

**National Children's Study  
Discussion of Potential Sampling Strategies for the Main Study  
May 29, 2012  
6100 Executive Boulevard  
Rockville, MD**

The National Children's Study (the Study) is led by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH) in collaboration with a consortium of federal government partners.

**Purpose of the Meeting**

The purpose of the meeting was to discuss in more detail the analytic strategies and potential efficiencies that can be achieved in all stages of sampling. The meeting participants explored desired characteristics of and questions about five strategies:

- Geographic area sampling
- Provider-based sampling
- Sampling within providers
- Supplemental sampling
- Missingness by design.

The discussion was moderated by Enrique Schisterman, Ph.D., Senior Investigator, Chief, Epidemiology Branch, Division of Epidemiology Statistics and Prevention Research, NICHD, NIH, Department of Health and Human Services (HHS).

**Opening Remarks from the Director of the National Children's Study**

*Steven Hirschfeld, M.D., Ph.D., Director, National Children's Study, NICHD, NIH, HHS*

Dr. Hirschfeld reviewed the following:

- **Children's Health Act of 2000.** The legislation states that the Director of the NICHD, together with a federal consortium, will (1) plan, develop, and implement a prospective cohort study, from birth to adulthood, to evaluate the effects of both chronic and intermittent exposures on child health and human development; and (2) investigate basic mechanisms of developmental disorders and environmental factors, both risk and protective, that influence health and developmental processes." The Study is required to "(1) incorporate behavioral, emotional, educational, and contextual consequences to enable a complete assessment of the physical, chemical, biological, and psychosocial environmental influences on children's well-being; (2) gather data on environmental influences and outcomes on diverse populations of children, which may include the consideration of prenatal exposures; and (3) consider health disparities among children, which may include the consideration of prenatal exposures."
- **Study's Structure.** The Study is a multi-component system that includes a Vanguard Study, a Main Study, and formative research. The goals of the Vanguard Study are to assess the feasibility, acceptability, and cost of recruitment, logistics, operations, and Study visits. The goals of the Main Study are to investigate exposure-response relationships of the growth, development, and health of children.

- **Main Study Objectives.** The primary objective of the Main Study is to collect information on and investigate the factors that determine children’s health and development. These factors include genetic context and environmental exposures with a broad definition of environment. The Study is not a study in a conventional sense. It will function as a high-quality data collection platform for researchers to access and analyze data, biospecimens, and environmental samples. For purposes of modeling a design, the Study’s Program Office proposes the use of surrogates for exposure and surrogates for response with the implied understanding that all the potential hypotheses and analyses cannot be anticipated in advance. Thus, the Study’s Program Office proposes a few examples to frame a design discussion with the expectation that if a design will support exposure-response analyses from the limited set of example exposures and outcomes, then the design can support many more analyses.
- **Core Physical Environmental Exposure.** The Study will use as examples of more comprehensive exposures to analyze heavy metals, pesticide residues, and semi-volatile organic compounds in household dust, blood, and urine as general surrogates for more specific exposures.
- **Core Biological Assessments.** The Study will use as examples of more comprehensive outcome measures some core biological assessments that will include:
  - Linear growth rate and body mass as a surrogate for general health
  - Metabolic screen of serum total protein, blood urea nitrogen, cholesterol, iron, and calcium for nutrition and dietary exposure
  - Frequency and duration of health system encounters for respiratory illness for pulmonary health
  - Timing of standard neurodevelopmental landmarks for any deviation from adjusted trajectory for cognitive and social development.

Thus, for discussion purposes, a small set of examples to frame a study design would include analyses such as exposure to pesticide residues, drawn from the sample exposure list, and timing of neurodevelopmental landmarks, drawn from the outcome measure list. Another example might be exposure to semi-volatile organic compounds, also drawn from the sample exposure list, and health care system encounters for pulmonary disease, drawn from the outcome measure list.

- **Main Study Framework.** The Main Study will recruit 100,000 children from more than 100,000 pregnant women primarily through health care providers. The geographic locations have not yet been determined. The operational issues of how to retain enrolled women who move out of the geographic areas have not yet been addressed.
- **Terms to Distinguish.** Key operational terms are defined as follows:
  - Sample—a subset of individuals from within a population to estimate characteristics of the whole population
  - Sample frame—set or list of items or people that can be measured
  - Probability sample—selection of a sample where every unit in the population has a chance greater than zero of being selected, and that chance can be determined
  - Recruitment strategy—mechanism to enroll participants.

The Study may use a hybrid sampling approach that includes (1) a probability sample frame with three stages—area sample, provider sample, and participant sample—and (2) a separate supplemental recruitment effort. Supplemental recruitment could focus on populations that may

be underrepresented on the basis of factors such as demographics, exposures, and access to health care.

Dr. Hirschfeld explained that this meeting is one of a series of opportunities to receive input on the Study's sampling strategies. Additional discussions are scheduled and will occur before any decisions are made. The draft proposal for sampling strategies will be presented to the National Children's Study Advisory Committee (NCSAC) on July 24, 2012.

### **Opening Remarks from the Director of NICHD**

*Alan E. Guttmacher, M.D., Director, NICHD, NIH, HHS*

Dr. Guttmacher thanked the meeting participants for their attendance. He said that this is an important time in terms of Main Study's design. The goal of this meeting is not to reach a consensus but to gather input from individuals with different expertise and perspectives. Their input will help the Study examine the various sampling strategy options and their scientific benefits and liabilities. This examination will help craft the optimal sampling plan that will be presented to the NCSAC in July 2012.

### **Geographic Area Sampling**

#### **Questions**

- Can the Study have a purely probabilistic geographic sample (that is, with no oversampling, weighting, or stratification)? What is gained or lost by this approach?
- Should the geographic samples be clustered within regions? Would this clustering need to be defined by population density or could it be defined by environmental characteristics? What is gained or lost by this approach?
- Should the geographic samples be selected using a stratified frame (for example, population density, demographic characteristics, happiness index)? What is gained or lost by this approach?
- How could frame deficiencies be identified, and how could they be backfilled?
- How many geographic areas need to be selected in order to generalize the findings of the study?
- What should the area of the sampled geographic units be (for example, state, county, zip code, census tract, census block group, census block)?

#### **Discussion**

- Jennifer Madans, Ph.D., said the questions need to be put in context. The Study could have a purely probabilistic geographic sample, and the Study could oversample. However, the questions cannot be answered unless the Study is trying to maximize something. The "something" that the Study is trying to maximize needs to be identified. It would be more useful to discuss some of the assumptions in the white paper—titled Potential Sampling Strategies: Main Study—because some may not be as accurate as others. An important issue is the way in which a supplemental sample would be combined with a probability sample.

- Edward J. Sondik, Ph.D., M.S.Hyg., noted that the Study is not a typical study; it is a data collection platform. Key issues are understanding what the Study is trying to identify and the precision for identifying particular causalities or particular relationships. The data collection platform would be used to conduct a variety of studies that have not yet been conceived.
- Dr. Hirschfeld explained that the goal of the Study is to investigate the effects of exposures on health outcomes. Because not all of the exposures are known, certain exposures have been selected as paradigms or surrogates, and certain specific outcomes have been selected for health surrogates. The Study could map some of the specific exposures to some of the specific outcomes or map a single exposure to a single outcome in order to make technical or statistical recommendations or statements. The Study's platform will allow many studies that have not yet been contemplated. A logistical and operational framework will be needed to permit these future studies. Therefore, the Study will need to collect data with a high degree of rigor. The Study has not yet determined the sampling strategy. However, data show support for the use of health care providers to bring pregnant women into the Study. The Study seeks to continue a dialogue to determine the best sampling strategy.
- Roderick Little, Ph.D., asked whether the Study is considering a hybrid of a probability sample and a nonprobability sample, which raises a fundamental issue of weighting the estimates to combine them.
- Dr. Schisterman replied that the Study's Program Office is seeking input on whether a hybrid approach can be used and if so, how to do it and what the analytical approaches would be to resolve issues such as weighting.
- Graham Kalton, Ph.D., commented that early discussions of a purely probabilistic approach included considerations of oversampling certain underrepresented groups. Because of the multipurpose nature of the sampling design, the discussions concluded that concentrating resources on oversampling would not be fruitful for the Study. With regard to exposures to environmental contaminants, participant mobility away from exposures becomes an issue. With an equal probability design, stratification must be considered as well as weighting due to nonresponse. Whether the Study can achieve an equal probability sample at a reasonable cost is an issue.
- Warren Strauss, Sc.M., said that in terms of identifying design optimality it would be helpful to understand whether the Study's primary goal is to understand the relationships between exposures and health outcomes and to understand whether estimating prevalence is secondary or relationships and prevalence are equally important.
- Dr. Hirschfeld confirmed that the Study's primary goal is to understand the relationships between exposures and health outcomes.
- Michael D. Sinclair, Ph.D., said the way in which providers are stratified will depend on what information can be collected about the providers to select them from the sample. Stratification may help to address issues of provider nonresponse.

- Dr. Madans explained that discussions of stratification are secondary to more major issues, such as the type of data collected and the timing of data collection. Certain decisions about data collection that are made early in the Study will affect the ability to determine relationships between exposures and outcomes. Not including certain women in the sample will also have a far-reaching effect in determining relationships. Therefore, the Study needs to maximize the ranges of exposures and outcomes by ensuring that all women that need to be in the sample are included by using a nonperson probability sample. The Study also needs to identify pregnant women who may be missing from the sample and the time that they should be identified. Understanding the limitations and objectives of provider-based sampling for identifying pregnant women will help with the provider-based design.
- Dr. Hirschfeld clarified that the Study would ideally like to enroll women by the 8th week of pregnancy. Enrolling at this time is the target.
- Dr. Little asked whether geographic area sampling is part of a hybrid design, which might include an equal probability design and provider-based sampling. Combining the two resources remains an issue.
- Dr. Hirschfeld explained that the Study was exploring an option for a three-stage sampling process. The first stage is an area sample, the second stage is a provider sample, and the third stage would be a participant sample.
- Mr. Strauss asked whether this three-stage sampling involves nested stages, where broad geographic areas are selected in the first stage and sampling frames within the geographic areas, which might include provider sampling and a supplemental sample. Dr. Hirschfeld said that such a hierarchical process is one possible sampling approach.
- Dr. Little asked whether the Study will sample from the original 105 Study locations. Dr. Hirschfeld replied that Study may use some of these original locations, but discussions should consider sampling approaches without pre-identified locations.
- Dr. Kalton said that the sampling design should consider some form of clustering because of the issues of data collection, even though mobility will be a factor. Clustering allows construction of frames of providers. Some form of provider clustering is needed, and it is important to effectively stratify the clusters. The goal of the original geographical stratification with 105 Study locations was to achieve a national representative sample.
- Dr. Sondik commented that the Study needs to understand the characteristics of providers, such as geographic location, the populations they do not serve, and the populations that would be left out of the Study using solely provider-based sampling. Once the providers' characteristics are understood, discussions could then focus on the fundamentals of sampling.
- In response to a question about retaining the original 105 Study locations, Mr. Strauss explained that the selected locations were probability proportional to size in terms of the number of children within each potential primary sampling unit (PSU). There were 13 certainty strata, which were the largest population centers in the United States. The National

Center for Health Statistics (NCHS) then created strata according to urban/rural areas, demographic factors, and populations. This approach may not be optimal for creating estimates of variance within the strata. More than one PSU may be selected within a stratum to assist with the estimation of variances. Mr. Strauss briefly described a National Heart, Lung, and Blood Institute study that used a hybrid approach to geographical sampling, which included a national probability-based sample with samples from certainty areas.

- Dr. Kalton commented on the origins of the original sampling design. He noted that the issues related to the number of PSUs are secondary to the sampling strategy.
- In response to a question from Dr. Little, Dr. Hirschfeld explained that Vanguard Study recruitment and data collection will continue. The Vanguard Study and Main Study are operationally separate studies. Their data are not intended to be combined at an aggregate level.
- Karol Krotki, Ph.D., said that if he were to create an index of complexity, difficulty, or challenges, scores for geographic issues would be low compared with screening issues and coverage. He said a larger number of geographic clusters might be considered. Instead of geographic sampling, another approach would be to use providers as clusters.

## **Overview of Provider-based Sampling**

*Graham Kalton, Ph.D., Westat*

The aim of provider-based sampling is to identify and enroll women as early in their pregnancies as possible. Provider-based sampling does not cover preconception women. Provider-based sampling can draw on some of the experience gained from provider-based recruitment, but it is a very different approach.

There are two versions of provider-based sampling. Version A samples pregnant women within the PSU via prenatal provider locations, which implies sampling some locations outside the PSU. Version B samples provider locations only from within the PSU, which implies selecting some pregnant women from outside the PSU. Version A is compatible with alternative area-based samples in other PSUs. Version B is not. Version A is more consistent with the measure of size (MOS) used in sampling PSUs. With Version A, birth certificate data for the PSU can be used to assist in constructing the frame of provider locations and MOS, and also for checking on sample coverage. For Version B, the birth certificate data from adjacent counties would also be needed. With Version A, there are more small-provider locations because size is measured in terms of pregnant women living in the county. This problem is avoided with Version B. Version A has been chosen for the pilot studies now under way.

There are four steps for the provider-based sampling process:

- Construct a list frame of prenatal care provider locations that provide care to pregnant women in the sampled PSU
- Populate the frame with estimated numbers of pregnant women from within the county seen in a year (that is, MOS for use in probability proportionate to estimated size [PPES] sampling) and variables that may be of use in stratified sampling

- Select a stratified PPES sample of providers
- Select a sample of pregnant women who are making their first visit to a prenatal care provider location for this pregnancy.

The provider-based sampling pilot studies are being conducted in three single-county PSUs. The studies will select 15–25 provider locations in each PSU and about 20 eligible pregnant women at each sampled location. Lists of prenatal providers will be compiled by the contract Study Centers for each location, complemented by birth certificate records. List frames will include prenatal care provider locations from outside the PSU. Provider locations with very small MOS will be dropped from the final frames. Hospitals and birthing centers will be included on the list to cover women with no prenatal care and women who receive prenatal care from a provider location not on the final frame (either dropped or missed). MOS and stratification data will be collected from birth certificate data and/or from a provider location questionnaire. The questionnaire will collect information such as the numbers of providers of different types, the type of practice, the number of prenatal visits in 2011, and the number of first prenatal visits in the three single-county PSUs.

Birth certificates can provide information on the MOS for each location for PPES sampling and can also provide information on characteristics such as the percentage Hispanic, percentage Black, and percentage of women on Medicaid that can be used for stratification.

The provider locations are to be sampled within the strata with PPES. The overall probability of sampling an eligible woman is to be a constant, implying that within selected provider locations women need to be sampled at rates that compensate for the differential sampling rates for the locations.

The sample of women is restricted to (1) pregnant women making their first prenatal visits to any of the provider locations on the final sampling frame, (2) women living in the sampled PSU, and (3) women 18 years of age and older. The sampling operation will be more efficient if location staff can accurately filter out ineligible women before the sampling is performed. Screening for visits to other provider locations will not be performed by location staff. Eligibility will also be determined at interview.

Some of the lessons to be learned from the pilot studies include:

- Cooperation rate for provider locations and reasons for noncooperation
- Response rates for sampled women (screening and enrollment)
- Stage of pregnancy at which enrolled
- Proportion of women enrolled at birth
- Accuracy of the MOS
- An estimate of sample coverage
- Kinetics and efficiency of enrollment overall and for various provider types.

## **Provider-based Sampling**

### **Questions**

- How can providers be enumerated efficiently?
- How can selected providers that choose not to engage in the Study be replaced or substituted in a way that preserves the probability sample?
- How can the method of selecting providers increase recruitment success (for example, restricted frame sampling)?
- Can the demographic characteristics of the provider's practice be determined before sampling?
- Can the sampling method be flexible in this stage in order to allow for regional differences in identification of providers (such as availability of birth certificate records)?
- Are there features of a provider practice (for example, practice type or practice size) that might bias the recruitment of participants?

### **Discussion of Dr. Kalton's Presentation and Provider-based Sampling Questions**

- Dr. Schisterman asked whether provider practices have multiple doctors working on the same day. Dr. Kalton replied that many practices have multiple doctors, but it is the practices, not the doctors, that are considered the practice location.
- Mr. Strauss asked whether a hierarchy of hospitals could be established as a first sampling stage within a location. The providers that deliver in those hospitals would serve as the second stage. Using this approach would reduce the number of hospitals that are engaged. Mr. Strauss asked about the methods of substitution for providers that do not agree to participate as well as the women who do not agree to participate. He also asked how the women who seek prenatal care before they are pregnant will be enrolled.
- Dr. Kalton said the number of women seeking preconception care is not known. Such women would not be part of a representative sample. For provider substitution, another provider would be selected from the same stratum. The providers could be ordered, for example, by practice size. Sampling from hospitals would essentially be cluster sampling. Clusters tend to be homogeneous. The Study should ideally sample from as many clusters as possible and not just a small number.
- Mr. Strauss noted that in a large urban location with many hospitals, arrangements might have to be made with all hospitals for those providers that refer to and deliver at multiple hospitals. A hierarchical approach could limit the number of hospitals that need to be engaged. A hierarchical approach would be less applicable in rural locations with fewer hospitals.
- John K. Gohagan, Ph.D., commented that it might be important to understand what is not known about women who did not enroll. Information about these women could provide insights on the limitations of provider-based sampling.

- Mr. Strauss said a hierarchical approach could maximize efficiency of data collection without violating any principles of probability-based sampling.
- Dr. Madans said an issue may be determining the quality of the provider lists. Birth certificates generally list the attending physician and certifier but not the prenatal care provider. A mechanism would be needed to match physicians back to providers as well as determine the MOS from birth certificates. Another issue is the quality of the MOS using the questionnaire. Some providers may not complete the questionnaire. Errors may be introduced in this early stage, and evaluating errors may be challenging.
- Dr. Kalton commented that the birth certificates are a source of data but not the only source. The birth certificate data are highly variable across locations. He noted that the provider-based questionnaire relies on a complete list of providers. One of the goals of the pilot studies is to compare the questionnaire data with the birth certificate data in locations where both exist and then evaluate the MOS.
- Dr. Madans said the quality of provider lists compiled by local staff may vary across locations and may require external verification. She asked whether the questionnaire will be able to collect data on the stage of pregnancy at first visit. Leaving out women who did not seek care early in pregnancy may be a source of bias.
- Randall J. Olsen, Ph.D., asked for clarification on the definition of “provider.” Dr. Hirschfeld responded a provider could be an obstetrician-gynecologist, a family practitioner, or any other licensed practitioner that provides prenatal care to women. Health care providers could be routing women to prenatal care providers based on certain conditions or nonrandom factors, which might lead to a choice-based sample.
- Dr. Kalton noted that a woman is sampled at first visit to one of the prenatal care providers in the sampling frame.
- Dr. Schisterman asked whether there are regional differences in when a woman seeks her first prenatal care visit.
- Dr. Kalton replied that data are available on the number of women who seek care during the first trimester. Some data may be available from birth certificates in some locations. However, the current provider-based design enrolls women at the first visit, regardless of stage of pregnancy, and these data are available. Dr. Kalton commented that it is important to know the analytic objectives for the prenatal data, how the prenatal data will fit into the overall Study, and how important the prenatal data are, given potential attrition issues. Attrition may be random or nonrandom. If there is “reverse attrition,” early prenatal data may be missing.
- Dr. Little explained that missing data from reverse attrition are a covariate in a regression analysis. Missingness that depends on a covariate does not bias the regression analysis.

- Dr. Krotki asked whether there is any evidence on the difficulty of constructing the provider frames in the three single-county PSUs.
- Dr. Hirschfeld described the frame construction as a work in progress. Because the Study has not yet received Office of Management and Budget (OMB) clearance, empiric data are not available. He noted that the quality of the sampling frame is an essential criterion for moving forward.
- Iris M. Shimizu, Ph.D., asked whether the Study has considered the American Medical Association as a source to compile physician lists. The Health Resources and Services Administration could be a source for community center lists.
- Dr. Madans commented that the Study should clarify the importance of preconception and early prenatal exposure data in order to determine whether the provider sampling frame is appropriate for collecting these data. Knowing how the data will be used will help determine decisions about sampling design. Preconception and early prenatal exposure data will be key for assessing health outcomes at birth and early childhood.
- Mr. Strauss noted that one of the Study's early goals was to have 30 percent of the sample preconception women and the rest as early as possible in pregnancy (ideally, less than 8 weeks) in order to capture early exposures.
- In response to a question from Dr. Schisterman about determining which providers should be included, Dr. Kalton explained that the criteria have been established, and there are some practical considerations. Although the providers will mostly be obstetrician-gynecologists, some locations will include family practitioners.
- Nigel Paneth, M.D., M.P.H., a member of the viewing audience, described the process of compiling the provider list in the Wayne County, MI, Study location. The first step was to examine birth certificate records to identify birth attendants, which include mostly obstetrician-gynecologists and family practitioners. Other sources such as hospital-delivery privilege licenses, county medical societies, and regional obstetric societies were then examined. Address lists for prenatal care practices that advertise were another source. Constructing the provider lists was not difficult. The provider lists can be further characterized using birth certificate records. It was determined that 15 practices could provide a representative sample for Wayne County. Because most practitioners do not deliver at multiple hospitals, the number of hospitals can be constrained in provider-based sampling. Dr. Paneth noted that in 2 years of effort, agreements were reached with about 70 percent of the Wayne County hospitals.
- In response to a question from Dr. Schisterman, Dr. Kalton explained that in Version A of provider-based sampling, women are within the PSU, whereas providers can be outside the PSU. In Version B, providers are within the PSU, but women can be located outside the PSU. In smaller rural counties, all providers may be outside the PSU, which would require some flexibility in sampling. Therefore, the PSU definition and MOS have to be reset. In some cases, women who reside in a PSU may deliver outside the PSU.

- Dr. Hirschfeld asked whether sampling a geographic area would need to be changed if all providers are outside the PSU.
- Mr. Strauss explained that the geographic areas should be selected probabilistically at the first stage. At the second stage, construction of a valid probability-based sample could be flexible from one location to another for operational efficiency and feasibility.
- Dr. Kalton said the PSU needs to be large enough to support the sample size (currently 1,000 women per PSU over the entire recruitment period). For Version A, the MOS can be determined on the required sample size. If a PSU does not have sufficient women, the PSU can be merged with another PSU. For Version B, the MOS is the number of first prenatal care visits to providers in the PSU. The PSUs need to be big enough to support the required sample size. The PSUs can then be stratified once they have been constructed.
- Anjene Addington, Ph.D., M.P.H., commented that the level of provider cooperation and participation is critical to the success of provider-based sampling.

## **Sampling within Providers**

### **Questions**

- What are effective ways to enroll women using systematic sampling? What could the basis of this sample be? Are there additional ways women could be sampled in an effective way that preserves the probability sample?
- Should women residing outside of the geographic sample, but seeking care from a selected provider, be included? If they are included, what would this do to the comparisons to extant natality or American Community Survey data?
- Should preconception women be eligible in the probability sample, or should they be from a separate cohort?
- How would the sample of pregnant women be evaluated for frame coverage or population representation? How could deficiencies be addressed in this stage of sampling?
- How could women who change providers be handled?
- How could women who move out of the geographic area be retained?
- How can additional women who reside within the geographic sample, but are not in the provider sample, be included in a “light touch” cohort (for example, could a provider recruit women from a practice location other than the selected one) as a supplement?

### **Discussion**

- Mr. Strauss said the pilot studies demonstrated systemic provider sampling in terms of identifying the MOS and target number of enrollees and working with providers in a practical way to meet these objectives. This systematic sampling is consistent with probability sampling.
- Dr. Kalton said a carefully balanced design is needed for a time sample.

- Mr. Strauss noted that the sample size can be increased in many of the PSUs with provider sampling Version A. In certain locations, the sample size could be increased by going outside the PSU, which might be operationally feasible but would not have the same statistical efficiency.
- Dr. Sinclair commented on the extant data comparison. The NCS Program Office has made a request for a probability-based sample evaluation system to look at whether the providers recruited in the PBS pilot are representative of the provider population as whole and whether the participants are representative of the population as a whole. For Version A, in which women within a PSU are being sampled, there is direct comparison between the birth data and participants to assess whether participants are representative. Whether the providers are representative will have to be determined by the local contract Study Centers. Assessing characteristics of providers within a PSU would not be an argument for preferring Version B over Version A.
- Mr. Strauss explained that if the preconception cohort can be identified as a random subset of the rest of the Study cohort, then information would be missing at random from women without preconception measures. This missing information could be leveraged across the entire cohort. Preconception exposure measures could then be compared with exposures later in pregnancy or closer to the time of birth. Moving away from a probability-based sample makes it more difficult to leverage the preconception information in an effective way. The preconception cohort should be considered in a probabilistic way and consistent with sampling of other women.
- Dr. Schisterman noted that there will be feasibility issues for recruiting a preconception cohort of 30,000 women. Current estimates for the percentage of couples who are planning pregnancy are much lower than previous estimates (0.01 percent versus 1–3 percent). In addition, although some pregnancies are planned, others are not, which are more challenging to capture.
- Mr. Strauss said preconception information could be leveraged using a smaller probability-based cohort, with restricted samples. Although not a random sample, women who are seeking assistance for infertility could be probabilistically sampled. The preconception cohort also could include subsequent siblings of children already enrolled in the Study. The preconception cohort should be located in the same geographic area as the rest of the cohort.
- Dr. Schisterman commented that probability sampling for a preconception cohort may not be as feasible as probability sampling of a pregnancy cohort. Data may have to be collected within short intervals around the time of conception.
- Dr. Madans commented that, ideally, the entire cohort should be preconception. However, recruiting a preconception cohort of 100,000 women would be very expensive, very challenging, and logistically not possible. What still needs to be resolved is the acceptable window for collecting data during pregnancy. Recruiting/enrolling women who are trying to become pregnant may be biased. Women who are already enrolled in the Study may have a

greater probability of becoming pregnant. A preconception cohort drawn from these women may be acceptable and should be considered.

- Dr. Hirschfeld said preliminary modeling has shown that about 40,000–50,000 women would need to be followed over a 3-year period to enroll 15,000 into the preconception cohort.
- Dr. Sinclair asked whether historic/extant environmental data and medical records have been considered in order to assess preconception exposures.
- Dr. Hirschfeld said the use of extant data has been considered but there are concerns about the reliability of all these data sources. Self-reported and recall data are particularly unreliable. Biosamples may be of some value but may not have the same degree of precision as other measures. Extant data for environmental exposures could be informative and could supplement Study data but would not replace data collected by the Study.
- With regard to women who change providers, Dr. Kalton said they need to be followed, because they—and not the providers—are the source of Study data. Following women who change providers should not be an issue.
- Dr. Madans said women who move need to be followed, regardless of how they are recruited into the Study.
- Dr. Hirschfeld described the “light touch” cohort in the supplemental sample questions as women who are interested in participating in the Study but may not be enrolled in the probability sample. No biospecimen or environmental samples would be collected from these women.
- Dr. Olsen said there may be an operational issue for women who change providers to a noncooperating provider, if data collection is needed from these providers. He also said that increasing the size of the PSUs may be one approach to following women who move.

## **Supplemental Sampling**

### **Questions**

- Example 1—women without prenatal care access:
  - Would it be adequate to recruit women from hospitals or birthing centers, excluding those who received prenatal care?
  - Could a sample like this be recruited as a convenience sample?
  - Would the women recruited in this way be considered a substudy, or could they be a part of the larger sample?
  - When should a supplemental sample become a separate sample frame?
  - Could meta-analysis techniques be used to combine a supplemental frame with the larger probability-based cohort? For example, if a hypothesis was posed about left-handedness and an exposure, could the information be pooled from both cohorts with regard to the exposure-outcome relationship?

- What sources of bias could be anticipated by introducing the supplemental frame?
- Example 2—deficiency of lower income women in the sample frame
  - Could Women, Infants, and Children (WIC) Program providers or other list frames be added to the provider frame, with exclusion criteria for women who have already seen one of the selected providers?
  - Could a sample like this be recruited as a convenience sample?
  - Would the women recruited in this way be considered a substudy, or could they be a part of the larger sample?
  - When should a supplemental sample become a separate sample frame?
  - Could meta-analysis techniques be used to combine a supplemental frame with the larger probability-based cohort? For example, if a hypothesis was posed about left-handedness and an exposure, could the information be pooled from both cohorts with regard to the exposure-outcome relationship?
  - What sources of bias could be anticipated by introducing the supplemental frame?

## Discussion

- Mr. Strauss said a preconception cohort could be considered as a supplemental sample because recruiting preconception women through provider-based sampling will be challenging. Recruiting other difficult-to-reach women (for example, women without access to health care) through providers will also be challenging.
- Dr. Madans commented that it may not be possible to incorporate data from supplemental sampling into the Main Study in a reasonable way. However, because the provider-based sampling design will leave out some unknown number of women with characteristics that are important to the Main Study, supplemental sampling is necessary. It will be particularly challenging to combine data from supplemental sampling and the Main Study for all Study locations for a nationally representative sample.
- Mr. Strauss noted that a proportion of women receive prenatal care through free clinics and WIC clinics, not through obstetrician-gynecologists. In PSUs with lower socioeconomic populations, the provider-based sampling frame will need to be expanded—using probability-based sampling—to include clinics and practices other than obstetrician-gynecologists. There is a role for creatively leveraging other practices that provide prenatal care.
- Dr. Kalton said women who do not receive prenatal care through obstetrician-gynecologists can be recruited/enrolled at the time of birth in hospitals.
- Dr. Krotki said oversampling of certain target populations could occur at the provider level.
- Dr. Schisterman asked whether the Study would oversample at certain locations to compensate for undersampling at other locations in order to achieve samples that are representative of the U.S. population.

- Dr. Madans replied that that this is not so much a sampling issue but that recruiting is missing women with certain characteristics. By adding these women, the sample would no longer be representative. She explained that in order to have a representative sample, the Study will have to sample at every place that women receive prenatal care and enroll them at the appropriate time (for example, at 8 weeks of pregnancy). If 30 percent of the women are enrolled in the last trimester, the cohort will not be an early pregnancy cohort. The data will not be consistent for every woman in the sample. The cohort will not be representative of the risks that the Study wants to measure (that is, exposures early in pregnancy). The Study may have to redefine its objectives if women are not enrolled until later in pregnancy.
- Mr. Strauss commented that the sample will have missing data that is nonrandom and nonresponse that is nonrandom.
- Dr. Hirschfeld asked how the Study could establish a complete sampling frame as the first step, that is, a provider list that is as complete as possible. Having a complete list would be prioritized as an essential component of design. If the list is incomplete, the Study will have to understand exactly what is missing and determine ways to handle it.
- Mr. Strauss replied that the providers list should not be restricted to obstetrician-gynecologists. Every physician that is providing prenatal care in reasonably large numbers should be included.
- Dr. Madans said that, for assessing data validity, there should be evidence that the provider list or some percentage of it is complete, that the percentage of providers that were willing to participate is known, that the percentage of women who were willing to participate was known, and that the time (that is, at first trimester, third trimester, or birth) of enrollment was known.
- Dr. Hirschfeld asked whether the ability to generate a complete provider list can be a stratification factor in establishing the area sample.
- Dr. Kalton said generating provider lists is not a key problem, as shown in the pilot studies. However, assessing noncoverage is an issue. Noncoverage can be determined by matching sample data with county birth records. The challenge with provider-based sampling is whether women can be enrolled early in pregnancy (for example, by the 8th week) and complete the first interview. If only a certain percentage can be enrolled early in pregnancy, the Study may have to reconsider its key objectives for measuring exposure-outcome relationships during pregnancy.
- Dr. Schisterman said he does not consider supplemental sampling as a substitute for the Main Study sample. Supplemental sampling could be used for existing frames so that collecting additional data is not as expensive and the data could be used for other purposes.
- Dr. Little commented that he is not in favor of trying to enroll additional women early in the first trimester through other sampling strategies. Adding additional operations could potentially add risk. Although some women may not be enrolled early in pregnancy, some

information could be collected retrospectively and still address some hypotheses. So it is important to include women who are enrolled at delivery.

- Dr. Hirschfeld asked whether supplementary cohorts (for example, to focus on particular exposures or biological characteristics) should be considered analytically separate studies and whether information from a supplementary cohort could be leveraged to inform the Main Study cohort.
- Dr. Madans said there are two ways to approach supplementary cohorts. The sample size in a PSU could be increased, using the same sampling design and infrastructure. Increasing the sample size would not be difficult but would increase costs. The women would not be part of the Main Study sample but would be a subset. Data from this subset and Main Study subset could be combined for analyzing the exposure or outcome of interest. However, if data collection for the supplementary sample is separate from the main sampling design (for example, outside a PSU), the same overall data collection protocol should be used. Data could be compared but would not be combined.
- Dr. Krotki noted that integrating supplementary samples into the Main Study would not be mathematically challenging, as long as it is a probability sample and not a convenience sample.
- Mr. Strauss commented that it would be challenging to integrate into the Study a supplementary sample of women who are seeking *in vitro* fertilization and appropriately combine the data analytically.

## **Missingness by Design**

### **Questions**

The Study is considering the use of a core questionnaire for all participants, with additional modules or data sets for subsets of the Study population.

- Can the Study have different questionnaire intensities within the larger frame?
- Can the Study have different questionnaire intensities between the supplemental frames and the probability sample?
- What would be the parameters for determining the sample sizes of women receiving low-intensity and high-intensity instruments?

### **Discussion**

- Mr. Strauss said the Study can have different questionnaire intensities within the larger frame, and he encouraged the Study to do so. With regard to the second question, one of the central tenets of missingness by design is the assumption of missing by randomness. There are ways of doing outcome-dependent and covariate-dependent sampling in a resource-efficient manner.

- With regard to the third question, Dr. Kalton commented that in order to get a good response, the burden should be equalized in some fashion. He noted that several years ago an expert panel concluded that matrix sampling for the Study would not be appropriate for the design proposed at the time.
- Mr. Strauss explained that the low-intensity/high-intensity approach captures surrogate information across the entire cohort and then is carefully designed to capture the more expensive gold-standard information on a subset of participants. Inferences about the low-intensity participants are made as if the gold standard information was measured on the entire cohort by doing a proper measurement error analysis adjustment. Probability-based sampling is used to determine which participants get gold standard measures. The correlation between the surrogate information vs. the gold standard data is a vital component of the analysis.
- Dr. Little cited an example from the Health and Retirement Study called the AHEAD (Asset and Health Dynamics among the Oldest Old) supplement, which assessed mental health. Simple measures were used for entire study cohort, whereas elaborate measures were used for the supplementary sample.
- Dr. Madans said other types of population-based sampling could be included if, after evaluation, the provider-based sampling cannot provide the full range of needed data. The population-based sampling would be less intensive but could still provide basic information to augment provider-based sampling information. Data from the two sources would not be combined.
- The discussion concluded with Drs. Schisterman, Hirschfeld, and Guttmacher thanking all the participants and noting that their input was informative and constructive.
- The next scheduled public discussion for the Main Study design will be the NCSAC meeting on July 24, 2012. The briefing document and other information for this meeting is available at [http://www.nationalchildrensstudy.gov/about/organization/advisorycommittee/Pages/July\\_2012.aspx](http://www.nationalchildrensstudy.gov/about/organization/advisorycommittee/Pages/July_2012.aspx).

## Panel

Anjene Addington, Ph.D., M.P.H., Booz Allen Hamilton Inc.  
 Ned English, M.S., NORC at the University of Chicago\*  
 Alan E. Guttmacher, M.D., NICHD, NIH, HHS\*  
 Brian Harris-Kojetin, Ph.D., Statistical and Science Policy Office, OMB\*  
 Steven Hirschfeld, M.D., Ph.D., National Children's Study, NICHD, NIH, HHS  
 Graham Kalton, Ph.D., Westat  
 Karol Krotki, Ph.D., RTI  
 Jennifer Kwan, Ph.D., National Children's Study, NICHD, NIH, HHS  
 Roderick Little, Ph.D., U.S. Census Bureau\*  
 Jennifer Madans, Ph.D., NCHS, CDC, HHS  
 Teri Manolio, MD, PhD, National Human Genome Research Institute, NIH, HHS\*

Randall J. Olsen, Ph.D., Ohio State University  
Charles Pierret, Bureau of Labor Statistics, U.S. Department of Labor  
Enrique Schisterman, Ph.D. (panel moderator), NICHD, NIH, HHS  
Iris M. Shimizu, Ph.D., NCHS, CDC, HHS\*  
Michael D. Sinclair, Ph.D., NORC at the University of Chicago  
Edward J. Sondik, Ph.D., M.S.Hyg., NCHS, CDC, HHS  
Warren Strauss, Sc.M. Battelle  
*\*Participated by phone*

## **Observers and Other Participants**

Hibest Assefa, M.P.H., Johns Hopkins University Bloomberg School of Public Health\*\*  
Dean Baker, M.D., M.P.H., University of California, Irvine\*\*  
Baylor University\*\*  
John Bourgeois, M.P.H., Tulane University School of Public Health and Tropical Medicine\*\*  
Mary Boyd, University of Louisville\*\*  
Ruth A. Brenner, M.D., M.P.H., NICHD, NIH, HHS  
Nicole Cederblom, University of Washington\*\*  
Tahleah Chappel, M.S., Johns Hopkins University Bloomberg School of Public Health\*\*  
Gilles Charest, University of Massachusetts Medical School\*\*  
Suzanne Cox, Ph.D., M.P.H., Northwestern University\*\*  
Michael J. Dellarco, Dr.P.H., National Children's Study, NICHD, NIH, HHS  
Melody Drnach, Brown University\*\*  
Ned English, M.S., Johns Hopkins University Bloomberg School of Public Health\*\*  
Barbara Entwisle, Ph.D., University of North Carolina, Chapel Hill\*\*  
Elaine Faustman, Ph.D., University of Washington\*\*  
Michele Forman, Ph.D., M.S., University of Texas at Austin\*\*  
Bruce Gale, M.S., University of Louisville\*\*  
John K. Gohagan, Ph.D., Office of the Director, NIH, HHS  
Jessica E. Graber, Ph.D., National Children's Study, NICHD, NIH, HHS  
Jay Greenfield, M.A., Ph.D., Booz Allen Hamilton Inc.\*\*  
Deborah Hendricks, R.N., M.P.H., University of Minnesota\*\*  
Charlotte Hobbs, M.D., Ph.D., University of Arkansas for Medical Sciences\*\*  
Lisa Kaeser, J.D., NICHD, NIH, HHS\*\*  
Carol H. Kasten, M.D., National Children's Study, NICHD, NIH, HHS  
Jean Kerver, Ph.D., M.S., Michigan State University\*\*  
Liz Langlois, M.S.P.H., Tulane University School of Public Health and Tropical Medicine\*\*  
Colleen Lee, M.S., National Children's Study, NICHD, NIH, HHS\*\*  
Maria Lopez-Class, Ph.D., M.P.H., National Children's Study, NICHD, NIH, HHS  
Eric Lorenzo, Ph.D., National Children's Study, NICHD, NIH, HHS  
John Lumpkin, M.S., M.B.A., National Children's Study, NICHD, NIH, HHS\*\*  
William Lyman, Ph.D., Wayne State University School of Medicine\*\*  
Maine Medical Center\*\*  
John McGrath, Ph.D., National Children's Study, NICHD, NIH, HHS\*\*  
Nolen Morton, J.D., (Contractor), National Children's Study, NICHD, NIH, HHS\*\*  
Jack Moye, Jr., M.D., National Children's Study, NICHD, NIH, HHS\*\*

Mignon Murray, M.B.A., University of Texas Southwestern Medical Center at Dallas\*\*  
Sharon Nuss, University of Louisville\*\*  
Nigel Paneth, M.D., M.P.H., Michigan State University  
Nancy Parfitt Hondros, National Children's Study, NICHD, NIH, HHS  
Christina H. Park, Ph.D., National Children's Study, NICHD, NIH, HHS  
Susan Schechter, M.A., NORC at the University of Chicago  
Deidre Sepavich, M.B.A., University of Massachusetts Medical School\*\*  
Becky Smith, University of New Mexico\*\*  
Gitanjali Taneja, Ph.D., National Children's Study, NICHD, NIH, HHS  
University of Miami\*\*  
Meredith Wadman, B.M.B.Ch., *Nature*\*\*  
Kevin Wafula, Johns Hopkins University Bloomberg School of Public Health\*\*  
Kate Winseck, M.S.W., National Children's Study, NICHD, NIH, HHS\*\*  
Sharon Wyatt, Ph.D., University of Mississippi Medical Center\*\*  
Yale University\*\*  
Kendall Ziegler, M.P.H., Michigan State University\*\*  
*\*\*attended by Webinar*