Paediatric and Perinatal Epidemiology
Affiliated to the Society for Pediatric and Perinatal Epidemiologic Research

doi: 10.1111/ppe.12005

Implementing Provider-based Sampling for the National Children's Study: Opportunities and Challenges

Kathleen Belanger,[a]* Stephen Buka,[b]* Debra C. Cherry,[c] Donald J. Dudley,[d] Michael R. Elliott,[f] Daniel E. Hale,[e] Irva Hertz-Picciotto,[i] Jessica L. Illuzzi,[a] Nigel Paneth,[g,h] James M. Robbins,[j] Elizabeth W. Triche,[b] Michael B. Bracken[a]

[a]*School of Public Health, Yale University Schools of Public Health and Medicine, New Haven, CT*
[b]*Department of Epidemiology, Brown University, Providence, RI*
[c]*University of Texas Health Sciences Center at Tyler, Tyler, TX*
[d]*Department of Obstetrics and Gynecology, and*
[e]*Department of Pediatrics, University of Texas Health Sciences Center at San Antonio, San Antonio, TX*
[f]*Biostatistics Department, University of Michigan School of Public Health, Ann Arbor*
[g]*Department of Epidemiology and Biostatistics, and*
[h]*Department of Pediatrics & Human Development, Michigan State University, East Lansing, MI*
[i]*Department of Public Health Sciences, University of California at Davis, Davis, CA, and*
[j]*Department of Pediatrics and Department of Psychiatry, University of Arkansas, Little Rock, AR*

Abstract

Background: The National Children's Study (NCS) was established as a national probability sample of births to prospectively study children's health starting from *in utero* to age 21. The primary sampling unit was 105 study locations (typically a county). The secondary sampling unit was the geographic unit (segment), but this was subsequently perceived to be an inefficient strategy.

Methods and Results: This paper proposes that second-stage sampling using prenatal care providers is an efficient and cost-effective method for deriving a national probability sample of births in the US. It offers a rationale for provider-based sampling and discusses a number of strategies for assembling a sampling frame of providers. Also presented are special challenges to provider-based sampling pregnancies, including optimising key sample parameters, retaining geographic diversity, determining the types of providers to include in the sample frame, recruiting women who do not receive prenatal care, and using community engagement to enrol women. There will also be substantial operational challenges to sampling provider groups.

Conclusion: We argue that probability sampling is mandatory to capture the full variation in exposure and outcomes expected in a national cohort study, to provide valid and generalisable risk estimates, and to accurately estimate policy (such as screening) benefits from associations reported in the NCS.

Keywords: *National Children's Study*, *sampling methods*, *probability sampling*, *multi-stage sampling*, *epidemiology methods*.

The National Children's Study (NCS) was initiated in response to the Children's Health Act of 2000 calling for a longitudinal study of the health of children in the United States. The study was initially designed to generate a national probability sample of children primarily recruited while their mothers were in early pregnancy. Seven vanguard centres implemented the study in 2007 using a household-based sampling approach.

In response to perceived difficulties in recruitment experienced by the seven centres, alternatives to household-based recruitment were tested.[1] In 2011, 10 additional study centres began recruiting women from prenatal care settings. This report draws on experiences from this pilot recruitment strategy to inform the design of the NCS and future pregnancy cohorts.

Rationale for a national probability sample in national cohorts like the NCS

There are four essential justifications for assembling birth cohorts such as the NCS based on a national

Correspondence: Michael B. Bracken, Center for Perinatal, Pediatric and Environmental Epidemiology, School of Public Health, Yale University Schools of Public Health and Medicine, One Church Street, 6th Floor, New Haven, CT 06510, USA.
E-mail: michael.bracken@yale.edu
*Co-first authors.

probability sample: (1) The national burden of childhood disease needs to be estimated accurately. Current national surveys (e.g. NHANES, NHIS) do not measure some important diseases, they are one or two orders of magnitude smaller than the NCS and less precise, and they cannot fully capture disparities in disease prevalence in smaller subgroups of the population. (2) Population-based exposure data are presently unavailable or imprecisely measured. This precludes hypothesis generation based on widespread environmental conditions linked to national disease incidence. (3) Treatment, screening, prognostic and diagnostic interventions must be based on accurate risk assessment in which selection bias is minimised. Estimates of the number-needed-to-treat (or harm), screening sensitivity and specificity and parameters for diagnostic and prognostic tests all depend on accurately estimating the prevalence of disease in the population of interest. Non-probability samples are prone to systematic error in estimating risk. (4) Probability samples help ensure that disease associations are accurately estimated when, as is often the case, the effect of the exposure of interest interacts with other (possibly unobserved) exposures. An example of this has been demonstrated in gene association studies, where the strength of the association with disease may depend on the extent of certain environmental exposures.

Rationale for provider-based sampling in national pregnancy cohorts like the NCS

In this paper, we propose an alternative to the geographically based secondary sampling initially proposed for the NCS, namely sampling prenatal care providers. We contend that this approach will both result in a national probability of births and will also efficiently achieve other design features central to the NCS and similar prenatal cohort studies. The following discussion is limited to secondary sampling (providers) and tertiary sampling (participants). The primary sampling unit (PSU) in the NCS is the County, details of which are described elsewhere.[2,3]

The NCS had other objectives besides obtaining national estimates of environmental exposures, disease and risk. These included: (1) first trimester biological collections; (2) clinical assessments including protocol ultrasounds conducted by the provider as standard care; (3) biomarker collections (cord blood, placenta, maternal and infant blood, meconium) at delivery, just after birth, and in childhood; and (4) environmental sampling around the home. This report examines whether probabilistically sampling provider practices can successfully and efficiently integrate clinical, population and cost objectives of studies like the NCS. For example, successful birth collections depend upon collaboration with birth hospitals, but in the household sampling plan there was no restriction on place of delivery resulting in major attendant cost and feasibility challenges. In large counties, this necessitated negotiating elaborate birth protocols with up to 25 birth hospitals, many of which delivered few study babies. Provider-based sampling cannot only produce a nationally representative sample but avoids many impractical consequences of household sampling.

Over a third of the sample of 105 counties selected to be included in the NCS had fewer than 4000 births per year, larger counties had over 50 000. We propose in under-populated counties to recruit a fixed fraction of pregnancies in all practices and in more populous counties to recruit in a stratified sample of practices. This makes it feasible to accommodate two critical design parameters, recruiting women early in pregnancy and constraining the number of birth hospitals to a manageable number.

Selecting a sample of provider groups (secondary sampling)

In counties with few provider groups (<10), the sampling fraction would be 100%. Within each provider group we would select a sample of pregnant women (see: Sampling women within selected provider group). Where there are many providers, sampling a subset of them is advantageous.

Assembling a sampling frame of provider groups

Providers will include obstetricians, midwives, family practice physicians or others responsible for prenatal care and delivery. Usually, the delivery attendant provides prenatal care, or is in the same practice group. Data from birth certificates provide the most comprehensive sampling frame for prenatal care providers. In some states, names of delivery attendants are computerised with other birth certificate data. After care providers have been identified, they can be grouped by practice group which is the usual site of prenatal care. Individual providers may change group, but characteristics of patients from a specific provider group are more stable. If individual providers cannot be identified from birth certificates, it is possible to identify provider groups using directories, hospital lists and county medical society records.

Stratification of provider groups

Provider groups can be stratified to insure that specific patient subgroups are represented in the sample. This is important when the county is heterogeneous and subgroups seek care from particular provider groups.

The number of stratification variables may vary across counties but should not be large. Birth certificates provide information about patient populations, including age, race/ethnicity, education, gestational age at entry into prenatal care, and payment type. Often, several characteristics of interest are correlated. Stratifying provider groups by government-funded vs. private insurance may result in strata having similar patients with respect to other important characteristics such as income.

Geographic location within a County may also be used to stratify provider groups. For example, if there were two cities within a county, each with two large hospitals, it might be desirable to ensure that the sample be stratified to include practice groups from each city.

Stratified sampling of provider groups

Preliminary work (see Appendix S1) suggests that a representative sample of births is obtainable, even in large counties, with a sample of 10–20 provider groups. The exact number per county depends on the number of provider group strata and the underlying homogeneity of the population. Sampling proportional to number of anticipated prenatal care recipients could be used to draw a relatively efficient probability-based sample.

Sampling women within selected provider groups (tertiary sampling)

Except in very small PSUs (counties), sampling provider groups will be followed by sampling women within provider groups because the number of pregnant women per provider group could otherwise outnumber the PSU target. For example, assuming a county-wide enrolment of 250 subjects per year, if 10 provider practices are selected only 25 enrolled women are needed per group. A larger number would be sampled to account for refusals, pregnancy loss and dropouts prior to birth. For example, if 60% of eligible women agree to enrol, approximately 420 women would need to be sampled. This would require 42 women per provider group with 10 groups or 21 women per group with 20 groups. Whether an equal number of women would be sampled from each provider, or the number sampled weighted according to size of provider practice, depends on study goals. Hypotheses requiring broad geographic distribution argue for equalising the number of women across provider practices.

Several methods are available to sample women within provider groups and these are described in detail below.

Probability-based sampling

Simple random sampling is not feasible for a national cohort study as women will enter prenatal care over time creating a ‘rolling’ sampling frame and the desired sampling fraction may vary during the recruitment period. Systematic sampling has been proposed as an alternative; for instance, taking every fifth patient to achieve a sampling fraction of 20%. However, systematic sampling in physician offices may be cumbersome and open to manipulation.

Women could be selected from those with appointments on randomly sampled days of the week, or randomly sampled weeks of the year. This option might be popular with providers since it limits the time study personnel would be in their office, perhaps to a few weeks per year. It is also efficient for study personnel who rotate from one provider group to another. To ensure integrity of sampling, it is critical that sub-sample selection within provider groups be conducted by an outside organisation.

Another option for sampling within provider groups is to maintain county geographic clusters (segments in the original NCS protocol). Utilising geographic sub-sampling may facilitate efficient school-based and environmental sample collections. However, predicting school affiliation and future residence during the prenatal period is difficult as at least 30% of pregnant women will not live at the same address seven years later.* This approach replaces the

*In two independent Yale University pregnancy cohorts followed for 7 years, 30.0% (n = 1760) and 30.1% (n = 2256) had changed addresses between birth and 7-year follow-up. This does not include additional addresses where a ‘change of address’ card permitted mail to be forwarded.

challenge of finding pregnant women during episodic visits to households within a geographic segment with the challenge of finding those residing within a specific geographic area among the pregnant women in a large prenatal care practice. Identifying these women and arranging to recruit them during prenatal appointments scheduled at disparate times and locations across the county proved to be resource intensive in the vanguard pilot phase.

Representative sampling of women within provider groups: not probability based

An alternative sampling method within provider groups would enrol a representative, but not probability-based, sample. The study would recruit women expressing interest within sampled provider groups. Screening information might include age, ethnicity and address, which are entered into a computer program; women are selected using a minimisation algorithm. The minimisation algorithm adapts the selection probability of the volunteers as they enrol to ensure the final sample is representative of the County. This is similar to minimisation allocation used in randomised clinical trials.[6] The minimisation algorithm is programmed to adjust for over- and under-ascertainment of population sub-samples as the study progresses. For example, the probability of a white, college-educated woman being selected early in recruitment might be 20%; however, if more of these women enrol than anticipated the probability of them being selected is reduced.

This selection process is not probability based because women not interested in NCS are not sampled. It is also not a purely volunteer sample since women are randomly selected among those interested in a manner designed to represent the county population of pregnant women.

Representative sampling at the tertiary stage has several advantages. First, it may appeal to provider groups and potential participants by allowing interested women an opportunity to be screened. Second, sampling interested women may result in a cohort more likely to be retained over time although we have only anecdotal evidence for this. Third, it would be cost-effective.

Non-response adjustment

Any covariate information available from both responding and non-responding units, either practices or women, can be used to develop estimates of the response probabilities whose reciprocal will provide non-response weights to adjust for bias resulting from differential response probabilities. For practices, birth certificate information used to develop strata (e.g. race/ethnic distributions of patients, geographic location) can also be used to predict probability of response via logistic regression models. For women, factors such as calendar age or gestational age at first visit could be similarly used, subject to restrictions on providing information from non-consenting women.

Calibration to known population distributions

Regardless of how the sample is selected, both bias and variance can be reduced through construction of calibration weights so that the calibration-weighted sample distribution matches known distributions in the population. Calibration can be done via post-stratification (when joint population distributions are known) or by raking (only marginal population distributions are known). Care is taken to assess key covariates that differ between sample and population, balancing between an overly coarse set of covariates that miss opportunities for bias and variance reduction and an overly refined set of covariates that inflate variance due to sampling variability in the calibration weights.[5] When convenience sampling is used, adjustment for selection bias is virtually impossible since wide swaths of the population are not provided an opportunity to participate and will not be represented at all, precluding post-randomisation stratification.

Special issues and challenges

Optimisation of key parameters of provider-based sampling design

There are several challenges for second-stage sampling in a national study: (1) balancing the number of provider groups sampled within a county or PSU; (2) establishing the number or proportion of patients sampled within each provider group; (3) deciding the approach used to sample patients within provider groups; (4) considering the impact of these decisions on the geographic distribution of participants; and (5) restricting the number and geographic dispersion of hospitals. There may be no single optimal solution for every county – each parameter has an effect on others.

For example, given a fixed goal of 250 births per year per PSU, if a smaller number of provider groups are selected a higher proportion of women would be sampled from each, increasing recruitment efficiency but likely diminishing geographical variation and statistical efficiency as the within-PSU correlation will be higher due to the smaller number of practices sampled. In contrast, selecting a larger number of provider groups requires sampling a smaller number of women per practice, permitting greater geographical clustering and statistical efficiency but at greater cost due to exclusion of readily identified potential participants from within provider offices. Optimisation of these parameters depends on many factors, including PSU size, number and diversity of providers, and scientific priorities.

Retaining geographic clustering in a provider-based sampling scheme

Retaining some level of geographic sub-sampling may offer advantages. Once children reach school age, their school environment could play a strong role in their health and development. To the extent that geographically clustered sampling occurs, it facilitates relatively efficient school-based data and environmental sample collections. However, limiting geographic distribution (such that enrolled children attend only a handful of schools in the county) may circumscribe the range of environmental and social exposures among children in the PSU. Ecological variables, including neighbourhood, school and air pollution index, are assigned to the participant. If clustered into large geographic areas, many participants will receive the same value of the variable. Thus, broadening geographic clusters could lose power to examine ecological effects by reducing the range of variance in examined exposures. When geographic segments are too small many practices may provide only several segment-eligible women per year. Therefore, the value and disadvantages of geographic-based sampling must be carefully considered.

Potential bias from enrolling women from provider practices: women who receive late or no prenatal care

The goal will be to recruit women at selected provider groups during the first trimester of pregnancy. Nonetheless, women who enter prenatal care late should be included. To include all women in the selection process but encourage early entry, eligibility could be restricted to within 6 weeks from the first prenatal visit (some practices may prefer study recruitment at the second prenatal visit because of pressure from clinical demands at the first visit). This would insure that women who enter care early are enrolled early, and women who enter care later remain eligible for 6 weeks from their first prenatal visit. This provides a representative distribution of prenatal care timing by women in the PSU.

Although the number of women who do not receive any prenatal care is very small (approximately 2%),[6] all birth hospitals should have a mechanism to identify them so that they may be invited to participate during the delivery hospitalisation.

Enrolling women through community engagement

Using a minimisation algorithm described above, community engagement could supplement provider group sampling. In this hybrid design, women who contact the study (email, phone or through community events) are screened. The same algorithm that determines eligibility in sampled provider groups is used to randomly select some women from the community who wish to participate. This procedure might also facilitate representation of women who are difficult to recruit in prenatal care settings. For example, undocumented immigrants may be reluctant to participate at their providers' office but more willing to be screened at a community meeting where trusted community leaders encouraged participation.

Operational challenges to sampling provider groups

An accurate sampling frame of provider groups necessitates generating complete lists of providers and practice groups. Developing size measures for provider groups requires estimates of practice size (i.e. delivery volume). This in turn needs timely access to complete birth certificate data and additional coding to group births by practice. Counties may vary in their ability to readily access birth records, including the birth attendant variable. Cross-coverage from another practice or resident service can lead to inaccuracies in the size of provider practices based on birth certificate data. Finally, practice lists must be routinely updated to sample newly created provider groups.

Stratification based on patient characteristics within provider groups (e.g. sociodemographic and geographic factors) requires analysis of birth registry data by practice or analysis of statistics self-reported by each provider group. It may also be desirable to stratify on provider practice characteristics, including practice size and volume.

Types of providers to include in sampling frame

In rural counties, residents often travel outside the PSU to obtain prenatal care. The sampling frame of providers obtained from birth certificates of PSU residents may be so geographically dispersed that enrolment from prenatal care providers and collection of birth data at hospitals in a wide radius outside the PSU is impractical. In rural areas, women may obtain prenatal services at agencies such as WIC and other social services. An alternative to sampling birth attendants dispersed over a broad area is to include social service providers supporting pregnant women located in the PSU. Their sampling fraction may need to be derived from state Medicaid data.

Non-participating provider groups

Provider groups may refuse participation, some providers will move during enrolment and new provider groups will form. Sampling provider groups rather than individual providers reduces some challenges. The sample of practices should be chosen so that a practice that refuses to participate is replaced with a practice from the same stratum.[7] The refusal should be tracked to maintain an accurate response rate.

Differentiating between the eligible sample and the enrolled sample

The strategy used to select the sample does not guarantee that women eligible for the study enrol. We have two essential tasks: to ensure the eligible pool is representative of the population and to ensure that enrolled women are representative of the eligible pool. The former does not ensure the latter.

Use of health maintenance organisations

It has been recently suggested the NCS use convenience sampling in large health maintenance organisations (HMOs) with additional sampling to include patients not typically found in an HMO. As the preceding discussion intimates, this design is flawed for several reasons: (1) for the great majority of women who become pregnant during the recruitment period, the probability for any of them to be in the study will be unknown, inestimable or zero; (2) convenience sampling from large HMOs reduces the chance of representing 'unknown confounders' in the sample with frequencies expected in the general pregnant population; (3) while some contend that adjustments on known population characteristics is adequate to create a 'representative' sample, *post hoc* adjustment can lead to disproportionate under- or over-representation of some population members in the sample and because HMOs do not cover many rural areas, complete absence of certain strata; (4) over-sampling on known clinical or environmental conditions may lead to under-sampling presently unknown factors; (5) the validity of exposure-disease associations that interact with other factors, including genotypes, can be substantially disrupted in non-probability samples and these are associations likely to dominate future research; and (6) it becomes problematic to calculate attributable risks necessary for estimating population benefits from social and clinical interventions that follow discoveries made during the NCS project.

Conclusion

We have argued that some degree of probabilistic sampling is necessary to fully capture the variation in exposures and outcomes needed in a national cohort study, particularly when such a study is expected to deliver definitive results that few other cohorts will have the statistical power to replicate.

A simple random sample is not practical for building a pregnancy cohort having to meet demands of efficiency. Some stratification, over-sampling, calibration or minimisation will be required. The fundamental goal of sampling should be a probability sample in which the likelihood of any woman in the US being selected for the study can be estimated. In this paper we discuss a variety of probabilistic and non-probabilistic sampling strategies and conclude there are formidable objections to using non-probabilistic sampling in a national cohort study. We support the view that household sampling is not a cost-effective and feasible option to obtain a probability sample of women early in pregnancy and we highlight the benefits and challenges of an alternative design using provider-based sampling.

References

1 Hirschfeld S, Songco D, Kramer BS, Guttmacher AE. National Children's Study: update in 2010. *Mount Sinai Journal of Medicine* 2011; 78:119–125.

2 Montaquila JM, Brick JM, Curtin LR. Statistical and practical issues in the design of a national probability sample of births for the Vanguard Study of the National Children's Study. *Statistics in Medicine* 2010; 29:1368–1376.

3 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; 171:465–480.

4 Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; 31:103–115.

5 Deville J-C SC-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association* 1992; 87:376–382.

6 Elam-Evans LD, Adams MM, Gargiullo PM, Kiely JL, Marks JS. Trends in the percentage of women who received no prenatal care in the United States, 1980–1992: contributions of the demographic and risk effects. *Obstetrics and Gynecology* 1996; 87:575–580.

7 Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; 56:645–646.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. A simulation study of provider-based sampling in Wayne County, MI.