Appendix A

Advantages and Limitations of Probability-based Sampling for the NCS

February 2004

White Paper

on

Advantages and Limitations of Alternative Sampling Methods for the National Children's Study

by

Steve Rust

with input from

Jeff Lehman, Nancy McMillan, John Menkedick, Colm O'Muircheartaigh, Louise Ryan, Lowell Sever and Warren Strauss

BATTELLE
505 King Avenue
Columbus, OH 43201-2693

Contract No. 282-98-0019

Dr. Peter Scheidt
Project Officer

National Children's Study Program Office
National Institute for Child Health and Human Development
6100 Executive Blvd - 5C01
Rockville, MD 20892

Table of Contents

GLOSSARY OF TERMS A-4
A-1. INTRODUCTION A-9
A-2. VALIDITY A-12
A-3. A HISTORICAL PERSPECTIVE A-14
A-4. ADVANTAGES AND LIMITATIONS OF ALTERNATIVE SAMPLING STRATEGIES A-16
A-4.1 Process for Drawing CONCLUSIONS from NCS Data A-16
A-4.2 Assumptions Related to Advantages and Limitations A-21
A-4.3 Advantages and Limitations of Probability-Based Random Sampling A-22
A-4.4 Advantages and Limitations of Non-Probabilistic Sampling A-25
A-5. NCS REQUIREMENTS AND HYBRID STRATEGIES A-27
A-5.1 Probability-Based Purposive Sampling A-28
A-5.2 Probability-Based Purposive Exclusion A-29
A-6. CONCLUSIONS A-31
A-7. REFERENCES A-32

Appendices

Appendix A. Rationale for Sampling from Subpopulations A-34

Tables

Table 1. Advantages and Limitations of Alternative Sampling Strategies A-17

Figures

Figure A-1. Context for NCS Sampling Design A-10

Glossary of Terms

Cluster Sampling: A method of sampling in which, at some stage, elements (e.g., children) are selected from the population in groups or clusters. In multi-stage cluster sampling, a sample of elements within a selected cluster may be taken during a subsequent stage of sampling.

Design Variables. The set of variables required to implement a *probability-based sampling* process, including stratification variables and any variables used to calculate probabilities of inclusion.

External Validity. Relationships identified in a study are considered to be externally valid if they are valid for the reference population associated with the study.

Internal Validity. Relationships are considered to be internally valid if they are statistically significant for the study sample, if the effects of extraneous variables, plausible confounders, and plausible effect modifiers have been properly accounted for, and if hypothesized causal factors precede the effect.

Multi-Stage Sampling. Multi-stage sampling methods allow selection of groups of elements from the sampling frame at one stage and then subsequent sampling from the selected groups of elements at a subsequent stage.

NCS Cohort. The *study sample* for the National Children's Study.

Probability-Based Purposive Exclusion: A probability-based method for assuring that specified elements of the study population are excluded from the target sample with certainty.

Probability-Based Purposive Sampling: A probability-based method of assuring that specified elements of the sampling frame are included in the target sample with certainty.

Probability-Based Random Sampling: A probability-based sampling method for which each element has a probability of being included in the target sample that is strictly greater than zero and strictly less than one, and that uses a random procedure to select elements into the *target sample* according to these probabilities.

Probability-Based Sampling: A method for selecting a *target sample* from a *sampling frame* in which the probability of occurrence for each and every possible *study sample* is a function of a set of *design variables*; an important property of a probability-based sampling process is that the probability of

inclusion in the *target sample* is known for each and every element (e.g., child) in the *sampling frame*.

Quota Sampling: A method of sampling in which certain characteristics of potential study participants are measured and participants are included in the *study sample* in such a manner as to obtain pre-determined numbers of participants in specified classes defined by values of the measured characteristics.

Recruitment Rate: The ratio of the number of subject initially enrolled in the NCS cohort divided by the number of subjects for which a recruitment attempt is made.

Reference Population: The population about which valid inferences are desired and to which study inferences will be extrapolated in one form or another.

Relative Risk Ratio: The proportion of diseased people among those exposed to a relevant risk factor divided by the proportion of diseased people among those not exposed to a relevant risk factor.

Response Rate: The ratio of the number cohort members providing sufficient data for a particular line of inquiry divided by the number of cohort members for which an attempt is made to collect such data.

Retention Rate: The ratio of the number of actively enrolled cohort members at a given point during the data collection phase of a study divided by the number of cohort members initially enrolled.

Sampling Frame: That portion of the *study population* that has a positive probability of being included in the *target sample*; in practice, the sampling frame is constructed to be as close to the *study population* as possible subject to the requirements that (1) the sampling frame can be fully enumerated and (2) *design variable* values are available for each element of the sampling frame.

Simple Random Sampling. Simple random sampling methods select the target sample from the sampling frame in a totally random fashion without replacement.

Stratified Sampling. Stratified random sampling methods control the sub-sample sizes for subsets (strata) of the sampling frame defined by one or more design variables.

Study Population. The population of elements that would be included in the *sampling frame* if full enumeration of the sampling frame and values for the design variables were not required.

Study Sample. All elements of the study population that are successfully recruited into the study, are successfully retained as study participants, and produce the required study data.

Target Sample. Those elements of the study population for which a recruitment attempt is made; the target sample is the union of the study sample, the set of recruitment failures, the set of retention failures, and the set of retained study participants that fail to produce the required data.

White Paper

on

Advantages and Limitations of Alternative Sampling Methods for the National Children's Study

Steve Rust[1]

with input from

Jeff Lehman[1], Nancy McMillan[1], John Menkedick[1], Colm O'Muircheartaigh[2], Louise Ryan[3], Lowell Sever[1], and Warren Strauss[1]

A-1. INTRODUCTION

The purpose of this white paper is to summarize and assess the advantages and limitations of employing probability-based and non-probabilistic sampling methods when selecting a cohort of children (NCS cohort) for long-term follow-up in the National Children's Study (NCS). As a background for this assessment, we first address a context for design of the NCS sampling protocol.

In the terminology of Cochran (1977), the NCS is primarily an "analytical" survey or study rather than a "descriptive" or enumerative survey or study. The term "descriptive study" is used here to refer to a study having the objective to describe a population of interest in terms of measurable characteristics. The term "analytical study" is used here to refer to a study having the objective to identify differences between subpopulations of a population of interest. For the NCS, subpopulations will be defined by levels of exposure and the differences will be characterized in terms health and developmental outcomes. Often the ultimate purpose of an analytical study is to take action on the cause-and-effect system(s) underlying identified relationships with the aim of improving future conditions (Hahn and Meeker (1993)). While the NCS will necessarily focus on a population of contemporary children, by the time relationships are identified in the NCS data, it will in most cases be too late to take effective action to improve the health and development of the children in this contemporary population.

Figure 1 illustrates a context within which to consider various sampling design options for the NCS. Under any design scenario, data from the NCS cohort will be

[1] Battelle
[2] University of Chicago
[3] Harvard University

analyzed to test multiple hypotheses regarding environmental exposures in the "broadest sense" to identify potential cause-and-effect relationships between environmental exposures and health and developmental outcomes. A long-term objective of the NCS is to influence public health policy and social behavior to bring about the application of effective environmental, behavioral and medical interventions. Such interventions, when applied to a future population of children in the US, should lead to improved health and developmental well-being.

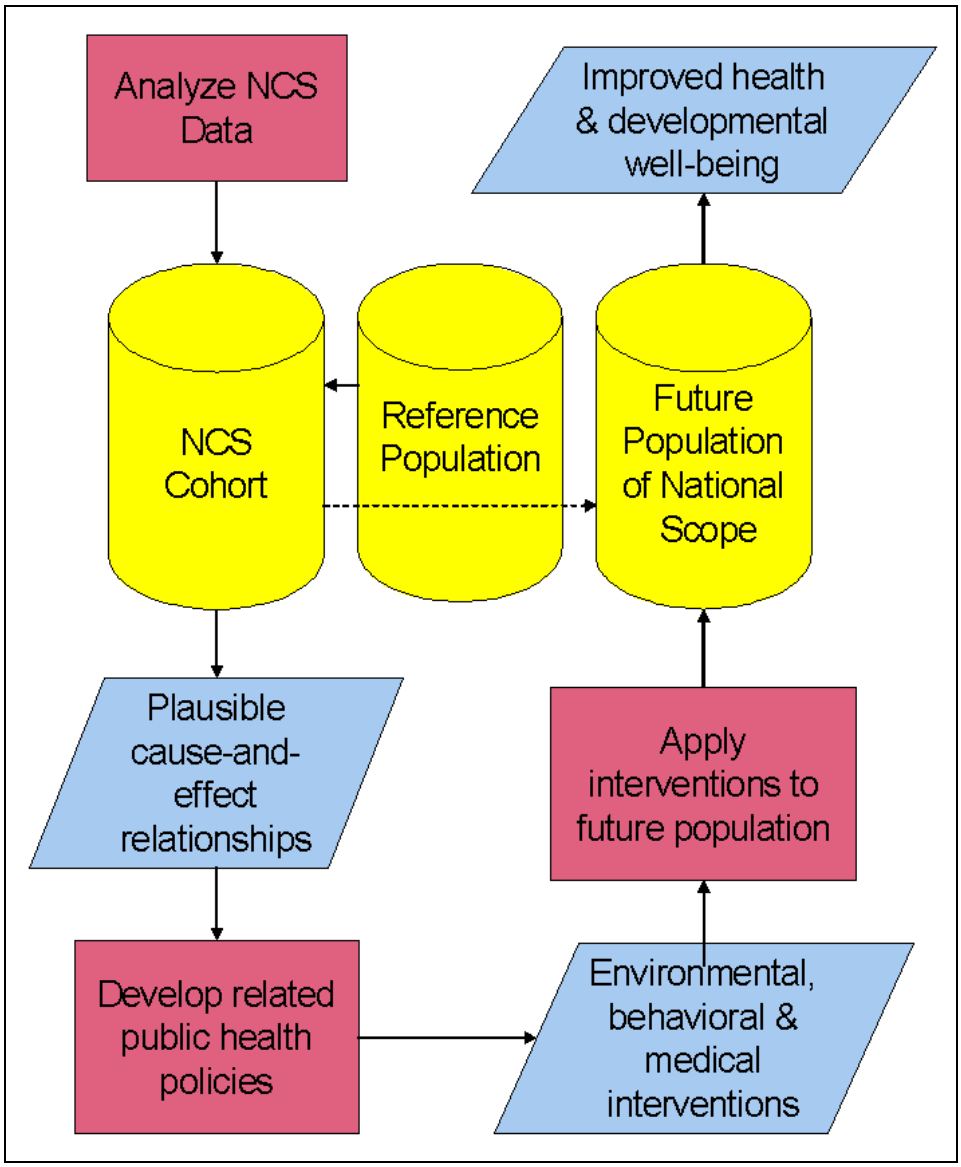

Figure A-1. Context for NCS Sampling Design

Consider the example of characterizing the relationship between pesticide exposure and the presence of autism at a specified stage of development. By the time any firm conclusions have been drawn from the NCS data regarding the relationship between

pesticide exposure and autism, all children who had an opportunity to participate in the NCS will have already experienced the pesticide exposure period in question and any increase in the likelihood of autism will have already taken its toll. Thus, it is logical in this example to focus one's ultimate attention on a future population of children for whom intervention to reduce pesticide exposure is possible.

In order to have an impact on the health and well-being of children in the US, the relationships identified in the NCS data will have to be valid for a future population of children in the US for which some form of intervention is possible. But given that this future population of children cannot be studied, it is logical to instead seek relationships that are valid for the current population of children in the US and rely on similarities between current and future populations of children as well as the external validity of models developed from the study to extend the validity of the relationships into the future. Therefore, in this paper, we will consider the population of children born in the US during the NCS enrollment phase to be the reference population for the study.

External validity (Campbell and Stanley, 1963) refers to the validity of relationships identified in the NCS data when extended to this reference population. As illustrated in Figure 1, the reference population serves as a stepping stone between the NCS cohort and the future population for which health and developmental benefits are sought.

Internal validity (Campbell and Stanley, 1963) refers to validity of identified relationships within the restricted context of the NCS cohort. Internal validity derives from several conditions:

- Identified relationships must be statistically significant,
- The cause must temporally precede the effect in the cause-and-effect relationship, and
- The effects of extraneous variables, plausible confounders, and plausible effect modifiers have been properly accounted for.

Imposing conditions of internal validity forces one to consider not just the statistical significance of hypothesized relationships but also the nature of those relationships. The concepts of external and internal validity as they apply to the NCS are explored in Section 2.

There are numerous sampling strategies that could be implemented to select the NCS cohort. In this paper we will first focus on two strategies that have received much attention during the early discussion of design options for the NCS:

- Probability-based random sampling
- Non-probabilistic sampling

In approximate terms, probability-based random sampling strategies select cohort members according to a structured random process that results in every child in the

reference population having a known probability between zero and one of being included in the NCS cohort Non-probabilistic strategies put few or no constraints on the selection of children for the NCS cohort.

In Section 3, we present a historical perspective by first summarizing arguments that have been presented to advocate use of probability-based sampling when selecting the NCS cohort followed by a summary of arguments that have been similarly presented to advocate the use of non-probabilistic sampling. Specific advantages and limitations of the two sampling approaches are discussed in Section 4. These advantages and limitations address issues including the feasibility of constructing a sampling frame, ability to obtain a range of exposures, subject recruitment rates, subject retention rates, data quality, cost-efficiency of data collection, internal validity and external validity.

It is not necessary to restrict the design of the NCS to either probability-based random sampling or non-probabilistic sampling as one of two choices. In fact, practical implementation of the NCS may inevitably result in some mix of the two approaches. In Section 5 we introduce two additional hybrid sampling strategies that can be customized to meet specific requirements of the NCS. The methods are:

- Probability-based purposive sampling
- Probability-based purposive exclusion

Probability-based purposive sampling allows specified members of the reference population to be included in the NCS cohort with certainty. In a complementary fashion, probability based purposive exclusion allows specified members of the reference population to be excluded from the NCS cohort with certainty.

The discussion of hybrid strategies in Section 5 serves as a backdrop for the conclusions presented in Section 6. Finally, references are provided in Section 7.

A-2. VALIDITY

When traditional survey sampling methods are applied during the design phase of an analytical study, the driving force behind decisions is the pursuit of external validity for relationships identified in the study data. Probability-based random sampling methods are employed to select a sample from a sampling frame constructed to include as much of the reference population as possible. If the sample is so selected and response rates are high, then relationships identified in the study data can be validly extended to the reference population based on the random sampling mechanism employed to select the sample. Thus, traditional survey sampling methods are based on the concept of external validity through probability sampling and, therefore, are the methods of choice if external validity is of primary importance. Traditional survey sampling methods have emphasized the statistical significance of relatively simple associations, with little attention to the subject matter nature of the relationships themselves and detailed modeling of the relationships.

In contrast, when traditional epidemiological methods are applied during the design phase of an analytical study, the primary driving force behind decisions is often pursuit of internal validity for relationships identified in the study data. While issues of statistical significance are certainly important drivers of traditional epidemiological studies, much more attention is given the subject matter nature of hypothesized relationships. Does the exposure temporally precede the outcome? Have potential confounders and effect modifiers been measured and ruled out? What is the hypothesized nature of the relationship and how might this affect timing and extent of measurement? Internally valid relationships between exposures and outcomes are more likely to lead directly to intervention concepts. While traditional epidemiological methods embrace external validity as an important objective, external validity is often a secondary objective with the majority of emphasis given to internal validity. Traditional epidemiologists may also rely on consistency with subsequent studies to confirm models developed from an internally valid study.

It is quite conceivable that a team of scientists trained in traditional survey sampling methods and a team of scientists trained in traditional epidemiological methods, when given exactly the same context for designing the sampling protocol for the NCS, would design studies that have vast differences. Survey sampling methods would emphasize coverage of the reference population and considerable study resources might very well be devoted to sampling hard-to-study elements of the reference population, conversion of reluctant participants, and proper handling of non-responders to maintain representativeness. By explicitly devoting considerable study resources in this manner, fewer study resources would be available for data collection, adversely affecting the level of detail possible in the data collection protocol. This effect on level of detail is likely to be somewhat implicit in nature.

In contrast, epidemiological methods would emphasize the level of detail in the data collection protocol because, without that detail it would be impossible to properly assess the role of potential confounders and effect modifiers. By devoting considerable study resources to the data collection protocol, fewer resources are available for assuring the representativeness of the study sample. This effect on representativeness is also likely to be implicit in nature.

To further explore the issue of internal versus external validity, consider a simple linear regression model that explores the relationship between an adverse health effect and some measure of exposure. In a study the size of the NCS, it is certainly possible to identify statistically significant relationships between disease and exposure without explaining a large percentage of the variability in the response variable (as characterized by the R^2 statistic). When attempting to extrapolate these results to a larger reference population, we may feel uncomfortable about the factors left unexplained by this model absent a probability-based sampling design which allows us to assume that the sample is unbiased relative to the reference population. In contrast, it is also possible to explore relationships in which a very large percentage of the variability is explained, perhaps using a more complex model that includes covariates, confounders and effect modifiers. In this situation, it may be reasonable for scientists to conclude, based on the defensibility

of the model, that the relationships observed are unbiased relative to the reference population regardless of the mechanism by which participants are recruited into the sample.

A-3. A HISTORICAL PERSPECTIVE

An influential group of scientists involved with the early planning and discussion on an optimal sampling design for the NCS, including the NCSAC, have strongly advocated a probability-based sampling approach for the NCS. Without a significant probability basis to the sampling, they maintain that 1) at the end of the study scientists will be forced to say that they have no idea what population the study results are generalizable to; 2) that this will undermine the scientific credibility of the entire study; and 3) that the unknown and unknowable biases that might be introduced through convenience sampling could lead to false conclusions. Their concern is particularly highlighted for the social environment and behavioral assessments in the NCS where relevant exposure and risk factors are less well-characterized and where unknown biases may be more likely.

There are several positive aspects of probability based sampling that have been stressed, including:

1. Each child in the NCS cohort represents a known number of children in the sampling frame

2. Each child in the sampling frame has a known and positive probability of being selected into the NCS, providing some sense of political fairness to the process of sample selection

3. Probability-based sampling provides a feasible and scientifically defensible mechanism for applying scientific results observed in the NCS to a reference population of children without needing to worry about unintentional systematic biases introduced through the sampling mechanism.

Advocates of probability-based sampling contend that, with the broad range of health outcomes, potential exposures, and critical stages of vulnerability covered in the NCS, it will be operationally infeasible (both from a cost perspective and from the perspective of the burden placed on study participants) to assess all of the important risk factors, covariates, effect modifiers, and confounders across the entire cohort over time. If important factors cannot feasibly be observed within this study, the potential for misleading inferences due to a biased sampling approach become much more likely, increasing the value of a probability-based sample for drawing externally valid inferences.

In addition, while the study design for the NCS is currently hypothesis driven, much of the value of the NCS is based on its ability to support scientific discovery. The NCS will provide a rich source of observational data that will be explored by scientific experts in a variety of disciplines for decades. To support this future research, potentially on topics

completely unrelated to the original core hypotheses that currently form the basis for the study design, it is of paramount importance that the NCS sample be generalizable to a reference population with known characteristics.

Finally, advocates of probability-based sampling suggest that limitations in response rates or other sampling deficiencies do not imply that it is acceptable to use a more convenient, less expensive, or less demanding sampling method. It must, however, be recognized that low recruitment and retention rates will make it difficult to calculate meaningful sample weights from a probabilistic point-of-view.

While there has been a influential component of the scientific community that has advocated probability sampling as a requirement for the NCS as discussed above, there has also been an equally influential component, led by, but not limited to, epidemiologists, who have strongly advocated a non-probability approach as the only feasible way to conduct the study. This viewpoint begins with the recognition that the primary objectives of the NCS are related to understanding relationships between risk factors and disease, including understanding the etiology of disease. In order to maximize the likelihood that the study will provide information on etiology, it is necessary to maximize the amount of information on exposure, effect modifiers, covariates, and outcomes that can be collected, maximize the retention rate to observe effects over time, and therefore to maximize the internal validity of the study.

This group feels strongly that a design that is primarily selecting participants at random 1) will not be able to get its participants to agree to the level of burden that will be required to collect the necessary scope of information to reasonably understand the exposure-outcome relationships, and 2) that such a sample will also inevitably lead to such high attrition rates as to jeopardize study objectives related to outcomes or exposures arising later in childhood. In other words, they believe that probability sampling will jeopardize the internal validity of the study. From their perspective, probability sampling leads to a Catch 22. If the study collects sufficient exposure, covariate, and effect modifier data to do a good job of understanding relationships and etiology, then external validity can be achieved through trust in the model and probability sampling was not necessary; on the other hand, if there are associations observed that do not have the exposure, effect modifier, and covariate information well captured, then the statistical validity of the relationship is of little additional value in the overall scheme of the usefulness of study results. They also believe that for most chemical, biological or physical exposures, there is little basis to assume that associations would be significantly biased, if the participants are chosen from well known quotas, or strata, defined by characteristics such as SES, age, sex, and race. There is also a sense that the effective response and retention rates for a randomly-selected cohort will be so low as to result in what amounts to a convenience sample in the long run anyway.

In summary, advocates for non-probability sampling believe that internal validity will be compromised by extensive probability sampling, that external validity can be achieved by collecting better information to construct more defensible models, and that probability sampling will not be able to match the response rates, agreement for burdensome

measures, and retention of a convenience based sample, and will cost more. It should be noted that most advocates of this position acknowledge that if their goals of measurement, response and retention could be met equally well with probability sampling, they would recognize the statistical advantages offered by this approach.

A-4. ADVANTAGES AND LIMITATIONS OF ALTERNATIVE SAMPLING STRATEGIES

This section is devoted to describing and examining the advantages and limitations of probability-based random sampling and non-probability based sampling, attempting to capture and place into context an ongoing debate between advocates of each approach. Before addressing each sampling approach on its own merits we describe a conceptual process for drawing conclusions from NCS data so that the process steps can be used as an organizational structure for summarizing advantages and limitations and. we review some assumptions that underlie that discussion.

A-4.1 PROCESS FOR DRAWING CONCLUSIONS FROM NCS DATA

The advantages and limitations of alternative sampling methodologies tend to exist in advantage-limitation pairs. For example, employing probability-based sampling methods has disadvantages with respect to recruitment and retention that result in an advantage when issues of external validity are later addressed. In contrast, employing non-probabilistic sampling methods has advantages with respect to the recruitment and retention that result in a limitation when issues of external validity are addressed. We have found it useful to organize the set of advantages and limitations for each sampling method according to steps in a conceptual process for drawing conclusions from the NCS data. The conceptual process has the following steps:

1. **Sampling Design:** Select the children that will be targeted for recruitment.

2. **Recruitment:** Recruit targeted children into the NCS cohort.

3. **Retention:** Maintain continued participation of the children in the NCS cohort.

4. **Data Collection:** Collect study data for the children in the NCS cohort.

5. **Internal Validity:** Identify cause-and-effect relationships that are valid for the NCS cohort.

6. **External Validity:** Validly extend cause-and-effect relationships identified in the NCS data to the reference population of all children born in the US during the NCS enrollment period.

The second column of Table 1 contains a summary of advantages and limitations of probability-based sampling within the context of the NCS, organized by the conceptual

Table A-1. Advantages and Limitations of Alternative Sampling Strategies (Continued)

Process Step	Probability-Based Random Sampling	Purposive Sampling	Purposive Exclusion	Non-Probabilistic Sampling
Sampling Design	• **Advantage:** No requirement for a "representative" sample; includes mechanisms for over- and under-sampling specifically defined subsets of the sampling frame – Minorities – Children known to represent extremes with respect to exposure to suspected risk factors for health/developmental outcomes • **Limitation:** Requires the construction of a sampling frame; could be difficult in the case of pre-natal and pre-conception sampling	• **Advantage:** Can incorporate effects of constraints imposed by the Federal procurement system • **Advantage:** Can include – Geographically-isolated areas of exposure – Individuals known to have been exposed (useful in the case of rare exposures)	• **Advantage:** Can be used to exclude portions of the reference population that are difficult to enumerate	• **Advantage:** Since there are no constraints on how participants are located, simplifies the process of pre-natal or pre-conception sampling • **Limitation:** Over- and under-sampling can produce an undesirable sample mix – Mitigator: Quota sampling may be used to control characteristic discrepancies between the NCS cohort and the reference population or to assure sufficient numbers of cohort members in various characteristic classes to support using the associated characteristics as effect modifiers; quota sampling is only useful if all important effect modifiers and confounders are known in advance of sampling; because of low prevalence of many outcomes of interest, there may not be sufficient knowledge to do this – Mitigator: Can post-stratify to adjust for measured effect modifiers
Recruitment	• **Limitation:** Anticipated high participant burden may lead to a low successful recruitment rate; must contact many more people than you would enroll – Mitigator: Can over-sample children known to have desirable properties (e.g., existing relationships) with respect to recruitment	• **Advantage:** Sampling design may assign certainty status to subsets of the reference population that are easier to recruit	• **Advantage:** Sampling frame may exclude subsets of population expected to be more difficult to recruit	• **Advantage:** Volunteer participants simplify recruitment effort thereby reducing recruitment costs

Table A-1. Advantages and Limitations of Alternative Sampling Strategies (Continued)

Process Step	Probability-Based Random Sampling	Purposive Sampling	Purposive Exclusion	Non-Probabilistic Sampling
Retention	• **Limitation:** Actual high participant burden will lead to high levels of attrition and high levels of sporadic non-response – Mitigator: If no attempt is made to convert reluctant participants, attrition and sporadic non-response rates may be no worse than for non-probabilistic sampling – Mitigator: Can over-sample children known to have desirable properties with respect to retention and/or reliable response – Mitigator: Can use techniques to maximize retention (e.g., develop inter-personal relationships between staff and participants) • **Limitation:** Inability to include volunteers and/or convenient participants may result in less community involvement and therefore have an indirect negative effect on retention/response rates for probabilistically-selected participants	• **Advantage:** Sampling design may assign certainty status to subsets of the reference population that are easier to retain	• **Advantage:** Sampling frame may exclude subsets of population expected to be more difficult to retain	• **Advantage:** Allows use of volunteer participants that should have better retention and reliable response rates because of carry-over effects from their initial positive attitude toward study participation • **Advantage:** Allows use of volunteer participants in cases where the respondent burden is extreme leading to higher retention rates

Table A-1. Advantages and Limitations of Alternative Sampling Strategies (Continued)

Process Step	Probability-Based Random Sampling	Purposive Sampling	Purposive Exclusion	Non-Probabilistic Sampling
Data Collection	• **Limitation:** Use of study resources elsewhere may require that a simplified data collection protocol be employed • **Limitation:** If cohort is geographically dispersed, may be difficult to maintain data quality and/or cost-efficiency – Mitigator: Provides mechanisms (e.g., cluster sampling) for controlling the geographic dispersion of the sample (can be said for all approaches) • **Limitation:** Inability to include volunteers and/or convenient participants may result in less cooperation from organizations charged with data collection	• **Advantage:** Can take advantage of specialized facilities and equipment that are available on a very limited basis; similarly can take advantage of existing environmental data	• **Advantage:** Sampling frame may exclude subsets of population for which data collection is expected to be more expensive	• **Advantage:** Allows selection of cohort members that have a relationship with data collection organizations, perhaps leading to very cost-efficient data collection • **Advantage:** May be easier to schedule volunteers at times that are convenient or available to staff; volunteers may be more willing to travel.
Internal Validity				• **Advantage:** Better cost efficiencies may perhaps allow a more detailed protocol to better address – Extraneous variables – Confounders – Effect modifiers • **Limitation:** Limited ability to empirically check the validity of the assumed statistical model

Table A-1. Advantages and Limitations of Alternative Sampling Strategies (Continued)

Process Step	Probability-Based Random Sampling	Purposive Sampling	Purposive Exclusion	Non-Probabilistic Sampling
External Validity	• **Advantage:** In theory, inferences may be extended to sampling frame based on the random sampling mechanism employed to select the sample (with assumptions regarding non-respondents) • **Advantage:** Enables the use of model-based inference procedures by assuring that the sampling method does not select cohort members who represent a biased sample with respect to health and developmental outcomes • **Advantage:** Provides externally valid prevalence information for exposures and outcomes; this information may be required to – Estimate the number of children affected by policy changes and intervention strategies – Perform a cost-benefit, economic-health impacts analysis • **Limitation:** Imperfect recruitment and retention complicates extension of inferences to sampling frame – Mitigator: Statistical methods exist for dealing appropriately with non-response; require assumption that responders and non-responders are the same conditional on model covariates and effect modifiers or follow-up with a sample of non-respondents • **Limitation:** Useful inferences for under-sampled population subsets may be impossible	• **Limitation:** Useful inferences for under-sampled population subsets may be impossible	• **Advantage:** Reduced sampling frame is completely defined allowing an explicit assessment of the potential for systematic bias (are exclusion factors likely to be related to associations of interest?) • **Limitation:** Inferences to the full reference population are subject to systematic bias because it is difficult to say how the exposure-response relationships might differ for excluded groups – Mitigator: If less cost-effective subsets of population have been excluded, benefit of additional information collected may outweigh risk associated with systematic bias	• **Limitation:** The statistical significance of hypothesized associations cannot be based on statistical inference to a larger population • **Limitation:** Impossible to characterize the sources of systematic bias introduced by volunteer participants or even the larger population to which one might attempt to generalize study findings because differences between participants and non-participants are not defined; volunteer effect may be particularly important for the NCS because of the social environment and behavioral aspects of the study – Mitigator: Associations, particularly those that are more biological, chemical or physical in nature, may be less subject to external validation problems • **Limitation:** Does not provide externally valid prevalence information for exposures and outcomes – Mitigator: Perhaps this information can be obtained via another probability-based sampling mechanism (e.g., NHANES)

steps for drawing inferences from NCS data. Similarly, the last column of Table 1 summarizes the advantages and limitations of non-probabilistic sampling.

A-4.2 ASSUMPTIONS RELATED TO ADVANTAGES AND LIMITATIONS

In listing the advantages and limitations for each sampling method, there are general assumptions about the distribution of study resources and the benefits associated with certain types of over-sampling. Sampling designs that employ probability-based methods for identifying potential cohort members are assumed to have two characteristics: (1) they require that more study resources be devoted to sampling design, recruitment, and retention and (2) they will have lower retention rates resulting in data collection resources being devoted to larger numbers of early cohort members that eventually withdraw from the study. Both of these characteristics result in reduced resources being devoted to direct data collection for cohort members that participate throughout the duration of the entire study. Under various assumptions about retention rates and the manner in which probability-based sampling is implemented, the resources diverted from direct data collection can be estimated to be as small as $100 million or as large as $1 billion.

It is assumed that the effect of having reduced data collection resources would manifest itself either in the form of fewer cohort members or a simplified data collection protocol that would be less likely to include the measurement of confounders and effect modifiers associated with hypothesized relationships. Thus, sampling designs that employ probability-based methods for identifying potential cohort members were assumed to have an advantage with respect to extending statistically significant associations present in the NCS data to the reference population but to have a limitation with respect to demonstrating that true cause-and-effect relationships underlie these associations.

It is also assumed that sampling designs that target children with a wide range of exposures will result in better statistical power when assessing the significance of hypothesized relationships. In a similar fashion it is assumed that sampling designs that target children with a wide range of values for potential confounders and effect modifiers will result in better statistical power when attempting to refine statistical associations down to plausible cause-and-effect relationships.

Finally, it is assumed that relationships that have been refined through a process of carefully examining potential confounders and effect modifiers are likely to be more robust when attempts are made to extend those relationships to the reference population. This assumption is strengthened if the differences between the study population and the reference population are characterized and the variables that characterize the difference are considered to be potential effect modifiers.

A-4.3 ADVANTAGES AND LIMITATIONS OF PROBABILITY-BASED RANDOM SAMPLING

Probability-based sampling methodologies are methods for selecting a study sample from a sampling frame such that the probability of occurrence for every possible study sample is a known function of a set of design variables. An important property of a probability-based sampling process is that the probability of inclusion in the study sample is known for each and every element (e.g., child) in the sampling frame. In this section, we focus on probability-based sampling methods for which each element of the sampling frame has a probability of being included in the study sample that is strictly greater than zero and strictly less than one. Probability-based sampling methods that relax this restriction are discussed in Section 5.

There are many forms of probability-based random sampling including:

- Simple random sampling,
- Stratified random sampling,
- Cluster-based random sampling, and
- Multi-stage random sampling.

Simple random sampling methods select the study sample from the sampling frame in a totally random fashion without replacement. Stratified random sampling methods control the sub-sample sizes for subsets (strata) of the sampling frame defined by one or more design variables. Cluster-based random sampling methods allow sample elements to be selected in groups to improve the cost-efficiency of the data collection process. Multi-stage random sampling methods allow selection of groups of elements from the sampling frame at one stage and then subsequent sampling of elements from the selected groups at a subsequent stage.

Probability-based random sampling methods allow a tremendous degree of flexibility in setting inclusion probabilities for elements of the sampling frame. Subsets of the sampling frame can be easily over- or under-sampled simply by adjusting the inclusion probabilities associated with the elements of these subsets. This flexibility could be exploited, for example, to over-sample children with desirable properties with respect to cost-efficiency and/or data quality such as children living within 50 miles of a qualified medical center or children living in a geographical area with existing data on sources of environmental exposure. In a similar fashion, over-sampling can be employed to ensure that sufficient numbers of target minorities are included in the NCS cohort as well as to ensure children with a wide range of exposures are included in the NCS cohort. Over-sampling children with “desirable” properties necessarily results in under-sampling of children with “undesirable” properties.

Probability-based sampling designs derive a large portion of their value from the following characteristic. If the NCS cohort is selected from the reference population according to a probability-based sampling design, then inferences drawn from the NCS

cohort will be valid for the reference population as long as a sampling frame corresponding to the reference population can be constructed and the vast majority of children targeted by the probability-based sampling process are successfully recruited into the NCS cohort and retained until required study data have been collected.

Probability-based sampling is often considered to be a potentially cost-inefficient method for selecting a study sample because of expectations of low recruitment levels, high levels of attrition, and prohibitive study costs. In reality, this criticism may be leveled more at the way in which probability-based sampling is often employed rather than at probability-based sampling itself. If the need for the study sample to be "representative" of the entire population is paramount, then a disproportionate share of study resources may be devoted to obtaining equitable coverage of elements of the population that are subject to higher rates of non-response, higher costs for obtaining study information, or lower-quality study information. However, it is quite possible to stay within a probability-based sampling framework and tailor the sampling in a more cost-efficient direction. If response rates can be predicted as a function of the design variables, then elements of the population that are predicted to have higher response rates may be over-sampled. For example, suppose that a dozen major medical centers have agreed to participate in the NCS and leaders of the communities in which these medical centers reside have agreed to sponsor programs designed to emphasize the importance of individual participation in the NCS. A probability-based sampling plan could be designed to over-sample the children in these communities where community involvement is expected to result in higher recruitment and retention rates. While such an approach will increase the cost-efficiency of the study, it requires an up-front acceptance of the fact that other elements of the population will be underrepresented, perhaps to a point where strong statistical inferences cannot reasonably be extended to these under-represented elements.

In a departure from traditional survey sampling practice, it may actually be detrimental to attempt to convert reluctant participants to join the NCS cohort under the theory that converted reluctant participants are very likely to drop out of the NCS cohort before all required study data has been collected. As such, the resources devoted to the successful recruitment of these children as well as the resources devoted to collection of data for these children before they drop out of the study would be largely wasted. A better strategy may be to include only those children among those targeted by the probability-based sampling plan that are enthusiastic about involvement in the study. While such an approach would likely result in a low initial recruitment rate and higher recruitment costs, it is likely to have a very positive effect on the retention rate and minimize the magnitude of resources wasted on children who eventually drop out of the NCS cohort. The overall participation rate (combining recruitment and retention) for this approach may very well be similar to the participation rate that would be achieved via a more aggressive recruitment strategy.

Strict adherence to probability-based sampling would require that volunteers who learn of the study and wish to be included be turned away. Such a policy could have a negative effect on community involvement in the NCS which in turn might have a

negative effect on recruitment/retention rates, and the level of cooperation from organizations charged with data collection.

The geographic dispersion of a sample can be a major factor affecting its cost-efficiency. A widely dispersed sample can result in significant increased costs with respect to training data collectors, standardizing data collection methods, maintaining quality control with respect to specimen processing and analysis, and other similar data quality issues. One solution is to over-sample children affiliated with a smaller, more dedicated data collection mechanism. However, within the probability-based sampling framework, cluster sampling provides a mechanism for controlling the geographic dispersion of the study sample and effectively dealing with a number of cost-efficiency issues.

Over the past three decades, the statistical literature contains numerous contributions to an ongoing debate concerning the correct basis for statistical inferences from sampling survey data. For a sample of contributions to this design-based versus model-based inference controversy, see Scott and Smith (1973), Rubin (1976), Smith (1976), Scott (1977), Little (1982), Smith (1983), and Sugden and Smith (1984), Smith (1994), Kish (1995), Valliant, Dorfman and Royall (2000) and Little (2003). It is interesting to note that this entire debate takes place within a context that assumes the sample being analyzed was selected by a probability-based sampling method. Even advocates of model-based inference need to assume that the sampling method or mechanism does not select cohort members who represent a biased sample with respect to health and developmental outcomes. "Otherwise, the sampling mechanism needs to be modeled, and appropriate modeling in such cases is at best difficult. (Little, 2003)" Probability-based sampling methods provide the needed assurance that the sampling method does not select cohort members who represent a biased sample with respect to health and developmental outcomes.

Less than perfect recruitment, retention, and response rates erode the ability of probability-based samples to guarantee external validity. Two options exist for dealing with small amounts of departure from the targeted probability-based sample. The first option is to perform a second near-perfect[4] study of a sample of initial non-responders. This is not likely to be a viable option for the NCS. It may be possible to obtain follow-up information to assess basic differences (e.g., demographics, housing characteristics) of non-responders, but it will likely not be possible to obtain exposure-response information. The second option is to assume that, conditional on any model covariates, the sub-population of initial responders is unbiased relative to the reference population and apply methods such as the multiple imputation (Rubin (1987) and Rubin (1996)) to properly treat the missing data. The rate of recruitment and retention failures associated with the NCS may be too high for such methods to be effective, thereby reducing imputation to a model-based inference approach for improving external validity.

While this paper is focused on the very important NCS objective of identifying cause-and-effect relationships between environmental exposures and

[4] Nearly 100% successful recruitment of targeted participants

health/developmental outcomes, it is important to note that only probability-based samples can provide externally valid estimates of prevalence for specific environmental exposures and health/developmental outcomes. However, prevalence estimates may be alternatively available from other sources such as NHANES. Such estimates may be important to support cost-benefit analyses that must accompany public health policy development.

A-4.4 ADVANTAGES AND LIMITATIONS OF NON-PROBABILISTIC SAMPLING

In straightforward terms, a sampling method is a non-probabilistic sampling method if it is not a probability-based sampling method. Common non-probabilistic sampling methods include:

- Convenience sampling (e.g., recruiting patients as they arrive at a medical facility for otherwise scheduled appointments)
- Volunteer sampling (e.g., recruiting potential participants that respond to an advertised request for volunteers to participate in a study)

It is also worth noting the ways in which a sampling method can be non-probabilistic which include:

- The children in the sampling frame cannot be enumerated, or
- The probability of inclusion in the study sample cannot be determined.

Non-probabilistic sampling methods are characterized by a minimization of constraints on the methods used to identify and select study participants. This minimal level of constraints produces tremendous benefits in terms of simplifying the sampling design process and improving recruitment and retention rates. Study resources that might otherwise have been consumed for sampling design, recruitment, and retention can be diverted to data collection resulting in a more comprehensive data collection protocol. With more resources available for data collection, a more complex data collection protocol could be employed increasing the likelihood that data for potential confounders and effect modifiers would be available.

Non-probabilistic recruitment can emphasize volunteer participants who would be expected to exhibit much higher recruitment and retention rates. The respondent burden associated with the NCS data collection protocol **may** be so great that only enthusiastic volunteers could be expected to remain a part of the NCS cohort for the entire study period of more than two decades. In this case, probability-based sampling might be abandoned entirely in favor of a volunteer sample.

The benefits of non-probabilistic sampling during the sampling design, recruitment, and retention phases result in limitations during the data analysis phase. The primary limitations associated with non-probabilistic sampling are:

- Uncontrolled over- and under-sampling can produce an undesirable sample mix

- Statistical models of relationships may only be valid for the NCS cohort; validity beyond the NCS cohort must be assumed based on some other scientific criteria rather than the known sampling and probability characteristics of the study, and
- Identifying and assessing the potential sources of systematic bias is made difficult by the fact that the true sampling frame and true study population cannot generally be identified; it is difficult to assess differences between participants and nonparticipants given limited (or no) information about those who could have volunteered.

The first limitation may be partially mitigated by employing quota sampling methods. Quota sampling may be used to alternatively (1) control characteristic discrepancies between the NCS cohort and the reference population or (2) assure sufficient numbers of cohort members in various characteristic classes to support using the associated characteristics as effect modifiers. The success of quota sampling depends entirely on the assumption that important confounders and effect modifiers are known. Because of the low prevalence of many of the NCS outcomes, there may not be sufficient knowledge to identify these important confounders and effect modifiers a priori. As an alternative to quota sampling, post-stratification methods can be used after the fact to attempt to adjust for the effects of over- and under-sampling.

Since statistical inferences for non-probabilistic samples cannot be based on a random sampling mechanism for selecting cohort members, assumptions associated with model-based inference procedures may be invalid. Unfortunately, there is also limited ability to empirically check the validity of model assumptions. A key assumption underlying most statistical models in this situation is that the sampling method does not select cohort members who represent a biased sample with respect to health and developmental outcomes. With volunteer and other forms of convenience sampling, this assumption may not be valid and an undetected systematic bias may accompany any conclusions drawn. For example, in the case of pre-natal or pre-conception sampling, women with a prior history of or risk factors for reproductive health problems (including adverse pregnancy outcomes) might be more likely to volunteer for the NCS. Systematic bias related to a volunteer effect may be particularly relevant to the NCS because of the social environment and behavioral aspects of the study. The behavior of volunteering may be correlated with unmeasured aspects of the social environment that have a direct effect on health and developmental outcomes of interest. This correlation is a potential source of systematic bias when relationships identified in the NCS data are extended to the reference population. Relationships that are based on actual physical exposures and biological consequences of such exposures may not be as subject to such systematic biases. However, relationships that have behavioral components (e.g., time-location profiles and activities that affect the extent of contact with contaminants) may be particularly susceptible to systematic bias.

Since one objective of the NCS is to provide a data set that can be used in the future to test hypotheses that are not currently anticipated, the data set would have to include explicit warnings about the degree to which the data from a non-probabilistic sample can be generalized to any population. While one would hope that such warnings

would be duly noted and included in reports and publications of findings from the NCS data set, it is quite conceivable that the warnings would be largely ignored by at least a portion of scientists and researchers who use the NCS data set in the future.

A final limitation of non-probabilistic sampling is the inability to provide externally valid prevalence information for exposures and outcomes. Because acquiring such information is only a secondary objective of the NCS, this must be considered only a minor limitation. It is possible that other federally-funded health survey mechanisms (e.g., NHANES) represent better vehicles for obtaining prevalence information.

A-5. NCS REQUIREMENTS AND HYBRID STRATEGIES

In designing the sampling protocol for the NCS, one should consider requirements for both external and internal validity. With limited resources it is generally difficult to simultaneously satisfy strong external and internal validity requirements. Thus, a well-designed sampling protocol for the NCS is likely to strike a balance between external and internal validity.

Applying different scientific perspectives as the basis for drawing conclusions, valid conclusions may be drawn from both probability-based samples (statistically-based) and non-probabilistic samples (statistical and model-based). Probability-based samples offer the very desirable property of basing statistical inferences on the random sampling mechanism employed to select the sample. However, imperfect recruitment, retention, and response rates may limit such inferences. Valid inferences from non-probability-based samples, on the other hand, require the assumption that the NCS cohort is unbiased relative to the reference population. Thus, while inferences based on probability-based samples can draw their validity from the manner in which the sample was selected, inferences based on non-probability-based samples are only as valid as the NCS cohort is unbiased with respect to the relationships of interest.

There are two problems that complicate a sampling design process that attempts to strike a balance between external and internal validity. These are:

- The tendency of probability-based sampling methods to emphasize external validity and de-emphasize internal validity, and
- The emphasis on internal validity in non-probability samples that leads to the abandonment of probability-based methods and a resulting de-emphasis of external validity.

Taken at face value, these problems appear to leave little in the way of middle ground where a compromise might be found. However, it is almost certainly true that a polar application of either probability-based sampling or non-probabilistic sampling that fails to recognize the strengths of and motivations for the opposing approach will fail to address issues that are critical to the success of the NCS. Realistic anticipation of the true magnitude of respondent burden for NCS cohort members places expected recruitment and retention rates for a probability-based sample at low levels. Further, efforts to

convert reluctant participants into cohort members, as are traditionally applied in survey sampling applications, may only result in retention and data collection resources being wasted on children who eventually drop out of the study prior to all required study data being collected. Probability-based efforts to obtain a representative sample that result in a geographically dispersed cohort may require unattainable data collection resources and adversely affect the quality of the data that is collected. The expenditure of limited study resources on sampling design, recruitment, and retention activities may require that a simplified data collection protocol be employed to control the data collection resources required.

On the other hand, adoption of a non-probabilistic approach requires faith that systematic biases will not limit the relevance of NCS conclusions to a limited and not specifically identifiable population. The exclusive use of non-probabilistic sampling results in considerable risk that relationships identified in the NCS data may simply not be valid when extended to the reference population of all children born in the US during the NCS enrollment phase. This risk may prevent conclusions drawn from NCS data from being widely accepted and thereby limit the value of the NCS for improving the health and development of future generations of children. The anticipated magnitude of resources that will be invested in the NCS as well as the one-time-opportunity nature of the NCS dictate that actions be taken to mitigate this risk.

There are inherent risks associated with both probability-based sampling methods and non-probabilistic sampling methods. Unfortunately, the raw data required to quantitatively estimate the risks does not exist. Therefore, efforts to choose one set of the methods over the other as optimal are frustrated by a lack of solid information. Within such an uncertain decision-making framework, it is logical to abandon the notion of choosing one set of methods over the other and instead plan for a study that implements both probability-based and non-probabilistic sampling as part of a hybrid sampling strategy. Within such a hybrid strategy, each set of methods acts as a hedge against the risks associated with the opposing set of methods. Motivation for such an approach can be found in the financial investment community where it is not uncommon to package collections of dissimilar investments so that each specific investment acts as a hedge against the risks associated with other investments in the package.

Before drawing final conclusions in Section 6, we present two particularly relevant sampling methods that may be used to provide added flexibility to a hybrid sampling approach. Both methods are completely compatible with a probability-based sampling approach.

A-5.1 PROBABILITY-BASED PURPOSIVE SAMPLING

Probability-based random sampling methods generally attempt to keep inclusion probabilities for all elements of the sampling frame strictly greater than zero and strictly less than one. It is quite acceptable, however, to employ inclusion probabilities of one for some elements of the sampling frame. One can view this as taking the concept of over-sampling to an extreme. For example, consider a two-stage probability-based random

sampling process that first selects counties proportional to size from a sampling frame of all counties in the US and then selects a simple random sample of children within each selected county. Suppose that there are a dozen medical centers across the US that have successfully negotiated contracts with NIH to participate in the NCS. It would be quite acceptable, without leaving the confines of probability-based sampling methods, to specify that the dozen counties within which the medical centers reside must be included in the set of counties selected for the NCS. This purposive selection of specific counties does have consequences regarding the external validity of the study results in that the children selected from these specific counties can only represent their own county. Thus, these study subjects will have limited value for weighted analyses conducted for the purpose of demonstrating external validity. However, all the analysis methods that accompany probability-based sampling methods and the external validity that they afford to relationships identified in the study data remain valid in the context of probability-based purposive sampling.

In standard multi-stage applications of probability-based sampling, it is not uncommon for the inclusion probabilities of some primary sampling units to be set to one. For example, this can happen for populous counties when counties as primary sampling units are sampled proportional to population size. Thus, even standard applications of probability-based sampling can involve inclusion probabilities equal to one.

Several aspects of the NCS might lead to the use of purposive sampling. For example, in order to control the overall cost of medical data collection and improve the quality of such data, NIH may choose to solicit proposals from qualified medical centers with the objective of successfully negotiating contracts with a network of medical centers that would collect a large portion of the NCS medical data. This constraint could be accommodated within a probability-based sampling framework by setting the inclusion probabilities for the primary sampling units in which the targeted medical centers reside equal to one. Purposive sampling could also be used to include geographical areas that represent isolated areas of exposure, that represent the extremes of exposure conditions, that contain specialized facilities or equipment, or for which existing environmental exposure data already exists.

The only real limitation associated with purposive sampling is that the extreme over-sampling of purposively targeted elements of the sampling frame necessarily results in the remainder of the sampling frame being under-sampled.

A-5.2 PROBABILITY-BASED PURPOSIVE EXCLUSION

If one takes the concept of under-sampling certain subsets of the sampling frame to an extreme, it leads to setting the inclusion probability to zero for specific subsets of the sampling frame. In this case, rather than these elements of the sampling frame being under-represented, these elements are simply not represented at all. Alternatively, and perhaps more intuitively, one can view this process as defining the sampling frame to

exclude certain subsets of the reference population. In this sense the sampling frame represents but a subpopulation of the reference population.

Purposive exclusion methods could be used, for example, to focus the NCS on a subpopulation with desirable properties with respect to cost-efficiency and/or data quality. The concern with this approach is the potential that relationships that are internally valid for the study population will be somehow systematically biased when extended to the larger reference population. Thus, purposive exclusion of elements of the reference population would raise questions about the external validity of relationships identified in the NCS data.

There are most definitely circumstances under which the purposive exclusion of elements of the reference population may be the statistically optimal sampling design approach. The reference population, because of its all-inclusive nature, may contain a sizable number of children that are hard to recruit, hard to retain, and/or more expensive with respect to data collection. Focusing on a study population of children that have desirable properties with respect to recruitment, retention, and cost-efficiency would allow more children to be included in the study or, alternatively, more information to be collected for the same number of children studied. In either case, more information would be available for identifying relationships between exposures and outcomes and, therefore this approach is attractive from the point of view of maximizing the amount of information produced by limited study resources. However, as the study population is narrowed to achieve better cost efficiencies, the exposure-response relationships in the study population may become systematically biased relative to the exposure-response relationships in the reference population. These trade-offs are often navigated as part of sample surveys when the sampling frame is constructed. For example, when household-based sampling frames are constructed for population surveys, homeless people, women living in battered women's shelters, and incarcerated people may be excluded from the sampling frame for practical reasons.

Potential approaches for targeting more cost-efficient sub-populations have varying degrees of specificity. Limiting study participants to those residing within 50 miles of a major medical center might offer some cost-efficiencies while yielding a sub-population that includes a significant percentage of the nation's children. At the other extreme, suppose that study participants were limited to the existing patients of a dozen medical centers that successfully negotiate NIH contracts to conduct portions of the NCS. In this case, the study population includes only the existing patients of the dozen medical centers, a very small percentage of the nation's children. All other things being equal, the smaller the study population relative to the reference population, the greater the potential for bias in exposure-response relationships.

The trade-off between improved cost-efficiency and systematic bias can be formulated quantitatively in terms of total error where total error includes the contributions of both systematic bias and random error. This trade-off is addressed in detail in Appendix A.

The advantages and limitations of purposive exclusion are fairly simply stated. The primary advantage of purposive exclusion methods is the ability to exclude subsets of the reference population that are expected to be difficult to recruit, difficult to retain, or more expensive with respect to data collection. The limitation that these exclusions impose is an inability to extend relationships identified in the NCS data to the full reference population on an empirical statistical basis. Empirical statistical arguments may be used to demonstrate external validity relative to the actual study population but not beyond the study population to the full reference population.

A-6. CONCLUSIONS

There are compelling arguments for the use of both probability-based and non-probabilistic approaches to sampling in the NCS. Probability-based sampling methods add value in terms of protection against unexpected systematic bias. The reality of anticipated low recruitment and retention rates diminishes but does not negate this value. In a similar fashion, the use of volunteer participants adds value assuming that they provide better assurance of continued participation throughout the duration of the NCS. The reality of systematic biases that are inevitably introduced by volunteer participants diminishes but does not negate this value.

It is likely not possible to reach any kind of scientific consensus on the clear superiority of either approach due to the uncertainty that surrounds implementation of a study as unprecedented as the NCS. In particular, there are no definitive data sources that allow the precise prediction of likely retention rates under competing sampling design options. Lacking precise retention rate predictions, it is hard to imagine a clear scientific consensus emerging for either a fully probability-based sampling design or a fully non-probabilistic sampling design

A hybrid sampling design that employs both probability-based sampling and non-probabilistic sampling would allow each set of methods to act as a hedge against the risks associated with the opposing approach. The resulting NCS database would address issues of both internal and external validity resulting in the identification of cause-and-effect relationships that can validly be extended to a population including most or all of the nation's children.

In order to derive maximum benefit from the probability-based methods employed within a hybrid strategy, it will likely be necessary to take advantage of all the flexibility that such methods provide. Important considerations include:

- It may be advisable to focus attention on a study population (or sampling frame) that represents only a cost-effective subset of the reference population

- Including a wide range of exposures in the NCS cohort may be much more important than having the NCS cohort reflect the demographic characteristics of the reference population

- Over-sampling and perhaps even purposive sampling may be necessary to assure that cohort members have a wide range of exposures

- Purposive sampling may play an important role in allowing targeted resources such as qualified medical centers, specialized facilities/equipment, and existing environmental databases to be employed in conducting the study

- The use of unequal inclusion probabilities to over-sample cost-effective subsets of the study population may be necessary to control data collection costs while maintaining a reasonable level of complexity in the data collection protocol

- Cluster sampling may play an important role in
  - Controlling data collection costs
  - Creating a data structure that is conducive to examining phenomenon that occur at the neighborhood or census tract level

- Unless the hypotheses requiring large sample sizes are related to outcomes that occur early in a child's lifetime, it may be advisable to enroll only enthusiastic participants in the NCS cohort in an attempt to maximize retention rates; active conversion of reluctant participants may be ill-advised

Fortunately, probability-based sampling methods that incorporate elements of purposive inclusion and exclusion have sufficient flexibility to at least partially achieve many of the objectives that motivate the consideration of non-probabilistic methods. That said, the inclusion of some proportion of volunteers in the NCS cohort offers a benefit that probability-based methods cannot provide, that being a self-motivated cohort member that has the highest likelihood of retention until all required study data have been collected.

A-7. REFERENCES

Campbell, D.T. and Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research***.** Chicago: Rand McNally & Company.

Cochran, W.G. (1977), *Sampling Techniques*, New York: John Wiley & Sons.

Hahn, G.J., and Meeker, W.Q. (1993), "Assumptions for Statistical Inference," The American Statistician, 47, 1-11.

Kish, L. (1995), "The Hundred Years' Wars of Survey Sampling," Statistics in Transition, 2, 813-830.

Little, R. (1982), "Models for Nonresponse in Sample Surveys," Journal of the American Statistical Association, 77, 237-250.

Little, R. (2003), “To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling,” U. of Michigan Dept. of Biostatistics Working Paper Series, Year 2003, Paper 4.

Rubin, D.B. (1976), “Inference and Missing Data,” Biometrika, 63, 581-92.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Rubin, D.B. (1996), “Multiple Imputation After 18+ Years,” Journal of the American Statistical Association, 91, 473-489.

Scott, A.J. (1977), “On the Problem of Randomization in Survey Sampling,” Sankhya C, 39, 1-9.

Scott, A.J. and Smith, T.M.F. (1973), “Survey Designs, Symmetry and Posterior Distributions,” J R Statist Soc B, 35, 57-60.

Smith, T.M.F. (1976), “The Foundations of Survey Sampling: A Review,” J R Statst Soc A, 139, 183-204.

Smith, T.M.F. (1983), “On the Validity of Inferences from Non-Random Samples,” J R Statist A, 146, 394-403.

Smith, T.M.F. (1994), “Sample Surveys 1975-1990: An Age of Reconciliation?” International Statistical Review, 62, 5-34.

Sugden, R.A. and Smith, T.M.F. (1984), “Ignorable and Informative Designs in Survey Sampling Inference,” Biometrika, 71, 495-506.

Valliant, R., Dorfman, A.H., and Royall, R. (2000), *Finite Population Sampling and Inference*, New York: John Wiley & Sons.

APPENDIX A-A

RATIONALE FOR SAMPLING FROM SUBPOPULATIONS

If valid statistical inferences for an entire reference population are sought, the logical sampling approach would seem to be a probability-based sample from the entire reference population. If unbiased statistical inferences are required, this might be the only valid approach. However, unbiased statistical inferences are seldom actually required of a study. Instead, the actual requirement is for statistical inferences that are approximately valid for the reference population. In cases where the cost of obtaining study information is lower for certain elements of the reference population and higher for others, it can be statistically optimal to study a biased portion of the reference population where cost-efficiencies are possible.

For estimation of a statistical parameter, such an approach can be justified in terms of minimizing the total error of estimation. The statistical parameter could be the prevalence of exposure to particular environmental contaminant or the odds ratio between an exposure variable and a health outcome. Thus, the total error of estimation argument applies equally well to descriptive parameters and parameters that characterize the strength of relationships.

Total error has two components, systematic error and random error. Systematic error is that part of the estimation error that derives from (1) differences between the study population and the reference population and (2) systematic biases in the measurement protocols employed to collect study data. Systematic error is unaffected by the size of the sample taken. Random error, on the other hand, derives from (A) differences between the sample and the study population and (B) random measurement error associated with the measurement protocols employed to collect study data. Random error is reduced as the sample size increases.

A widely-employed version of the total error concept is embodied in the mean square error (MSE) of a parameter estimator. The MSE of the estimator is the expected squared deviation of the estimator from the true parameter value. The corresponding systematic error component is represented by the bias of the estimator squared, while the random error component is represented by the estimator variance. With these definitions, we have

$$\text{MSE} = (\text{Bias})^2 + \text{Estimator Variance}$$

or

$$\text{Total Error} = \text{Systematic Error} + \text{Random Error}.$$

Consider the case where the study population is equivalent to the reference population. In this case, systematic error is minimized because there are no differences between the study and reference populations. Because the entire reference population is

studied, the sample must include sample elements for which data collection is more expensive. Alternatively, consider the case where the study population is defined to exclude elements of the reference population for which data collection is more expensive. Because the study population is now more cost-efficient, it is possible to observe a larger sample using the same data collection resources. The larger sample size results in a reduction in random error. However, since there are differences between the study and reference population, systematic error has now been introduced. If the reduction in the random error component is larger than the increase in the systematic error component, then it is statistically optimal to study the smaller, more cost-efficient subpopulation because the total error of estimation is smaller.

The preceding paragraph demonstrates that it is possible to justify studying a cost-efficient sub-population as statistically optimal. While it is possible to do so, this justification is rarely formally completed for various reasons. It is generally difficult to quantify the systematic errors and increased cost-efficiencies associated with a proposed sub-population, and therefore difficult to formally compare the increase in systematic error to the expected reduction in random error. Instead, this comparison is made in an approximate manner at a more general level by asking and answering in a general fashion questions such as the following:

- Will statistical conclusions drawn from the proposed subpopulation be approximately valid for the entire subpopulation?
- Will improved cost-efficiencies associated with the proposed sub-population provide an opportunity to make much stronger statistical inferences about the sub-population than would be possible for the entire population?
- Will the increase in strength of the statistical inferences for the subpopulation be large enough to counteract any systematic error associated with the sub-population relative to the reference population as a whole?

Final Report from the
National Children's Study Sampling Design Workshop
March 21-22, 2004
Arlington, Virginia
May 9, 2004

Introduction

The National Children's Study Panel on Sample Selection was charged with 1) Providing an approach to the sampling design that would reconcile competing priorities, needs, and limitations; 2) Assessing the background papers provided by Battelle for addressing the design decisions; 3) Addressing the strengths and weaknesses of selected design options; and 4) Identifying options that require pilot testing to reach a final decision. The panel consisted of nine researchers with a diverse range of disciplinary backgrounds and research experiences, listed at the end of this report. We were provided with the detailed Battelle "Draft White Paper on Evaluation of Sampling Design Options for the National Children's Study" along with appendices. The panel met for two days, March 21-22, 2004, in Arlington, Virginia, with the first day devoted to hearing from selected key individuals involved in the planning of the study from both within and outside the federal government. The second day was set aside for panel deliberations and an oral summary to Dr. Alexander and other leaders of the planning effort for the study.

Those who helped to prepare us for the workshop, particularly Drs. Quackenboss and Scheidt, were extremely responsive to our needs, offering candid insights before and during the panel meeting. The presenters gave succinct, informative talks on a range of issues bearing on the approach to sampling and were able and willing to respond to all the questions that we posed. While we prepared and deliberated over a relatively short period of time, and cannot claim the depth of knowledge of those who have been engaged over several years, we believe we can offer a useful perspective of informed outside experts free of entrenched, longstanding positions regarding the study. The panel was chosen to have research backgrounds that would enable them to appreciate the goals and methods of the study, but there was little previous involvement of panel members in the study, and we were able to approach the issues objectively.

Points of Agreement Regarding Sampling Plan

We discussed a number of issues that bear on the approach to recruiting participants into the National Children's Study, which set the stage for more detailed consideration of two competing selection plans, a national household probability sample and a center-based design in which recruitment is conducted by academic medical centers working in targeted communities. The panel agreed unanimously on the following points:

1) A national probability sample is preferred to other sampling approaches based on a number of specific reasons as described in detail below. All panel members recognize the challenges in implementing this approach successfully, with varying views regarding the feasibility for such an approach to generate acceptably high participation and retention proportions and its feasibility relative to a center-based design. However, we are all in agreement that it would offer distinct benefits. Such a national probability sample would call for incorporating extensive biomedical and clinical detail into the design, well beyond simple

biospecimen collection, which has become common in such surveys. The alternative, a center-based model, would require extension in the other direction, moving from the traditional convenience sample based solely on recruiting patients towards a more complete community representation through collaboration and outreach, which would include women outside the medical system, some of whom would be recruited prior to conception. Both approaches would seek to integrate the strengths of biomedical and population research, and each poses real challenges in deviating from the ways such studies have been done in the past.

2) We do not see advantages in allocating proportions of the study sample across recruitment approaches, unless there are explicit goals regarding what can be learned from each subset. While having a larger cohort that provides core information and a subset that is followed more intensively should be considered, simply recruiting individuals through different mechanisms into the overall cohort does not offer any apparent advantages over expanding the best approach to include the entire sample.

3) Under any approach to sampling participants, many of the key activities of the National Children's Study will need to be centralized in order to maintain standard methods and quality control and to ensure that the most capable groups are performing key tasks. The formulation and conduct of interviews, the collection of environmental samples, specimen receipt, processing, storage, and assays, and follow up of children over the extended study period will require central planning and management regardless of whether the pregnancies are initially identified for the study through a national probability sample design or independently by multiple centers. The continued follow up of children, however they are initially recruited, will occur throughout the country (given the mobility of the population) and require ongoing decisions regarding the data to be collected and hypotheses to be tested. Contrary to the citations provided in the Battelle report, several long-duration national probability samples (e.g., the National Longitudinal Survey of Youth) have generated high recruitment and retention proportions through such an approach.

4) A mechanism is needed to fully engage the energy and intellectual resources of the research community, balanced against the need for a centrally managed study. The strengths of central planning are consistency and quality control, but the potential weakness is the loss of the creativity, energy, and full buy-in of the broad research community concerned with children's health. There would clearly need to be early access to the data that are generated for public use and a way to entertain proposals for use of biospecimens. At an earlier stage, it would be desirable to consider ways to solicit and evaluate competing ideas for specific research proposals that might be incorporated into the study.

5) Both social and biomedical aspects of children's health are of central importance to this effort, with a need to consider how social factors affect behavior and biological pathways, as well as discover more basic mechanisms of disease causation. The approach to sampling needs to accommodate both themes.

6) All pregnancies, and all the fetuses and infants resulting from a given pregnancy occurring during the recruitment period to a given woman should be included to

optimize recruitment and retention, and to allow for study of siblings. This would facilitate special studies of genetic and gene-environment effects.

7) The potential for including a fraction of women who would be enrolled and monitored prior to conception was seen as highly desirable and can perhaps be done by monitoring non-pregnant women of reproductive age for the onset of pregnancy. Inclusion of such women would permit study of infertility and pregnancy loss. In principle, such preconception enrollment would also allow for specimen collection to be done before or very early in pregnancy, in the period during which structural malformations and perhaps other pediatric health problems have their origins. The importance of this effort depends on the priority given to such outcomes as congenital malformations, for which the rarity of individual types may make even a sample of 100,000 marginally adequate.

8) To address some of the study goals, it may be necessary to overrepresent selected geographic locations and subgroups, such as locations with certain environmental exposures of interest, e.g., specific forms of air and water pollution, and participants of certain race or ethnicity or with specific socioeconomic conditions. This overrepresentation is attainable under either sampling approach. Views of panel members vary on the extent to which such decisions to optimize one goal may compromise other study goals.

9) A streamlined approach is needed for design and implementation decisions in order to allow for our suggestions regarding design and further pilot work to be of value in moving the study forward. While the benefits of having multiple agencies, diverse sources of scientific input, and careful deliberations regarding the conduct of this study of unprecedented size, cost, and importance are clear, a mechanism needs to acknowledge competing considerations, reach decisions, and move forward. We came to quickly appreciate the strong views held regarding how this study should be done as well as the magnitude of challenge it poses. Given the complex array of committees in place, we have real concern that there may not be a clear, widely understood plan for exactly who will make the hard decisions required and ensure that the benefits of outside influences can be realized without preventing progress. We would hope that our panel offers useful insights to accelerate progress towards firm decisions about the design and not result in any unnecessary delays. We propose below specific, limited pilot activities that should provide key information for making a decision regarding the design and may even warrant a prespecified decision algorithm in response to the data that are generated in order to avoid indecision at the end.

Option 1: National Probability Sample

The national probability sample design we considered calls for a full national probability sample of households, recruitment of reproductive age potentially fertile women residing in those households, and prospective monitoring of those women over some period of years for pregnancy and births. There would be a clear need for geographic clustering in the sampling strategy, but the approach would involve many such geographic units, well over 150, and be widely dispersed. The sampling plan could be weighted to achieve the desired diversity of geographic location and ethnic composition for estimating prevalence and attaining sufficient precision for measuring associations of

particular interest, but allowing for generation of weighted nationally representative estimates.

The reasons for preferring this approach were notably diverse across panel members, not all of who value each benefit similarly. Nonetheless, the multiple perceived advantages build on one another and collectively make a case that the panel was unanimous in supporting. The key points are as follows, not necessarily in order of importance:

1) The National Children's Study needs to be able to contribute to understanding of the impact of social, economic, and environmental factors, not just biomedical factors, providing guidance to public policy decisions affecting the health of children. Given the desire to address both individual-level and population-level effects on behavior and ultimately on health, the use of a probability sample offers substantial advantages. While biomedical influences may also be vulnerable to biases as a result of recruitment using non-probability based sampling, social and economic influences are likely to be more directly linked to access to and selection of health care providers. Therefore, associations between such factors and child health outcomes are particularly susceptible to varying in relation to the source of participants. Furthermore, a true national probability sample is likely to enhance the perceived value of the study's findings among the public at large as well as policy makers, including those who will be called upon to consider continued funding for the research effort. Links to agencies concerned with education, housing, and a range of other policy arenas are more likely to appreciate and support the study if it is a national probability sample.

2) The environmental and social influences on child health that are of interest vary both within and between geographic locations, and it would be advantageous to be able to describe such variation in explicit and quantitative terms as can be done in a probability sample that is appropriately geographically dispersed.

3) As influences on children's health are identified through this research, the information to generate estimates of attributable fractions, which require population-level information on the distribution of determinants and relative risks, would be available directly from a study with a probability sample provided such sampling has sufficient recruitment and retention rates.

4) Having access to medical care or seeking medical care would not be prerequisites for enrollment in the study, allowing inclusion of women who do not seek prenatal care or who only seek care late in pregnancy. While other designs could sample within households to achieve this goal, a probability sample provides a scientifically valid means of doing so. This is particularly so when the sampling unit is the woman and not the pregnancy, since probability-based sampling of prevalent pregnancies would be subject to length bias (in which pregnancies that persist for longer periods are more readily identified and included).

5) A household probability sample would involve sampling of women, not pregnancies, and allow preconception enrollment to be done in an unbiased way and fully within the context of the study of pregnancy outcome and subsequent

children's health. Recruitment through prenatal clinics, for example, would not include all such women and would not enroll women at the beginning of their pregnancy, and risks biases associated with the many characteristics influencing enrollment in prenatal care. Under a center-based model, preconception recruitment would have to be done in a separate arm of the study.

6) With relatively few exceptions, the desired data and specimens can be collected in the home, not requiring collection within the setting where prenatal care is obtained. Home-based rather than clinic-based collection would likely be more convenient for the participants.

7) A probability sampling plan would be rigorous and explicit, allowing replication of the sampling process and the analytical results, at least in principle, presuming recruitment and retention rates are sufficiently high. Bias due to non-participation could be assessed by comparing the study participants to population-based data from the census and from birth certificates.

8) Separation from the medical care system by sampling households may offer advantages in marketing the study and recruiting women, freeing the study of any perceived negative aspects of medical research, sometimes an issue of particular concern in minority and poor communities. In addition, such a separation may help to facilitate the needed standardization of methods and centralized follow-up through an experienced survey organization.

9) The desire to ensure a unique contribution of the National Children's Study, above and beyond what is already being done within medical centers and the large cohorts that have been assembled in other settings. Current efforts in Norway and Denmark do not involve probability samples, and only some of the studies in England have had this characteristic. None of the studies in North America have previously attempted to develop a true population-based probability sample.

10) A geographically clustered probability sample of households may facilitate collection of data on the social, physical, and chemical environment in which the participants reside since the data collection would be done at home. Studies based in health care facilities have a physical separation of the data collection site (clinic) and the environment of interest (home).

While the assessed *desirability* of such an approach was universally supported, the assessed *feasibility* of this approach was judged differently among panel members. The lack of precedent for a study of this size and complexity at least in the United States dictate the need for pilot efforts to assess the feasibility of this approach. The pilot study needs to be carefully planned and implemented so that the most effective methods of recruitment and data collection are identified.

We were able to identify two key concerns, that, if answered affirmatively, would persuade even the most skeptical of panel members that study participants can be identified and recruited as a household probability sample of the population and that the required array of data can be collected. While not all issues are unique to the probability sampling design, all are important considerations in assessing the feasibility of this approach. They are as follows:

1) Feasibility of identifying and recruiting women in a household survey and identifying and recruiting those who become pregnant: There are several steps required, with the most experience in obtaining respectable response rates for household surveys to the point of enumerating the household members and identifying any of them who are currently pregnant. Proceeding from that point, the study would require the initial and continued involvement of reproductive age women not currently pregnant and not known to be sterile (regardless of whether they report being sexually active or using contraception) in order to identify new pregnancies as they occur. The issues of unplanned pregnancies, induced abortion, and general sensitivity surrounding pregnancy would all pose a challenge in that the study would be focused on women who have not self-identified by seeking prenatal care, for example. The feasibility of recruiting pregnant women in this manner could be tested by determining the yield, in terms of response rates and numbers of pregnancies that can be identified for a given cost, through a pilot study conducted in a small number of carefully chosen diverse geographic locations.

2) Feasibility of obtaining access to hospitals serving recruited women: Once enrolled in the study, it is generally believed that the desired data can be collected without active involvement of prenatal care providers. What is desired, however, is access to the woman and her infant at delivery in order to collect cord blood, the placenta, and conduct research-quality neonatal examinations. This would require cooperation at every hospital or other location at which the sampled women deliver with either hospital staff, recruited and trained to collect these data, or study staff attending each delivery to collect the needed specimens. There are both access issues, in the sense of requiring a high degree of cooperation of health care providers, and logistical issues, in having an appropriate person collecting the right information and specimens at the needed time. It is recognized that obtaining the desired extent of access and assistance would not be easily obtained in the center-based approach either, but there would be greater familiarity with the investigators seeking such material, making this a special challenge for the national probability sample design. The degree to which this is needed for all study participants rather than a subset is not entirely clear.

In addition, three other important issues could be usefully examined in such a pilot study, addressing issues important to but not necessarily unique to a national probability sample:

3) Feasibility of and need for fostering a community commitment in the randomly selected areas: Views of the panel varied on how important they thought it was to have the National Children's Study be a visible, collective activity on the part of the community versus simply identifying and recruiting individual participants in isolation. If in fact community engagement is advantageous, it was unclear how feasible it would be to foster this commitment in a widely dispersed array of randomly selected locations, i.e., locations not chosen for making this aspect of the study feasible. While it is not clear that this issue can be effectively resolved through pilot studies, perhaps there would be an opportunity to compare across at least two communities, one of which could have the study promoted energetically in the media, through local civic groups, etc., in a manner that

would be feasible for any chosen location, and the other half pursued without such efforts in order to assess the impact on response.

4) Feasibility of collecting reproductive tract specimens outside medical settings: While there is abundant evidence that most biospecimens can be collected in the home, such as blood, urine, saliva, or hair, there would be an interest in having some reproductive tract specimens collected. Further work would be needed to determine if self-collection would be adequate for these purposes, and second, whether women would be compliant with doing so in their homes. It is presumed that a complete pelvic examination as required to collect cervical specimens, for example, would not be feasible under this design, unless special arrangements were made with all prenatal care providers to collect the specimens within prenatal care settings or trained clinicians (nurses or physicians) were sent to the home.

5) Feasibility of collecting biological specimens very early in gestation: With the planned identification of some fraction of pregnancies prior to conception, there is the potential for collecting biological and environmental specimens in the first weeks of gestation, a period of special interest. In practice, what is unclear is just how burdensome this approach would be with regard to participant tolerance and cost, and how early such specimens could in fact be collected. For even a small number of such women identified before conception, it would be helpful to undertake specimen collection to more accurately weigh the feasibility of this desirable component of the study.

Option 1A: Probability Sample with Investigator-Initiated Components

A potential limitation in the fully centralized national household probability sample approach was seen as the limited opportunity to engage fully the ideas and talent of university-based investigators. While there would be a need for the medical care community to support in-hospital data and specimen collection, there is not under this approach a direct mechanism for investigators with promising ideas to have the opportunity to compete for their incorporation into the National Children's Study.

One way for this to be done would be to set aside some resources, including funds, interview time, specimen allocation, and respondent burden more generally, that would be open for competition to the research community. This would need to be done *before* finalizing the study plans to allow for the most promising ideas to be incorporated. The scientific promise would need to be evaluated in balance with the burdens imposed on study participants and staff, but the process should generate more useful information than would be provided by having the entire set of hypotheses and data needs determined centrally.

Option 2: Center-Based Recruitment

The center-based approach that we considered is based on academic medical centers working to identify and recruit participants within targeted communities. In this model, the natural base of these academic centers would need to be expanded in most cases in order to be more broadly representative of the population residing in defined

communities. This extension could be done through health care providers and through active outreach to the population. Potential centers would be invited to compete for this opportunity and judged on such criteria as their ability to accurately reflect pregnancies in the population residing in a defined geographic area and inclusion of the full spectrum of socioeconomic and ethnic groups, as well as the ability to provide a sufficient number of participants. If such centers had a means of generating probability samples of the catchment area, that would of course be most appealing and competitive for selection as a site, but it seems more likely that they would instead seek to generate populations reflective of the composition of the community in a less formal manner, working with other health care facilities. Birth records would allow for making comparisons of those included with live births in the area, at least with regard to those characteristics available from the birth certificate. In addition to centers being selected in part based on their ability to meet specified desires for special populations targeted by the study planners, e.g., agricultural workers or ethnic minorities, optimal geographic locations could be sought out for encouragement to compete. There would still need to be centralized planning and administration of the study and a standardized approach to interviewing, biological specimen collection and processing, environmental measurements, and follow-up of the children. A sizable number of sites would be needed to generate the desired study size of 100,000, with a tradeoff such that more sites provide for greater geographic diversity and more extensive involvement of the clinical and research community, but also are more challenging and expensive to coordinate.
The key strengths of this approach include:

1) The scientific community would be fully engaged through the development of proposals and competition for participation in the National Children's Study. The most creative approaches of the nation's best researchers would be brought to bear from beginning to end under this model, within the practical constraints of cost and respondent burden. Supplementary research would undoubtedly evolve from the centers and networks of collaborating centers.

2) An active community engagement would be attainable under this approach in that centers could be selected in part based on having such a component. An identity would be established for the study locally, with a committed leader or set of leaders, and a carefully fostered sense of the value of this study. This model has worked successfully in selected locations and would be applied for the first time in a large number of sites with central coordination and assistance.

3) By building out from existing research centers into the community, a beneficial change in the perspective of these research centers would result, more fully embracing a population perspective on the health of the local community.

4) Collection of specimens and medical examination data at the time of delivery may be enhanced by the involvement of those medical centers at which many of the study participants would deliver.

With this model, there are also real concerns that would need to be addressed to ensure its success. While there is little doubt that academic medical centers can do what they have already been doing through the collaborative research networks, in which common goals and protocols are pursued, there are features of expanding this model for the National Children's Study that are unproven.

1) Sufficient response from needed number and scope of centers to conduct a national survey of the desired size: While a handful of academic medical centers with strong records of funded research of this type would undoubtedly respond to an invitation to develop proposals, it is unclear whether a sufficiently large number of well-qualified centers covering a diverse enough population would in fact step forward. Many academic medical centers are based in urban areas and may have less access to and experience with other populations in their region. It is difficult to resolve this question with much confidence in advance, but perhaps by generating selected locations or populations of interest and surveying candidate center leaders, informative insights could be gained about the scope of candidate sites and whether there would be a sufficiently broad menu from which to select.

2) Whether academic medical centers are capable of expanding beyond their traditional patient clinical base is uncertain. There is much competition for delivery of clinical services, and relatively little tradition for such clinical centers to extend beyond their traditional boundaries. In many cases, local providers compete with one another rather than cooperate. In addition, economically disadvantaged populations, an essential component of the planned study, have particular considerations with regard to where they obtain medical care and whether they can be successfully enrolled in the study. The ability to enlist the support of other health care providers in their communities to obtain a sufficiently broad patient base would need to be evaluated. To establish the feasibility of a center-based approach, there would be a need for a pilot study comparable to that proposed for the probability sampling strategy and ideally done at the same time. A small number of centers would need to implement the process from identifying women from the community at large, not just relying on their patient base, obtaining a sufficient response rate for enrolling such women, and collecting required interview data and specimens, up to and including the collection of placenta, cord blood, and standardized neonatal examinations at delivery.

3) Relying on specific centers would imply relying on the institution to follow through for a sustained period, more than 20 years, even if the principal investigator were to change. The most intensive involvement would be early in the recruitment phase, with more and more centralized work needed in the follow-up period, but nonetheless, some sustained linkage to the medical center would be required over the life span of the National Children's Study, at minimum for Institutional Review Board purposes. For follow-up purposes, the population's mobility would result in a need for national efforts and any initial advantages in having center-based recruitment would diminish over time.

4) Centers may not be universally capable of generating desired recruitment rates or fully complying with a standardized recruitment protocol. Although some centers would have experienced research teams in place to achieve the desired rates of recruitment of eligible participants, many would not. Lacking a centralized mechanism of recruitment through a survey research organization, there would be significant variability in the level of success, with some centers falling below desirable rates. Furthermore, such centers would need to be capable of recruiting women early in pregnancy, collecting the required biological specimens and environmental samples, and arranging for newborn

examinations. While there would be some transition to the centralized research organization following recruitment, the early components would have to be facilitated by the individual centers. Another concern with the decentralization to multiple centers as opposed to a national survey research organization is that the desired standardization with regard to defining eligibility and recruiting and enrolling participants would be more difficult to sustain.

Conclusions

Both the national probability sample and center model have successful precedents to draw upon, but in neither case has there been a study that combines the scope, size, and detail intended for the National Children's Study. In fact, it is probably appropriate that the study set its sights on making a contribution that is far more ambitious than could ever be done through ongoing mechanisms of developing research proposals. Expanding the scope of either approach into uncharted territory will be very challenging, and for that reason, we would encourage the simultaneous pilot testing of key issues affecting the feasibility of both approaches. We recommend proceeding rapidly with targeted, efficient pilot efforts to better inform this key decision regarding sampling design.

We recognize that the product of the pilot data collection effort is certain to be informative, but unlikely to be definitive unless one approach or the other (or both) is shown to be completely infeasible. Assuming that instead the relative feasibility of each is measured and quantified, a panel such as ours or some other appropriately constituted group should be configured to balance the strengths and limitations and quickly put forward a plan to conduct the best study possible and recommend a specific sampling approach for implementation.

National Children's Study Sampling Design Workshop Panel Members

David A. Savitz, Ph.D., Panel Chair, Professor and Chair, Department of Epidemiology, University of North Carolina School of Public Health, Chapel Hill, North Carolina

Scott Zeger, Ph.D., Panel Vice Chair, Professor and Chair, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

Jean Golding, Ph.D., Professor of Paediatric and Perinatal Epidemiology, Scientific and Executive Director, Avon Longitudinal Study of Parents and Children (ALSPAC), Department of Community Based Medicine, University of Bristol, Bristol, England.

Barry I. Graubard, Ph.D., Senior Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland

Graham Kalton, Ph.D., Senior Vice President and Senior Statistician, Westat, Rockville, Maryland

Michael S. Kramer, MD, M.Sc., Professor of Pediatrics and Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada, and Scientific Director, Canadian Institutes of Health Research, Institute of Human Development, Child and Youth Health

Anne R. Pebley, Ph.D., Professor of Population Studies, UCLA School of Public Health, University of California, Los Angeles

John Spengler, Ph.D., Professor of Environmental Health and Human Habitation, Department of Environmental Health, Harvard School of Public Health

Clarice R. Weinberg, Ph.D., Chief, Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

Expressing Your Priorities for the National Children's Study

created March 8, 2004

R.T. Michael

This exercise is intended to provide the FAC some sense of the prevailing priorities about the NCS that might guide judgments regarding the sampling design of the study. The exercise should not require more than ten minutes of your time; it will be more successful if you respond with your initial instincts rather than ponder the implicit complexities of the study before you respond and if you do not attempt "to game" the outcome by overstating your real views to influence the averages. The exercise has two separate parts; both explore the same few issues and the repetition is intended to give different perspectives on essentially the same few issues that may affect the sampling design of the NCS.

PART 1: In this exercise, assume that reasonably sensible decisions will be made about all the issues listed, since all are undoubtedly important to the success of NCS. The question for you is where you place your greatest interest in behalf of the study. To indicate your priorities, you have 100 points to allocate to any one or any combination of the seven domains listed below. Put your points where your passions lie.

There are seven domains here, described as follows:

I am most interested in or passionate about:

E the study's insights about one or a few of the **environments** that are a focus of NCS

O the study's insights about one or a few of the child **health outcomes** of focus of NCS

M the study's **mechanisms** (medical, familial, social…) that connect the environments and outcomes of focus in the NCS

L the study's **long-term research potential**, such as focus on selecting issues in infancy that are most likely to have payoff in adult health.

I the study's insights for the **immediate future**, those pertaining to the pregnancy and the neonatal period.

G the **generalizability** of the study's results to a wide spectrum of children

S the insights or results that pertain to **specific or particular groups of children**, such as those in poor families, African-American, or those served by medical centers of excellence.

E_____

O_____

M_____

L_____

I_____

G_____

S_____

Total: 100.

[As an example, if you think a pivotally important focus that will be a big factor in the ultimate payoff from NCS should be the findings about the effects on pregnancy of certain chemical environmental insults on all children, you might allocate 30 points to E, 20 points to O, 40 points to I and 10 points to G.]

PART 2: Here you are confronted with four separate pairs and for each of the four, please indicate where you stand, in terms of the trade-offs to be made by NCS. These four choices are independent of each other. Express your priority on each separate issue by placing an "X" along the line of each of the four continua.

2a: Hypothesis-driven v serendipity in NCS potential

Here, the issue is not how to craft a particular investigation with the data, it is instead how to think about the nature of the data to be collected. If you think the NCS's potential lies mostly with the specified "core" hypotheses, put your priority for the hypothesis end of the continuum which will imply a heavy weight to capturing the specific pieces of information critical to those core hypotheses. If, on the other hand, you think the NCS's potential lies mostly with the omnibus character of the wide-ranging data set that will provide opportunity for inquiries not currently envisioned, then express your priority for the "serendipity" end of this continuum which will imply placing a heavy weight on capturing information more broadly so those research opportunities that come from unanticipated changes in environments and new knowledge can be exploited.

Hypothesis Driven Inquiry	|_____|	Serendipity Enhanced Inquiry

2b: Generalizability

Here, the issue is how important it is to you that the findings from the NCS are applicable to *at least fifty percent* of *all* children born in the U.S. in the time interval of the NCS's selection of live births for the NCS. (Some sampling schemes yield samples that can project to large populations, other schemes yield samples that project to none or to few . others than those actually in the group of observations. The question here is how widely do you think it is important for the NCS findings to be applicable.)

Not Important To me	|_____|	Of Paramount Importance To me

2c: Universality of the key findings.

Some "findings" from the NCS are likely to apply to all children because those findings are universal, as are chemical reactions and many in-the-body environment-outcome mechanisms. Other likely "findings" from the NCS are probably dependent on the circumstances and behavioral responses that accompany the exposure to those environments, so these "findings" are not universal but instead highly context specific. The sample of pregnancies or children needs to be consistent with the judgment about how universal the important findings from NCS are: if those key findings are in-the-body or chemical relationships, for example, it may not matter who the observations are or whether they "represent" a larger population of children, but if those key findings involve social circumstances or varied responses, then that lack of universality calls for a

probability sample. So this continuum asks you how invariant, universal you think the NCS's key findings probably are.

Most Key NCS Findings are Universal	|_____|	Most Key NCS Findings are NOT Universal

2d: The Trade-off of data precision and generalizability of NCS Findings

Here, like exercise 2b, you are asked to think about the population of children to whom you think the CNS findings should apply, but here the "trade-off" of generalizable and data quality is confronted. It would of course be ideal if the findings pertained to "all children" and if the data in the data set were perfectly measured, captured, and characterized, but both these ideals will be sacrificed by any real study done at any realistic expense. Thus the trade-off this exercise asks you to confront. The topics you hold most dear will influence your choice here.

quality, detail, precision of *measurement* of captured data is of highest priority to me	|_____|	g*eneralizability* to a wide, known, population of children is of highest priority to me

Thank you.

Some overarching design issues

- What is being measured, where, when and on whom?
- To what extent is sampling frame centered around medical center enrollees, medical center catchment areas, or the broader US population?
- To what extent should units be selected probabilistically, or by volunteer samples?
- To what extent do different design options differ in recruitment and retention rates, and how much should these differences drive the choice of design?