Appendix D

Statistical Documentation

In the following appendix sections we provide more detailed discussions of some of the topics discussed in Chapters 2, 3, 5 and 9. In particular, Section D-1 provides additional notes on dual and multi-frame samples and inference for these samples. Section D-2 provides further details of the qualified Centers utilized in the family of designs and presents one example of sampling qualified Centers. Section D-3 provides further details of the sample selection for the family of designs described in Chapters 3 and 5. Sections D-4 and D-5 discuss the calculation of design effects for the estimation of population means and the estimation of exposure/outcome relationships, respectively (note that the results of D-5 on design effects for relationships are preliminary and may undergo revision). Section D-6 provides several detailed figures presenting the power to detect relationships of interest for a simple random sample, and Section D-7 discusses further details of the methods for simulating correlated binary data and calculating power via simulation.

D-1 INFERENCE FOR DUAL AND MULTI-FRAME SAMPLES

As discussed in Chapter 2 of this report, dual- and multi-frame samples have a long history, going back to Hartley (1962) who noted that such designs can result in considerable cost savings over a single-frame design with similar precision. The general idea is that by drawing a study population using a combination of several different sampling frames, one can benefit from the advantages while minimizing disadvantages associated with each individual frame. Lohr and Rao (2000) provide some excellent examples of how dual-frame sampling might work. For example, a sample of individuals with Alzheimer's disease might be constructed by drawing some individuals from the general population (in order to ensure representativeness) and drawing others from senior care facilities (in order to reduce costs by sampling from a high prevalence population). They cite this Alzheimer's example as an illustration of the general principle of generating a sample of individuals with a rare disease by augmenting a population-based sample with one drawn from a high prevalence, yet incomplete population. Lohr and Rao also describe an example of particular relevance to the National Children's Study, namely Canada's National Longitudinal Survey of Children and Youth which was based on three different sampling frames. Two of the frames correspond to one used for the Labour Force Survey, before and after a redesign in 1995, while the third frame is the one used by the National Population Health Survey.

For the NCS, it is anticipated that a dual- or multi-frame sampling strategy would combine a broad probability-based population-wide sample (call this frame A) with a sample based on a Center-based recruitment strategy (e.g., recruitment through university-based medical Centers). For ease of discussion, call the latter frame B. By incorporating a sampling strategy based on frame A, the NCS will have a greater chance of being truly representative of the entire United States. For instance, such a sample will ensure appropriate representation of low-income subjects or subjects from minority ethnicities. However, the downside is that some of the subjects sampled from frame A might be more likely to refuse to participate in the study, or might be more difficult to retain (i.e., being more likely to drop out before study completion). A careful choice of frame B can potentially identify a more compliant population (lower refusal rates, easier tracking, greater cooperation with follow-up appointments, etc.). For example, study subjects recruited through a university-based medical Center already have built-in alternative tracking and contact mechanisms, as well as incentives to maintain contact with study

staff as part of receiving ongoing care for their child. While the use of dual-frame sampling is appealing from a heuristic perspective in terms of enhancing study validity, statistical analysis of data collected in such a manner poses considerable challenges. The purpose of this discussion is to present an overview of the various possible approaches to inference for dual- and multi-frame samples and to make some recommendations regarding the approach to be taken for the NCS.

In actuality, frame A is likely to be reasonably complex in itself, for example, involving clustering at the county or possibly census tract level, as well as stratification with respect to ethnicity, socioeconomic status, and other factors. In addition, sampling may be done in a multi-stage (or multi-phase) manner, for example, oversampling of cases for the purpose of measuring certain expensive or difficult to measure covariates. For the purpose of discussion, however, we begin with the assumption that sampling frames A and B are relatively simple.

In the more standard single-frame setting, a great deal has been written on the topic of model-based versus design-based statistical inference. Good reviews on the topic can be found in Rao (1999), as well as in the introductory chapters of Chambers and Skinner (2000). In brief, model-based approaches involve making distributional assumptions regarding the joint distribution of sampling indicators and observed data, while design-based approaches generally use more robust approaches based on weighted estimating equations. Chapter 1 of Chambers and Skinner provides a particularly lucid discussion which we describe briefly here for the purpose of framing our later discussion. Let y_i represent the value associated with the ith member of some population of interest, denoted by U. Suppose we wish to take a sample of size N from this population and that our goal is to estimate some characteristic of this population, for example, the population mean,

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \,. \qquad \text{(D-1)}$$

Let π_i be the probability that subject i is selected for sampling, let δ_i be an indicator of whether (δ_i =1) or not (δ_i =0) this same individual was actually selected and let $\{i : i \,\varepsilon\, S\}$ represent the set of N sampled individuals. As discussed by Chambers and Skinner, inference needs to consider the distribution of the sampling indicators themselves, as well as the actual sampled data. Although there are other approaches (e.g., Bayesian methods), most common methods are divided into two broad classes, namely design-based and model-based. The design-based approach is often described as focusing on the randomness of the sampling indicators, treating the population values, $Y_1, \ldots Y_N$, as fixed quantities. Estimators derived under a design-based approach are often found to be equivalent to the solutions to weighted estimating equations, with weights that reflect the sampling probabilities. For example, a design-based approach to estimating the population mean based on data observed in the sample can be found by solving the following equation:

$$\sum_{i \in S} w_i (y_i - \mu) = 0 \qquad \text{(D-2)}$$

where $y_1, \ldots. y_n$ represent the observed values for the sampled subjects (i.e., those with δ_i =1), and w_i is the inverse of the sampling probability, π_i=Pr(δ_i =1). Chapter 6 of Chambers and Skinner provides an excellent discussion of the link between familiar design-based estimators that combine strata-specific summary statistics with analysis based on unit level data, which will

generally involve solving estimating equations. An advantage of formulating at the unit level is that additional covariates are relatively easy to incorporate as well.

Model-based approaches involve making distributional assumptions on the Y_i's as well. While classical model-based inference tends to ignore the distribution of the sampling indicators, inference based on such an approach can lead to bias unless the sampling indicators are ignorable. A correct model-based approach should consider the joint distribution of the (δ_i, Y_i) pairs. In addition, a full model-based approach will consider the distribution of associated "design variables," which will generally correspond to individual characteristics (such as age, gender, ethnicity, SES, etc.) that might be used in specifying the design itself or which might affect the distribution of the outcomes of interest. The full joint distribution can be factored as follows:

$$f(Y_i, Z_i, \delta_i) = f(\delta_i | Y_i, Z_i) f(Y_i | Z_i)\, f(Z_i). \qquad \text{(D-3)}$$

Since Y_i and Z_i can be observed only if $\delta_i = 1$ (i.e., the subject is actually sampled), the observed data likelihood will involve integrating over the distribution of these random variables. In some cases (e.g., when the selection probabilities do not depend on the unobserved data) the sampling mechanism is said to be ignorable and relatively standard methods can be applied. More generally, it will be necessary to consider the sampling mechanism as part of the likelihood construction in order to assure valid inference on the parameters of interest. While it can be argued that a model-based approach might gain in statistical efficiency when the modeling assumptions are correct, the downside to such an approach is that it can lead to bias if the distributional assumptions are wrong. Rao (1998) provides a good discussion on this topic and explains that for this reason, many researchers prefer a design-based approach. In particular, Rao explains that the weighted estimating equation above will still lead to valid inference, even if the distributional assumptions on the population data are incorrect. Use of the weighted estimating equation approach is in fact quite consistent with the increasing popularity among statisticians of generalized estimating equations as an alternative to likelihood-based inference.

Lohr and Rao (2000) discuss extensions of design-based estimators to the dual-frame setting. To do so requires extending notation to reflect the existence of two different frames. For example, we need to specify two sets of sampling probabilities, π_i^A being the probability of being sampled into frame A and π_i^B being the probability of being sampled into frame B. Let w_i^A and w_i^B be the inverse of these sampling probabilities. Similarly, we need notation for two different sampled populations, S_A referring to those sampled from frame A and S_B those sampled from frame B. Lohr and Rao consider that the population of interest (U) can be divided into three mutually exclusive domains, $a = A \cap B^c$ reflecting members of the population who are in frame A but not frame B, $b = A^c \cap B$ or those who are in frame B but not frame A and $ab = A \cap B$ or those who are in both frames. Lohr and Rao go on to describe a variety of different estimators that combine estimates corresponding to the three domains *a, b,* and *ab*. The estimates and inference based on them differ according to whether or not the sizes of the various domains can be assumed known. We describe here the relatively simple case where these are known and equal to N_a, N_b and N_{ab}, respectively. Defining the domain specific estimators as follows,

$$\hat{Y}_a^A = \frac{1}{N_a} \sum_{i \in S_A} w_i^A (1 - \delta_i^B) Y_i \tag{D-4}$$

$$\hat{Y}_{ab}^A = \frac{1}{N_{ab}} \sum_{i \in S_A} w_i^A \delta_i^B Y_i \tag{D-5}$$

$$\hat{Y}_{ab}^B = \frac{1}{N_{ab}} \sum_{i \in S_B} w_i^B \delta_i^A Y_i \tag{D-6}$$

$$\hat{Y}_b^B = \frac{1}{N_b} \sum_{i \in S_B} w_i^B (1 - \delta_i^A) Y_i \tag{D-7}$$

Lohr and Rao describe various proposals for estimating the population mean using combinations of these four estimates. A particularly appealing approach is one that estimates population parameters of interest by treating all observations as though they were drawn from the same frame, but using modified weights for observations that fall into the intersection (Lohr and Rao cite papers by Bankier, 1986, Kalton and Anderson, 1986 and Skinner, 1991). An additional advantage of this approach is that it lends itself to formulating estimators based on unit level data (see Chapter 6 of Chambers and Skinner). Defining $w_i = 1/\pi_i^A$ for subjects in frame *a,* $w_i = 1/\pi_i^B$ for subjects in frame *b,* and $w_i = 1/(\pi_i^A + \pi_i^B)$ for subjects in *ab,* it follows that a simple but valid design-based estimator is given by the solution to the following equation:

$$\sum_{i \in S_A} w_i (y_i - \mu) + \sum_{i \in S_B} w_i (y_i - \mu) = 0 . \tag{D-8}$$

Note that there is no requirement here to identify individuals who happen to appear in both samples. Of course in practice when sampling from large populations, it is unlikely that there would be many such individuals. A particularly appealing aspect of this formulation is that it can be easily generalized to handle covariates. Such models have not been extensively explored in the statistical literature at this time, and represent a promising avenue for future research relevant to the NCS.

D-2 SAMPLING OF QUALIFIED CENTERS

As described in Chapter 5 of this report, Centers capable of performing the tasks associated with the NCS would likely be selected through a competitive procurement process, in which the Centers would demonstrate their ability and capacity to perform appropriate data collection activities. Since this formal process is not available at this stage, a surrogate list that includes 105 medical research institutions with affiliated hospitals and their total dollar amount of National Institutes of Health (NIH) research grants (averaged over 2001 and 2002) was obtained from *U.S. News and World Report*. Table D-1 displays these 105 Centers along with their amount of NIH funding in millions of dollars, and Figure 5-1 in Chapter 5 illustrates the distribution of these Centers across the U.S. (one of the institutions is in Puerto Rico and is not displayed in Figure 5-1).

It should be noted that this list includes only research centers that have had NIH grant funding, so that some large medical Centers not affiliated with a university, such as the Cleveland Clinic, and some research universities not affiliated with particular hospitals, like the University of California, Berkeley are excluded from the list. This is not meant to suggest that we are implicitly assuming that these respected institutions would be excluded from participating in the NCS. However, as a starting point for understanding the portion of the population that might be within the geographical area of a Center, this list, although not a comprehensive enumeration of every institution that might compete to participate in the NCS, was utilized.

Table D-1. Table of 105 Centers and their average 2001-2002 NIH research funding (in millions of $)

Centers	Grant Funding	Centers	Grant Funding	Centers	Grant Funding
Harvard University (MA)	957.8	Mayo Medical School (MN)	126.4	University of Louisville	29.2
University of Washington	431.5	University of Minnesota–Twin Cities	124.9	Medical College of Georgia	25.8
University of Pennsylvania	431.4	University of Rochester (NY)	118.4	New York Medical College	22.1
Baylor College of Medicine (TX)	382.8	University of Virginia	108.6	University of Nebraska College of Medicine	22
Johns Hopkins University (MD)	372.6	University of Maryland	108.5	Uniformed Services Univ. of the Health Sciences	17.8
University of California–San Francisco	368.7	Wake Forest University (NC)	95.9	Drexel University	16.6
University of California–Los Angeles (Geffen)	340.5	University at Buffalo (NY)	91.2	University of North Dakota	14.5
Washington University in St. Louis	320.4	Indiana University–Indianapolis	90.6	Medical College of Ohio	11.9
Columbia U. College of Phys. and Surgeons (NY)	260.5	University of Miami (FL)	90.1	University of Mississippi	11.6
University of Michigan–Ann Arbor	255.7	University of Massachusetts–Worcester	88.1	University of Missouri--Columbia	10.9
University of Pittsburgh	247.8	Brown University (RI)	86.4	West Virginia University	10.3
Yale University (CT)	245.4	Tufts University (MA)	83.3	Texas A&M Univ. System Health Science Center	9.9
Cornell University (Weill) (NY)	227	Dartmouth Medical School (NH)	80.2	Michigan State University	9.8
Duke University (NC)	225	Jefferson Medical College (PA)	76.3	Wright State University	8.2
Stanford University (CA)	198.4	University of Utah	72.1	Michigan State Univ. College of Osteopathic Med.	8
University of California–San Diego	195.7	University of Florida	71.8	Creighton University	7.9
University of Alabama–Birmingham	195.7	Medical College of Wisconsin	69.4	University of South Dakota	7.1
Vanderbilt University (TN)	180.2	Georgetown University (DC)	64.2	Ponce School of Medicine	6.6
Case Western Reserve University (OH)	177.5	Univ. of Texas Health Science Center–Houston	59.5	U. of North Texas Health Sci. Center (Texas Col. of Osteopathic Medicine)	6.3
University of North Carolina–Chapel Hill	162.1	Stony Brook University	59.4	University of South Carolina	5.4
Yeshiva University (Albert Einstein) (NY)	157.7	Medical University of South Carolina	59	Eastern Virginia Medical School	5
Boston University	155.9	Va. Commonwealth U.–Medical Col. of Va.	55.5	Texas Tech University Health Sciences Center	4.9
University of Southern California	148.9	Wayne State University (MI)	50.4	UMDNJ--School of Osteopathic Medicine	4.1
U. of Texas Southwestern Med. Center–Dallas	148.8	University of Vermont	49.4	East Tennessee State University (J.H. Quillen)	3.2
Oregon Health & Science University	146.3	UMDNJ-Robert Wood Johnson Medical School	49.4	Southern Illinois University--Springfield	2.7
University of Cincinnati	145.8	University of California–Davis	47.3	Mercer University	2.2
Univ. of Colorado Health Sciences Center	143	UMDNJ-New Jersey Medical School	40.3	Northeastern Ohio Universities College of Medicine	1.9
Northwestern University (Feinberg) (IL)	141.1	University of New Mexico	38.1	University of Minnesota--Duluth	1
Univ. of Iowa (Roy J. & Lucille A. Carver)	136.5	Tulane University (LA)	37.4	Ohio University College of Osteopathic Medicine	0.9
Emory University (GA)	134.5	University of Kansas Medical Center	36.8	Univ. of New England College of Osteopathic Med.	0.7
University of Wisconsin–Madison	133.3	University of California–Irvine	35.9	Chicago College of Osteopathic Medicine	0.6
Ohio State University	133.3	George Washington University (DC)	32.9	Kirksville College of Osteopathic Medicine	0.4
Mount Sinai School of Medicine (NY)	132.3	University of Oklahoma	31.6	Philadelphia College of Osteopathic Medicine	0.3
New York University	128	University of South Florida	30.1	Oklahoma State Univ. College of Osteopathic Med.	0.3
University of Chicago	126.4	St. Louis University	29.9	Arizona College of Osteopathic Medicine	0.1

As displayed in the table, for each of these 105 Centers the total dollar amount of NIH research grants awarded to the medical school and its affiliated hospitals (averaged over 2001 and 2002) was available. In addition to these data, Census data were merged with the list of Centers to provide counts of the total population, the population of children 0 to 3 years of age, the population of females of childbearing age, and the total number of households in the respective counties and metropolitan statistical areas (MSAs) where the Centers are located. Finally, data on the annual number of births at each Center were obtained directly from each Center for all but 34 of the Centers. For example, Figure D-1 displays a map of the 105 Centers along with the number of children between the ages of zero and three years for their corresponding counties.

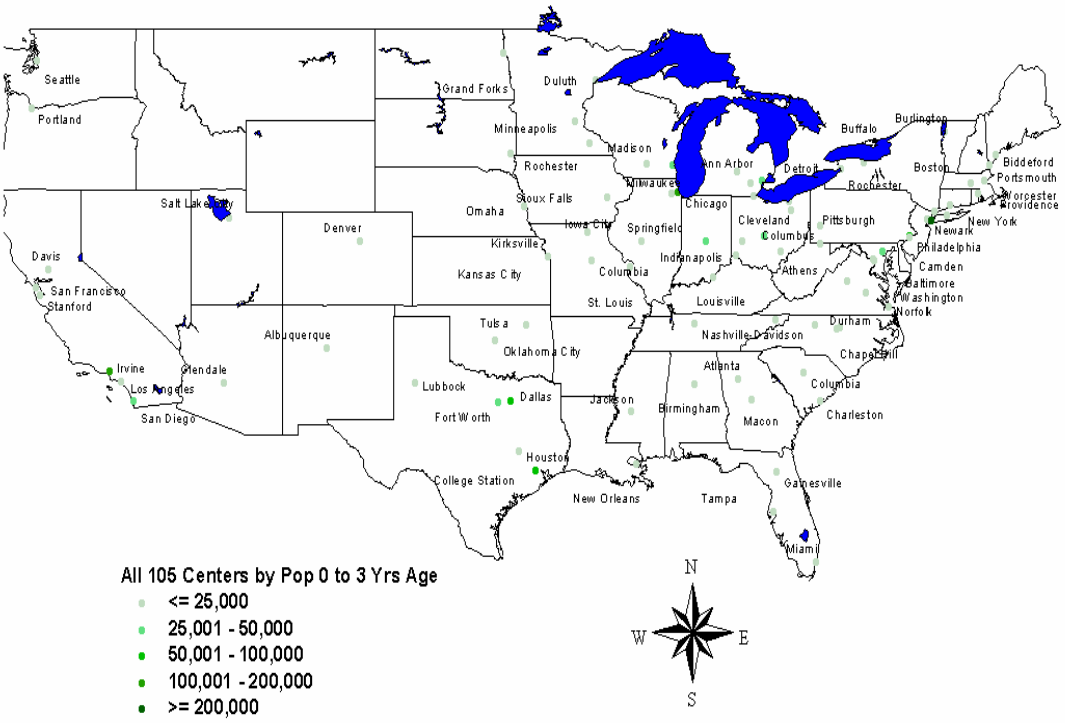

Figure D-1. Map of U.S. Indicating Locations of All the 105 Centers With County Population of Children 0 to 3 Years Old.

As mentioned in Chapter 5, several methods for selecting Centers from the list of Centers were considered. Probabilistic approaches included probability proportional to size (PPS) sampling from the list of Centers, with size defined by 1) total amount of NIH funding, 2) number of households or children aged 0 to 3 years in the geographical area, and 3) number of births annually at the Center. Here we demonstrate some of the results of applying a PPS sampling approach where the size of the Center is defined as their total amount of NIH funding. For example, Table D-2 displays the number of children between the ages of zero and three (i.e., representing the number of births in the county for a four-year recruitment period) as a function of the number of Centers and the proportion of the population that is selected in each of the counties corresponding to the selected Centers. Note that to achieve at least 100,000 children, at least 6 to 7 percent of the children would need to be selected in each of the counties and up to 50 Centers would need to be selected. Thus, Figure D-2 displays a realization of the counties that are selected if PPS sampling of 50 Centers is conducted. (Note that some of the Centers end up being selected with certainty since their relative size exceeds the inverse of the Centers sample size.) These are denoted as certainty counties in the figure, and all other counties are denoted as uncertainty counties.

Table D-2. Total number of children aged 0 to 3 sampled from counties where medical research centers are located.

Number of Centers	Percent of Sampled Population from Counties (Average over 20 Trials)									
	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
30	16,947	27,218	36,228	44,759	52,801	60,410	67,138	72,925	77,865	81,764
31	18,653	29,809	39,113	47,832	56,048	63,852	70,735	76,579	81,540	85,412
32	18,126	29,386	38,879	47,868	56,386	64,460	71,687	77,909	83,230	87,264
33	18,674	30,314	40,271	49,529	58,303	66,653	74,107	80,579	86,124	90,339
34	18,998	31,266	41,599	51,292	60,494	69,302	77,222	84,078	89,961	94,568
35	19,247	31,324	41,886	51,808	61,197	70,139	78,203	85,166	91,069	95,562
36	20,342	33,218	44,327	54,674	64,319	73,354	81,434	88,471	94,441	99,022
37	20,853	33,902	45,354	56,102	66,131	75,576	84,075	91,448	97,665	102,399
38	20,821	34,376	46,164	57,034	67,150	76,680	85,177	92,621	98,910	103,782
39	22,062	36,173	47,759	58,609	68,762	78,391	87,030	94,552	100,966	105,861
40	22,938	37,756	50,134	61,613	72,205	82,108	91,014	98,908	105,629	110,810
41	22,683	37,373	49,959	61,593	72,423	82,569	91,715	99,792	106,753	112,143
42	23,491	38,262	50,593	62,134	72,956	83,179	92,407	100,484	107,346	112,607
43	24,478	40,310	53,365	65,521	76,752	87,214	96,690	105,018	112,151	117,752
44	24,905	40,764	53,846	65,975	77,353	88,166	97,954	106,519	113,774	119,377
45	23,883	39,951	53,327	65,858	77,562	88,617	98,661	107,556	115,101	120,976
46	25,262	41,477	54,916	67,672	79,708	91,161	101,643	111,049	119,018	125,125
47	26,119	43,160	57,340	70,568	82,861	94,384	104,927	114,345	122,377	128,578
48	26,348	43,555	57,732	71,011	83,458	95,212	105,960	115,583	123,790	130,227
49	26,952	44,699	59,248	72,845	85,543	97,483	108,423	118,142	126,421	132,915
50	27,378	45,550	60,525	74,332	87,186	99,246	110,280	120,092	128,458	135,005

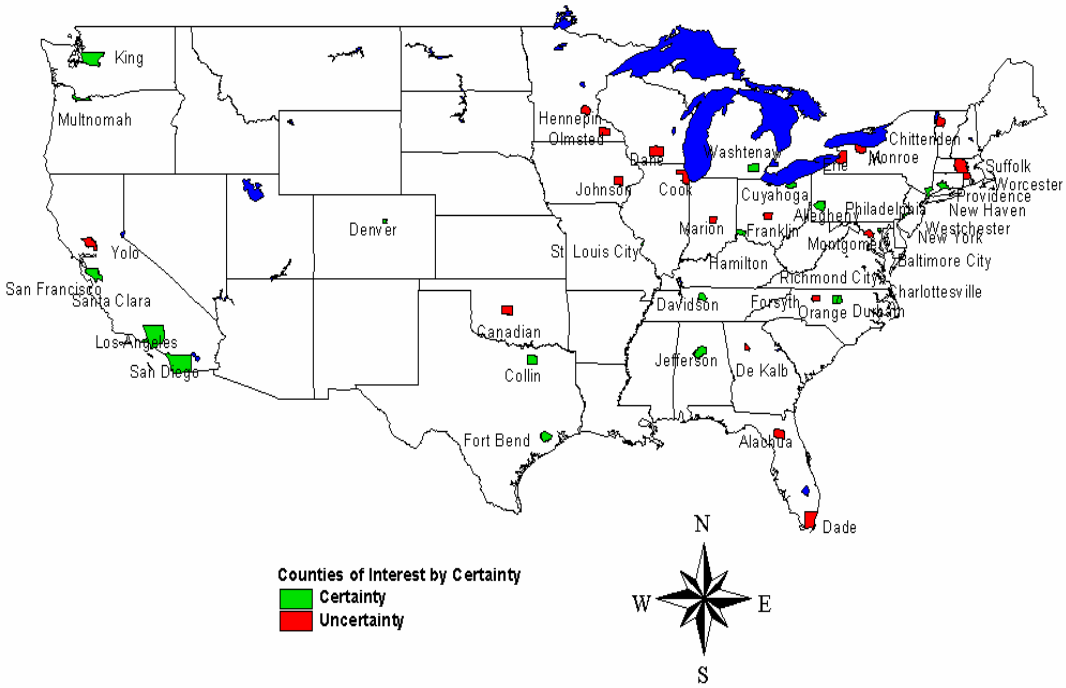

Figure D-2. Map of U.S. indicating counties corresponding to 50 sampled Centers (using PPS sampling with size defined as the amount of NIH funding).

Table D-3, on the other hand, displays the number of children between the ages of zero and three (i.e., representing the number of births in the county for a four-year recruitment period) as a function of the number of Centers and the proportion of the population that is selected in each of the MSAs corresponding to the selected Centers. Here we see that a smaller percentage of the children could be selected in each MSA and a smaller number of Centers could be selected. Thus, Figure D-3 displays a realization of the MSAs that are selected if PPS sampling of 43 Centers is conducted. (Again, note that some of the Centers end up being selected with certainty since their relative size exceeds the inverse of the Centers sample size.)

Table D-3. Total number of children aged 0 to 3 sampled from MSAs where medical research centers are located.

Number of Centers	Percent of Sampled Population from MSAs (Average over 20 Trials)				
	1%	2%	3%	4%	5%
30	55,883	78,060	88,057	94,798	99,194
31	56,676	79,807	90,382	97,586	102,225
32	57,822	81,163	92,070	99,561	104,691
33	60,061	84,407	95,623	103,622	108,959
34	59,519	85,143	97,399	105,829	111,537
35	63,671	88,953	101,459	109,908	115,388
36	64,095	90,829	104,273	113,523	119,598
37	67,135	94,502	107,967	117,199	123,395
38	66,050	93,399	107,171	117,040	123,873
39	67,854	96,221	110,494	120,606	127,490
40	69,523	97,801	112,448	122,726	129,671
41	71,269	100,506	115,507	125,875	132,935
42	73,148	103,338	118,476	128,989	136,086
43	75,625	105,332	120,591	131,380	138,441
44	76,112	106,781	122,681	133,813	141,389
45	77,981	109,699	126,422	138,283	146,339
46	79,769	111,855	128,517	140,334	148,330
47	81,016	113,091	129,872	141,980	150,270
48	83,323	115,391	132,521	144,806	153,261
49	84,091	117,464	135,190	147,858	156,676
50	85,210	119,494	137,594	150,529	159,619

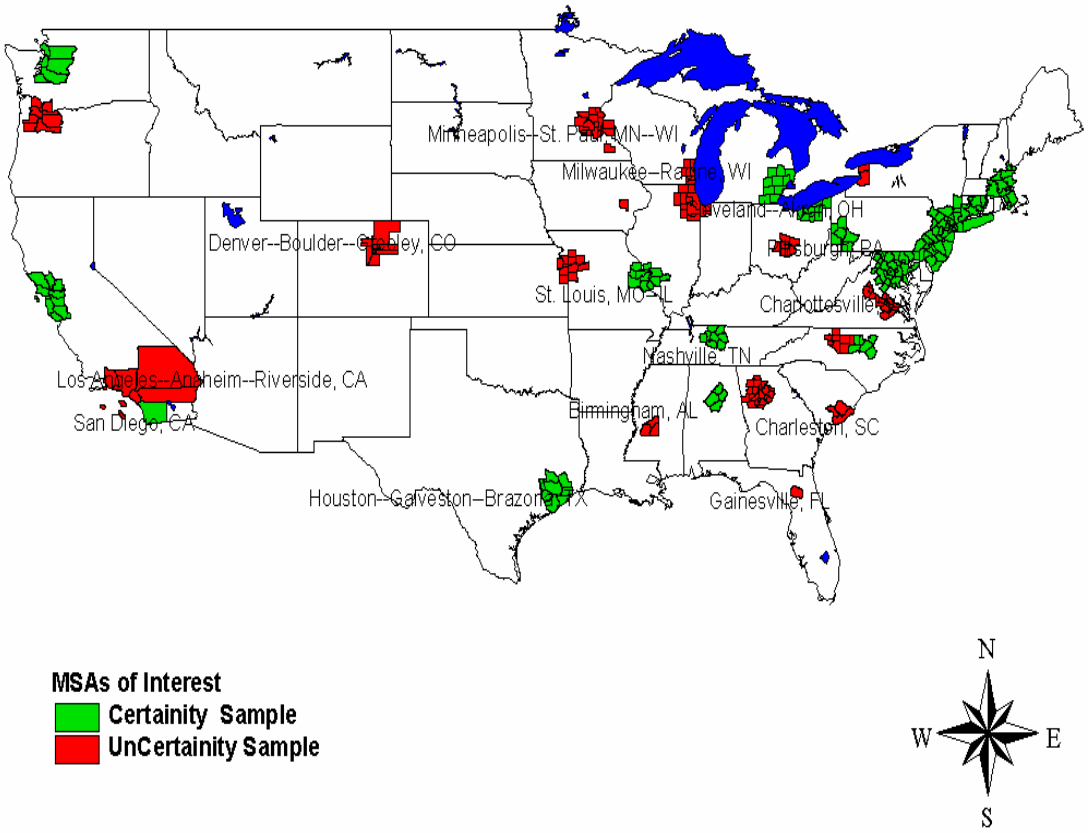

Figure D-3. Map of U.S. indicating MSAs corresponding to 43 sampled Centers (using PPS sampling with size defined as the amount of NIH funding).

These examples provide insight into some of the plausible properties of a design that is based on a Centers approach to sampling. For example, the results may aid in identifying appropriate numbers of Centers to include in the NCS, and illustrate the potential that a Centers design has in providing geographic coverage of the nation. However, it should be noted that the Centers approach may suffer from selection of a disproportionate number of subjects from urban areas since typically respected medical Centers are in these more urban areas. To mitigate this, Chapter 5 suggests that it may be possible to establish one or several Centers in more rural areas, such as the Children's Health Center operated by UC Berkeley in the Salinas Valley Farm Community.

The examples above illustrate a probabilistic approach to sampling Centers from the list of Centers (admittedly this list is limited at this stage). Of course, as noted in Chapter 5 of this report, it is more likely that Centers would be selected in a purposive manner through a

procurement process, and thus, a purposive approach to sampling Centers was implemented in the designs described in Chapter 5 of this report (and utilized in Chapter 9).

D-3 DETAILS OF SAMPLE SELECTION

Figure 3-1 of Chapter 3 provides a view of the structure of the samples used in the simulation. Each sample is the result of a mix of two sampling strategies that together provide about 100,000 eligible women. A portion, P_1, of the sample comes through the use of a national probability-based sample (NPBS) and the remainder, $1-P_1$, comes through the use of a purposively selected Centers sample. The NPBS selects households through a probability sample that first selects a fixed number of counties and then households within the selected counties. We will refer to the resulting NPBS sample as the two-stage sample (or two-phase sample). The second strategy begins by choosing a predetermined number of medical Centers that provide services to pregnant women. The number of potentially eligible women served by a Center dictates whether or not a Center is chosen. Once the Centers are selected, each center is associated with a random sample of eligible women referred to as a Center sample.

Center samples in turn come from two sources. A portion of the Center sample, P_2, comes through a probability sample of households from the metropolitan area served by the Center (the Center's MSA). The remaining portion, $1-P_2$, is obtained through random sampling of the Center's list of patients.

Each sample used in the simulation was derived using one of 21 parameter settings. The variations are listed in the first three columns of Table 5-4 in Chapter 5. The number of counties selected as part of the two-stage sample was set to 0, 50, or 100. A setting of 0 meant that the entire sample was gathered through Centers. In this case, P_1 was set to 0 and P_2 to 0.25, 0.50, or 0.75. When 50 or 100 counties were selected, both P_1 and P_2 could vary from 0.25 to 0.75 in increments of 0.25. The simulation included 50 samples of each of the 21 sample types.

In the following paragraphs we describe in more detail the two-stage NPBS and Center samples.

D-3.1 THE TWO-STAGE (NPBS) SAMPLE

Counties were selected as the primary sampling units (PSUs) of the two-stage sample. A stratified sample based on region and degree of urbanization was used to ensure representation of different kinds of counties. The number of counties entering the sample from a given stratum was proportional to the relative size of the stratum as measured by the number of households. Within a stratum, probability proportional to size (PPS) sampling was used to select the stratum's counties. If a county's chance of selection met or exceeded 1, it was removed from its assigned stratum to form its own stratum in the PSU sample (i.e., a certainty strata).

Households were the second stage sampling units (SSUs). An attempt was made to choose the same number of households from each county. If a county had fewer households than the number required, it was presumed that all households in the county would be sampled. In

order to maintain the overall required number of sampled households in the face of counties with insufficient numbers, sampling from counties with sufficient household numbers was uniformly increased.

As outlined in Chapter 5, we presumed that 1 of 12 selected households would result in the recruitment of an eligible woman. This meant that P_1*100,000*12 households had to come from the two-stage sample.

D-3.2 THE CENTER SAMPLE

Centers were selected from a list of 104 medical research institutions (note that the Center in Puerto Rico was excluded from the list of Centers for the purposive sampling of Centers). The institutions had been recent recipients of National Institutes of Health (NIH) grant funding. An annual number of births was associated with most of the centers, derived through contacting centers or through web-based resources. Centers for which birth rates could not be found were assigned the median value of the non-missing rates.

The number of women entering the sample through a Center sample was fixed at 2000. Given that $(1-P_1)$* 100,000 of the desired sample of women was to come through the Centers, the number of Centers that had to be selected was given by the smallest integer (n_c) bigger than $(1-P_1)$*100,000 /2000. To pick the needed number of Centers, we simply selected the top n_c Centers in terms of annual birth rate.

A Center sample was derived through two mechanisms: list sampling and area sampling. List sampling gathered from each Center $(1-P_2)$*2000 women through a random sample of eligible women (i.e., women that give birth) from the list of patients served by the Center. In this sampling, all eligible women at a Center were presumed to have an equal chance of entry into the sample.

Center area sampling involved random sampling of households in a center's MSA. From each center, 12* P_2*2000 households were sampled, with each household having an equal chance of selection. The factor of 12 is derived from the presumption that 1 of 12 selected households will result in the recruitment of an eligible woman (i.e., a live birth).

For Center area sampling, counties selected as part of the two-stage sample were excluded from Center areas. When all counties in a Center MSA were selected as part of the two-stage sample, no restrictions were placed on the counties targeted for that Center area. If a county participated in both the two-stage and Center area sampling, the chance of a woman in that county being selected into the sample was set to the chance of being selected into the two-stage or Center-area sample.

D-4 DESIGN EFFECTS FOR ESTIMATION OF POPULATION MEANS AND PERCENTAGES

As suggested in Chapter 5 of this report, there is a trade-off between the cost and complication of coordinating a long-term longitudinal cohort study spread over many PSUs, and the loss of information inherent in highly clustered data when the study is not spread over many PSUs. Additionally, if weighted analyses are to be utilized, there may also be a loss of information due to unequal sample weights for the selected units. In this section, we consider this issue of information loss by discussing design effects associated with the estimation of *population means and percentages*. Note that the primary purpose of the NCS is estimation of relationships between exposures and outcomes, not estimation of population means and percentages, suggesting that design effects for these estimates are not of primary interest. However, since design effects for population means and percentages are relatively straightforward to calculate, they offer an important starting point for this discussion. In the following section of this appendix we provide more relevant results on the design effects associated with estimation of relationships.

In general, the design effect measures the loss of information by computing the ratio of the variance under the selected sample design (e.g., a clustered and unequally weighted design) to the variance that would be realized under a design that involves simple random sampling. Thus, the design effect can be thought of in terms of the impact of different weighting and clustering schemes on the estimated variances of parameters of interest. For estimation of population averages, calculation of the design effect involves consideration of both the clustering associated with the design and the unequal weighting that is realized in the design. In the following paragraphs we outline how these two factors are incorporated into the design effect.

The effect of PSU level clustering on estimating a *population mean* can be expressed in terms of the design effect (DE_{CL}) using the following formula

$$DE_{CL} = 1 + \delta_{PSU}(\bar{n}_{PSU} - 1) \tag{D-9}$$

where $\bar{n}_{PSU}$ is the average number of participants per PSU, and δ_{PSU} is the PSU intraclass correlation coefficient. Intraclass correlation coefficients measure the homogeneity of the parameter of interest within a cluster, in this case within a PSU, and, for a particular population parameter and class (PSU), can be estimated using a variance component breakdown of the variability in measurements of the parameter. In particular, letting σ_B^2 be the mean square deviation among the classes, average subject-level measurement of the parameter and letting σ_W^2 be the average across classes of each PSU's individual mean square deviation in measurement error across subjects, an estimate of the PSU intraclass correlation coefficient is

$$\delta_{PSU} = \frac{\sigma_B^2 - \dfrac{\sigma_W^2}{\bar{n}_{PSU} - 1}}{\sigma_B^2 + \sigma_W^2}. \tag{D-10}$$

In the Westat report (2002) a general (or average) PSU (recall that PSUs are counties here) intraclass correlation coefficient of 0.01 is presented. It is based on data from a wide variety of items studied in Cycles II and IV of the National Survey of Family Growth (NSFG). Here we use generalized estimating equations and the NHANES III data to estimate the intraclass correlation coefficients for asthma, injury, obesity, and low birth rate outcomes as surrogates for the responses of interest in the NCS core hypotheses. Table D-4 presents the NHANES III intraclass correlation coefficient estimates (note that some of these estimates are negative indicating that they are essentially zero). Based on these estimates, we assumed the range of reasonable intraclass correlations was from 0.005 to 0.02, and use values from this range in the results of Chapter 9 (e.g., 0.01 to 0.02 for asthma and injury, and 0.005 for outcomes that occur less frequently such as autism). (Of course, note that correlations exhibited in the NCS data may be higher or lower than those observed in the NHANES III data.)

Table D-4. Intraclass correlation coefficient estimates based on NHANES III data.

Condition	Age Group	Intraclass Correlation Coefficient
Asthma	2-11 months	0.011
	12-35 months	0.000
	3-5 years	0.010
	6-11 years	0.013
	12-19 years	0.020
Injury	2-11 months	-0.003
	12-35 months	0.001
	3-5 years	0.004
	6-11 years	0.018
	12-19 years	0.009
Obesity	2-11 months	-0.006
	12-35 months	-0.004
	3-5 years	0.020
	6-11 years	0.007
	12-19 years	0.004
Low Birth Weight	2-11 months	0.0152
	12-35 months	0.0029
	3-5 years	0.0049
	6-11 years	0.0070
	12-19 years	-0.0037

As noted previously, in addition to clustering, unequal sample weights have an effect on the variability of population mean (or percentage) estimates. This effect can also be expressed in terms of a design effect (DE_W) by

$$DE_W = \frac{n\sum_j w_j^2}{\left(\sum_j w_j\right)^2}, \qquad \text{(D-11)}$$

where the w_j are the sample weights (Kalton et al., 2003). Note that the form of this equation suggests that increased variability in the weights leads to increased variability in overall population estimates. Recall that Section 5.1 discusses implementation of a national probability-based sample, and provides examples of selecting a national probability-based sample with 50 and 100 PSUs. For the weight distributions summarized in Figure 5-2 (single realizations of national probability designs with 50 and 100 PSUs), the value of DE_W for the sample with 50 PSUs is 1.40 and the value of DE_W for the sample with 100 PSUs is 1.16. This is not surprising as two of the issues impacting weights, small counties and integer allocation of PSUs to strata, are mitigated with more PSUs (i.e., the definition of a small county includes fewer counties when more PSUs are available, and integer proportional allocation strays less from proportional allocation when more PSUs are allocated).

Table D-5 presents estimated clustering contributions to design effects for PSU sample sizes of 50 and 100 under intraclass correlation coefficient assumptions deemed appropriate for each core hypothesis measure (see Table D-4). Also presented are estimated design effects due to unequal weighting for PSU sample sizes of 50 and 100. The effects provided in the table were estimated using 50 simulated samples of each design, and assuming that the entire cohort is selected in a national probability-based sample. Design effects for the multi-frame (hybrid) designs described in Chapter 5, where a portion of the cohort is selected in a national probability-based sample and a portion is selected through a set of purposively selected Centers, are described below. Note from the table that there appears to be a large design effect due to clustering of the data, and a somewhat smaller design effect for the unequal weighting resulting from the design.

Table D-5. Effect of PSU sample size on clustering and weighting contributions to the design effect

Number of PSUs	Clustering Design Effect				Weighting Design Effect
	Asthma	Injury	Obesity	Low Birth Weight	
50	21.0	11.0	21.0	11.0	1.24
100	11.0	6.0	11.0	6.0	1.09

For the 21 hybrid designs described in Chapter 5, and utilized in Chapter 9, Table D-6 provides the estimated design effect *for estimating a population average* (or percentage) due to weighting for each of the hybrid designs defined by the levels of P_1 and P_2 considered (see Chapter 5 or Chapter 3). Since the number of clusters and the weighted average sample size within clusters are important to assessing design effects due to clustering within a design (see

formulas above), these values are also provided in Table D-6 for each of the 21 designs considered. Recall that we assume that each Center can recruit and follow a total of 2000 subjects, and thus, given the number of PSUs, and given the value of P_1, the number of clusters is determined by simply adding the number of national probability sample PSUs and the number of Centers. (Note that in Chapter 9, the number of clusters is defined slightly differently as the number of NPBS PSUs plus two times the number of Centers.)

Note from Table D-6 that the design effect due to weighting generally decreases as the portion of the cohort selected in the NPBS increases. Additionally, note that as the number of PSUs increases from 50 to 100, the design effect also decreases. However, we would again like to note that these design effects are relevant to estimation of a population average (not a relationship) when conducting a weighted analysis. As discussed below, they may not reflect the design effect associated with estimating a relationship, and certainly do not apply to the case of conducting an unweighted (or model-based analysis).

Table D-6. Number of clusters, weighted average sample size within cluster and weighting contributions to design effects for 21 hybrid designs.

P_1	Number of PSUs	P_2	Number of Clusters	Weighted Average Sample Size Within Cluster	Design Effect Due to Unequal Weighting
0		25	50	2000	5.275
		50	50	2000	2.751
		75	50	2000	2.061
25	50	25	89	1601	2.448
		50	89	1596	2.276
		75	89	1600	2.205
	100	25	76	1497	1.604
		50	76	1498	1.419
		75	76	1502	1.372
50	50	25	64	1652	1.42
		50	64	1650	1.229
		75	64	1660	1.204
	100	25	138	1530	2.562
		50	138	1524	2.428
		75	138	1527	2.418
75	50	25	125	1241	1.564
		50	125	1238	1.442
		75	125	1239	1.439
	100	25	113	1046	1.256
		50	113	1044	1.167
		75	113	1047	1.165

For the NCS, one major issue with the above approach to design effects is that the formulas provide design effects for population means (or percentages), which is not the primary goal in the NCS. Computing design effects for parameters of a regression relationship (i.e., design effects for estimates of a relationship) may depend on a number of additional factors. For example, the design effect due to clustering might be reduced if clustering of the response variable is partially explained by clustering of an explanatory variable. Additionally, the values of explanatory variables impact the effect that weights have on design effects. Finally, design effects for relationships that are constant across clusters may be significantly smaller than design

effects for relationships that vary by cluster. In the following section we provide further discussion of the calculation of design effects when estimating relationships.

D-5 DESIGN EFFECTS FOR ESTIMATION OF RELATIONSHIPS

The purpose of this section is to explore the use of design effects for estimating relationships of interest in the presence of clustering and/or unequal weighting. While quite a lot has been written in the sample survey literature, much of this has been in the relatively simple context where the goal is to assess the precision that a planned study might have to estimate a summary quantity such as a mean (see previous section). In the context of the NCS, however, the situation is substantially more complicated, since estimation of relationships is of primary interest. To address this, we begin from first principles.

Suppose we are interested in exploring the relationship between an exposure and an outcome, based on data from clusters of individuals, each of whom has a binary response. Let X_{ij} be the exposure for individual j in cluster i and let Y_{ij} be this individual's corresponding response. Suppose also that we are interested in fitting the following marginal logistic model:

$$\text{Logit}[\Pr(Y_i=1|X_i=1)] = \text{Logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij}. \qquad \text{(D-12)}$$

In practice, of course, there will also be interest in including additional covariates and risk factors. For the purpose of power and sample size considerations, however, it is enough to consider just the main effect of interest. As discussed elsewhere in the report (see Chapter 9), generalized estimating equations (GEEs) provide an appropriate basis for analysis that accounts for both non-constant sampling probabilities, as well as for clustering of individuals (see Diggle et al., 2002). Let w_{ij} be the sampling weight for the jth individual in the ith cluster (generally, this will be the inverse of their selection probability). Then, a suitable estimating equation for the unknown parameter $\beta=(\beta_0,\beta_1)^T$ is

$$U(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\begin{pmatrix}1\\ X_{ij}\end{pmatrix}(Y_{ij}-\mu_{ij}) = 0\,, \qquad \text{(D-13)}$$

where n is the number of clusters, m_i is the number of subjects in cluster i, and μ_{ij} is the mean response for individual j in the ith cluster. The introduction of the weights, w_{ij}, into the estimating equation (D-13) complicates the estimation of the variance of the parameter estimates, $\hat{\beta}$. However, standard estimating equations theory can be used to establish that

$$Var(\hat{\beta}) = B^{-1}A(B^T)^{-1}, \qquad \text{(D-14)}$$

where B is the matrix of partial derivatives of $U(\beta)$ and A is the variance of $U(\beta)$. This is the calculation automatically performed in software such as SUDAAN or SAS PROC GENMOD (if the empirical variance option is invoked). It is relatively straightforward to show that

$$B = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij} \mu_{ij} (1-\mu_{ij}) \begin{pmatrix} 1 & x_{ij} \\ x_{ij} & x_{ij}^2 \end{pmatrix} \tag{D-15}$$

and

$$A = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m_i} w_{ij}^2 \mu_{ij} (1-\mu_{ij}) \begin{pmatrix} 1 & x_{ij} \\ x_{ij} & x_{ij}^2 \end{pmatrix} + \sum_{j \neq j'} w_{ij} w_{ij'} \sqrt{\mu_{ij}(1-\mu_{ij})\mu_{ij'}(1-\mu_{ij'})} \begin{pmatrix} 1 & x_{ij'} \\ x_{ij} & x_{ij} x_{ij'} \end{pmatrix} \right\}, \tag{D-16}$$

where j and j' represent two arbitrarily chosen individuals from the ith cluster. In certain cases, the expression for the $Var(\hat{\beta})$ simplifies. For example, suppose that the covariate of interest, X, is cluster specific so that x_{ij} is the same for all members of the same cluster. Then, A and B simplify and in large samples will approximate the following:

$$B = \sum m_i E(w_i) \mu_i (1-\mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}, \tag{D-17}$$

where $E(w_i)$ refers to the average of the weights for the i^{t}h cluster, and

$$A = \sum_{i=1}^{n} m_i \mu_i (1-\mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \left\{ E(w_{ij}^2) + \rho (m_i - 1) E(w_{ij} w_{ij'}) \right\}, \tag{D-18}$$

with $E(.)$ again referring to an average over the cluster and ρ referring to the within-cluster correlation with respect to the outcome, Y. A few more special case considerations are helpful. First, consider the case where there is no within-cluster correlation (ρ=0) and also assume that the weights are independent of cluster membership and exposure, hence can be pulled out of the sums. It follows in this special case that

$$Var(\hat{\beta}) = \frac{E(w^2)}{\left(E(w)\right)^2} \left\{ \sum_{i=1}^{n} m_i \mu_i (1-\mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \right\}^{-1}. \tag{D-19}$$

Note that this expression corresponds to the standard variance estimate based on a logistic regression, multiplied by a factor that involves the weights. The factor can be re-expressed as:

$$\frac{E(w^2)}{\left(E(w)\right)^2} = \frac{Var(w) + \left(E(w)\right)^2}{\left(E(w)\right)^2} = 1 + CV^2, \tag{D-20}$$

or 1 plus the squared coefficient of variation of the weights. When the weights are constant, this factor equals 1 and the standard logistic regression variance formula applies. When the weights vary, then this factor will always exceed 1; hence the variance of parameters estimated using weighted estimating equations will always exceed those based on a simple logistic regression. This is a well known result among sample survey statisticians. The extra term is often referred to as a *design effect.* As described in the previous section, these design effects provide a very

useful tool when it comes to study planning and design, since one can think in terms of the impact of various different weighting schemes on the estimated variances of parameters of interest, and adjust accordingly.

Now consider the slightly more complex setting where the intra-cluster correlation, ρ, is non-zero. Using a similar logic, it is relatively straightforward to show the *design effect* (or the factor that multiplies the usual logistic regression variance) is:

$$1+\rho(m-1)+CV^2+\rho(m-1)\operatorname{cov}(w_{ij}w_{ij'}), \tag{D-21}$$

where m is the average cluster size and the covariance term refers to the covariance between weights within the same cluster. In general, we would expect this covariance term to be 0. In the special case where the weights are all equal (variance and covariance of the weights equal 0), the design effect reduces to (1+ρ(m-1)), which is the usual inflation factor for a variance based on cluster data (see Diggle et al., 2002).

When the covariate of interest, X, is allowed to vary within-cluster, all these calculations become considerably more complicated. To facilitate our discussion here, consider the case where exposure is binary and let p_1 denote the probability that an individual is exposed, and p_0=(1-p_1) the probability that an individual is not exposed. Also, for simplicity, we define μ_1 to denote the response probability for exposed individuals and μ_0 the response probability for an unexposed individual. Then, it is relatively straightforward to show that the derivative of the estimating equation (see Equation (D-15)) will, in large samples, be approximately

$$B = nm\left[p_1\mu_1(1-\mu_1)\begin{pmatrix}11\\11\end{pmatrix}+p_0\mu_0(1-\mu_0)\begin{pmatrix}1\,0\\0\,0\end{pmatrix}\right]E(w)=nm\Delta_1E(w) \tag{D-22}$$

where, as before, m is the average cluster size and Δ_1 refers to the term in square brackets. Similarly, the variance of the estimating Equation (D-16) will be approximately:

$$A = nm\left[\Delta_1E(w^2)+\rho(m-1)\Delta_2E(w_{ij},w_{ij'})\right], \tag{D-23}$$

where Δ_2 is more complicated and equal to the following:

$$\begin{aligned}\Delta_2 = {} & p_{11}\mu_1(1-\mu_1)\begin{pmatrix}11\\11\end{pmatrix}+p_{00}\mu_0(1-\mu_0)\begin{pmatrix}1\,0\\0\,0\end{pmatrix}p_{11}\mu_1(1-\mu_1)\begin{pmatrix}11\\11\end{pmatrix}+\\ & p_{10}\sqrt{\mu_1(1-\mu_1)\mu_0(1-\mu_0)}\begin{pmatrix}10\\10\end{pmatrix}+p_{01}\sqrt{\mu_1(1-\mu_1)\mu_0(1-\mu_0)}\begin{pmatrix}1\,1\\0\,0\end{pmatrix}\end{aligned} \tag{D-24}$$

where p_{11} is the probability that two members of a cluster are both exposed, p_{00} is the probability that they are both unexposed, and p_{01} and p_{10} refer to the probability that one is exposed and the other is not exposed. Note that we have made a simplifying assumption here that the intra-class correlation (ρ) is constant for all subjects, and not dependent on the value of covariates. In this

complicated setting, it is not so straightforward to specify a design effect. However, consideration of some special cases for Δ_2 is worthwhile and allows us to explore the impact of various correlation patterns on estimated variances. First, consider the special case where there is perfect within-cluster correlation with respect to exposure values, meaning that p_{01} and p_{10} are both zero, $p_{11}=p_1$ and $p_{00}=p_0$. In other words this is once again the cluster-specific covariate case, where all subjects in a cluster are either exposed, or all subjects in a cluster are unexposed. In this case, Δ_2 is identical to Δ_1, and the design effect is once again given by Equation D-21. Alternatively, for the case where there is no within-cluster correlation with respect to X, we have $p_{11}=(p_1)^2$,$p_{00}=(p_0)^2$ and $p_{10}=p_{00}=p_0p_1$. In this case, there is no simple way to compute a design effect; however, it is easy to use a computer package such as R or Splus to compute the variance of the estimated parameter under various assumptions on the degree of clustering and weighting. In particular, the variance of the estimated parameter under the clustered and weighted design is given by:

$$V_{wc} = \left[\Delta_1^{-1}E(w^2) + \rho(m-1)E(w_{ij}, w_{ij'})\Delta_1^{-1}\Delta_2\Delta_1^{-1}\right]/[nmE(w)]^2 . \qquad \text{(D-25)}$$

In contrast, the variance under simple random sampling is:

$$V_s = \left[\Delta_1^{-1}\right]/[nmE(w)]^2 . \qquad \text{(D-26)}$$

Note that in general, there is no simple multiplicative relationship here, as we saw in the simpler setting of cluster-specific covariates. Indeed, the relationship between variance estimates under simple and complex sampling differs according to which component of the parameter vector is being examined. To examine the ratio of the variances for the coefficient β_1, we simply pull off the (2,2) elements of these two variance expressions and take their ratio. Figure D-4 displays how this ratio varies as a function of μ_0, the intraclass correlation in Y, and whether or not the exposure covariate is cluster-specific or varies within-cluster. Note that this plot assumes equal weights so that we can focus on just the effect of clustering. Additionally, note that the ratio displayed is the ratio of the variance under simple random sampling to the variance under the clustered design, which can be thought of as the inverse of the design effect.

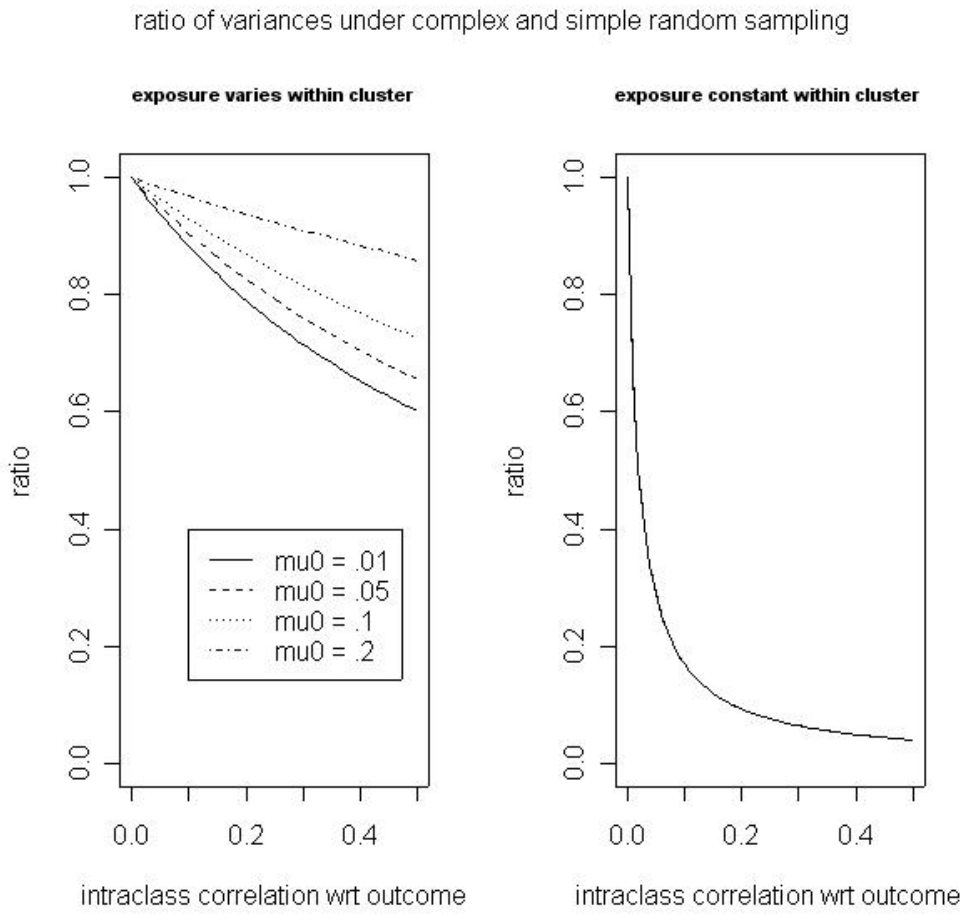

Figure D-4. Ratio of the variance of the parameter (relationship between *Y* and *X*) estimate for a simple random sample to that for the clustered design (assuming equal weights).

The left-hand panel of the figure corresponds to the case of a non-zero within-cluster correlation with respect to exposure. This means that each cluster is likely to have a mix of exposed and unexposed individuals. The right-hand panel corresponds to the case where there is perfect within-cluster correlation with respect to exposure – that is, either all individuals in a cluster are exposed, or all the individuals in a cluster are unexposed. Note that the inverses of the "design effects" are much closer to 1 in the left-hand panel, suggesting that the effect of clustering is not nearly as severe when we have a within-cluster varying covariate. In other words, the impact of clustering on the estimated variances of parameter estimates is moderate compared to the more familiar case where covariates are constant within-cluster. The figure suggests that use of standard "design effects" arguments can lead to misleading results when designing a cohort study such as the NCS. In fact, we suspect that there may be cases where the inverse of the design effect is greater than 1 (i.e., the clustering is actually allowing more accurate estimates of the relationship). Further work is needed to verify this suspicion, and to

better lay down the framework and assumptions that are inherent in calculation of design effects when estimating relationships.

For these reasons, the power results in Chapter 9 of this report estimate power via simulation under a number of assumptions regarding the specific regression relationship between response and explanatory variables. In other words, since design effects for relationships are not easily calculated, the power calculations conducted in this report are done via simulation, rather than through the use of design effects.

D-6 POWER FOR A SIMPLE RANDOM SAMPLE

For simple inferences that treat the cohort as a simple random sample (SRS), analytical formulas are available for computing the power to detect a specified effect. Consider the simple 2-by-2 table of disease presence by exposure presence displayed in Table D-7. Based on this table, an estimate of the odds ratio is given by

$$\hat{OR} = \frac{n_{11}n_{00}}{n_{01}n_{10}}, \tag{D-27}$$

and an approximate variance of the log of this estimate (Agresti, 1990) is

$$V\left(\ln(\hat{OR})\right) = \frac{1}{n_{11}} + \frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} \tag{D-28}$$

Table D-7. Example 2-by-2 table of disease presence by exposure presence.

Disease Presence	Exposure Presence		Total
	Present (1)	Absent (0)	
Present (1)	n_{11}	n_{10}	$n_{1.}$
Absent (0)	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	N

Under the assumption that the estimate of the log-odds ratio is approximately normally distributed (it is asymptotically normally distributed), an analytical formula relating the total sample size (n), the Type 1 error rate (α), the power ($1-\theta$) for a two-sided test, the prevalence of the risk factor ($n_{.1}/n$), the prevalence of the outcome ($n_{1.}/n$), and the odds ratio describing the relationship between the disease and the risk factor (OR) can be derived. The formula is:

$$\text{Power} = 1 - \Phi\left(z_{1-\alpha/2} - \frac{\ln(OR)}{\sqrt{V\left(\ln(OR)\right)}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\ln(OR)}{\sqrt{V\left(\ln(OR)\right)}}\right), \tag{D-29}$$

where Φ is the cumulative distribution factor (CDF) of a standard normal distribution, $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of a normal distribution, and the formula for $V(\ln(OR))$ can be found above.

(Note that given the odds ratio, the sample size, the prevalence of disease, and the prevalence of exposure, the values for n_{ij} can be derived.)

As mentioned in Chapter 9, there are certainly other methods for evaluating the significance of the relationship between a single binary exposure factor X and a single binary health outcome Y, and these other methods would lead to alternative formulas for the power to detect a relationship of interest. For example, Whittemore (1981) also provides a formula that relates sample size, Type 1 error rate, power for a one-sided test, prevalence of the risk factor, prevalence of the outcome, and the strength of the relationship between the risk factor and the outcome for a simple univariate logistic regression model with a single dichotomous covariate. The formula is based on the sampling distribution of the Wald statistic for the estimate of the logistic regression coefficient (i.e., β_1), and is as follows

$$n = (1+2p_0)\times\frac{\left(z_{1-\alpha}\sqrt{\frac{1}{1-\pi}+\frac{1}{\pi}}+z_{1-\theta}\sqrt{\frac{1}{1-\pi}+\frac{1}{\pi e^{\beta_1}}}\right)^2}{p_0\beta_1^2} \qquad \text{(D-30)}$$

where $\pi=P(X=0)$, p_0=P(Y=1|X=0), β_1 is the true log-odds ratio, $z_{1-\alpha}$ and $z_{1-\theta}$ are the upper α and θ percentiles of a normal distribution, and $1-\theta$ is the desired power (see Hosmer and Lemeshow, 2000).

Finally, a formula based on simply comparing the proportion of individuals with the disease in the unexposed group, P_0=P(Y=1|X=0), to that of the exposed group, P_1=P(Y=1|X=1), could be derived. For example, as described by Rosner (1999) and many others, (and similar to Equation D-29 above) the power associated with a study of N individuals (assuming a 2-sided test at significance level 0.05) can be calculated as

$$\text{Power} = 1-\Phi\left(z_{1-\alpha/2}-K\right)+\Phi\left(-z_{1-\alpha/2}-K\right), \qquad \text{(D-31)}$$

where $K=(P_1-P_0)\sqrt{Np_x(1-p_x)/[p(1-p)]}$, p_x is the probability of being in the exposed group, and p represents the weighted average of P_1 and P_0, $p=p_xP_1+(1-p_x)P_0$ (i.e., the marginal probability of disease). Since it is envisioned that these alternative formulas would generally provide qualitatively similar results, we will present results corresponding to the power calculation provided in Equation D-29.

Thus, analytical formulas for the power of detecting a specified relationship are available when the data are selected as a simple random sample. For this reason, power can be calculated for a large number of scenarios; however, the power values and resulting conclusions must be interpreted in light of this *simple random sample* assumption (see discussion in Section 9.2). More particularly, the power values do not account for the effect of clustering and unequal weighting that will likely be elements of any feasible NCS design. Therefore, these calculations likely represent optimistic values for the power to detect the relationships of interest; however, their ease of computation allows investigation of a large number of scenarios and provides insight into:

- The effect of sample size (which is influenced by retention rates associated with a selected hypothesis and the costs associated with different designs),
- The effect of differing levels of disease and exposure occurrence rates, and
- The odds ratios that can be detected for the different scenarios.

In the following, we provide a set of general SRS power results that may be applicable to a variety of situations, and discuss several examples of the usefulness of these results. Recall that the factors affecting the simple random sample power of detecting a significant relationship between a categorical outcome *Y*, and a binary risk factor *X*, are: sample size, strength of the relationship, rate of occurrence of *X* and *Y*, and the desired significance level of the hypothesis tests. For these factors, we evaluate the following scenarios:

- The sample size *n* takes values of 5000, 10000, 25000, 35000, 50000, 75000, and 100000 individuals. Comparing the power associated with different sample sizes could allow comparison of different designs in terms of their retention rates, cost, and/or availability of the information relevant to the hypothesis of interest. For example, comparing a design that would result in a retention rate (retention through the time period associated with a selected hypothesis) of 50% versus a design that would result in a retention rate of 75% could be done by comparing the n=50000 and the n=75000 sample size power estimates. Similarly, comparison of a design that costs twice as much per individual as another design could be done by comparing the n=50000 and n=100000 sample size power estimates. (Note that this may be useful when comparing designs assuming a fixed cost. Under a fixed cost, the cheaper design would allow the study to include 100000 individuals, whereas the more expensive design would only allow the study to include 50000 individuals.)
- The strength of the relationship between *X* and *Y* is parameterized using the odds ratio (see Section 9.2), and takes a range of values between 1 and 11 so that a full range of odds ratios (displayed on the log-scale in the figures below) can be considered. (Note that an odds ratio of 11 represents a *very* strong relationship between *X* and *Y*.)
- The rate of occurrence of *X*, or exposure prevalence, takes values of 0.01, 0.05, 0.10, and 0.20. This range of exposure prevalence is meant to span the reasonable range of possible exposures that are of interest in the core hypotheses. For higher exposures that are more prevalent than this range (e.g., exposure prevalence up to 0.50), the resulting power would be even higher than is realized for the lower exposure prevalences considered.
- The rate of occurrence of *Y*, or disease prevalence, takes values of 0.0025, 0.005, 0.01, 0.05, and 0.10. This range of disease prevalence is meant to span the disease prevalences associated with the core hypotheses (e.g., autism and cerebral palsy occur with prevalences of approximately 0.0030 and 0.0020, respectively, and asthma or injury occurs with a prevalence on the order of 0.05 to 0.10).
- The significance level of all tests is assumed to be α=0.05.

For each of the above settings, figures that display the power as a function of the odds ratio are constructed. For example, Figure D-5 displays the power for detecting the relationship

between a rare health outcome (disease prevalence of 0.0025 or 25 out of every 10,000 individuals), such as autism and cerebral palsy, and a binary risk factor. The upper left panel of the figure displays the power as a function of sample size (the different lines) and the odds ratio (the horizontal axis) for a binary exposure risk factor with a prevalence of 0.01 or 1 out of every 100 individuals (the horizontal line is drawn at a power of 0.80). As expected, power increases as a function of sample size and as a function of the strength of the relationship between the outcome and the risk factor. The graph demonstrates that even for a sample size of *n*=100,000, an odds ratio of close to 3 is required in order to detect a significant relationship with 80% power. This picture becomes somewhat more promising as the prevalence of the exposure (or exposure occurrence rate) increases, with odds ratios on the order of 1.5 being detectable with 80% power when the exposure prevalence is 0.20 (20 out of every 100 people experience the exposure). From a design perspective this generally implies that for diseases with very low prevalence, weak relationships (e.g., odds ratios of 1.1) will be difficult to detect even with a sample size of 100,000 individuals. The other panels of the figure display the same type of information but for different values for the prevalence of the risk factor (0.05, 0.10, and 0.20, respectively). Figures D-6 through D-9 display similar figures for the other levels of disease prevalence investigated.

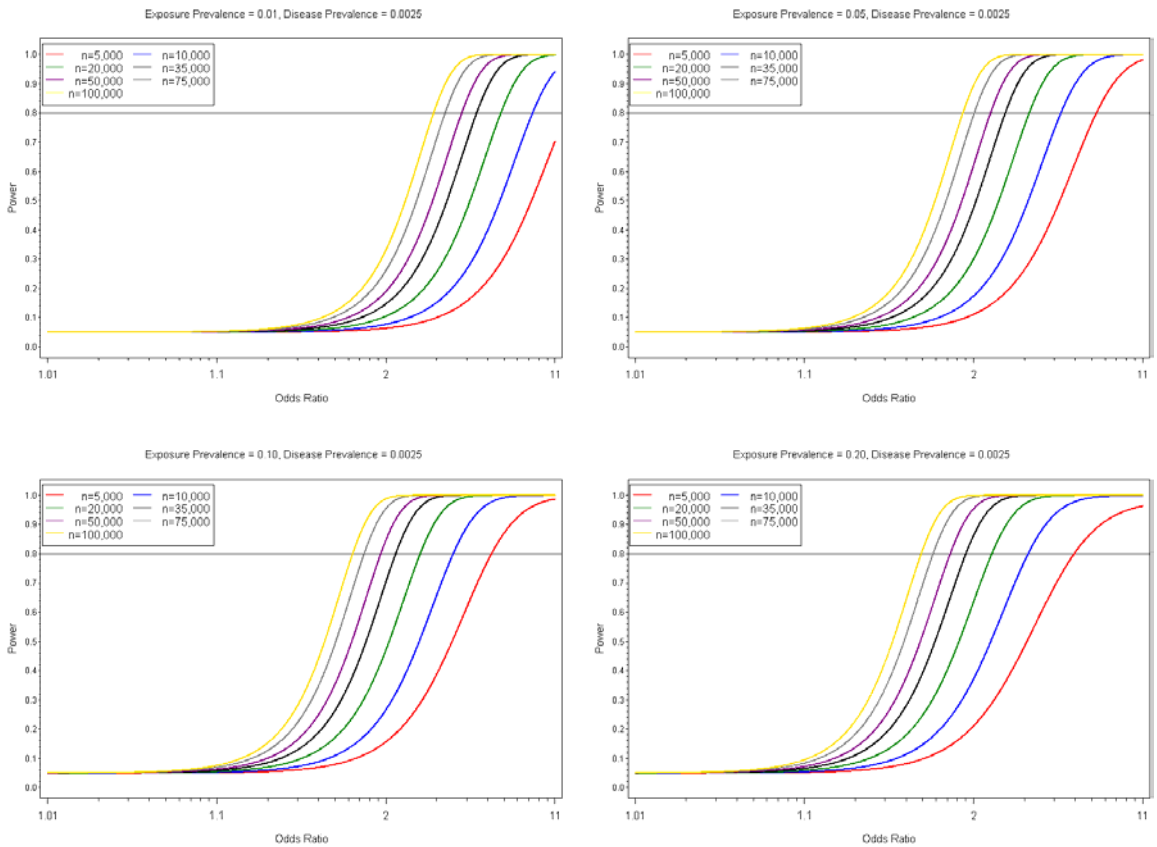

Figure D-5. SRS power for detecting a significant relationship between a health outcome with a prevalence of 0.0025 and a binary risk factor.

One possible use of these figures is to compare different designs through their potential realized sample size. For example, suppose one design involves a retention rate of 50% for the time period associated with the hypothesis of interest and another design involves a retention rate of 75% over the same time period. Then comparing the *n*=50,000 and *n*=75,000 power curves can provide a means of comparing the power associated with the different designs. Similarly, comparing a design that costs twice as much per individual as another design can be accomplished by comparing the *n*=50,000 and *n*=100,000 power curves. While these curves seem relatively similar in the figures it should be noted that there can be significant differences in terms of their power. For example, if we consider the top right panel of Figure D-5, an odds ratio of 2.0 corresponds to a power of over 0.85 for a sample size of 100,000 individuals; however, for a sample size of 50,000 the resulting power is on the order of 0.60. Thus, despite the relative similarity of these curves, there can be significant differences in their corresponding power.

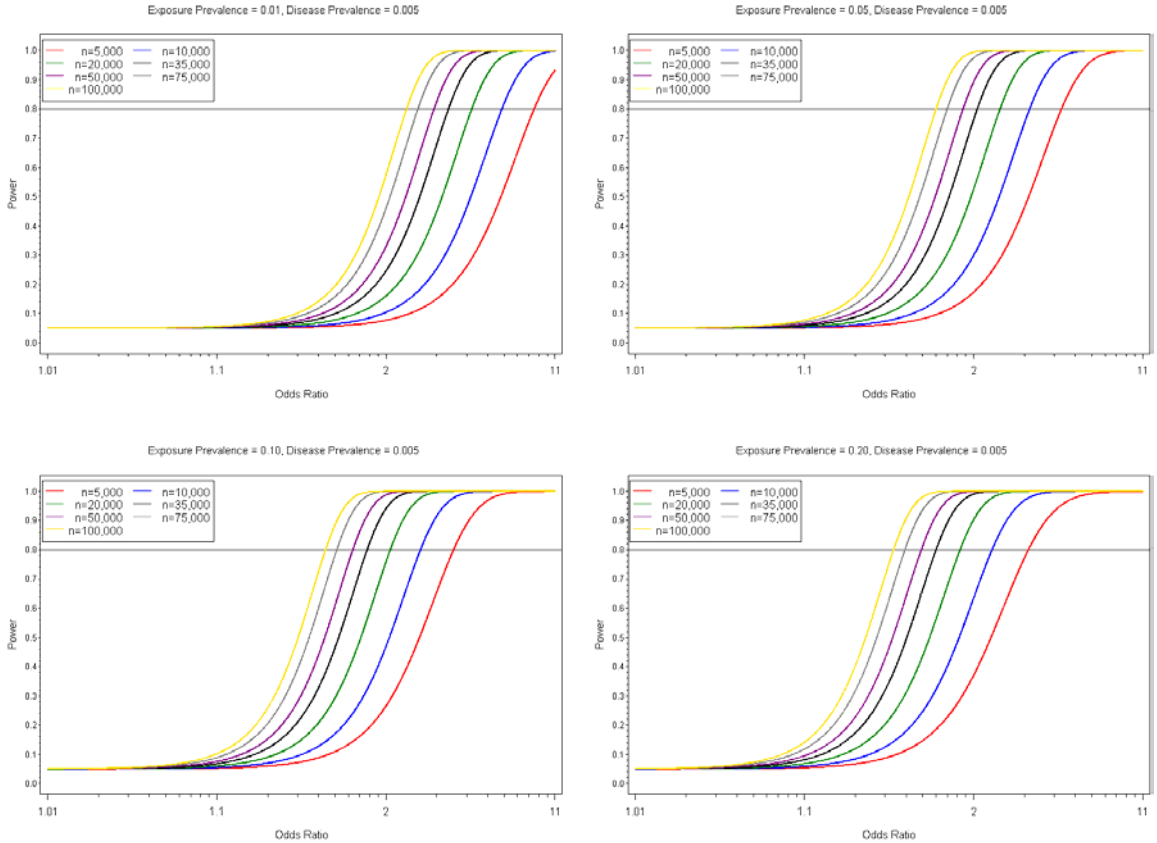

Figure D-6. SRS power for detecting a significant relationship between a health outcome with a prevalence of 0.005 and a binary risk factor.

As a more concrete example, consider the core hypothesis concerned with the relationship between major congenital malformations of the heart (birth defect health outcome) and impaired glucose metabolism during pregnancy for women without diabetes before pregnancy. Congenital heart defects occur in approximately 0.60% of individuals, or 6 out of 1000 individuals, and impaired glucose metabolism occurs in approximately 5% of pregnant

women (see Chapter 6). For these disease and exposure characteristics, the top right panel of Figure D-6 displays an estimate of the power to detect relationships of specified sizes. For example, a sample size of 100,000 individuals would provide approximately 80% power to detect an odds ratio of approximately 1.6, and a sample size of 50,000 individuals would provide greater than 80% power to detect an odds ratio of 2.0. Thus, for this hypothesis, assuming that the necessary information (i.e., presence/absence of congenital heart defects and presence/absence of impaired glucose metabolism in the mother) is available for most, if not all, of the cohort, there will likely be sufficient power to detect odds ratios of 1.6 and greater. Additionally, note that this result is likely robust to the retention rate associated with a design since the information for evaluating this hypothesis would presumably be collected very early in the study (i.e., prior to any significant portion of the cohort dropping out).

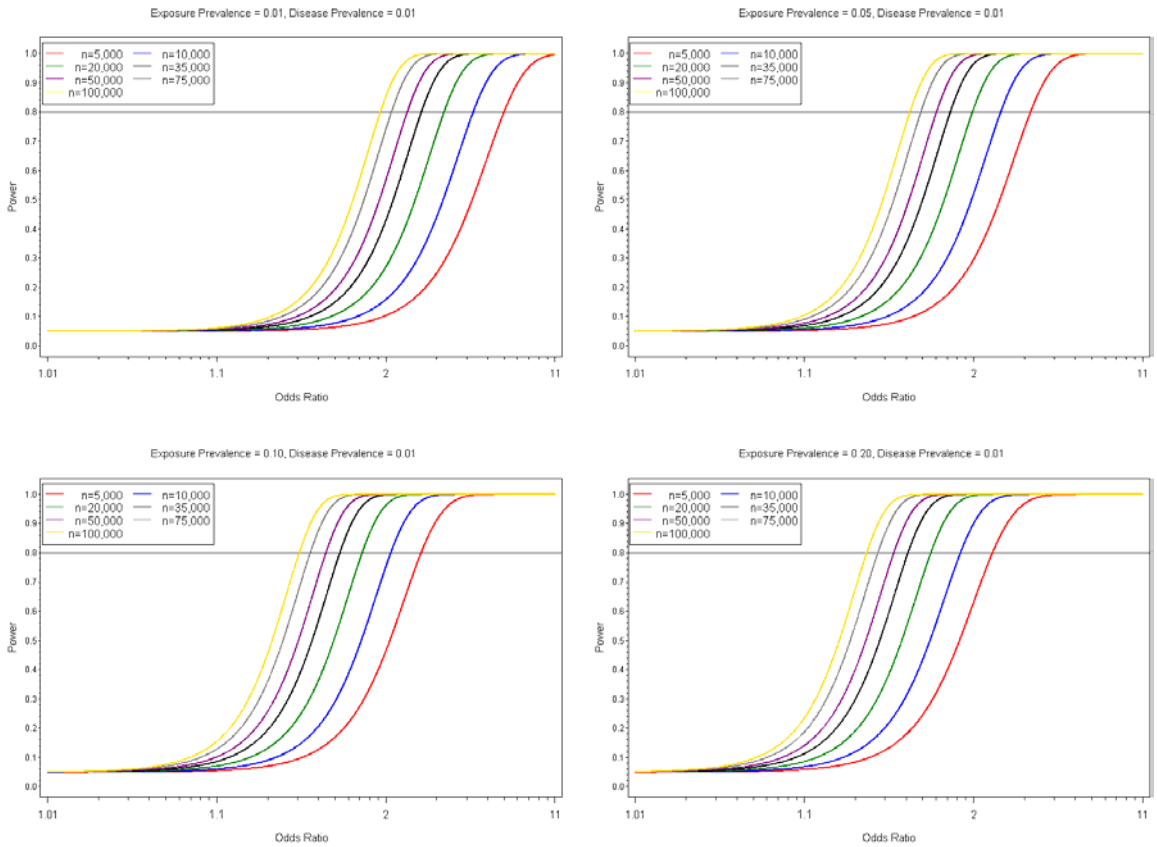

Figure D-7. SRS power for detecting a significant relationship between a health outcome with a prevalence of 0.01 and a binary risk factor.

As another example, consider the hypothesis concerned with whether infection and mediators of inflammation during pregnancy and the perinatal period are associated with increased risk of schizophrenia. Approximately 1 percent of the population develops schizophrenia during their lifetime; however, the disease rarely develops prior to adolescence. From a retention standpoint this may be one of the most difficult diseases to investigate, since evaluation of the presence/absence of disease will necessarily occur at some point far along in the study (e.g., after 10 years or even up to the full length of the study). Presumably, the

retention rates will decrease as a function of time, perhaps significantly, meaning that there may be a significant reduction in the available sample size to evaluate this hypothesis. The bottom right panel of Figure D-7 displays the power to detect specified odds ratios for a disease that occurs in approximately 1 percent of the population and for a risk factor that occurs in 20 percent of the population (i.e., we nominally assume that mediators of inflammation during pregnancy and the perinatal period occur in 20 percent of the population). If a design were to result in a retention rate of 20 percent over the 20-year period, then at a maximum, 20,000 individuals would be available for evaluating this hypothesis, and an odds ratio around 1.6 would be detectable with 80 percent power. On the other hand, if a design were to result in a retention rate of 50 percent over the 20-year period, then at a maximum, 50,000 individuals would be available for evaluating this hypothesis, and an odds ratio of approximately 1.3 would be detectable with 80 percent power.

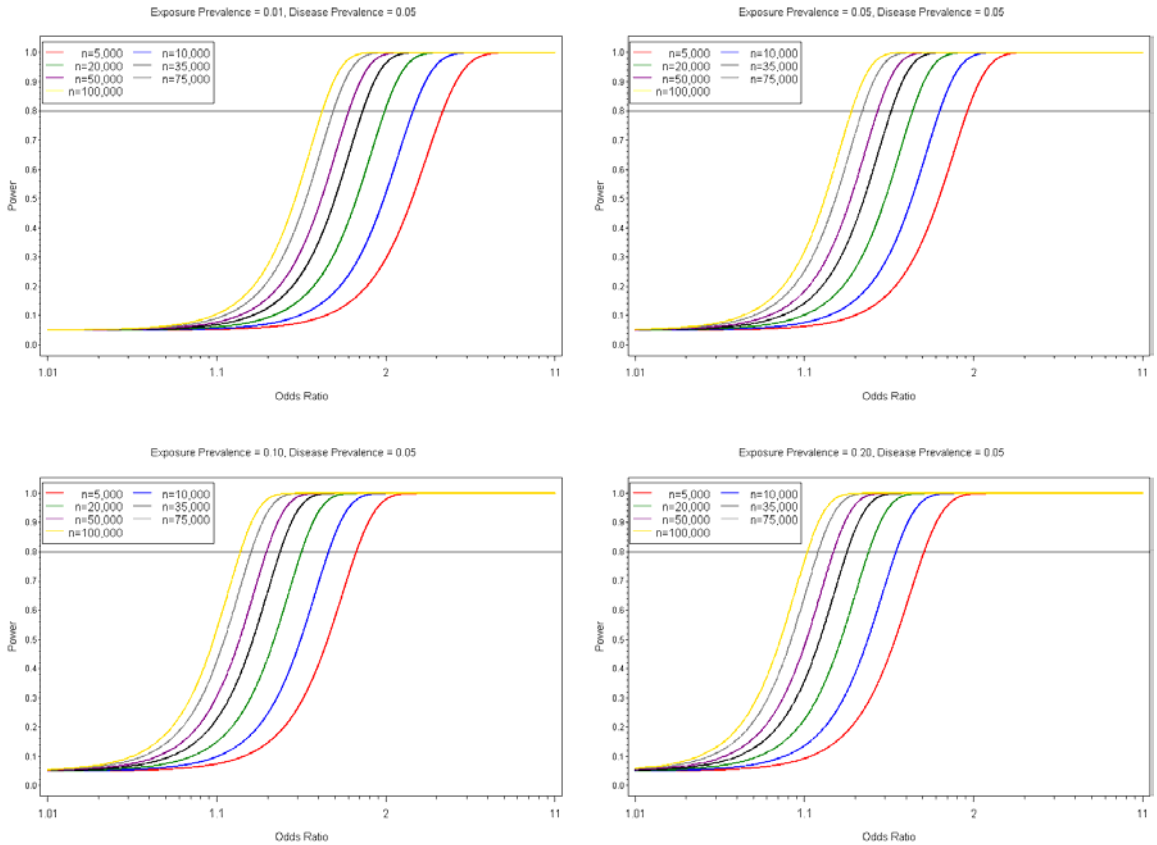

Figure D-8. SRS power for detecting a significant relationship between a health outcome with a prevalence of 0.05 and a binary risk factor.

These figures can also be used to evaluate and interpret the sample sizes necessary to detect relationships of interest. For example, suppose we desire to detect an odds ratio of 2.0 with 80% power for a disease that occurs in approximately 5 out of every 1000 individuals and an exposure that occurs in 20 out of every 100 individuals (i.e., lower right panel of Figure D-6). Then, a sample size somewhere between 10,000 and 20,000 individuals would be necessary. However, if we desire to detect an odds ratio of 1.1 for this scenario, even a sample size of 100,000

individuals would not provide sufficient information to achieve 80% power. Thus, as expected, for rare diseases it will be difficult to detect weak relationships between the disease and the risk factor of interest implying that it will be necessary to collect the data needed to evaluate the hypothesis of interest for as many individuals in the cohort as possible.

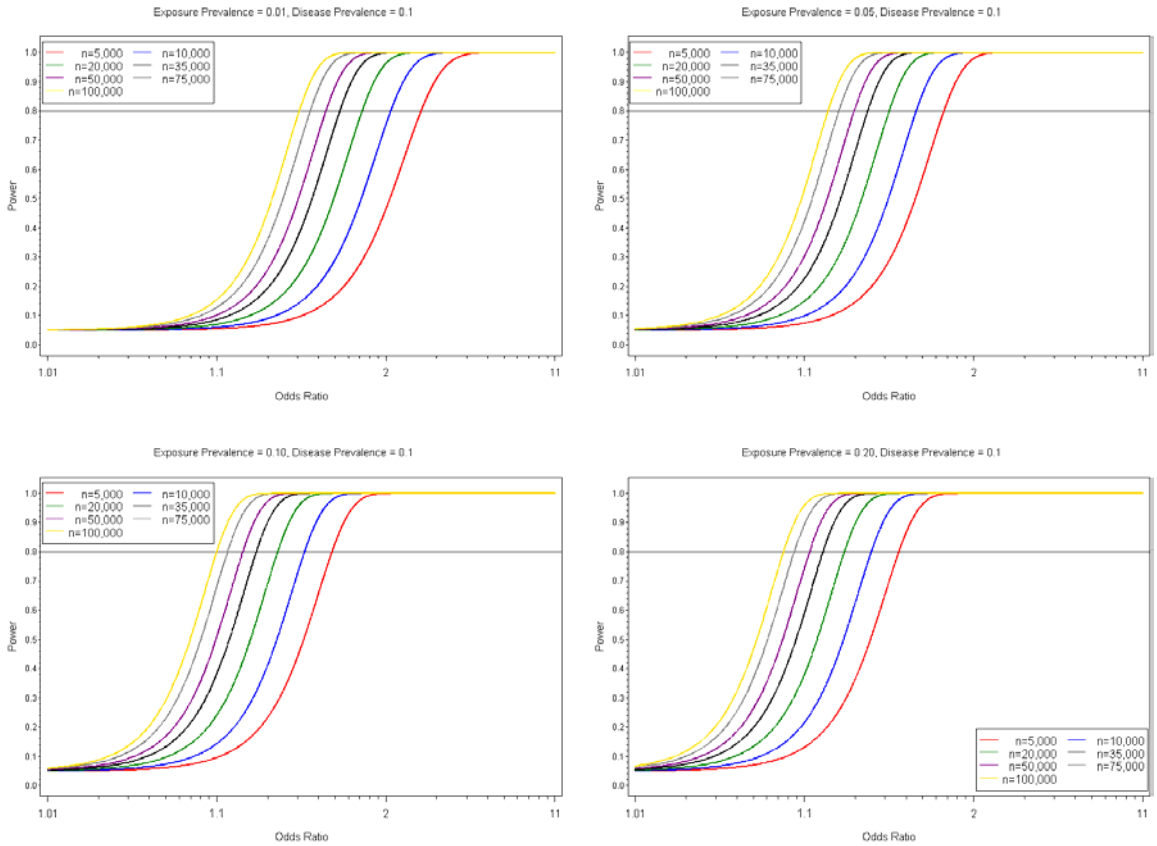

Figure D-9. SRS power for detecting a significant relationship between a health outcome with a prevalence of 0.1 and a binary risk factor.

On the other hand, for adverse health outcomes that are more common, such as asthma or injury, it may be necessary to collect the necessary information on only a subset of the cohort. For example, for a disease that occurs in approximately 5 out of every 100 individuals and for a risk factor that occurs in 20 out of every 100 individuals (bottom right panel of Figure D-8) it may only be necessary to collect the desired information for a subset of the NCS cohort (e.g., even with 20,000 individuals there is sufficient power to detect an odds ratio of 1.3). From a cost perspective, these types of implications could lead to significant cost savings when/if evaluation of the risk factor or disease of interest is expensive. In other words, for hypotheses involving collection of expensive information the power analysis suggests that sample sizes smaller than 100,000 will result in significant power to detect even weak relationships, then it may not be necessary to collect this expensive information across the entire cohort. Rather, it may be more cost efficient to collect the expensive information for only a subset of the cohort, perhaps chosen at random or perhaps chosen based on other demographic or subject-specific

characteristics (e.g., family health history). These types of design choices (i.e., collecting some information for only a portion of the cohort) represent an important consideration in the design of the NCS from both a cost perspective as well as a subject-burden perspective.

D-7 FURTHER DETAILS OF SIMULATION APPROACH TO CALCULATING POWER

D-7.1 SOME NOTES ON SIMULATING CORRELATED BINARY DATA

The National Children's Study (NCS) will involve correlated data, not only because children will be measured and assessed repeatedly over time, but also because recruitment is likely to be based on a clustered design. Here we focus on the challenging issue of assessing power for the NCS when the data are correlated. For simplicity, we start out by ignoring the longitudinal aspect of the design and consider power calculations for the setting where subjects are clustered within groups and assessed for the presence/absence of a disease or condition of interest by a specified age. By varying the assumed parameters, we will be able to consider power for different hypothetical situations, for example involving relatively rare conditions such as whether or not a child is diagnosed with autism, or more common conditions, such as whether or not the child develops asthma.

The analysis of correlated binary data is complex and has been the focus of numerous statistical papers, especially over the past several decades. Many different approaches have been suggested, mostly based on natural generalizations of the kinds of normal random effects models that are so widely used for correlated continuous data. The main complication is that different approaches to incorporating random effects lead to fundamentally different models for binary data. Although many different approaches have been suggested, the most popular methods tend to fall into one of the three following classes:

1. Logistic-normal models (which entail incorporating a cluster-specific normal random effect into a standard logistic model)
2. Beta-binomial models (which assume that cluster-specific response rates are generated from a beta distribution)
3. Generalized estimating equation (GEE) models (which avoid the need to assume any particular random effects distribution by specifying the mean and covariance structure of the observed data).

Heagerty and Zeger (2000) provide an excellent overview of recent work related to the analysis of correlated data. They distinguish between conditional and marginal model specifications. The conditional approach generally involves formulating the mean outcome as a function of covariates as well as an unobserved subject or cluster-specific random effect, while the marginal approach specifies the mean outcome only as a function of observed covariates. Heagerty and Zeger (2000) discuss some of the reasons why maximum likelihood methods have traditionally been used to fit the conditional models, while marginal models have often been fit using GEEs. They argue that maximum likelihood methods could also be used to fit marginal models and propose a method for doing so. For the purpose of the power considerations here,

we will assume that analysis of data from the NCS will be based on GEE models. As discussed in Chapter 9, the Sudaan software will be used for analysis, since this program provides a particularly convenient way to allow for sampling weights, as well as to adjust for clustering effects. We also propose the use of GEEs based on an independence working correlation. The reasons for this are several-fold. First, our simulations suggest that the independence working assumptions will provide more stable numerical results. Also, there is good reason to believe that this approach will be more robust to nonrandom selection problems and subject dropout.

In order to describe the approaches in more detail, and to consider their implication for these power analyses, it is useful to introduce some notation reflecting outcomes measured on clusters of individuals, each of whom has a binary response. Let X_{ij} be the exposure for individual j in cluster i, and let Y_{ij} be the individual's corresponding response. Random effects logistic models assume that conditional on a cluster-specific random effect (suppose that this is b_i for the ith cluster) that the response probability follows a logistic model:

$$\text{Logit}[\Pr(Y_i=1|X_i=1, b_i)] = \beta_0 + \beta_1 X_{ij} + b_i, \qquad \text{(D-32)}$$

where b_i is a random effect, assumed to be normally distributed with mean 0 and variance σ^2. The parameter β_1 is the conditional log-odds ratio associated with exposure for the ith cluster, given the value of that cluster's random effect. It is sometimes also called a subject-specific log-odds ratio and can be interpreted as reflecting the impact of a one-unit change in the covariate X on the response rates for an individual whose cluster membership remains unchanged. One disadvantage of this random effects logistic model is that it can be difficult to interpret for the general population. In particular, the odds ratio associated with the response rates of two arbitrarily chosen individuals is a fairly complex function of the parameters β_0, β_1 , and the random effects variance σ^2. Another disadvantage of the logistic normal approach is that there is no easy correspondence between the model parameters and the marginal response rates. For example, the average response rate among unexposed individuals is **not** $e^{\beta_0}/(1+ e^{\beta_0})$, as might be expected, but rather it is:

$$P_0 = \int e^{\beta_0+b}/(1+ e^{\beta_0+b})\, f(b)db, \qquad \text{(D-33)}$$

where f(b) is the density of a normal random variable with mean 0 and variance σ^2. Additionally, if we use the same logic to define

$$P_1= \Pr(Y=1|X=1) = E[\Pr(Y=1|X=1, b)], \qquad \text{(D-34)}$$

and then compute the log-odds, $\log[P_1(1-P_0)/(P_0(1-P_1))]$, this is not equal to the conditional log odds, β_1. Furthermore, the nonlinearity of the logit function implies that the marginal response probability, $\Pr(Y=1|X)$, no longer follows a logistic model that is linear in X (see Liang et al., 1992 for further discussion about the relationship between marginal and conditional regression models). Finally, although it is not necessarily a problem, it is interesting to note that the intraclass correlation is not constant (even though there is a constant random effect). The intraclass correlation needs to be calculated by integrating over the random effects distribution. In particular,

$$\mathrm{Cov}(Y_{ij}Y_{ij'}) = \mathrm{E}(Y_{ij}Y_{ij'}) - \mathrm{E}(Y_{ij})\mathrm{E}(Y_{ij'}). \qquad \text{(D-35)}$$

Aside from the problems with interpretation, fitting the logistic normal model is also a challenge, due to the fact that the likelihood does not exist in closed form. A variety of approaches have been suggested, including numerical methods based on quadrature as well as Laplace approximations (see Breslow and Clayton, 1994, for a good discussion of this topic).

To avoid these problems with the logistic normal, a number of authors have explored the beta-binomial model for the analysis of correlated binary data (Williams, 1988, was one of the earlier authors to suggest this method). This model has been mostly explored in the context where the covariate (exposure) is measured at the cluster level (e.g., a litter of mice is either exposed or unexposed) in which case we have $X_{ij}=X_i$ for all j. In that setting, the model is:

$$\Pr(Y_{ij}=1|p_i) = p_i \qquad \text{(D-36)}$$

where p_i is the cluster-specific probability of response, and the p_i's in turn are generated from a beta distribution:

$$p_i \sim \mathrm{beta}(\alpha_1, \alpha_2), \qquad \text{(D-37)}$$

where the beta parameters α_1 and α_2 are chosen so as to produce the desired marginal expectations and correlations. In the case of a dichotomous exposure, for example, these parameters could be chosen to yield an expected value of $e^{\beta_0}/(1+ e^{\beta_0})$ for the unexposed group and $e^{\beta_0+\beta_1}/(1+ e^{\beta_0+\beta_1})$ for the exposed group, along with the required intraclass correlation.

When individuals vary with respect to their exposure levels, the beta-binomial is more complicated and much less has been written on the topic. A natural extension goes as follows. For simplicity, consider the case of a dichotomous exposure. Let p_{i0} be the baseline response rate for unexposed subjects in the ith cluster. Suppose we specify the response rate for an exposed individual, p_{i1}, so as to ensure that the conditional odds ratio associated with exposure within the ith cluster is given by Δ_c. Simple algebra establishes that this is easily done by specifying

$$p_{i1} = \Delta_c\, p_{i0} /[1-p_{i0}\,(1-\Delta_c)]. \qquad \text{(D-38)}$$

Note, however, that since p_{i1} is now a nonlinear function of p_{i0}, this formulation loses the simple correspondence between the marginal response probabilities, P_0 and P_1. To see this, suppose that we choose a beta distribution with parameters that ensure the desired marginal response rate among unexposed subjects, $P_0 = \Pr(Y_{ij}=1|X_{ij}=0)$. Iterated expectation establishes that

$$\begin{aligned} P_1 &= \Pr(Y_{ij}=1|X_{ij}=1) \\ &= \mathrm{E}(\Pr(Y_{ij}=1|X_{ij}=1, p_{i0})] \qquad \text{(D-39)} \\ &= \mathrm{E}[\Delta_c p_{i0} /[1-p_{i0}(1-\Delta_c)]]. \end{aligned}$$

Because this is an expectation of a nonlinear function of p_{i0}, Equation D-39 will not in general equal $\Delta_c P_0 /[1-P_0(1-\Delta_c)]]=P_1$. In other words, imposing a constant within-cluster odds ratio of Δ_c will necessarily imply that the marginal odds ratio between unexposed and exposed subjects will be something different than Δ_c, though a standard Taylor series argument establishes that the result will be close in certain cases. Just as in the logistic normal case, the nonlinearity of the logit function implies that if we assume the within-cluster response rates follow a logistic model with respect to a covariate X, then the marginal response rates cannot also have a logistic form, except in the degenerate case where the intraclass correlation is 0 (in which case the p_{i0} are all constant and equal to P_0). To avoid this problem, one might consider formulations other than a logistic model for relating within-cluster response rates as a function of an exposure X. For example, one might specify the within-cluster effect on a linear scale, $p_{i1} = p_{i0} + \Delta$, so that the marginal response rates would also satisfy the same relationship: $P_1 = P_0 + \Delta$. However, the disadvantage of this formulation is that p_{i1} is no longer constrained to lie between 0 and 1. A practical solution is to use the implied value of P_1 (Equation D-39) to determine the within-cluster odds ratio that will yield a desired marginal odds ratio.

D-7.2 IMPLICATIONS FOR POWER ANALYSIS

In the simple (nonclustered) setting, power analysis for binomial data is relatively straightforward (see SRS results described above). The complication is that once we start considering clustering, then individual response rates vary by cluster, as well as by X, with respect to their response probabilities. The simple argument used above for the binomial setting no longer applies. Indeed, it is impossible to simultaneously force all the desired conditions (specified intra-class correlation, specified odds ratio, specified response rate among unexposed subjects) to hold. Depending on how the true underlying model is specified (logistic normal, beta-binomial etc), power has to be considered for settings where either response rates vary by cluster or odds ratios vary by cluster.

One approach is to base power calculations on asymptotic considerations, based on either a logistic normal or a beta-binomial model. Another approach is to use simulation. An advantage of the latter is that data can be simulated under more general settings (e.g., allowing for dropout, etc.). In particular, data can be generated from a parametric model (such as the beta-binomial), but then analyzed in the simulations using a GEE-based method. This is the general approach we have taken in the power analyses presented in this report.

D-7.3 SIMULATING CORRELATED BINARY DATA

As indicated above, the very same factors that complicate the analysis of correlated binary data make their simulation difficult as well. A number of authors have suggested simulation methods (see Qoqish, 2003; Lee, 1993). In practice, a popular choice is to use the logistic normal model to generate correlated binary data. The major advantage of this approach is that it is simple – one simply generates a normal random variable for each cluster, adds it to the logistic regression equation as in (2), then generates random bernoulli random variables with the corresponding probabilities. The disadvantage, as indicated above, is that the model parameters (β_0, β_1 and σ^2) are not easily matched to marginal response rates for exposed and unexposed individuals. In the case of a dichotomous exposure, a commonly used approach is to

start out by specifying P_0 and P_1, the average response rates among exposed and unexposed individuals, then, given the value of σ^2 , determine the corresponding values of β_0 and β_1 by inverting Equations D-33 and D-34. Unfortunately, this inversion involves numerically solving an integral that is not in closed form.

On the other hand, correlated binary data are relatively easy to simulate from the beta-binomial distribution in the context where the covariate (exposure) is measured at the cluster level. The first step is to generate cluster-specific response rates (p_i) from a beta distribution whose parameters have been chosen so as to ensure the desired marginal means and covariances. To be precise, consider a beta distribution with parameters α_0 and α_1, so that the corresponding mean and variance are:

$$E(p_i) = \alpha_0/(\alpha_0+\alpha_1) \quad \text{and} \quad Var(p_i) = \alpha_0\alpha_1/[(\alpha_0+\alpha_1)^2(1+\alpha_0+\alpha_1)]. \tag{D-40}$$

Then, it is straightforward to show the following with regard to the means, variances and covariances of random variables generated from the beta-binomial:

$$E(Y_{ij})=E[E(Y_{ij}|p_{ij})]=E(p_{ij})= \alpha_0/(\alpha_0+\alpha_1) \tag{D-41}$$

$$Var(Y_{ij})=E[Var(Y_{ij}|p_{ij})]+ Var[E(Y_{ij}|p_{ij})]= \alpha_0\alpha_1/(\alpha_0+\alpha_1)^2 \tag{D-42}$$

$$Corr(Y_{ij}, Y_{ij'})=1/(\alpha_0+\alpha_1+1), \tag{D-43}$$

where Y_{ij} and $Y_{ij'}$ represent responses from two different individuals in the ith cluster. Suppose we wish to have the following marginal response rates:

$$Pr(Y=1|X=0) = P_0 \quad \text{and} \quad Pr(Y=1|X=1)=P_1, \tag{D-44}$$

and also want to have an intraclass correlation of ρ. Then, it is straightforward to invert these equations in the following manner. For unexposed clusters (X=0), we have

$$\alpha_0 = P_0*(1/\rho-1) \qquad \alpha_1= (1-P_0)*(1/\rho-1) \tag{D-45}$$

while for exposed clusters, we have:

$$\alpha_0 = P_1*(1/\rho-1) \qquad \alpha_1= (1-P_1)*(1/\rho-1). \tag{D-46}$$

Simulating clustered binary data where the covariate varies within-cluster is more complicated. One natural extension proceeds as follows. Let p_{i0} be the baseline response rate for unexposed subjects in the ith cluster and suppose that we would like to have a correlation of ρ between two unexposed individuals from the same cluster. We need to generate p_{i0} from a beta distribution with parameters α_0 and α_1, chosen (as above) so that $E(p_i)=P_0$, the desired response rate for unexposed individuals, and so that $Corr(Y_{ij}, Y_{ij'})=\rho$, where Y_{ij} and $Y_{ij'}$ represent responses from two different unexposed individuals from the ith cluster. Now we need to consider the response rate for an exposed individual in cluster i. As described above, we propose to specify p_{i1} so as to

ensure that the odds ratio associated with exposure within the ith cluster is given by Δ_c (the 'c' denotes conditional):

$$p_{i1} = \Delta_c\, p_{i0} /[1-p_{i0}\,(1-\Delta_c)]. \qquad \text{(D-47)}$$

Note that as discussed above, since p_{i1} is now a nonlinear function of p_{i0}, this formulation loses the simple correspondence between the marginal response probabilities, P_0 and P_1. In fact, using iterated expectation, it follows that

$$P_1 = E[\Delta_c p_{i0} /[1-p_{i0}(1-\Delta_c)]]. \qquad \text{(D-48)}$$

A standard Taylor series argument establishes that this last expression will be approximately, but not exactly, equal to $\Delta_m P_0 /[1-P_0(1-\Delta_m)]]=P_1$, where Δ_m denotes the desired marginal odds ratio. In other words, imposing a constant within-cluster odds ratio of Δ_c will necessarily imply that the marginal odds ratio between control and exposed subjects, Δ_m, will be something slightly different from Δ_c. Using a numerical integration procedure, it is straightforward to calculate the marginal odds ratio associated with a specified conditional odds ratio. We have written a function in R to do this calculation.

D-7.4 DISCUSSION

We have discussed some of the complexities of sample size and power considerations for correlated binomial data, especially in the setting where covariates are expected to vary within-cluster. We have suggested the use of simulation-based methods for power estimation. While there have been a number of proposals in the recent statistical literature for methods to generate correlated binary data, there are advantages and disadvantages to all approaches. We have used the relatively simple approach of a beta-binomial, generating baseline (i.e., in the absence of exposure) rates from a beta distribution for each cluster, then generating a conditional odds ratio chosen so as to ensure the desired marginal odds ratio associated with exposure. Our experience suggests that considerable care is needed in choosing the appropriate beta parameters to generate the baseline response rates. In particular, it is important to keep in mind that not all correlation values between -1 and 1 are possible for binomial data. We have found that it is better to specify a feasible range for the response rates from cluster to cluster, rather than specifying the within-cluster correlation directly.

Thus, given a design scenario and a hypothesis of interest, the steps to calculating power are as follows:

1. Obtain a realization of the proposed design (i.e., sample 100,000 individuals and compute their probability of selection according to the specified design scheme).
2. Simulate the binary exposure and binary disease variables according to the assumed conditions for the hypothesis of interest. This will depend on the prevalence of the exposure, the prevalence of the disease, the amount of within-cluster correlation in the *X*s and the *Y*s, and the assumed odds ratio.

3. Fit a logistic regression GEE model that accounts for the possible clustering and unequal weighting of the observations. This provides an estimate of the log-odds ratio, and its corresponding standard error and statistical significance under the design.

These steps are repeated a large number of times to obtain an estimate of the power to detect the odds ratio of interest for the selected design characteristics and the selected hypothesis of interest.

Finally, we would like to note that the use of the beta-binomial distribution to generate data works well for the setting that we have considered, namely a dichotomous exposure variable. However, things become more complicated for the setting where exposure is considered to be continuous. In that case, specifying a linear logistic regression relationship for the conditional within-cluster odds ratio will necessarily induce a marginal relationship that is no longer linear in the logit scale. This is due to the fact that the marginal odds ratio will correspond to the expectation of a nonlinear function of the cluster-specific response rates, as well as the conditional odds ratio.

REFERENCES

Agresti, A. (1990), Categorical Data Analysis, New York: John Wiley and Sons, Inc.

Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. Journal of the American Statistical Association, 81, 1074-1079.

Breslow NE, Clayton DG (1993). Approximate inference in generalized linear mixed models. JASA 88: 9-25.

Canada Labour Force Survey

Canada National Longitudinal Survey of Children and Youth

Canada National Population Health Survey

Chambers, R.L., and Skinner, C.J., eds (2003). Analysis of Survey Data. John Wiley and Sons, Chichester.

Diggle, P., Heagerty, P., Liang, K-Y., Zeger, S. (2002). Analysis of Longitudinal Data. Oxford University Press.

Emrich LJ, Piedmonte MR (1991). A method for generating high-dimensional multivariate binary variates. American Statistician 45: 302-304.

Hartley, H.O. (1962). Multiple frame surveys. Proceedings of the Social Statistics Section, American Statistical Association, 203-206.

Heagerty PJ, Zeger SL (2000). Marginalized multilevel models and likelihood inference. Statistical Science 15: 1-19.

Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). John Wiley and Sons, Inc. New York.

Kalton, G., and Anderson, D.W. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, Ser. A, 149, 65-82.

Kalton, G., Brick, J.M., and Le, T. (2003). Household Surveys in Developing and Transition Countries: Design, Implementation and Analysis, United Nations.

Lee AJ (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. American Statistician 47: 209-215.

Liang KY, Zeger SL, Qaqish B (1992). Multivariate regression-analyses for categorical-data. Journal Of The Royal Statistical Society Series B, 54: 3-40.

Lohr, S.L. and Rao, J.N.K. (2000). Inference from Dual Frame Surveys. Journal of the American Statistical Association, 95: 271-280.

National Survey of Family Growth, Cycles II and IV

NHANES III

Qaqish BF (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika 90 (2): 455-463.

Rao (1998)

Rao (1999)

Rosner, B. (1999), Fundamentals of Biostatistics (5th edition), Boston: Duxbury Press.

Skinner, C.J. (1991). On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys. Journal of the American Statistical Association, 86, 779-784.

U.S. News and World Report

Westat. (2002). "Sampling Strategies for the Proposed National Children's Study," technical report to the National Center for Health Statistics at the U.S. Centers for Disease Control and Prevention, October 25, 2002.

Whittemore, A.S. (1981). Sample size for logistic regression with small response probability. JASA 76: 27-32.

Williams DA (1988). Estimation bias using the beta-binomial distribution in teratology. Biometrics 44: 305-308.