Published in final edited form as:
Stat Med. 2010 June 15; 29(13): 1368–1376. doi:10.1002/sim.3891.

Statistical and Practical Issues in the Design of a National Probability Sample of Births for the Vanguard Study of the National Children's Study

Jill M. Montaquila[1], **J. Michael Brick**[1], and **Lester R. Curtin**[2]

[1]Westat, 1600 Research Blvd., Rockville, Maryland 20850

[2]National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Rd., Hyattsville, MD 20782

Summary

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 used in a pilot study called the Vanguard 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

Multistage sample; area probability sample; primary sampling unit; segment; listing

1. Introduction

The National Children's Study (NCS) is the largest and most comprehensive longitudinal study of children ever undertaken. The study's goal is to sample and obtain data for 100,000 births in the United States, and then follow the sampled children for 21 years. The study will support the examination of numerous hypotheses concerning genetic and environmental effects on the health and development of children. More specifically, the NCS will be used to assess how health outcomes are related to environmental exposures. The “environmental” exposures are broadly defined and include natural and man-made environmental factors, biological and chemical factors, physical surroundings, social factors, genetics, culture and geography. Ellenberg and Hirschfeld [1] (this issue) and Michael and O'Muircheartaigh [2] describe in much greater detail the genesis of the NCS, the plans for its implementation, and its goals. Additional information on the NCS is available at the study's web site (www.nationalchildrensstudy.gov).

A study of this size and importance presents many substantive and statistical challenges. This paper examines the statistical issues associated with sampling and the approaches developed to deal with these. The sample design for the NCS is heavily influenced by two

study design decisions. The first is that the goals of the study are analytic, meaning they focus on relationships or associations rather than a description of the current population. The second involves the method used to obtain representativeness of the sample. Two different structures were considered for the NCS – a household probability sample approach and a medical center-based approach. Other structures such as an office-based probability sample were also considered, but did not support all the key requirements such as collecting preconception data.[1] The household structure, resulting in a national probability sample of all U.S. births, was adopted for the NCS Vanguard Study, a large-scale pilot study. In the next section, we review the process that resulted in choosing the household model for the NCS, and how this affected the sample design. The NCS design described here is based on the household model; however, evaluation of the Vanguard Study that is currently underway as well as other pilot studies that are being contemplated may lead to changes in the design for the NCS Main Study.

While the choice of the household probability sample fundamentally required the sample design for NCS be a multistage probability sample, it still left many unresolved design elements regarding the sampling stages. The first stage of sampling is the selection of primary sampling units (PSUs) or large geographic areas of the country. The third section describes the how PSUs were formed and selected for the NCS. The second sampling stage is the selection of smaller geographic areas within the PSUs, called “segments.” Sampling segments is described in the fourth section.

In most PSUs, all of the households in the sampled segments will be canvassed and monitored to identify women who become pregnant to enroll them into the study. In a few very densely populated segments where the number of expected births is greater than desired sample size, a third stage of sampling will be employed. The within-segment procedures for listing and enumerating households and the subsampling of households in the densely populated segments are discussed in the fifth section.

The last section summarizes some of the statistical challenges that must be addressed as the data collection stage, beginning with a the Vanguard Study in 2009, approaches.

2. General Study Design Structure

As noted above, the NCS is designed to provide data appropriate for analyzing relationships between health outcomes and exposures and genetic factors. The sampling unit is the birth, and the target population is all births occurring during the birth enrollment period to women who are U.S. residents at the time of the birth. A set of key hypotheses was formulated to make the research objectives more specific and to help guide the design of the NCS. In addition to these hypotheses, the study design also must consider a myriad of other research questions that will be addressed using the data. These include current research questions that are of substantive interests, as well as research questions that have not been developed (e.g., those based on associations that will emerge from future research).

The study hypotheses had a direct influence on the design of the study. Some hypotheses require preconception and prenatal measures, which means the study must enroll at least some women into the study prior to conception and most women as soon after conception as possible. Most of the hypotheses demand postnatal and childhood measures, so a prospective, longitudinal study of births is necessary. Other hypotheses call for

[1]The Household Model involves “screening a large sample of households to identify and locate pregnant women”; the Office Model is also a probability sample but involves “selecting a large sample of physicians and medical offices and securing their cooperation to recruit pregnant women seen in their practices”; and the Center Model involves “selecting and funding a small number of large health care centers to recruit volunteer samples of pregnant women” (Westat [3]).

environmental and ecological measurements, implying the data collection areas where the women reside should be as contiguous and as compact as possible to reduce costs. The full set of research hypotheses provided the rationale for a long-term, prospective study of approximately 100,000 children. As described below, all of these requirements are dealt with in the sample design.

The problem of determining the necessary sample size for the NCS was extremely complex. For each hypothesis considered, the required sample size depends on the type of statistic required, the level of confidence needed for point estimates, and the statistical power for analytic comparisons. Because the NCS is an environmental study, the proportion exposed to risk must also be considered. For hypotheses to be investigated over time, the potential for loss to follow-up has to be incorporated; that is, the sample size for a longitudinal objective must be inflated to a larger number for the initial birth cohort to allow for non-retention over time. To ensure sufficient sample sizes for all hypotheses, the required sample size was set for the most demanding analytic hypothesis.

As part of the planning process for the NCS, a range of detectable odds ratios with varying prevalence, varying levels of precision, and varying exposure levels were examined. This extensive investigation determined that a sample size of 2,000 births would be required for the most demanding hypotheses. The expected prevalence associated with these hypotheses were as low as 2 percent; thus, a total of 100,000 births is required to obtain the 2,000 births with expected characteristics of the outcome variable conditional on the exposure level. Once the 100,000 totatl sample size was established, potential hypotheses can be examined relative to the sample size to determine whether these hypotheses should be included in the study. As such, it was the full set of multiple research hypotheses that provided the rationale for a longer-term prospective study of 100,000 children. As described below, all of these requirements are dealt with in the sample design.

While different analytic methods will be used to address the hypotheses and other research questions that will be supported by the data, we expect generalized regression modeling (e.g., see McCullagh and Nelder [4]) will be the primary approach used with NCS data. Linear regression models for continuous outcome variables and logistic regression models for dichotomous outcome variables are two well-known regression models in this class that will be used often with NCS data. Other nonlinear models such as Poisson regression models for count data, polytomous logistic regression models for nominal outcome variables, proportional hazard models for ordered categorical outcomes, and survival or time-to-event models may also be appropriate for some analyses. Structural equation models and latent variable models are other analytic techniques that will be used to examine important research questions, especially those related to causality. Multilevel or hierarchical linear models may also be needed to address hypotheses related to community or neighborhood effects. All of these modeling efforts will be aided by collecting key confounding and mediating variables in the study.

Given the breadth of the research questions, disproportional sampling of some groups to improve the precision of estimates for those groups must be weighed against detrimental effects this would have on the research for undersampled groups. The NCS has the advantage that proportional allocation with 100,000 sampled children provides adequate sample sizes for most of the groups that are traditionally oversampled in national studies.

Another advantage of proportional allocation is that the sampling weights are approximately equal for the units in the sample. These weights are used to estimate population parameters in design-based analysis, where design-based analysis means statistical inferences are based on the distribution induced by the sample design (and, as such, incorporate aspects of the

sample design such as probabilities of selection, sample unit stratification, sample unit clustering, and adjustments for differential nonresponse and population coverage). An alternative approach to inference is the model-based approach, where the analyst assumes a statistical model and draws inferences based on that model. The sampling weights are not used in model-based analysis. When the weights are approximately equal, design-based and model-based parameter estimates are generally very similar, and most of the differences between these approaches are related to the precision of the estimates rather than the level of the estimates. Design-based estimates generally have larger variances due to the clustering of the sample, but with many analytic methods, such as linear and logistic regression analysis, even these differences are not as pronounced as they are with descriptive statistics such as means and totals. (See Korn and Graubard [5] for a more complete discussion of these approaches to inference and the effect on the estimates.)

As noted above, both household probability samples and center-based sampling approaches have been used in longitudinal surveys of health in the U.S. and internationally, and both have advantages and disadvantages. The rationale for choosing a structure depends on a number of factors that are often specific to the goals and resources of the study.

The household probability sample approach was chosen for the NCS, and is being used for the Vanguard Study, after much deliberation because it seemed most likely to support key objectives of the study. The primary advantage of the household probability structure is that it has greater potential to sample women early – prior to conception or soon after conception – thus providing preconception and prenatal data needed for some key hypotheses. Another advantage of the household structure relative to the center model is the more complete coverage of women, particularly those women who are outside the traditional medical system. The children of these women could have health outcomes that differ from children borne by mothers within the medical system, and the relationships of interest may vary as a result of the differences. The household approach may be less susceptible to these selection biases. A related advantage is that probability sampling may be more robust for making population-level inferences, as well as for making individual-level estimates (see Curtin and Feinleib [6]). Additionally, the multi-stage area probability sample design facilitates linkage of the sampled address to external data sources; e.g., merging census block-level data (environmental, socio-economic, crime statistics, etc.) to the NCS data for analytic purposes. Although this linkage to external data sources is also possible in the center-based and office models, the household model allows for much greater control over the degree of clustering of the sample within areas such as the census block; this control of clustering is beneficial in the estimation of neighborhood effects.

While there are good reasons for choosing the household structure, the choice was far from simple. There are distinct advantages for the center-based structure. A blue-ribbon panel consisting of national experts in sampling, study design, and epidemiology were asked to consider this question and to make a recommendation on the design. The final report of the panel (see www.nationalchildrensstudy.gov/research/workshops/Pages/samplingdesign032004.aspx, last accessed January 18, 2010) contains a complete accounting of the issues. Their recommendation was accepted and led to the adoption of the household sampling approach that is discussed in detail below.

The NCS organizational structure consists of a Program Office, a Coordinating Center, and a set of Study Centers. The Program Office oversees the operations of the study. The Coordinating Center is responsible for information management, sampling, data collection and analysis, and quality control. The Study Centers are the organizations responsible for recruitment of participants and data collection within the primary sampling units (PSUs).

The Study Centers collaborate with community representatives to tailor their outreach and data collection approaches, within the guidelines of the study protocol, to the needs, characteristics, and interests of their communities. Additionally, the Study Centers provide the community perspective that is critical to the segment formation process described in section 4.

3. Primary Sampling Unit (PSU) Selection

For the NCS, a PSU is defined as a single county or a small number (but limited to no more than 4) of geographically contiguous counties. In the preliminary evaluation of alternative sample designs the number of PSUs to be sampled ranged from 30 to 800. The largest figure was based on the need for geographic coverage for a rare environmental exposure. The smallest figure was based on minimizing operational costs. Typically the optimum number of PSUs is a function of the relative costs of the first and second stages of selection along with the proportion of sampling error attributed to the between-PSU component of sampling variation. However, the optimum number differs greatly for different study objectives, which makes it difficult to base the optimum number on sampling principles alone. For the NCS, an important consideration was the workload per PSU. Cost modeling indicated sharp increases when the number of PSUs exceeded 100. Similarly, issues of data quality, staff effectiveness, and cost efficiency arise when the workload decreases much below 1,000 persons per PSU. The design decision was to set the minimum number of births per PSU to 1,000 births per PSU, which led to the initial target of 100 PSUs. Given the trade-offs between costs and coverage, the initial target of 100 sampled PSUs was deemed acceptable for costs, geographic coverage for environmental exposures, and expected sample design effects due to clustering. Also, based on operational and budgetary considerations, each PSU selected (with the exception of very small PSUs) is to have the same number of participants (births enrolled in the Study), regardless of population size. However, the six smallest PSUs could support a sample size of only 600 participants each.

Because the enrollment period for recruiting births into the study will be four years in a given PSU, the sampling frame for the first stage of selection was based on Vital Statistics Data for 1999-2002 (the latest four years available at the time the PSUs were selected). Given about 16 million births will occur in the United States over the four years of sample enrollment, a stratum size of approximately 160,000 births was set. Any county with over 120,000 expected births over the four year period (i.e., over 75 percent of the target stratum size) was considered to be self-representing (SR), i.e., was brought into the sample with probability 1. Each SR county is so large in terms of expected births that it should always be selected, and so it is simpler to set these aside before drawing a sample of the smaller counties. A total of 12 counties met the criterion; to geographically balance the sample, a 13th county was classified as SR. Based on the number of resident births, and to maintain an approximately self-weighting (or equal probability) design, Los Angeles is targeted to have 4,000 enrolled births[2] (or 4 PSUs), and Cook County (Chicago), IL, and Harris County (Houston), TX, are targeted for 2,000 enrolled births (or 2 PSUs) each. Thus, the NCS sample can be considered to have 18 SR PSUs located in 13 counties.

The remaining 3,128 counties in the United States were grouped into 100 approximately equal-sized strata, with the goal of selecting one PSU per stratum. For 1999-2002 there were a total of 16,056,890 births; thus the expected stratum size was set at 160,569 births. To ensure representativeness of the sample with respect to geography and urban/rural characteristics, 18 major strata were formed using the cross-classification of the 9 Census Divisions and the 1990 metropolitan/non-metropolitan classification (as shown in the “Non-

[2]An *enrolled birth* is an NCS-eligible birth for which the birth mother is enrolled in the NCS at the time of the birth.

self-representing" columns of table 1); at the time of the selection the Census 2000 metropolitan classification was not yet available. Each county was classified into one of the major strata.

The non-self-representing (NSR) counties within each of the 18 major strata were then stratified according to a set of general design variables. Specifically, within each major stratum, minor strata were formed to have roughly equal numbers of expected births. The selection of stratification variables is based on the NCS being a study with multiple objectives. Rather than design for a single objective or a single environmental exposure, the sample is designed to meet the multiple objectives by (1) having a sufficient number of PSUs and (2) ensuring the sample's representativeness by geography, urbanicity, and demographic characteristics. As such, a total of 92 minor strata were formed by considering size (number of resident births) of the PSU, percent minority births (American Indian, Asian, Hispanic, Black), and/or percent low-birth-weight births. Note that, for the NCS, this stratification by percent minority births was done to ensure proportionate representation of the different subpopulations; it was not done to "over-sample" the subpopulations. On the average, each minor stratum had about 160,000 expected births over 4 years; but due to variations in major stratum sizes, there is some variability in the final stratum (minor stratum) sizes.

From each stratum, a single county was selected with probability proportional to size (number of resident births) using SAS PROC SURVEYSELECT. If the county selected did not meet the minimum sample size requirements (discussed below), adjacent counties in the same minor stratum were combined to form the single PSU for that stratum. The 101 selected PSU design was published by the National Institutes of Health (NIH) as part of the contract solicitation process for the Vanguard Study sites[3]. The 8 Vanguard Study sites were selected at random, subject to the constraint of having 2 sites in each of the four Census Regions[4], and having 2 large metropolitan areas, 4 medium size metropolitan areas, and 2 non-metropolitan areas. Seven of the eight randomly selected areas were awarded.

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. For the initial 101 PSU design, a PSU minimum size measure was established as the number required to obtain a target sample of 1,000 enrolled births (or 250 enrolled births per year). Because of the concern for participation rates, after the initial first design had been published, the minimum size was increased to 500 births per year, or 2,000 births per four years, in each PSU. Many counties in the U.S. (primarily the non-metropolitan counties) have very small numbers of resident births and it was not always possible to form a PSU that met the two criteria for number of births and no more than four counties combined. In the initial first stage sample, five of the selected PSUs did not meet the minimum size criterion and several others had birth counts close to the minimum. An additional concern is that for a design with 101 PSUs, it is not practical to attempt to control for fine levels of demographic and geographic detail. For example, one issue was the expected number of sample births to mothers who were low income, rural, black, and in the South.

To address the issue of increased minimum size and to accommodate the issues of more detailed demographic and geographic subdomains, an additional PSU was selected (again a

[3]The Vanguard Study sites are the study locations (PSUs) chosen as Vanguard Pilot sites. "Teams from the Vanguard Centers will be the first to work with their communities to recruit participants, collect and process data, and pilot new research methods for incorporation into the full Study" (www.nationalchildrensstudy.gov/studylocations/pages/overview.aspx, accessed January 18, 2010).
[4]For a map of the States included in each Census Region, see www.census.gov/geo/www/us_regdiv.pdf, accessed January 18, 2010.

random selection with probability proportional to size) from each stratum where there was a specific issue. This augmentation was done by first dividing each such stratum into two substrata, and randomly sampling from the substratum that did not contain an originally sampled PSU. As such, 9 additional PSUs were selected, resulting in the final sample of 110 PSUs.

Because most of the additional PSUs were in non-metropolitan strata, the 110 PSU design provides a slight over-sample for non-metropolitan areas. The target sample size of 1,000 enrolled births per PSU was maintained for all but a few small PSUs, as this also serves as a protection against lower than expected response rates. The final first stage design for the 110 PSU design includes 18 SR strata, 66 NSR Metropolitan PSUs and 26 NSR non-metropolitan PSUs. The distribution of the 110 PSUs is given in Table 1. A total of 43 States have at least one PSU in the sample (see http://www.nationalchildrensstudy.gov/studylocations/Pages/default.aspx, last accessed January 18, 2010).

4. Segment Selection

The second stage of selection is the selection of area segments within the sampled PSUs. A consequence of the requirement to collect environmental and ecological measures (or link them in from other sources) is that segments should be contiguous and as compact (i.e., tightly clustered) as possible. Additionally, it was decided that they should be constructed with the census block as the elemental unit to facilitate linking to other data sources, such as environmental databases. Thus, for the NCS, segments are formed by combining contiguous census blocks in an effort to create units with measures of size (for segments, the size measure is the expected number of births) as close as possible to the target segment size (discussed in section 4.1).

In this section, we describe the process of constructing and sampling segments in NCS PSUs. Preparing for segment selection is covered in section 4.1. Sections 4.2 and 4.3 describe the single-stage and two-stage selection approaches, respectively.

4.1 Preparing for Segment Selection

In preparing for segment selection for NCS, the key considerations are stratification, the size measure, and finalizing the segment sampling frame and selecting the sample. We will address each of these considerations in turn.

Segment stratification—Within each PSU, geographic stratification of segments is used frequently because many of the characteristics that differentiate subpopulations (such as income, race/ethnicity, and educational attainment), 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.

Within each stratum in a PSU (stratification is used in most PSUs), exactly one segment is 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; the Study Center provides input on factors such as operational concerns, important sub-PSU regions and other potential stratifiers.

Size measure—The measure of size (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 (i.e., birth records from the Vital Statistics System, geocoded to the block level) 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 directly, adjustments are made to the birth counts (based on aggregate data, e.g., ZIP code-level rates) in order to account for out-of-area births in the computation of the MOS.

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 dictate that within sampled segments, every birth should be eligible. Sampling all births in a segment facilitates estimation of neighborhood effects in multi-level models, provides an opportunity to engage entire communities (rather than subsamples within communities), and avoids data collection costs associated with listing households that are not in sample. Thus, the segment is intended to be the final stage of selection, and the segments are to be constructed to yield the target number of enrolled births in the PSU (1,000 over four years) and are to be sampled to yield an equal probability sample of births.

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, so that the number of segments per stratum is constant across strata within a given PSU. That is, for a given PSU, a constant number of segments (say, $\bar{N}$) is formed within each stratum. If one stratum has a MOS that is 25 percent larger than that of another stratum, then the $\bar{N}$ segments in the former stratum will be 25 percent larger, on average, than the $\bar{N}$ segments in the latter stratum. In each case, exactly one of the $\bar{N}$ segments will be randomly sampled, with each segment having an equal probability of selection.

Finalizing the segment sampling frame and selecting the sample—Initially, segment selection was designed to be a single-stage selection process, but for operational reasons this was modified for larger PSUs. Following the creation of the segment frame, the Study Center 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 (as identified by local community boards and authorities). Any such changes are made to the segment sampling frame prior to the selection of a sample of segments, so as to avoid changes after sampling that would alter the randomization distribution.

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 stage of sampling. In these large PSUs, geographic units (e.g., block groups or contiguous blocks that are considerably larger than an individual segment, abbreviated *GU*) are formed within strata, and one GU is sampled with probability of selection proportionate to the size (expected number of births) of the GU. Segment formation and sampling then proceed as for the PSUs, but all of the work is done

only within the sampled GUs. Segments within the sampled GU are constructed to be approximately equal in size and one segment is randomly selected within each sampled GU.

In the following sections, we demonstrate (first for single-stage segment selection, then for two-stage segment selection) that the approach described above results in an approximately equal probability sample of births and is designed to yield the target of 1,000 enrolled births in each sampled PSU.

4.2 Single-Stage Selection

As mentioned in the previous section, in most PSUs, the sampling of segments involves a single stage of selection. Here, we describe the statistical considerations in selecting that single-stage sample.

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

Let B_{hj} denote the MOS of segment j in stratum h. The number of segments formed is constant across strata, say $\bar{N}$, but the target MOS of the segments vary across strata, proportionate to the stratum MOS B_h. The target MOS is set such that if the strata were all equal in size, it would be equal to B^* / H, where B^* is the number of births needed in order to result in a total of 1,000 enrolled births in the PSU (after accounting for losses such as nonresponse, attrition, and mobility). Within each stratum, exactly one segment is sampled (using equal probability selection), with the intention that all births occurring in the sampled segment during the enrollment period will be eligible. (An exception that applies to some dense urban areas is discussed in section 5.)

Under this approach, the conditional probability of selection of segment hj in sampled PSU h is equal to $\bar{N}^{-1}$, a constant. The expected number of births in the sample is

$$E\left\{\sum_{hj\in S} B_{hj}\right\}=\sum_{h} E\left\{\sum_{j\in S} B_{hj}\right\}=\sum_{h}\frac{B_h}{\bar{B}}\frac{B^*}{H}=B^*\left[\frac{1}{\bar{B}}\frac{\sum_h B_h}{H}\right]=B^*. \quad (1)$$

4.3 Two-Stage Selection

Section 4.2 discussed the statistical considerations for selecting a single-stage sample of segments. As discussed in section 4.1, in some densely populated urban areas, segments are selected in two stages. Here, we consider aspects of the two stages of selection, particularly, sufficient conditions for obtaining an equal probability sample of segments designed to yield the target number of births in a PSU.

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 being sampled) 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

$$N_{hi}=k\frac{B_{hi}}{B_h}, \tag{2}$$

where *k* is a constant.

Expression [2] 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 the 4-year enrollment period, letting $hij\in S$ denote the sampled segments, then

$$\begin{aligned} B^* &= \sum_{hij\in S} B_{hij} \\ &\approx \sum_{hij\in S}\frac{B_{hi}}{N_{hi}} \\ &= \sum_{hij\in S}\frac{B_h}{k} \\ &= \frac{B}{k} \end{aligned} \tag{3}$$

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

The above shows that (a) the condition that $N_{hi}=k\frac{B_{hi}}{B_h}$ is sufficient for obtaining an equal-probability sample of segments, and (b) 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 enrolled births in each sampled PSU is required to

the extent possible and having only an expected value equal to 1,000 births is not the goal. 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. Thus, in order to ensure as close to 1,000 enrolled births as possible, it is necessary for the segments to be as close in size to $\frac{B_{hi}}{N_{hi}}$ as possible.

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 unanticipated 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.

5. Subsampling, Listing, and Enumeration of Dwelling Units

In most selected segments, household screening is attempted in all households or dwelling units (DUs) in the segment. 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, but the segment could not be split into two roughly equal-sized areas while still using Census blocks as the basic building block. That segment will be assigned twice the probability of selection as other segments in the stratum. If that segment is selected, a “chunking” operation will be done in the field to subdivide the segment into two roughly equal-sized areas (where size is based on the number of dwelling units), and one of the two chunks will be randomly selected to be retained in the sample.

For the NCS, once the sample of segments has been selected, lists of all residential addresses in each sampled segment must be obtained. Traditionally, this has been done in area probability samples 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 (USPS) residential delivery files in place of listing. (See O'Muircheartaigh, Eckman, and Weiss [7]; Iannacchione, Staab, and Redden [8]; Dohrmann, Han, and Mohadjer [9].) Evaluations of the USPS residential delivery files have examined two aspects of the quality of the lists for sampling purposes: coverage and geocoding errors. Such evaluations 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 the placements of boundaries and features 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. In light of the concerns with undercoverage and geocoding error, using USPS-based address lists in place of listing does not, at present, appear to be a viable approach for the NCS. A key issue for the NCS is that because the segments are designed to yield the target numbers of births, undercoverage and geocoding

errors could have substantial adverse effects on sample yield. Although the USPS-based address lists are not being used in place of listing, they are currently being used in NCS to evaluate the listings and make corrections to the listed addresses.

6. 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 sample PSUs were chosen to be Vanguard Study sites—sites in which a pilot study is being conducted prior to the start of the Main sSudy. As a result, nearly all procedures are being implemented in these Vanguard Study 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 Study 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 (Johnson, Montaquila, and Heller [11]). Second, a two-stage selection approach was implemented in the larger PSUs to reduce the amount of review required. (See Section 4.)

Another development was a refinement of the MOS 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 and to assess the quality of the birth data.

A third development was in the approach used for listing. The use of the USPS-based address lists to evaluate the listings is expected to result in more complete and accurate address data. Additionally, the ability to examine characteristics that appear on one list but not the other should serve to further inform survey practitioners about the applicability of these address lists and the characteristics associated with undercoverage of the lists.

The analytic needs of the NCS also pose a number of challenges. 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 propensity” 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 propensity classification. At each visit, her pregnancy propensity 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.

Another operational 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. Methods to avoid sample loss due to attrition are being developed to address this challenge. These methods include, for example, the transfer of cases between Study Centers, and outreach to communities to communicate the importance of the study.

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. The segments will need to be monitored for changes within the segment throughout the enrollment period. Additionally, households will need to be periodically re-enumerated to determine whether there have been any changes in household membership. As the study proceeds, it is expected that a variety of sampling and operational advances will be necessary.

Acknowledgments

This work was supported by The National Institute for Child Health and Human Development, NIH, contract numbers : HHSN275200503395C N01-HD-5-3395. The authors gratefully acknowledge the contributions of the NCS Sampling Working Group, a team of experts from the National Children's Study Program Office, Coordinating Center, and Study Centers, whose input to the sample design and implementation has been instrumental; in addition to the authors, team members included Ruth Brenner, Dean Baker, Jonas H. Ellenberg, Barbara Entwisle, Bryce Johnson, and Colm O'Muircheartaigh.

References

1. Ellenberg JH, Hirschfeld S. Proceedings of “The challenges and promises of a follow-up study of a randomly selected cohort of 100,000 pre and post conception women and their offspring through 21 years of life: Design, implementation and analysis issues of the National Children's Study”: Forward and study update. Statistics in Medicine. this issue.
2. Michael RT, O'Muircheartaigh CA. Design priorities and disciplinary perspectives: The case of the US National Children's Study. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2008; 171(2):465–480.
3. Westat. Sampling strategies for the proposed National Children's Study. Report prepared for the National Institute of Child Health and Human Development. October 25.2002
4. McCullagh, P.; Nelder, JA. Generalized Linear Models. New York: Chapman & Hall; 1989.
5. Korn, EL.; Graubard, BI. Analysis of Health Surveys. Hoboken, NJ: John Wiley & Sons; 1999.
6. Curtin, LR.; Feinleib, M. Considerations in the Design of Longitudinal Surveys of Health. In: Dwyer, JH.; Feinleib, M.; Lippert, P.; Hoffmeister, H., editors. Statistical Models for Longitudinal Studies of Health. New York: Oxford University Press; 1992. p. 49-87.Chapter 2
7. O'Muircheartaigh C, Eckman S, Weiss C. Traditional and enhanced field listing for probability sampling. Proceedings of the Social Statistics Section of the American Statistical Association. 2002:2563–2567.
8. Iannacchione VG, Staab JM, Redden DT. Evaluating the use of residential mailing addresses in a metropolitan household survey. Public Opinion Quarterly. 2003; 67:202–210.
9. Dohrmann S, Han D, Mohadjer L. Residential address lists versus traditional listing: Enumerating households and group quarters. Proceedings of the Survey Research Methods Section of the American Statistical Association. 2006:2959–2964.

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

11. Johnson B, Montaquila J, Heller A. An automated procedure for forming contiguous sampling units for area probability samples. Proceedings of the Survey Research Methods Section of the American Statistical Association. 2007

Table 1

Number of Sampled PSUs by Census Division and by Metropolitan and non-Metropolitan (1990) Classification

Census Division	Self-representing (Metro)	Non-self-representing		Total
		Metro	Non-metro	
New England	0	4	1	5
Middle Atlantic	2	10	1	13
East North Central	4	8	3	15
West North Central	0	6	3	9
South Atlantic	1	13	5	19
East South Central	0	5	3	8
West South Central	3	8	4	15
Mountain	1	5	3	9
Pacific	7	8	2	17

NOTE: The States included in each Census Division are as follows: New England: CT, ME, MA, NH, RI, VT; Middle Atlantic: NJ, NY, PA; East North Central: IL, IN, MI, OH, WI; West North Central: IA, KS, MN, MO, NE, ND, SD; South Atlantic: DE, DC, FL, GA, MD, NC, SC, VA, WV; East South Central: AL, KY, MS, TN; West South Central: AR, LA, OK, TX; Mountain: AZ, CO, ID, MT, NV, NM, UT, WY; Pacific: AK, CA, HI, OR, WA.