DUs in a segment are included in the NCS sample (with a few exceptions), this approach also permits listing and enumeration to be done concurrently.

Thus, out of both opportunity and necessity, the sample design and selection for NCS have evolved in important ways. We expect many further developments because we are still very early in the overall cycle of the survey. For example, the survey anticipates a four-year enrollment period for births and changes may occur within the segments during that four-year period (e.g., moves, new constructions, demolition, etc.) that will require new procedures be developed to deal with these changes. One such procedure that is planned is called segment vigilance. The SC will be required to monitor and identify changes within the segment throughout the enrollment period. Additionally, households will be periodically re-enumerated to determine whether there have been any changes in household membership. Other procedures, such as highly effective methods to retain the sampled children, will also be required. As the study proceeds, a variety of sampling and operational advances will be necessary.

References

Dohrmann, S., Han, D., and Mohadjer, L. (2006). Residential address lists versus traditional listing: Enumerating households and group quarters. *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 2959-2964).

Iannacchione, V.G., Staab, J.M., and Redden, D.T. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. *Public Opinion Quarterly*, 67, 202-210.

O'Muircheartaigh, C., Eckman, S., and Weiss, C. (2002). Traditional and enhanced field listing for probability sampling. *Proceedings of the Social Statistics Section of the American Statistical Association* (pp. 2563-2567).

O'Muircheartaigh, C., English, N., Eckman, S., Upchurch, H., Garcia, E., and Lepkowski, J. (2006). Validating a sampling revolution: Benchmarking address lists against traditional listing. *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 4189-4196).