The National Children's Study: Rationale for Increasing the Study Sample Size

Summary

The National Children's Study is a longitudinal observational birth cohort with a planned 21 years of follow-up for each enrolled child. Effective retention will be required to ensure statistical validity and generalizability of the Study's findings. Analyses of the initial Vanguard Study revealed a lower than expected enrollment of eligible pregnancies, and short-term attrition of these enrollees to be approximately 20 percent at birth. Several computer-based modeling strategies of long-term retention rates suggested that after 21 years, the Study could expect to retain only 30–40 percent of the initial enrollees. This retention rate would compromise the analysis for many conditions of potential interest. The National Children's Study Program Office initiated a retreat and five working sessions to explore this issue and define the parameters for an increased sample size. The resultant analyses suggested that 225,000 to 250,000 women should be enrolled to produce a final Study sample of about 100,000 children with adequate longitudinal data after 21 years. The increase in sample size can be attained by both increasing the size of the eligible geographic segments within each of the current Primary Sampling Units, and increasing the total number of Primary Sampling. New locations would be added using the same random selection process used to identify the initial locations. A new larger sampling frame is thus under development, with a probable total of 185 to 200 locations.

Introduction

The National Children's Study (NCS or "the Study") is a longitudinal observational birth cohort study consisting of several components with the overall goal of systematically collecting data, biospecimens, and environmental samples to examine the relationships between environment and genetics on growth, development, and health. Environment is broadly defined to include factors such as air, water, soil, dust, noise, diet, social and cultural influences, access to health care, socioeconomic status, and learning.

Recruitment and Short Term Retention Must Be Optimized for the Main Study

A successful Main Study with the planned 21 years of follow-up for each enrolled child will require effective recruitment and retention of the cohort to ensure statistical validity and generalizability of the Study's findings.

The original Study approach was to use household-based recruitment by sending field workers door-to-door in designated geographic locations within a Primary Sampling Unit. A Primary Sampling Unit is a geographic area that in general corresponds to a county in the United States that was selected as part of a national probability sample. The initial probability sample had about 100 Primary Sampling Units (PSUs). The PSUs were then either combined or further subdivided to yield segments that had comparable birth rates. The target for the initial National Children's Study Vanguard Study experience, implemented in 7 PSUs, was 250 annual births per geographic segment. Over the course of about 18 months, about 1500 families were recruited into the Vanguard Study from the 7 locations.

Analysis of the initial Vanguard Study experience revealed several findings related to retention:

- The household-based recruitment method enrolled participants with the same racial and ethnic demographics as the general population in the selected geographic areas and, as a composite, the 7 areas reflected the racial and ethnic demographics of the United States, with the exception of underrepresentation of non-English-speaking Asians. The National Children's Study Program Office interpreted this deficiency as due to the small number of geographic areas, which were selected randomly and not intended in the small initial pilot to be generalizable to the whole country.
- The proportion of all likely births in a given geographic area that enrolled in the Study was slightly above 40 percent, although this lower-than-expected rate is attributable in part to some of the logistical limitations during the initial Vanguard Study.
- About 20 percent of the women who enrolled were lost to follow-up prior to giving birth. The most common reason cited for leaving was moving out of the designated Study geographic area. An additional 3 percent of attrition was due to pregnancy loss.

Solutions to Enhance Recruitment and Short-term Retention

The additional data analysis identified a few critical areas for potential improvement to increase efficiency of recruitment efforts. The first was additional outreach to pregnant women through targeted messaging and engaging health care providers, because nationwide more than 90 percent of women have some form of prenatal care. The second was to improve the retention rate between enrollment and birth by relaxing the stringency of the qualifying geographic areas, so that moving short distances does not change eligibility, and focusing on retention strategies. The National Children's Study target, which appears realistic based upon the experience of other studies, is to improve the short-term retention of pregnant women between time of enrollment to the birth of their child from the current 80 percent to over 90 percent.

Projections of Long-term Retention and Study Visit Completion Suggest a Change in the Sampling Design

The analysis of short term retention prompted a re-examination of long term retention. In the second quarter of calendar year 2011, the Program Office organized a retreat to discuss determination of Study sample size and subsequently initiated weekly discussions with sampling statisticians from around the country to determine appropriate and reasonable assumptions for retention and sample size (see Table 1). Teams from the National Center for Health Statistics of the Centers for Disease Control and Prevention, Research Triangle Institute, Westat, Battelle Institute, and Booz Allen Hamilton all produced models and simulations using different techniques and assumptions on retention and attrition. While attrition can vary among different subpopulations, the models were constructed using overall estimates for attrition. Using average annual attrition rates of between 1 percent and 5 percent and data collection duration of 21 years, collectively the models projected the proportion of original participants remaining in the Study at the end of scheduled data acquisition would be in the high 30 percent to low 40 percent range. Thus, if the Study enrolled 100,000 as initially projected, the number of participants remaining after 21 years would be approximately 40,000.

Table 1: Summary of Meetings to Plan Sampling Approach for the Main Study

Date	Meeting Topics
May 27, 2011*	NCS Sampling Retreat: Defining Study Locations, Secondary Sampling Frames, and Recruitment Targets *See Appendices A (Agenda) and B (Meeting Overview and List of Attendees)
June 3, 2011	Discussion of combining geographic and provider probability sampling models within or across primary sampling units in the NCS, discussion of increasing the number of primary sampling units
June 10, 2011	Discussed options to model and estimate the following parameters and include the findings in the NCS sampling strategy: attrition, movement of participants, demographic trends, trends in the incidence and prevalence of various conditions, the size of secondary sampling units, segment structure, expanding segment boundaries, compressing the duration of the active recruitment, and other possible options
June 17, 2011	Discussed the potential goal of examining exposure response relationships in populations or conditions with a smaller than 5 percent prevalence. Discussed options for shortened recruitment duration, increased size of secondary sampling units, increased numbers of primary sampling units, and increasing the size of primary sampling units, particularly in non-metropolitan areas
June 24, 2011	Discussed the enrollment target range to ensure a retained cohort of 100,000 after 21 years, factoring in attrition and Study visit compliance
July 1, 2011	Concluded with a discussion of a Retention and Compliance Model (Appendix E), the number of primary sampling units to add in order to reach the target range, and the parameters for selecting those primary sampling units

A further level of modeling projected how many participants would complete all scheduled data collections. Because estimates for the proportion of all scheduled collections that would be completed ranged from 50 percent to 75 percent (see Appendix E), the ability to analyze meaningful exposure response relationships for exposures or conditions with 5 percent or lower prevalence became tenuous. The Program Office became concerned that with the greater specificity of population and individual characterization, a sample size of less than half of those enrolled, combined with the likelihood of missing data points, would compromise the analysis of many conditions of potential interest. Table 2 displays estimates of current prevalence of conditions of potential interest.

Table 2: Estimated Prevalence of Various Conditions in the U.S. in 2010

Of 100,000 children, an estimated:
30,000 will be overweight (17,000 will be obese)
5,000 will have learning disorders
5,000 will have asthma
1,000 will have autism spectrum disorders
800 will have congenital heart disease
470 will have epilepsy
320 will have childhood cancers
125 will have Down syndrome
50 will have Fragile X syndrome

The study of diseases or conditions with a prevalence of 0.5 percent (500/100,000) is of interest for this longitudinal study. If we expect about 80 percent compliance with Study procedures, then of the 100,000 participants remaining in the Study at 21 years, about 400 would be sufficiently informative for analysis.

A scientific rationale for establishing a retention threshold is that the data resources for the National Children's Study are designed for and available for case control analyses. Given the trend to more granular characterization of diseases and conditions and personalized medicine, analyses of specific well-characterized populations in the larger Study database may provide essential informative longitudinal data.

Exploring a Larger Sample Size

The Program Office requested sampling statisticians who participated in the retreat and follow-up discussions to calculate the initial sample size needed to retain a sample of 100,000 at the end of the Study.

To initiate these discussions, Dr. L. Randy Curtin (National Center for Health Statistics, Centers for Disease Control and Prevention) developed a methodology for defining an increased cohort, including the set of information that would be required to use the existing strata and primary sampling units. He identified a set of parameters and documented how changes in each affected the design of strata, primary sampling units, and the development of a new design that incorporates existing primary sampling units. The parameters are as follows:

- The desired analytic sample size at 21 years
- The sample design effect (DEFF)
- The expected initial enrollment rate
- The expected 21 year attrition rate
- The length of the enrollment period
- Whether or not to allow variable sample size per PSU, especially for small PSUs.

After setting a small range of assumptions for the above parameters, a sampling statistician can determine the impact of each parameter on minimal measures of size for PSUs, average strata size, minimal measure of sizes for self-representing PSUs, and the allocation of sample across PSUs. Using this information, strata in the current design can then be reconfigured and additional PSUs randomly selected from the "new" strata.

Drs. David Hubble and Graham Kalton (Westat) further refined the methodology to incorporate the field experience of the initial Vanguard Study and produced sample sizes calculations using alternate scenarios of parameter estimates (Appendix D). Other factors that were included in the sample size calculations were breaking out of retention into two factors, retention of enrolled pregnant women to the birth, and retention of enrolled children from birth to 21 years. The rationale for this refinement is that the Study data as well as data from similar longitudinal cohort studies suggest that the greatest short-term attrition can be expected to occur during the earliest stages of the Study. Relatively low yield rates of enrolling pregnant women (0.37 or 0.5) shown in Columns 2 to 5 of Appendix D reflect the experience of a series of household-based recruitment approach employed in the initial Vanguard Study.

Scenarios 2 through 5 suggested targeting 1,140 to 1,540 pregnant women to be recruited and 460 to 600 births per PSU, per year for a 2-year enrollment period.

Discussion of these scenarios identified a need for more exploration of the expected retention rates from birth through age 21. Drs. Juergen Klenk, Anjene Addington, and Brendan Murphy (Booz Allen Hamilton) led the development of a predictive model to understand long-term retention and compliance and the effects of these on the final effective cohort size (Appendix E). Their model incorporated differing distributions of individual compliance rates and individual retention rates across varying Study visit schedules to model the outputs of per visit completion rates and the distribution of total visits completed. They illustrate how dropout rates (attrition rates) might have a relatively small effect on overall Study visit compliance in the early stages of a study, but in later years, dropout rates and individual Study visit compliance both have larger effects on overall Study visit compliance (Appendix E, page 9). These modeled compliance distribution effects were then applied to 3 initial target birth cohort sizes (100,000, 250,000, and 400,000) with 3 yearly dropout rates (5 percent, 2 percent, and 1 percent) in order to estimate the number of children completing all Study visits, more than 80 percent of Study visits, or more than 50 percent of Study visits (Appendix E, page 10). With a cohort size of 250,000 births and assuming average case Study visit compliance distributions, between 97,000 and 130,000 children can be expected to complete more than half of the planned Study visits, depending on 5 percent to 1 percent dropout rate. Using this model and the Vanguard experience, the Study can plan for a cohort of sufficient size and sufficient completion of Study visit assessments.

These analyses in aggregate suggest that 225,000 to 250,000 women should be enrolled to produce a final Study sample of about 100,000 children with adequate longitudinal data after 21 years (see Appendices C, D, and E). Additional preliminary analyses (data and strategy not shown) from others were consistent with the three detailed analyses.

The increase in sample size can be attained by both increasing the size of the eligible geographic segments within each of the current Primary Sampling Units, and increasing the total number of Primary Sampling Units (adding new locations to the current planned portfolio of locations). New locations would be added using the same strata and random selection process used to identify the initial locations. A new larger sampling frame is currently under development, with a probable total of 185 to 200 locations.

Key Terms:

Main Study: The longitudinal National Children's Study, measuring exposures and outcomes from before birth to age 21, to assess effects of the environment on child health and development. Environment is defined broadly, to include the biological, physical, chemical, and psycho-social cultural environments. Gene-environment interaction is a key factor as well.

Primary Sampling Units (PSUs): Individual components into which the target population is divided for the first sampling stage. In the National Children's Study, the PSUs roughly correspond to U.S. counties.

Recruitment Strategy/Recruitment Schema: A description of a recruitment methodology that will be subjected to analysis regarding feasibility, acceptability, and cost. The three alternate recruitment strategies undergoing assessment are:

- Provider-based Recruitment Strategy
- Enhanced Household-based Recruitment Strategy
- Two-tiered Recruitment Strategy (High Intensity/Low Intensity)

Secondary Sampling Units: In a multi-stage probability sample, primary sampling units (PSUs) are divided into smaller geographic areas called secondary sampling units (SSUs). In the National Children's Study, the boundaries for these SSUs correspond to "neighborhoods or communities."

Study Locations: The counties (or in sparsely populated areas, groups of contiguous counties) that were selected as primary sampling units in the first stage of sampling.

Appendices

Appendix A: Agenda for NCS Sampling Retreat 8

Appendix B: NCS Sampling Retreat Overview 9

Appendix C: Sample Size Considerations for a Revision of the Sample Design
for the National Children's Study 12

Appendix D: Rough NCS Sample Design Calculations 16

D1: Example Rough NCS Sample Size/Design Calculations 16
D2: Rough NCS Sample Size/Design Calculations with Alternate Scenarios 17

Appendix E: Retention and Compliance Model Results 18

Appendix A: Agenda for NCS Sampling Retreat

NATIONAL CHILDREN'S STUDY
DEFINING STUDY LOCATIONS, SECONDARY SAMPLING FRAMES, AND RECRUITMENT TARGETS
May 27, 2011
9am – 4:30pm
6100 Executive Blvd – 5C01
Bethesda, MD 20892

Call-in Information: Dial In: 877-459-1243; Participant code: 493570#

9:00	**WELCOME AND OVERVIEW** Introductions Purpose and Goals of the Retreat	J. Graber
9:15	**CURRENT NCS SAMPLE DESIGN**	R. Curtin
10:00	**NCS EXPERIENCE TO DATE**	B. Haugen
10:30	**BREAK**	
10:45	**DISCUSSION TOPIC: Study Locations** Discussion	 Attendees
11:15	**DISCUSSION TOPIC: Alternatives Secondary Stages of Sampling** Discussion	 Attendees
12:30	**LUNCH** (on your own)	
1:30	**DISCUSSION TOPIC: Sample Size Requirements** Discussion	 Attendees
2:15	**DISCUSSION TOPIC: Length of Recruitment Period** Discussion	 Attendees
3:00	**BREAK**	
3:15	**SUMMARY – PRIORITIES AND ACTION ITEMS**	S. Hirschfeld
4:30	**ADJOURN**	

Appendix B: NCS Sampling Retreat Overview

National Children's Study
Defining Study Locations, Secondary Sampling Frames, and Recruitment Targets

May 27, 2011
9am – 4:30pm
6100 Executive Blvd – 5C01
Bethesda, MD 20892

Call-in Information: Dial In: 877-459-1243; Participant code: 493570#

Purpose & Goals: The goal of the meeting is to solicit technical guidance leading to concrete recommendations related to sample design, enrollment requirements, and timing of participant accrual.

To best maximize field efficiencies, the NCS is considering alternative secondary sampling stages that reflect the characteristics of specific Primary Sampling Units. For example, recruitment efforts in areas with small numbers of health care providers may be more efficient if a sample of providers was selected instead of housing units.

Participants are asked to consider these issues in advance and submit additional topics or questions for consideration. We are considering this a working meeting designed to address key technical issues and identify solutions that may be quickly implemented.

Proposed Concept for Consideration: Field two separate arms of the NCS

1. Probability sample of unknown size to receive a high intensity protocol
2. Census of all eligible women in Primary Sampling Unit to receive a low intensity protocol

Key Questions for Consideration:

- **Study Locations**
  - Should the current number of Primary Sampling Units be retained? Should additional PSUs be added?

- **Secondary Sampling Stage**
    - Can alternative secondary sampling units be developed and still maintain a probability sample? Can this stage vary across Study locations? How do we define the reference population if multiple frames are used?
    - How is unit nonresponse handled with alternative secondary sampling units? Is replacement allowable and still maintain a probability sample?
    - How can the Study best manage ongoing development of weights from complex sampling frames?
    - Do we require a clustered sample to do hierarchical modeling by community?
    - If a geographic sample is retained, should SSUs be expanded?

- **Sample Size**
    - What are sample size options for the two separate arms of the Study? What are the implications for the effective sample size?
    - Is it feasible to increase the sample size within the known Primary Sampling Units? Should additional PSUs be added or should the Secondary Sampling Units be enlarged? What is the most effective approach to increasing the sample size?

- **Recruitment Period**
    - Can the length of recruitment vary from Study location to Study location? What are the implications of this?
    - What are the implications for phasing in recruitment across Study locations?
    - What are the advantages and disadvantages of shorter or longer recruitment durations?

Invitees

Randy Curtin (NCHS);	lester.curtin@cdc.hhs.gov
Jennifer Madans (NCHS);	jennifer.madans@cdc.hhs.gov
Chris Moriarity (NCHS);	christopher.moriarity@cdc.hhs.gov
Graham Kalton (Westat);	grahamkalton@westat.com
David Hubble (Westat)	DavidHubble@westat.com
Michael Brick (Westat)	MikeBrick@westat.com
Jill Montiquila (Westat);	jillmontaquila@westat.com
Colm O'Muircheartaigh (U of Chicago);	colm@uchicago.edu
Robert Michael (U of Chicago);	r-michael@uchicago.edu
Jonas Ellenberg (U of Pennsylvania);	jellenbe@mail.med.upenn.edu
Marc Berk (S-3);	MBerk@s-3.com
Claudia Schur (S-3);	cschur@s-3.com
Deborah Bitner (S-3);	dbittner@s-3.com

Vincent Iannacchione (RTI);	vince@rti.org
Phillip S. Kott (RTI);	pkott@rti.org
Jim Quackenboss (EPA);	quackenj@mail.nih.gov
Marty Frankel (CUNY);	martin_frankel@baruch.cuny.edu
Alan Guttmacher (NICHD)	
John Jarman (NICHD)	
Barry Portnoy (NIH OD)	
John Gohagan (NIH OD)	
NCS Program Office	
NCS PO Support/Multi-Contactor List	

Reference Materials: Current and historical reference materials related to the NCS sample design can be found at the Web site below:

http://www.circlesolutions.com/ncs/sampling

Appendix C: Sample Size Considerations for a Revision of the Sample Design for the National Children's Study

Prepared by L. Randy Curtin, NCHS/CDC

Some decisions need to be reached on the specifications, or range of specifications, to be used in adding PSUs and/or births to the NCS sample.

It is recognized that there is not a single sample size that can be attached to the National Children's Study and the various sample sizes result from the interaction of a number of factors. To progress further, assumptions, or ranges for assumptions, need to be made as illustrated in the examples below.

For a study, sample size determination can work forward (select a sample at baseline and monitor the attrition over time) or backward (set the final sample size at the 20-year mark and determine what initial sample size is needed to meet this). Suppose statistical considerations of precision and/or power determine a (SRS) size required. This srs sample size is often called the analytic or effective sample size. For a clustered data collection, sample or clinical trial, this effective sample size is then inflated by a DEFF because cluster samples tend to have DEFF greater than 1 for prevalence estimates (in general, model parameter estimates have DEFF closer to 1.0). This is the respondent/actual sample size for which data is to be collected. Because there is always non-response in a survey sample, the respondent/actual sample size must be inflated by the reciprocal of the overall sample response rate. For a 20-year longitudinal sample, the cumulative attrition rate over the 20 years must also be used to inflate the sample to determine the sample size required at the baseline time period.

The variables that feed into the discussion are: based on statistical considerations, the simple random sample size needed at baseline, the simple random sample size needed at the end of the 20 years, the expected DEFF for the sample design, the expected initial response rate for participation, and the 20-year attrition rate. For example, suppose a sample of 100,000 live births was selected. Suppose the initial response rate would be 65 percent (given information to date this seems very optimistic), suppose the cumulative retention rate over 20 years is 67 percent (only an average of 2 percent per year attrition), then the respondent/actual sample will be about 43,400 at the end of 20 years. If an analyst is interested in estimating a parameter with a DEFF of only 1.1, this leaves an effective sample size of about 39,450. It might be that the 100,000 sample size is mostly needed for the hypothesis to be tested in the first few years of life and perhaps the hypothesis to be tested in year 20 may still have sufficient power with a final sample of slightly under 40,000.

However, if it is determined that an analytic or effective sample size of 100,000 is needed in year 20, then the initial sample size is determined by fixing the DEFF, initial response rate, and 20-year attrition rate and calculating an initial sample size that will result in a final sample of 100,000. If the DEFF is 1.1, the initial RR is 65 percent and the 20-year attrition rate is 0.67, then the initial sample to be selected should be about 253,300 to still have 100,000 participants at year 20. For a less optimistic set of assumptions, say a DEFF of 1.1, an initial response rate of 0.4 and a cumulative 20-year RR of 50 percent then an initial sample of 600,000 would have to be selected to end with a sample of 100,000 in year 20.

These options are summarized in the following three tables. An excel spreadsheet is attached where assumptions can be changed to see how the final sample size changes.

Target NCS Sample Size 2011 Considerations

Goal - end with 100,000 (pessimistic)	factor	Sample number
Analytic sample at 20 years		100,000.00
Actual sample at 20 years (DEFF)	1.20	120,000.00
account for 20 years attrition	0.50	240,000.00
account for initial response rate	0.40	600,000.00

Goal - end with 100,000 (optimistic)	factor	Sample number
Analytic sample at 20 years		100,000.00
Actual sample at 20 years	1.10	110,000.00
account for 20 years attrition	0.67	164,767.36
account for initial response rate	0.65	253,488.24

Goal start with 100,000	factor	Sample number
Analytic sample at 20 years		39,450.00
Actual sample at 20 years	1.10	43,395.00
account for 20 years attrition	0.67	65,000.72
account for initial response rate	0.65	100,001.11

The next table, using the optimistic assumptions, shows a selected sample size of about 250,000 would yield a final sample after 20 years of close to 100,000.

Goal start with	250,000	factor	Sample number
Analytic sample at 20 years			98,500.00
Actual sample at 20 years		1.10	108,350.00
account for 20 years attrition		0.67	162,295.85
account for initial response rate		0.65	249,685.92

Changing the number of PSUs and the enrollment period will have a rather dramatic effect on the required annual number of births per PSU (and this impacts the number of smaller counties that need to be aggregated into a single PSU). Currently, the generic requirement is 250 births per PSU per year. If the enrollment period stays at 4 years, and the initial sample stays at 100,000, but the number of PSUs is increased to 185, then only 135 births per year per PSU are required. Increasing the total sample size to 250,000 requires 338 births per PSU per year. Then changing this last example to 2 years of enrollment implies a need for 676 births per PSU per year – this would cause considerable trouble if the NCS tried to maintain a roughly equal sample size per PSU. Alternatively, the number of PSUs in strata with small counties could be increased even more while maintaining roughly 250 births per small PSU. I would expect an extra 10 to 15 small, non-metro PSUs could be added giving a total of close to 200 PSUs for the NCS – but this has cost implications.

In addition, changing the number of PSUs and the total number of births to be selected affects the average stratum size and the definition of the MOS required for a PSU to be considered self-representing. The following table summarizes the impact.

number of PSUs	100	185.00	185.00	185.00
Initial sample size selected	100,000.00	100,000.00	250,000.00	250,000.00
Avg total sample per PSU	1,000.00	540.54	1,351.35	1,351.35
enrollment period	4 years	4 years	4 years	2 years
Annual sample per PSU	250.00	135.14	337.84	675.68
Total national births	16,065,000	16,065,000.00	16,065,000.00	8,035,000.00
Avg stratum size	160,650.00	86,837.84	86,837.84	43,432.43
Def of SR strata/PSU (75%)	120,487.50	65,128.38	65,128.38	32,574.32
Annual avg SR size	30,121.88	16,282.09	16,282.09	16,287.16

In summary, in order to proceed, the following specifications are required.

- The desired analytic sample size at baseline
- The sample design DEFF
- The desired analytic sample size at 20 years
- The expected response rate for the initial enrollment
- The expected 20-year attrition rate
- The length of the enrollment period.
- Whether or not to allow variable sample size per PSU, especially for small PSUs

Based on a single set of assumptions or a small range of assumptions for the above, the impact on minimal MOS for PSUs, average strata size, minimal MOS for SR PSUs, and the allocation of sample across PSUs can be determined. Strata in the current design can then be reconfigured and additional PSUs randomly selected with PPS from the “new” strata.

Appendix D: Rough NCS Sample Size Calculations

Prepared by David Hubble and Grahm Kalton, 6/29/2011

D.1: Example Rough NCS Sample Size/Design Calculations

Rough NCS Sample Size/Design Calculations

100,000	E0	Effective Sample Size at End of 21 Year Study
1	F1	Assumed Design Effect (DEFF)
100,000	E1	Actual Sample Size at End of 21 Year Study (E0*F1)
0.6	F2	Cumulative Retention Rate from Completed Birth Visit to 21 years
167,000	E2	Number of Births/Birth Visit Data Collections (E1/F2)
0.8	F3	Retention Rate of Enrolled/Consented Pregnant Women to Birth Visit
209,000	E3	Number of Enrolled/Consented Pregnant Women (E2/F3)
0.37	F4	Yield Rate of Enrolling Pregnant Women
565,000	E4	Potential Number of Pregnant Women Expected to Be Selected in Sample PSUs/Segments (E3/F4)
2	F5	Recruitment Period (in Years)
282,500	E5	Annualized Potential Number of Expected Pregnant Women (E4/F5)
184	F6	Number of Sample PSUs
1,540	E6	Per PSU Annualized Potential Number of Expected Pregnant Women (E5/F6)
460		Numbers of Births/Birth Visits per PSU per Year (E6*F3*F4)

D.2: Rough NCS Sample Size/Design Calculations with Alternate Scenarios

	Estimates (E) and Factors (F)	Alternate Scenarios				
		Column 1*	Column 2	Column 3	Column 4	Column 5
Effective Sample Size at End of 21 Year Study	E0	75,000	100,000	100,000	100,000	100,000
Assumed Design Effect (DEFF)	F1	1	1	1	1	1.1
Cumulative Retention Rate from Completed Birth Visit to 21 years	F2	0.75	0.6	0.6	0.5	0.5
Retention Rate of Enrolled Pregnant Women to Birth Visit	F3	1	0.8	0.8	0.8	0.8
Yield Rate of Enrolling Pregnant Women	F4	0.7	0.37	0.5	0.5	0.5
Recruitment Period (in Years)	F5	4	2	2	2	2
Number of Sample PSUs	F6	101	184	184	184	184
Actual Sample Size at End of 21 Year Study (E0*F1)	E1	75,000	100,000	100,000	100,000	110,000
Number of Births/Birth Visit Data Collections (E1/F2)	E2	100,000	167,000	167,000	200,000	220,000
Number of Enrolled/Consented Pregnant Women (E2/F3)	E3	100,000	209,000	209,000	250,000	275,000
Potential Number of Pregnant Women Expected (E3/F4)	E4	143,000	565,000	418,000	500,000	550,000
Annualized Potential Number of Expected Pregnant Women (E4/F5)	E5	35,800	282,500	209,000	250,000	275,000
Per PSU Annualized Potential # of Expected Pregnant Women (E5/F6)	E6	350	1,540	1,140	1,360	1,490
Numbers of Births/Birth Visits per PSU per Year (E6*F3*F4)		250	460	460	540	600

* This set of design factors roughly emulates the "Original" design assumptions

Shaded cell indicates change in factor value from the preceding column

Appendix E: Retention and Compliance Model Results

Prepared by Booz Allen Hamilton, 7/2011

Retention and Compliant Model Results Presentation.pdf