10. STATISTICAL ANALYSIS PLAN

10.1 Introduction

10.1.1 Study Design

The design for the National Children's Study (NCS) is based on a nationally representative sample of about 100,000 births to be sampled in 105 geographic areas, called either primary sampling units (PSUs) or Study sites. The pregnancy status of all eligible women of child-bearing age in these areas will be monitored for 4 years, and prepregnancy survey data will be collected for those trying to get pregnant. All women living within the Study sites who become pregnant during the 4-year period will be enrolled in the Study as early in pregnancy as possible in order to measure in utero exposures. The pregnant women will be followed through birth; their children will be followed for 21 years. Throughout this period, the NCS will collect extensive data on a variety of health outcomes and environmental measures and social, demographic, economic, and neighborhood characteristics.

10.1.2 Objectives

The NCS is designed to address hypotheses developed over several years by a variety of stakeholders following a review of the current state of the art in the many areas related to child development and environmental exposures. These primary or "core" hypotheses relate to multiple diseases and developmental outcomes, including asthma, physical and neurological development, diabetes, adverse pregnancy outcomes, obesity, and behavior and mental health problems, such as autistic spectrum disorders.

The NCS will collect data on the children's exposure to chemical, physical, biological, psychosocial, and behavioral environments and their communities, child care, and schools. It will also collect data about the parents' workplaces concerning exposures that might affect their children and data on the children's health from their physicians. Thus, there will be multiple levels of data collection: individual, household, immediate neighborhood, community (e.g., community air quality), and county (e.g., schooling and sociodemographic characteristics).

The study will have the power to examine gene-environment interactions from a developmental perspective in a way that has not previously been done. The NCS will provide a rich source of data with which to investigate the genetic mechanisms associated with rare diseases such as autism; the quantitative contribution of genetic variation to common conditions such as obesity; and the impact of gene and environment interactions on complex diseases and conditions, such as asthma and depression. Multiple gene-environment and gene-gene interactions will play a key role, creating the need for highly complex, computer-intensive forms of analysis. An important goal of the NCS is to provide data to support such analyses.

10.1.3 Overview of the Chapter

This chapter describes statistical methods that will be employed in analyzing NCS data and important issues for these analyses. One primary consideration, of course, is the sample size and power that can be expected in the NCS. This is discussed in Section 10.2. In Section 10.3, we discuss a number of issues relevant to all statistical analysis of NCS data, such as design-based versus model-based analysis, confounding, measurement error, and missing data. Section 10.4 illustrates the range of methods that will likely be used in analyzing NCS data, and Section 10.5 discusses analysis of genomic data.

10.2 Sample Size and Power

10.2.1 Overall Sample Size and Key Subgroups

As noted earlier, the overall sample size for the NCS is about 100,000 sampled children at birth. However, this number is expected to decrease by about 2 percent per year so that, for example, the sample size at age 18 will be reduced to about 69,000 children remaining in the study. Furthermore, the sample size will be smaller for some endpoints. For example, for schizophrenia, the sample size will be reduced because the postulated analyses require placental data and serum from early in pregnancy that are assumed to be available for only 80 percent of the sampled children.

In addition, some hypotheses apply to selected subgroups, defined by characteristics such as sex, race, ethnicity, living area, genotype, or combinations of these characteristics. Examples include the following outcomes: age at puberty, which requires separate analyses for boys and girls; asthma among breast-fed children; and IQ score among “at-risk” children. Subgroup sample sizes are often small, leading to substantially less power.

10.2.2 Impact of Complex Sample Design

The sample design for the NCS is a complex clustered design involving unequal selection probabilities, stratification, and multi-stage sampling. Complex sample designs, particularly clustered designs, have a substantial impact on standard errors and power. The impact of the complex design is measured by the design effect for a given survey estimate. A design effect greater than 1.0 indicates the estimate is less precise than the corresponding estimate computed from a simple random sample of the same size.

Much empirical research has shown that design effects for complex clustered sample designs are generally lower for analytic statistics, such as odds ratios and regression coefficients, than for descriptive statistics, such as means and proportions (see, for example, Kish 1995). The design effects for regression coefficients are discussed in Scott and Holt (1982). The estimates of power for the odds ratios presented in Section 10.2.3 incorporate an allowance for estimated design effects associated with the complex NCS sample design. For most of the calculations, the homogeneity of the exposures in the PSUs is assumed to be modest. However, for hypotheses relating to infant mortality and rate of developmental disabilities, the exposures are assumed to be highly homogeneous within PSUs. The reason for the high level of homogeneity in these cases is that the exposures of interest are neighborhood or community characteristics and policies that will be the same for all children in the neighborhood or community.

10.2.3 Power for Subgroups/Primary Objectives

In hypothesis-driven studies, there are two types of errors. A type I error (generally denoted as α) occurs when the null hypothesis is true but is rejected; a type II error (generally denoted as β) occurs when the null hypothesis is false but is not rejected. For example, if the null hypothesis is that a given factor is not associated with an outcome, then a type I error occurs when there is in fact no association but the study concludes that there is one. Type II error, failing to reject the null hypothesis when a given factor is actually associated with an outcome, is the complement of statistical power; thus, the higher the power, the smaller the chance of making a type II error. While there are no universally accepted error rates, the values of $\alpha = 0.05$ and $\beta = 0.20$ (i.e., power = 80 percent), respectively, are most frequently used when designing studies.

A range of medically important outcomes will be used here to illustrate the ability of the NCS to test exposure-outcome associations involved in the primary hypotheses with power of 80 percent. These outcomes exhibit the range of prevalence that NCS outcomes are likely to have. While some outcomes are common, most are rare and some are very rare. Many of these outcomes are relevant for a single primary hypothesis, but some are relevant for more than one. For example, several hypotheses address different possible predictors of childhood asthma, including environmental factors, exposure to bacteria and microbial products, maternal stress during pregnancy, and diet. For each outcome, a set of different exposures is considered. In each case, power has been calculated for exposure prevalence of 1.0 percent, 2.5 percent, 5 percent, 25 percent, and 50 percent (this range is based on hypotheses developed for the NCS).

Using cerebral palsy (CP) as an example, the results on power displayed in Table 10-1 can be interpreted as follows: Since CP has a prevalence of about 0.2 percent in the general population, that is the rate to be expected in the NCS. Table 10-1 gives the odds ratio (OR) that can be detected with 80 percent power for exposures (i.e., risk factors) with a 5 percent significance level and a prevalence ranging from 1 percent to 50 percent. For very rare exposures (e.g., 1 percent), only those that have a dramatic impact on the occurrence of cerebral palsy (OR greater than or equal to 5.0) can be reliably detected in the NCS. However, for more common exposures, such as those with 5 percent prevalence or greater, factors with more modest effects (OR greater than or equal to 2.6) can be detected with 80 percent power.

Two simplifications were made in these power calculations. First, the analyses consider only the simple bivariate relationships between the exposures and outcomes without addressing the need to control for confounders. The inclusion of confounders likely results in a reduction in the power for detecting the effects of exposures, but often the reduction will be modest. Second, all outcomes and exposures are assumed to be dichotomous variables. This assumption is again made to simplify the table. In fact, most of the NCS outcomes and exposures will be continuous variables. As a result, the power estimates in the tables are likely to be conservative since dose-response analyses with continuous outcome and/or exposure variables would likely lead to greater power.

Table 10-1 displays the magnitude of the minimum odds ratios that can be detected with 80 percent power for the selected outcomes and the range of exposures for analyses. The sample sizes for Table 10-1 assumed to be the full sample for which data are available. As noted above, the sample available is reduced through attrition and, for some outcomes like schizophrenia, by availability of special data required for analysis. As Table 10-1 shows, the magnitude of the detectable odds ratio depends on the prevalence of both the outcome and the exposure. For a given outcome, the closer the prevalence of the exposed group is to 50 percent, the smaller the detectable odds ratio and the greater the power. Similarly, in general, the closer the prevalence of the outcome is to 50 percent, the smaller the detectable odds ratio, i.e., the detectable odds ratios are small when the exposure prevalence is reasonably high. All the ratios are less than 2 when the exposure prevalence is between 25 percent and 50 percent. The bold line in the tables separates the detectable odds ratios into those above and those below 2.

Table 10-1. Detectable Odds Ratio When Analyzing the Total Sample

Outcome	Age	Prevalence of outcome (%)	Prevalence of exposure				
			1%	3%	5%	25%	50%
Infant mortality*	1	0.7	6.01	3.87	2.95	1.97	1.94
Type I diabetes	18	0.2	5.71	3.72	2.86	1.93	1.89
Musculoskeletal defects	1	0.2	5.00	3.33	2.60	1.80	1.75
Cerebral palsy	1	0.2	5.00	3.33	2.60	1.80	1.75
Schizophrenia#	18	0.3	5.06	3.36	2.62	1.81	1.76
Nervous system defects	1	0.3	4.09	2.82	2.25	1.62	1.58
Metabolic syndrome	18	0.4	4.03	2.78	2.23	1.61	1.56
Autism spectrum disorder	4	0.4	3.66	2.57	2.09	1.54	1.49
Heart defects	1	0.6	3.03	2.21	1.84	1.42	1.38
Type 2 diabetes	18	1	2.75	2.05	1.73	1.36	1.32
Major birth defects	1	3.5	1.76	1.47	1.33	1.16	1.14
Adolescent aggressive behavior	18	4	1.82	1.50	1.35	1.17	1.15
Chronic physical aggression (CPA)	10	4	1.76	1.47	1.33	1.16	1.14
IQ score less than 75	18	5	1.73	1.45	1.31	1.16	1.14
Asthma	4	7.5	1.53	1.33	1.23	1.11	1.10
Neurocognitive development	12	8	1.55	1.34	1.24	1.12	1.10
Depression	18	8.3	1.57	1.35	1.25	1.12	1.11
Asthma	7	8.5	1.51	1.32	1.22	1.11	1.10
Neurodevelopmental disabilities	18	10	1.52	1.32	1.22	1.11	1.10
Preterm birth < 37 weeks	0	12	1.41	1.26	1.18	1.09	1.08
Asthma	18	12.5	1.47	1.29	1.20	1.10	1.09
Adverse pregnancy outcomes	0	15	1.38	1.23	1.16	1.08	1.07
Developmental disabilities*	18	17	1.92	1.54	1.37	1.18	1.16
Developmental disabilities	18	17	1.41	1.25	1.18	1.09	1.08
Obesity	12	17.1	1.39	1.24	1.17	1.08	1.07
IQ score less than 100	18	50	1.32	1.20	1.14	1.07	1.06

* The exposure for this hypothesis is a community rather than an individual level characteristic.

# This analysis is restricted to children for whom placental data and serum from early gestation are available (assumed 80 percent).

To illustrate the increase in the magnitudes of detectable odds ratios for subgroup analyses, Table 10-2 presents results comparable to those in Table 10-1, but with the sample size reduced to a 20 percent subgroup. The results in this table could be applied to case-control studies or other analyses based on subsets of the overall NCS sample. It is assumed that the geographic distribution of the subgroup is proportionate to the general population, which would generally be true in case-control studies and other subset analyses. The detectable odds ratio remains below 2 when the outcome prevalence is 3.5 percent or higher and the exposure prevalence is 5 percent or more, but for rarer outcomes and exposures, it exceeds 2. Many subgroups of interest will comprise less than 20 percent of the population and will thus have larger detectable odds ratios.

Table 10-2. Detectable Odds Ratio When Analyzing a 20 Percent Subsample

		Prevalence of	Prevalence of exposure				
Outcome	Age	outcome (%)	1%	3%	5%	25%	50%
Infant mortality*	1	0.7	17.91	10.13	7.08	4.32	5.35
Type I diabetes	18	0.2	16.13	9.42	6.68	4.12	4.99
Musculoskeletal defects	1	0.2	13.44	7.96	5.69	3.50	3.93
Cerebral palsy	1	0.2	13.44	7.96	5.69	3.50	3.93
Schizophrenia#	18	0.3	13.73	8.08	5.77	3.54	3.99
Nervous system defects	1	0.3	10.23	6.19	4.51	2.81	2.93
Metabolic syndrome	18	0.4	10.04	6.07	4.42	2.76	2.86
Autism spectrum disorder	4	0.4	8.77	5.39	3.97	2.51	2.55
Heart defects	1	0.6	6.73	4.26	3.22	2.11	2.08
Type II diabetes	18	1	5.87	3.78	2.89	1.94	1.90
Major birth defects	1	3.5	2.96	2.15	1.79	1.39	1.35
Adolescent aggressive behavior	18	4	3.13	2.24	1.85	1.42	1.38
Chronic physical aggression (CPA)	10	4	2.97	2.16	1.80	1.39	1.35
IQ score less than 75	18	5	2.89	2.10	1.76	1.37	1.33
Asthma	4	7.5	2.35	1.80	1.55	1.27	1.24
Neurocognitive development	12	8	2.40	1.83	1.57	1.28	1.25
Depression	18	8.3	2.46	1.85	1.59	1.29	1.25
Asthma	7	8.5	2.30	1.77	1.53	1.26	1.23
Neurodevelopmental disabilities	18	10	2.33	1.78	1.54	1.26	1.23
Asthma	18	12.5	2.21	1.71	1.49	1.24	1.21
Preterm birth < 37 weeks	0	12	2.04	1.62	1.43	1.21	1.18
Adverse pregnancy outcomes	0	15	1.95	1.56	1.39	1.19	1.17
Developmental disabilities*	18	17	3.79	2.46	1.96	1.44	1.39
Developmental disabilities	18	17	2.07	1.62	1.43	1.21	1.18
Obesity	12	17.1	2.00	1.59	1.40	1.20	1.17
IQ score less than 100	18	50	1.90	1.49	1.33	1.15	1.13

* The exposure for this hypothesis is a community rather than an individual level characteristic.

# This analysis is restricted to children for whom placental data and serum from early gestation are available (assumed 80 percent).

10.3 Statistical Inference

10.3.1 Design-Based vs. Model-Based Inference

Statistical theory provides the basis for drawing inferences about a population based on a sample taken from that population. One approach to statistical inference that could be applied when analyzing NCS data is based on the sample design (design-based inference), i.e., the randomized procedures used to select the sample. An alternative approach is based on a statistical model that the underlying data are assumed to follow (model-based inference). Design-based inference provides the basis for most published descriptive estimates. However, model-based inference is often used when statistical methods are more complex. This section discusses these two analytical frameworks.

10.3.1.1 Design-Based (Randomization) Inference

In design-based inference, a randomly selected sample is used to estimate parameters that would have been obtained had all members of the population under study been included in the sample and provided data. These parameters may be termed census parameters (see, for example, Chambers & Skinner, 2003; Kalton, 2002). A given individual's data are considered fixed, however. The randomization comes about through the sampling used to select the individual. The statistical theory for the designed-based approach to inference from population-based survey data was developed in the late 1940s and discussions of this topic are available from many sources (e.g., see Cochran, 1977; Kish, 1965).

In regression analysis, the census parameters to be estimated consist of the census regression coefficients and the census squared multiple correlation coefficient. If the regression model is correctly specified and the population is large, the census parameters would be virtually the same as the model-based parameters. However, when the model is not correctly specified, the census parameters will differ from the model-based parameters. In this case, the model-based parameters are problematic, but the census parameters are still interpretable (at least under mild misspecification). The census parameters provide the best fitting model of the given structure for the population under study. In that sense, the design-based approach is somewhat robust.

There are two distinctive features of design-based inference: the need to use sampling weights when analyzing the data to estimate the census parameters and the need to take the sample design into account in estimating the standard error of estimates derived from sample data. The weights reflect the unequal selection probabilities with which sample units are selected and also weighting adjustments to compensate for nonresponse and noncoverage and to calibrate the sample to conform to known population distributions. Standard errors, p-values, and confidence intervals for survey estimates must be calculated using special procedures that reflect both sampling weights and any stratification and clustering used in the sample design.

The robustness gained from using weights in making sample estimates of census parameters comes at a price of a loss in precision as compared with correctly specified model-based estimates. With large samples and limited variation in the weights, that loss of precision is generally acceptable. There are, however, cases when the loss of precision is very large, such as when units are sampled with very unequal selection probabilities. In such cases, alternative estimation approaches may be required, for example, incorporating the sample design features into the analytic model (see, for example, Korn & Graubard, 1999, with examples from health surveys, and Chambers & Skinner, 2003).

10.3.1.2 Model-Based Inference

Model-based inference assumes a model for the population data *Y* as a function of a set of parameters θ. One version of this approach is superpopulation modeling (Royall, 1970; Thompson, 1988) where values of θ are considered fixed, and the observed population values are assumed to be drawn from a superpopulation whose distribution is given by $f(Y \mid \theta)$. Inferences about θ are based on the joint distribution of *Y* and the sampling mechanism *S*.

An alternative modeling procedure is Bayesian population inference (Little, 2004). As in design-based approaches, Bayesian population inference focuses on population quantities of interest $Q(Y)$. However, inference is made about $Q(Y)$ by considering the marginal posterior predictive distribution (Ericson, 1969; Holt & Smith, 1979; Skinner et al., 1989), which requires postulating a prior distribution for the model parameters $p(\theta)$ in addition to the model for the data. This is similar to the missing data formulation in which all of the population not observed is considered to be missing and values are multiply imputed via the posterior predictive distribution of the data (Little & Rubin, 2002), although in practice the actual step of imputing values for the entire population can usually be avoided. Probability samples that are "noninformative" in the sense of Rubin (1987) in that the distribution of Y and S are independent (possibly conditional on fixed covariates X) so that the parameters θ and ϕ that govern the data and sampling mechanisms are distinct, allow inference to be made using a posterior predictive distribution based only on the model for the data. This is equivalent to the ignorable missingness assumption (see Section 10.3.4.1 for a discussion of this assumption) that allows inference about θ to be made conditional on observed data in item-missingness situations. However, to maintain the noninformative sampling assumption, the model must be formulated in a fashion that accounts for the sample design. Thus, for example, models being utilized in sample designs with unequal probabilities of selection might stratify based on the probability of selection to account for any associations between the parameters of interest and the probability of selection.

The model-based approach provides a framework in which point estimation and inference can be made in the same fashion as in other areas of statistics. As discussed in the previous section, the greatest disadvantage of the model-based approach is that, if the model is seriously misspecified, it can yield inferences that are worse—perhaps much worse—than design-based analysis. Careful model development and consideration of how and why models are likely to fail can serve as some protection against this outcome.

10.3.2 Confounding and Mediating Variables

Since the NCS is an observational study and not a randomized trial, the main challenge to making causal inferences from NCS data will be to control for confounding variables. Confounding variables are factors related both to an outcome variable and to exposure variables that are being evaluated as risk factors for that outcome, but that are not themselves dependent on the risk factors. The relationship between potential confounders and the outcome variable is not itself of analytical interest. However, the validity of estimated effects of exposures obtained from analyses depends critically on the inclusion of all the important confounders in the analysis.

When choosing potential confounders to be controlled for in an analysis, care must be taken to distinguish them from mediators. Confounders and mediators are each related to both the exposure and the outcome under study. However, confounders are causally prior to the exposure whereas mediators are on the causal path between the exposure and the outcome. Controlling for mediators will lead to a reduced or nonexistent relationship between the exposure and the outcome, thus providing a false impression of the full effect of the exposure on the outcome.

Confounders should also be distinguished from effect modifiers, sometimes called moderators. Effect modifiers partition an independent variable into subgroups where the effects of the independent variable on the dependent variable differ within each subgroup (Baron & Kenny, 1986). For example, Simons and Wood (2004) found that response to ozone exposure varies both by age and gender, with older persons (and particularly older women) experiencing less reduction in FEV_1 than younger persons. Thus, age and gender are effect modifiers for response to ozone exposure.

Appropriate control for confounders, whether by regression methods or propensity scoring, is essential with the NCS data. The NCS will collect information on a wide range of covariates that may be considered as potential confounders for a given analysis. In studying the possible effects of environmental pollutants on asthma and wheezing, for example, there are a number of confounding variables related to environmental and genetic factors as well as to the risk of asthma and wheezing in children. In general, the approach utilized will be to review the scientific literature that describes previously observed factors associated with environmental and genetic factors and the increased risk of asthma and wheezing in order to select a set of covariates that are potential confounders for a specific analysis. In the asthma example, potential confounders in an analysis of the possible effects of environmental pollutants and genetic variation on asthma severity might include maternal gestational factors, such as premature birth and stress and infection during pregnancy; childhood infections; diet and nutrition; socioeconomic variables, such as parents' education, household composition, and housing characteristics; demographic characteristics such as race/ethnicity; access to health care; and so forth.

Sections 10.4.2.1 and 10.4.2.2 discuss linear and nonlinear regression and propensity scoring as methods for controlling for confounding variables. Matching is another method used to control for confounding variables. In matching, the individuals in the comparison group are selected to match the target group on a potentially confounding variable. This holds the effect of the confounding variable constant across groups in analyses. For example, in a study of the influence of prenatal drug exposure on children's cognitive development, the nonexposed comparison group would need to be matched to the drug-exposed target group on premature birth status to rule out a potential alternate explanation for cognitive delays in the drug-exposed group. Case-control studies, which represent a particular example of matching, are discussed in Section 10.4.6.

10.3.3 Measurement Error

10.3.3.1 Impact of Measurement Error

The role of measurement error in the analysis of epidemiologic data (both environmental and other study data) is multifaceted. As with any other types of data, there is the potential for bias and increased uncertainty in predicting outcomes when outcomes or covariates are measured with error. Measurement error in environmental exposures often results from data collection decisions. For example, there is potential bias and increased uncertainty in ecological designs where the required individual-level exposures are measured at the population or group level rather than at the individual level.

Individual measurements for each subject at each time period may contain measurement errors. The extent of these errors, such as those caused by equipment limitations, may be constant across subjects and time, but they may also vary due to collection or processing methods across laboratories or study sites, particularly for measures based on environmental samples or biospecimens. A further complication is that NCS analyses will not be restricted to estimating mean levels or correlations. Some analyses will require sampled individuals to be classified according to whether or not they have been exposed to chemical levels above certain cutoffs; other analyses will want to use continuous measurements to investigate whether threshold levels are the same at different developmental stages.

The effects of measurement error, which depend on the measurement error distribution (Carroll et al., 1995), that are possible include: (1) attenuation of a regression coefficient or other effect measure to the null; (2) hidden effects; and (3) a sign reversal in estimated coefficients. Thus, measurement errors can introduce both variability and bias into data analysis and must be accounted for.

10.3.3.2 Types of Measurement Error

The two general types of covariate measurement error are classical and Berkson error. Let *X* represent the true covariate measure that cannot be observed for all study participants. If *X* is fixed and the surrogate measure *W* varies due to error, then the classical measurement error model is appropriate, $W = X + U$, where *U* represents measurement error. For example, the biological samples of phthalates obtained from blood/urine, cord blood, infant urine, and meconium samples are potentially measured with classical measurement error. Conversely, if *W* is fixed and *X* varies due to error, the regression calibration or Berkson model is appropriate, $X = \mathrm{W} + U$. For example, if the variable of interest is the actual amount of chemical absorbed by the body, the measurement of the chemical level in drinking water or air particles may be the fixed surrogate with the true level absorbed by the body varying as a function of the surrogate.

10.3.3.3 Assessing Measurement Error

There are several sources of data that can be used to evaluate the extent of measurement error. In some cases, study data can be validated for a subsample of cases. For example, it may be possible to obtain more accurate observations for a subset of the primary data or, at an aggregate level, from external sources. For air quality measurements, it is often too expensive to take a personal measurement for every study participant. It is more reasonable to randomly select a subset of the sample to measure personal air quality and collect more general measures such as room air quality for the entire study. The personal measurements can then be used to assess the extent of measurement error due to using room air data.

With replication or reliability assessment, multiple measures of the surrogate variable are observed via internal or external sources; the variation in the replicated measurements gives an indication of extent of variable measurement error.

Another approach uses instrumental variables. An instrumental variable must be (1) correlated with *X,* the covariate being measured; (2) independent of $W – X$ (i.e., measurement error: the true value minus observed value); and (3) independent of the outcome (*Y*) given *X* and any additional covariates that are measured without error (*Z*) (Carroll et al., 1995). For example, in the Faroese Mercury Study of a birth cohort of children, neither validation data nor replication data were available to estimate the cord blood mercury measurement error (Budtz-Jorgensen, et al., 2003). Instead, secondary exposure variables such as the concentration in maternal hair and the average number of whale dinners per month were used as instrumental variables.

10.3.3.4 Modeling Approaches

Regression calibration is essentially the replacement of the true covariate *X* by the regression of *X* on (*Z*, *W*) using replication, validation, or instrumental data (Carroll & Stefanski, 1990). It follows an algorithm of the following three steps: (1) use validation, replication, or instrumental data to estimate the regression of *X* on (*Z*, *W*); (2) replace the unobserved *X* by the estimate from step 1 and rerun the standard analysis to obtain parameter estimates; and (3) adjust the resulting standard errors to account for the estimation. In some cases, a simulation extrapolation approach can be used if validation or replication data are not available to model the calibration function. Heuristically, this approach is a self-contained simulation study that illustrates the effect of measurement error on parameter estimates (Carroll et al., 1995). There are commands available in Stata version 8 as well as macros for SAS that will fit generalized linear models when one or more covariates are measured with error.

In the context of structural equations models, it has been shown that latent variable models can be used to adjust for measurement error in the predictor and response with multiple measures on all subjects (Palta & Lin, 1999).

10.3.4 Missing Data

This section discusses types of missing data that will be encountered in the NCS, and methods that can be used either to adjust for them or to analyze data in the presence of missing data. In any survey, there are data losses due to noncoverage and nonresponse. Noncoverage occurs when individuals are missed in the listing process resulting in some members of the target population having no chance of selection. Total nonresponse (also called unit nonresponse) refers to eligible individuals who are sampled but do not provide any usable survey data. Item nonresponse refers to missing data items for eligible individuals who participate in the study and provide most of the required survey data. Partial nonresponse refers to eligible individuals who are sampled for the study but who provide only a portion of the survey data. This can occur, for example, when data collection involves multiple components (e.g., lab tests, questionnaires, etc.). Wave nonresponse occurs in longitudinal studies in which a sampled individual fails to provide data for one or more of the required waves of data collection. This type of nonresponse can be due to attrition, in which an individual who participated in early waves of data collection drops out of all subsequent waves. Wave nonresponse can also be intermittent rather than attritive, where a participant misses one or more waves of data collection but returns in a subsequent wave.

10.3.4.1 The Missingness Mechanism

When developing methods to account for missing data, it is important to evaluate the process that gave rise to the missing values. This process is called the missingness mechanism. Rubin (1976) and Little and Rubin (2002) define a typology of missingness mechanisms. Data are missing completely at random (MCAR) if the missing data are essentially a simple random sample of the underlying complete data. MCAR is unlikely to hold across an entire sample, but it may hold within strata or classes defined by race, sex, geographic location, or other variables.

The second mechanism is called missing at random (MAR). Data are said to be MAR if the probability that an observation is missing depends on the underlying complete data only through elements of the data that are fully observed. For example, if the probability that a subject drops out depends on classes defined by race, sex, or geographic location and class membership is known for all sampled persons, then the data are MAR.

The third type of missingness is the nonignorable (NI) mechanism. When data are NI, the probability of missingness depends on unobserved data even after adjusting for all observed data. With NI data, the application of standard approaches for handling missing data in the analysis are not valid. Since it is not possible to distinguish NI from MAR using observed data, the only way to identify NI missing data with any confidence is to gather the missing data from a fraction of those not responding. Thus, attempts to model NI missing data are generally speculative and have a limited role in applications.

10.3.4.2 Compensating for Missing Data in Design-Based Analysis

Design-based methods to compensate for missing data consist primarily of weight adjustment and imputation. The following sections describe methods that will be used to weight the NCS data as well as alternative methods for compensating for missing data under model-based inference. The model-based approach for handling missing data in the National Children's Study is discussed in Section 10.3.4.3.

Weighting adjustments

The primary method of compensating for unit nonresponse in survey data consists of adjusting the sampling weights. The initial sampling weight for each respondent is the inverse of the original selection probability for that respondent. These initial weights (often called base weights) can be adjusted for unit nonresponse and, in many cases, noncoverage using methods described below (see Brick & Kalton, 1996, for example, for a more detailed discussion). The NCS will use these procedures to adjust for unit nonresponse.

To compensate for nonresponse, adjustment factors are calculated within selected weighting classes formed by demographic or other data. These data must be available for both respondents and nonrespondents. The adjustment factors are then used to inflate the base weights. The weighting classes are typically based on information from the sampling frame. The underlying assumption is that the nonrespondents are missing completely at random (MCAR) within the weighting classes.

Adjustments for noncoverage are based on external data sources, typically the Census of population or, in this case, of National Children's Study birth certificate counts. The adjustment process consists of calibrating nonresponse adjusted weights so that sample estimates of key characteristics conform to the known population characteristics from the external source. This calibration compensates for noncoverage and it also reduces variance of estimates associated with the characteristics involved in the calibration. Birth certificate data are a likely external data source that can be used for making nonconvergence adjustments in the NCS. These data can, for instance, provide data on the numbers of births to mothers resident in a county and on the characteristics of those births (e.g., birth weight, APGAR results) and of the families (e.g., mother's age, race, education).

Imputation

Imputation is widely used in survey research to assign values to missing survey items and thus compensate for item nonresponse. The imputed values are derived using data from other items available for the respondent that serve as predictors of the missing values. The "hot deck" method is the simplest and probably most frequently used imputation method. In this method, missing data are assigned values from another respondent who is judged to have similar characteristics. For example, missing income data might be replaced with the income of another respondent of similar age, education, and gender. (See Brick, Kalton, & Kim, 2004, for more discussion.)

Numerous methods have been developed for imputation, ranging from the fairly simple and nonparametric hot deck to Bayesian model-based imputation (Little & Rubin, 2002). Multiple imputation is another frequently cited approach (Rubin 1987). As noted by Kalton and Kasprzyk (1986) and Brick and Kalton (1996), most of these methods fall within the general multiple regression framework.

Regardless of the method used, it is advantageous for a project like the NCS to produce filled-in, public-use data sets, as these can be analyzed by researchers without the need for sophisticated statistical modeling and missing-data adjustments.

Compensating for wave nonresponse in the NCS

Some NCS participants will fail to provide data for one or more of the survey waves. While some persons may drop out of the study at one wave and be lost for all subsequent waves (attritors), others may miss one wave but return to the study at a subsequent wave (nonattritors). The choice between weighting adjustments and imputation to handle wave nonresponse is not clear cut. In attrition nonresponse, however, weighting adjustments are usually preferred to a mass imputation of all variables for each of the missing waves.

Weighting adjustment for attrition nonresponse is relatively straightforward since this type of nonresponse is "monotone." Each successive wave adds an additional set of nonrespondents on top of those who dropped out in previous periods. Weighting adjustments at the current wave can be applied to the nonresponse-adjusted weights from the previous wave.

10.3.4.3 Compensating for Missing Data Under the Model-Based Approach

Missing data create a number of problems in statistical analysis. First, multivariate analysis methods typically assume complete data for all subjects. If some items are missing for a given subject, then the subject must be dropped from the analysis unless some form of adjustment is used (Vonesh & Chinchilli, 1997). A second issue concerns statistical efficiency since unavailability of some data elements decreases the effective sample size for statistical analyses resulting in wider confidence intervals and underpowered tests. A third issue is bias, since individuals with missing data may differ systematically from those with complete data. For example, in studies of health-related quality of life, subjects with the poorest quality of life are also most likely to be lost. Thus, analyses of the available data may be biased toward higher quality of life values.

However, a number of analysis methods can be used when missing data are present. The central tool for model-based analyses of longitudinal measurements is the linear mixed model (Diggle et al., 2002) as implemented in the SAS procedure Proc Mixed. The assumptions of MAR and parameter distinctness are sufficient to guarantee that likelihood-based analyses accomplished in Proc Mixed or similar software are correct even with substantial amounts of missing data.

Another analytic approach that avoids some of the stronger assumptions of likelihood-based analysis but still allows considerable flexibility in modeling is the generalized estimating equation (GEE) method (Liang & Zeger, 1986; Zeger & Liang, 1986; Diggle et al., 2002). With this so-called *marginal modeling* approach, one estimates a generalized linear model for the outcome variables (which can be either continuous or discrete), accounting for correlation within subjects (or larger units) by computing an adjusted variance matrix. Validity is robust to assumptions about the within-unit correlation. GEE modeling is valid under the assumption of MCAR though not generally under MAR. However, one can correct this by estimating a model for dropouts given observed data, and then weighting observations by the inverse dropout probability (Robins et al., 1995a, 1995b). A potential disadvantage is that GEE models describe only the marginal distributions of outcomes and therefore fail to capture within-subject correlation, which may be a critical feature of the phenomenon under study (Lindsey & Lambert, 1998).

Sensitivity analysis can be used to evaluate the impact of assumptions about the missingness mechanism. Pioneering work includes articles by Copas and Li (1997), Verbeke et al. (2001), and Troxel et al., (2004). A paper by Ma, Troxel, and Heitjan (2005) describes a method for local sensitivity analysis in a longitudinal model.

Another approach to handling missing observations is to impute them (as discussed in Section 10.3.4.3) and then analyze the filled-in data set as though it were complete. However, analyzing the filled-in data set without some accommodation for the imputation generally overstates precision. Rubin (1978, 1987) proposed multiple imputation (MI) as a way to avoid this difficulty. In MI analysis, one creates not one but several sets of filled-in data. The analyst then analyzes each filled-in data set separately, combining these results into a single overall analysis. Some MI algorithms are now available in commercial software, including the SAS procedure PROC MI.

10.3.5 Variance Estimation under the Design-Based Approach

The NCS is based on a complex sample design involving stratification and clustering by PSUs and by segments within PSUs. Under the design-based approach to inference, these features need to be taken into account in estimating the precision of estimates, whether they are basic descriptive statistics (means, percentages, totals in the total population or in subgroups) or analytic statistics (regression coefficients, odds ratios, etc). Failure to take account of the sample design in analysis can lead to invalid tests and erroneous conclusions (Skinner et al., 1989). This section briefly reviews the methods available for computing variance estimates for survey estimates based on complex sample designs.

Two approaches are commonly used for estimating sampling errors from complex sample designs: (1) Taylor series linearization methods, and (2) replication procedures. With the Taylor series method, estimates of variance are derived using a first-order Taylor series approximation of the deviations of estimates from their expected values. The required sums of squared deviations are then computed using the “ultimate cluster” approach described by Hansen, Hurwitz, and Madow (1953a, Chapter 6) and Kalton (1979). Replication methods, on the other hand, use subsamples of the full sample to obtain the standard errors of estimates (Rust & Rao, 1996). The subsamples, called “replicates,” can take on a variety of forms including balanced repeated replicates, jackknife replicates, and bootstrap subsamples. A statistic of interest is calculated for the full sample and for each replicate, and the variability of the replicate estimates is used to estimate the variance of the statistic. An advantage of replication methods is they eliminate the need to specify complicated variance formulas (e.g., see McCarthy, 1966).

Both Taylor series and replication methods require appropriate variance stratum and variance unit codes in order to calculate the sampling errors. The variance units correspond to the actual first-stage sampling units within a stratum; thus, for the noncertainty PSUs, the variance units are the PSUs themselves, whereas within the certainty PSUs (which are, in reality, strata), the variance units are the sampled segments. (If a segment in a certainty PSU is also selected with certainty, the variance units would be the dwelling units.) Moreover, both methods require at least two variance units per variance stratum. In order to satisfy this condition, it may be necessary to collapse some of the noncertainty sampling strata for variance calculations. If collapsing is required, the resulting variances will tend to be overstated (Hansen, Hurwitz, & Madow, 1953b, page 218). To accommodate analyses that take the sample design into account, software such as WesVar, SUDAAN, and STATA can be used (LaVange et al., 1996; Research Triangle Institute, 2004; Westat, 2002).

10.4 Major Types of Analysis

10.4.1 Overview

This section discusses statistical methods that would be used as tools in the major types of analysis that will be undertaken in the NCS. The analyses are:

10.4.2 Cross-sectional exposure on an outcome;

10.4.3 Identifying causal pathways;

10.4.4 Analysis of neighborhood effects; and

10.4.5 Evaluating temporal effects.

Sections 10.4.2 through 10.4.5 discuss the specific statistical methods that will be used to achieve these analytical objectives. Before these methods are undertaken, however, certain basic, descriptive analyses will be carried out for all the variables being considered for inclusion in the planned analysis. For example, in the case of a dichotomous outcome, the prevalence of exposures (e.g., percent with detectable levels) from various sources (e.g., air, urine, blood, cord blood) will be compared between groups with and without the outcome. The distributions of quantitative exposures (continuous data) from these sources will be assessed with logarithmic or other transformations carried out as necessary to meet assumptions for statistical modeling. Finally, exposure levels measured from these sources will be characterized by the mean, standard deviation, median, and interquartile range. After these exploratory steps, detailed analysis will be done using the methods described in Sections 10.4.2 through 10.4.5.

10.4.2 Individual Cross-Sectional Exposure/Outcome Analysis

10.4.2.1 Linear and Nonlinear Regression

An important part of the data analysis in the NCS will be to investigate the association between an outcome of interest and some exposure measurement, controlling for possible confounders. For example, researchers may be interested in evaluating the association between prenatal exposure to polychlorinated biphenyls (PCBs) and cognitive and motor development in young children. Regression models, linear or nonlinear, are important analytical tools to address such scientific questions.

Linear regression methods are associated primarily with continuous outcomes. Many outcome measures in the NCS can be regarded as continuous in nature, including fetal growth, children's fine and gross motor skills, bone density, and heath care utilizations. Linear regression models can be used to study the relationships between such continuous outcomes and exposures of interest while controlling for confounders.

As an illustration, Daniels et al. (2003) examined the association between mother's prenatal exposure to PCBs and children's cognitive and motor development using data from the Collaborative Perinatal Project. The Bayley Scales of Infant Development were used to assess the infants' mental and psychomotor development at 8 months of age. PCB exposure represented the sum of 11 measured congeners from maternal nonfasting blood sampling collected at the third trimester. Possible confounders included maternal education, socioeconomic index, intelligence quotient, marital status, prenatal smoking status, prepregnancy body mass index, third trimester serum triglyceride, total cholesterol, and dichlorodiphenyldichloroethylene levels, the child's birth order, gestational age, and whether the child ever breast fed. The investigators treated the Bayley Scales as continuous and fitted a linear regression on the exposure and the confounders.

On the other hand, many other outcomes in the NCS will be discrete. For example, congenital malformations can be categorized as present or absent. Outcomes may also take more than two levels of values. For example, a measure of activities of daily life may be coded as very good, good, bad, or very bad. Other outcome measures can be count variables, such as number of hospitalizations in a month. For these outcomes, nonlinear models are appropriate. In particular, logistic regressions (Hosmer & Lemeshow, 1989) are useful for modeling the association between a binary outcome and the exposure, controlling for the confounders. When the outcome is polytomous, some generalizations of the logistic regression can be applied. For example, when the levels of the outcome have no meaningful order, polytomous logistic models can be used (Hosmer & Lemeshow, 1989; McCullagh & Nelder, 1989). When the levels of the outcome follow a natural order, either the adjacent-category logistic model (Agresti, 1984), the proportional odds model (McCullagh & Nelder, 1989), or the continuation-ratio model may be used. When the outcome is a count measure, Poisson regression models or extensions of Poisson models are appropriate (McCullagh & Nelder, 1989).

It is likely that survival or time-to-event analysis will be appropriate for some outcomes, such as child development milestones or occurrence of childhood illnesses. The proportional hazards model is essentially a nonlinear regression model where the time to an event (e.g., infant's first steps or first holding a spoon) is the dependent variable. The potential effects of genetic and environmental factors can be modeled as predictors of delayed or accelerated development using the proportional hazards model (Marubini & Valsecchi, 1995). Variables that change over time (time dependent covariates) can also be included in the model. For example, environmental or other exposures that change with time can be included in the model.

Interaction terms can be added into both linear and nonlinear regression models when the exposure effect may vary with the level of a moderating variable. For example, if the decrease in motor development index for one unit increase in prenatal PCBs exposure is larger in breast-feeding children than in non-breast-feeding ones, then an interaction between the PCBs exposure term and the variable for breast-feeding can be added in the model. Variables such as gender, race/ethnicity, and age are likely to modify the effects of exposures in many NCS analyses.

The standard practice is to use forward or backward selection procedures in determining whether a confounder should enter the model. With these procedures, the main effects are selected first followed by the interaction terms. Likelihood or deviance measures are used to assess overall model fitting. Residuals and Pearson's residuals are used to address model diagnostics (Cook & Weisberg, 1982; McCullagh & Nelder, 1989).

10.4.2.2 Propensity Scoring

Propensity scoring provides an alternative method for controlling on confounder variables. Since its introduction by Rosenbaum and Rubin (1983; 1984), the method has become widely used for this purpose, particularly in biostatistical applications (D'Agostino, 1998).

With propensity scoring, exposure is viewed as a chance event where the chance of being exposed at a given level depends on the confounder variables. For example, breast-feeding rates differ substantially by race, socioeconomic level, and other demographic factors (Li & Grummer-Strawn, 2002). Thus, an infant's chance of being exposed to breast feeding depends on these factors, which may act as confounding variables in an analysis of true effects of breast feeding on healthy growth and development outcomes. The first step of the analysis is to develop a model for predicting the exposure level given the confounders. In the simplest case of a dichotomous exposure (exposed vs. unexposed), a logistic model can be used. With several levels of exposure, an ordinal logistic model can be used (Joffe & Rosenbaum, 1999). An attractive feature of this approach is that a large number of potential confounders and interactions can be introduced into the models. An individual's propensity score for a particular exposure level is then given by his or her predicted probability of experiencing that exposure level.

Comparing groups at different exposure levels controlled on the same propensity scores removes the effects of the confounders included in the propensity model. Often, the propensity score distribution is divided into a number (5 to 10) of subclasses, and confounder control is carried out by performing the analysis within each subclass (see, for example, Rubin, 2007). The subclass results can then be combined in a weighted analysis. Tests of balance for each of the individual confounders can be carried out to check that the distributions of the confounders are equated across the exposure levels. Further balance can be achieved for key confounders by additional calibration procedures (see, for example, Judkins et al., 2006).

Propensity scoring bears a close resemblance to survey weighting. Survey weighting aims to achieve a weighted sample that mirrors the total population of inference. Propensity score weighting can be applied to make the weighted sample at each exposure level have the same standardized propensity score distribution. That standardized distribution could be the distribution for any one of the exposure levels or for the full population. The latter (standardized population) was used in an examination of the effect of the Youth Anti-Drug Media Campaign based on a National Survey of Parents and Youth (Orwin et al., 2006). The propensity score weighting was applied after the survey weighting, and it was reflected in the survey sampling variance estimation procedures used for testing the effects of different levels of exposure to the media campaign. Propensity scoring can similarly be applied in the NCS to examine the effects of environmental, socioeconomic, or other exposures on health and other outcomes, controlling for many potential confounders.

10.4.2.3 Exposure/Outcome Analysis

Overview

The NCS will collect data on many types of environmental exposures. In some cases, different exposures can have related human effects; in other cases, the same type of environmental exposure can produce multiple effects, each of which can be measured at multiple times during the course of the NCS. Thus, the analysis of relationships between exposures and health and other outcomes is necessarily complex.

For example, neurotoxins, such as lead, mercury, and persistent pesticides, and nonpersistent pesticides have similar health effects (Weiss, 2000). Ozone, allergens, endotoxins, indoor air contaminants, and mold affect asthma in similar ways (Gold, 2000; Sunyer, 2001). Potential endocrine disruptors, which appear to cause reproductive problems in birds and fish, can include insecticides, herbicides, industrial chemicals, and heavy metals (Landrigan, Garg, & Droller, 2003).

Some outcomes are accentuated by the interaction between exposures over and above the additive affects of the individual exposures. Across the large number of study participants and geographic locations in the NCS, many different combinations of exposures will be observed. The relative importance of each of the analytes and their interactions can be determined by including all the different analytes in a multivariate analysis. Pathways relating different exposures to each other, to mediators, and to an outcome can be analyzed using structural equation modeling. This method is especially useful for creating a more comprehensive model of mediating mechanisms based on the results of univariate exposure/outcome analyses in the cohort. In this instance, structural equation modeling serves as a confirmatory method, assessing the fit or appropriateness of a proffered causal model rather than as a method used to develop a causal model.

The determination of the measure or measures of exposure to use in an exposure/outcome analyses will often be far from straightforward. Since exposure may derive from different sources at any point in time, it may be measured in several different ways and vary over time. Many of the exposures being measured in the NCS will arise in multiple media. For example, pesticides can be found in air, dust, soil, water, and food, and hormonally active agents can be found in drinking water, indoor air, food, soil, dust, and commercial products. Each of these exposures and media will be measured at multiple times during the course of the study. Exposure data can be collected with such measurement instruments as interviews, medical records, diaries, chemical environmental samples, biomarkers from blood or urine, and community level assessments. Measurement error in the various exposure measures also must be taken into account (see Section 10.3.3).

The sections that follow discuss, in turn, the analysis of exposure/outcome relationships that involve one exposure in multiple media with one outcome; multiple exposures with one outcome; and one exposure with multiple outcomes.

One outcome with multiple sources of exposure

In many cases, the NCS will collect data on exposures to contaminants that exist in multiple media or sources. The example given above is of pesticides, which can occur in air, dust, soil, water, and food. Since adverse outcomes may be accentuated by the interaction between multiple sources, it is important to include all the various sources in statistical models. Statistical models for exposure interaction will be similar to those discussed in the context of gene by environment interactions (Section 10.5.2), though environmental exposures are more likely to be continuously measured. For such models define:

E1 = exposure measure from source 1 or exposure type 1;

E2 = exposure measure from source 2 or exposure type 2;

Then the association with the outcome of interest may be modeled as,

$$\text{Logit}[\Pr(\text{outcome})] = b_0 + b_1 E1 + b_2 E2 + b_3 (E1 * E2).$$

As an example, consider “exposure” to phthalate esters. Phthalates have been shown to produce male reproductive tract malformations, including cryptorchidism or undescended testes (UDT), in rats when administered during sexual differentiation (Wilson et al., 2004). The quantification of phthalate levels in the human environment is crucial to determine whether exposures are sufficient to produce UDT or other “outcomes” in humans (Fisher, 2004). Phthalates are of particular concern because exposure is ongoing and ubiquitous (CDC, 2003; Silva, Barr, et al., 2004). They are widely used as softeners of plastics, solvents in perfumes, and additives to hairsprays, lubricants, and insect repellents. These exposures can be measured using air samples. Phthalates and its metabolized forms can be measured in biological samples such as maternal blood/urine, cord blood and infant urine, meconium, and amniotic fluid (Silva, Slakman, et al., 2004).

Swan et al. (2005) developed a global score for the assessment of the phthalates as a group that took into consideration exposures from multiple sources and diverse parent compounds. They constructed their global score by categorizing individual metabolite concentrations into low (below the 25th percentile), intermediate (between the 25th and 75th percentiles) and high (above the 75th percentile) groups, assigning a score to each group and then summing these scores across the metabolites measured in each urine sample. The use of a global score allows for assessment of the phthalates as a group and takes into consideration exposures from multiple sources and diverse parent compounds.

Once a subset of exposure summaries is constructed (e.g., cumulative exposure over time from different sources or peak exposure), correlations among the various summaries can be evaluated. This information along with estimates of the individual exposure associations with outcomes of interest can then be used to develop additional models. Latent variable models (Jöreskog & Sörbom, 1996) and generalizations of such models (Sammel & Ryan, 1996; Sammel, Ryan, & Legler, 1997; Muthén & Muthén, 2004) can be used to assess and validate proposed models of exposure and outcome relationships. (Latent variable models are discussed in more detail in Section 10.4.3.) These models should be used in a confirmatory setting for inferences regarding structure to be meaningful.

As an example, consider Figure 10-1, which illustrates how a latent variable for organophosphate (OP) exposure is derived from three observed sources of information. In this structural equations framework, the underlying, unobserved true OP exposure (represented by an oval) gives rise to the observed measured exposures (rectangles) from the various sources: two body compartment measurements and one indirect personal air monitor measurement. Each λ represents the contribution of a particular exposure type to a composite OP measure. The impact of "true" OP exposure on birth weight is estimated by *θ*. Statistical tests for *θ* are a type of global test for the impact of all the observed types of OP exposures on birth weight.

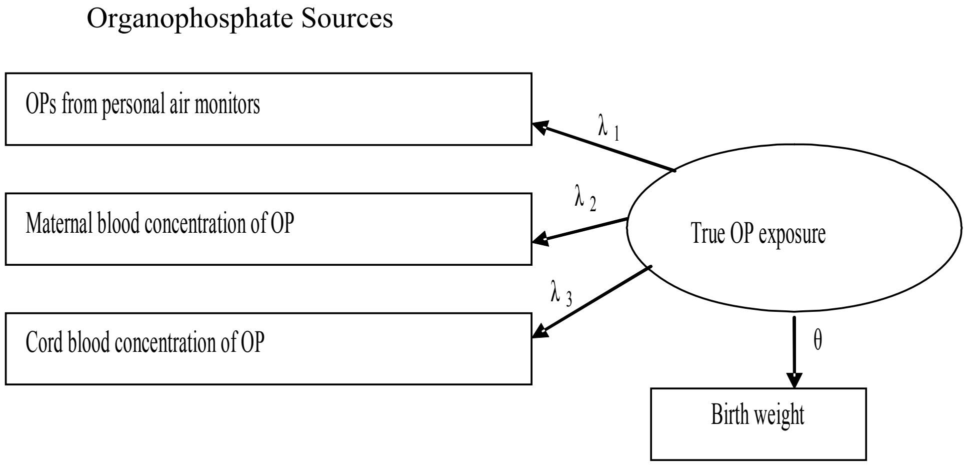

Organophosphate Sources
OPs from personal air monitors
Maternal blood concentration of OP
Cord blood concentration of OP
λ 1
λ 2
λ 3
True OP exposure
θ
Birth weight

Figure 10-1. Path Diagram for Multiple Exposures

One outcome with multiple exposures

Another type of exposure/outcome relationship occurs when one outcome is associated with multiple exposures, which may act independently or in combination to influence the outcome. The issue of timing further complicates evaluating the effect of an exposure, since the time of exposure may significantly modify its influence on an outcome. Logistic regression can be used to estimate both the independent and interactive effects of multiple exposures on the overall risk of a specific dichotomous outcome as well as the impact of timing of exposures.

Consider the example of preterm birth, which will be assessed in the NCS. Preterm birth is influenced by environmental, psychological, social, physical, and genetic factors. Two important mediators of preterm birth are inflammation and intrauterine growth restriction (Steer, 2006). The inflammatory response of the human body as it relates to preterm birth can result from bacterial vaginosis (Hartville, Hatch, & Zeng, 2005) and stress (Ruiz, Fullerton, & Dudley, 2003). In addition, stress can be the result of several factors, such as socioeconomic status (Misra, O'Campo, & Strobino, 2001) or lack of social support (Sheehan, 1998). Stress response itself is also mediated by endocrine function and corticotrophin-releasing hormone, which are related to the risk of preterm birth (Gennaro & Hennessey, 2003). The use of logistic regression modeling techniques will allow investigators to use NCS data to analyze how the interaction of these exposures affects the overall risk of preterm birth.

The impact of multiple interactive exposures on preterm births can be analyzed using a structural equation modeling framework (See Section 10.4.3). For example, Sheehan (1998) used structural equation models to show how economic stress, family stress, and lack of social support influence low birth weight.

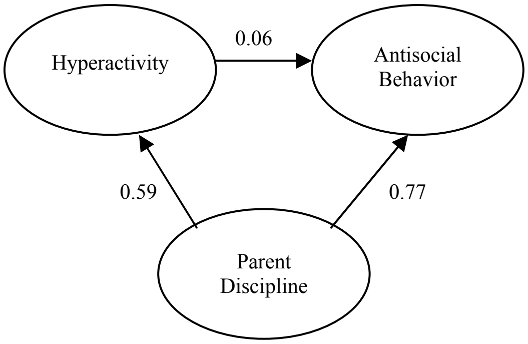
Multiple outcomes with a single exposure

Although some exposures lead predictably to a single identifiable outcome, both theory and empirical evidence suggest that a given exposure can lead to various, sometimes alternative, outcomes (Rutter, 1989). Elements of the environment, individual characteristics, and genetic predispositions can modify the trajectory such that the same exposure leads to a diversity of outcomes across, or even within, individuals. Both the various outcomes associated with an exposure and the moderating conditions that produce the different outcomes can be modeled using multivariate statistical techniques.

A set of trajectories with a unified starting point but multiple endpoints is described in the psychological and biological literature as multifinality (Cicchetti & Rogosch, 1996). Within the NCS, multifinality will arise with regard to processes of resilience. Resilience occurs when a child has an exposure that would logically lead to a negative outcome but does not due to a moderating influence. These individuals have outcomes substantially better than their history of exposure would predict, thus leading to a multifinality analysis. The resulting multiple endpoints consist of different and sometimes unrelated outcomes not simply different levels within the same outcome. For example, some children who experience physical abuse become highly aggressive toward others; some become vulnerable and open to subsequent exploitation; and others demonstrate remarkable resilience and exhibit high levels of behavioral and psychological competence.

Multifinality can be analyzed using one of several multivariate techniques that permit modeling of multiple, simultaneous dependent variables. One such technique is multivariate analysis of variance (MANOVA), which permits the prediction of multiple interval or ratio-level dependent variables from a common set of categorical independent variables. In the initial stages of analysis, MANOVA yields a multivariate *F*-statistic that indicates the significant effect of the independent variable across a multivariate response vector. Subsequent univariate tests are used to uncover the specific relations of the independent variable and any tested interactions with each of the dependent variables. The technique permits direct comparisons of strength of prediction to the multiple outcomes and the differential role of moderators across outcomes.

Multifinality can also be analyzed using structural equation models and path modeling where a single exposure can be modeled as resulting simultaneously in more than one outcome, and the relations between the exposure and each outcome can be compared. As an example, using structural equation modeling, Patterson, DeGarmo, and Knutson (2000) found that poor parental discipline predicted both child hyperactivity and child antisocial behavior in boys although the two outcomes were not significantly associated with each other (see Figure 10-2). The numbers on the arrows in the figure below are the estimated standardized coefficients for the indicated paths. Based on a separate analysis of the data, Patterson and colleagues suggested that parental antisocial behavior, a construct with a strong heritable component, might be the moderating factor that leads to this divergence in outcomes from parental discipline. The NCS would be well suited to test such multivariate models of multifinality.



Source: Patterson et al., 2000.

Figure 10-2. Model of Multifinality of Parent Discipline

10.4.3 Identifying Causal Pathways

The Children's Health Act of 2000, which authorized the planning and implementation of the NCS, also directed the Study to "investigate basic mechanisms of developmental disorders and environmental factors, both risk and protective, that influence health and developmental processes." This means that, in addition to establishing cause/effect relationships, the NCS has been directed to investigate the mechanisms mediating these associations. As an example, a number of factors have been shown to be associated with preterm birth: infection/inflammation, environmental toxins, and behavioral/psychosocial factors. What are the mechanisms through which these factors influence prematurity? Do they involve separate, independent pathways? Do they cumulatively influence the same pathway, such as prostaglandin synthesis, or do they interact in some other way? The objective of the NCS is not only to establish which risk or protective factors are associated with which outcomes but also to increase our understanding of how this occurs so that preventive efforts can be more specifically focused.

Structural equation modeling (SEM) can be used to address such issues. This method allows significant pairwise associations found between a set of single factors to be placed in a larger, theoretically derived contextual model with other significant associations to examine the interrelationships.

Thus, SEM can be used to examine relationships between exposures, mediators, and outcomes in a single overall model. The technique combines multiple regression, path analysis, and factor analysis to assist in causal inference. SEM extends the general linear model since it estimates simultaneous relationships between multiple covariates and responses (outcomes), possibly including unknown latent variables.

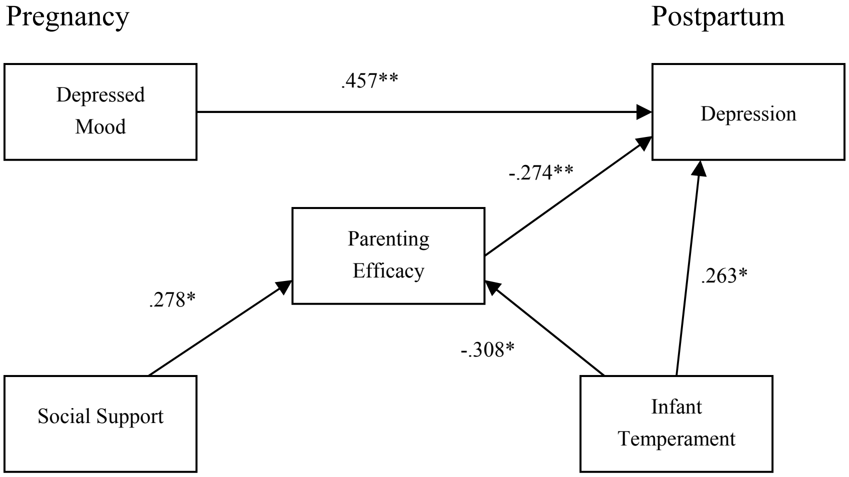
A major strength of SEM is its ability to also model constructs as latent variables simultaneously by means of separate regression equations (Bollen, 1989). Latent variables are unobserved variables or constructs (e.g., IQ) estimated indirectly in the model using measured variables (indicators) that affect them. They can be used as either independent or dependent variables.

SEM/LISREL[1] modeling focuses on two steps (Jöreskog & Sörbom, 1996): validating the measurement model and fitting the structural model. The measurement model specifies how latent variables/hypothetical constructs depend upon the observed variables, and how the association between observed variables is mediated through other observed variables. The structural model specifies the causal relationships between latent and/or observed variables, describes the causal effects, and assigns the explained and unexplained variance. At the outset, one specifies a model based on the underlying theory.

Some of the analyses in the NCS will involve SEM models that do not include latent variables. SEM also works very well in these cases (Bollen, 1989) where the objective is often to understand mechanisms that mediate the association between multiple exposures and a single outcome. As noted earlier, many of the conditions (e.g., premature birth) that the NCS plans to investigate will show associations with multiple psychosocial and physiological factors, and the aim will be to explain the causal mechanisms. For example, if infection and inflammation are shown to explain a significant amount of the variance in premature births and this is also the case with environmental toxins and stress, what are the causal pathways and how are they mediated?

A study of maternal postpartum depression by Cutrona and Troutman (1986) provides an illustration of this type of analysis. The study sought to identify maternal and child qualities associated with changes in depression from pregnancy to the postpartum period. The model was based on theoretically predicted relations from previous research on bivariate relations and significant associations resulting from subsets of analyses where social support, infant difficult temperament, and parenting efficacy were related. Thus, the model was built on theory and preliminary data analyses. Combining these multiple variables into a single model demonstrates, for example, that while having a temperamentally difficult infant predicts greater increases in depression directly, part of this effect is also mediated through feelings of efficacy about parenting (see Figure 10-3 below). Such analytic models inform both more in-depth research questions and process avenues for intervention.

[1] The terms SEM and Linear Structural relationships (LISREL) will be used interchangeably from here onwards.

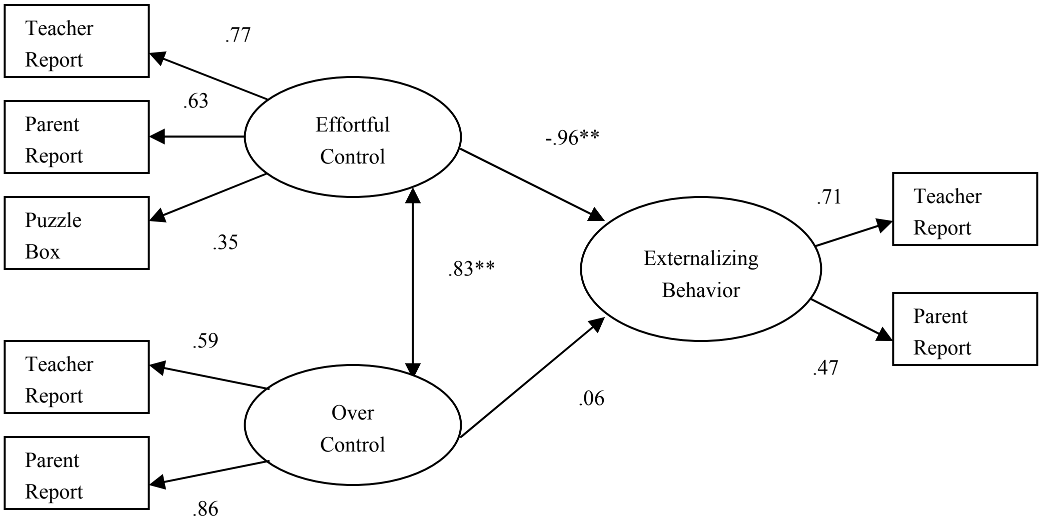


* significant path, p < .05
** significant path, p < .01
Source: Cutrona & Troutman, 1986.

Figure 10-3. Social Support, Infant Temperament, and Parenting Self-Efficacy as Predictors of Postpartum Depression

Other analyses in the NCS will involve multiple observed indicators of latent constructs, and, consequently, latent variable models will be applicable. The latent variable model of child temperamental self-control and aggressive behavior problems developed by Valiente et al. (2003) provides an example (see Figure 10-4). The latent independent variables in their model are temperamental effortful control and temperamental over-control (ellipses in the figure below). The observed variables (rectangles) associated with effortful control are parent, teacher, and task observation ratings of relevant child attention skills and persistent behavior. The observed variables (rectangles) associated with over-control are parent and teacher reports of relevant over -controlled behavior. The latent dependent variable is behavior problems (ellipse). The observed indicator variables are parent and teacher reports of child aggressive behavior. The arrows for the latent variables point toward the observed variables associated with them. The numbers on these arrows are the estimated standardized regression coefficients for the structural equation and measurement models.

This model indicates a direct relation between effortful control and child externalizing behavior. No significant direct relation is found between over-control and externalizing behavior with effortful control simultaneously accounted for in the model. The use of latent variables in this model permits more robust measurement of the latent constructs than would an analysis including only one observed assessment of each construct from a single reporter.



** significant path, p < .01
Source: Valiente et al., 2003.

Figure 10-4. The Relation Between Child Self-Control and Aggressive Behavior Problems

10.4.4 Analysis of Neighborhood Effects

The design of the NCS involves clustered sampling with clustering at the county level, and, within the county, at segment level. This type of design has the analytic benefit of providing a structured set of geographically defined neighborhoods for evaluating the effects of exposures that occur at the neighborhood level. Neighborhoods may be defined in a number of ways depending on the type of exposure and hypothesis being tested. For example, school districts might define a neighborhood for school performance assessments while geographic or administrative boundaries might define a neighborhood for examination of environmental exposures related to the water supply.

Data arising from neighborhoods or clustered structures have a hierarchical form since individual-level data can be grouped within a higher category. Since hierarchical data are nested within a higher structure, there is generally some degree of correlation between observations. For example, two individuals within one community generally are slightly more similar with regard to such factors as religiosity, socioeconomic status, education, and environmental exposures than two individuals sampled from different communities. Multilevel modeling (MLM) is the primary analytic method for analyzing the effects of the hierarchical structure.

MLM is also known as the random coefficient model (Rosenberg, 1973) and as the hierarchical linear model (HLM) (see Bryk & Raudenbush, 1992). MLM is based on the mixed-effects model with both fixed and random components. The regression coefficients are treated as random variables that can vary depending on the higher level unit (e.g., neighborhood). Consequently, MLM can be used to develop regression models with intercepts and regression weights across higher level units as outcomes and other higher-level variables as covariates.

Hierarchical models play an increasingly important role in epidemiology. For example, Juhn et al. (2005) assess the influence of neighborhood and individual-level factors on the incidence of childhood asthma among children born in Rochester, MN, between 1976 and 1979. The neighborhood-level variables considered in this study include collective efficacy, social cohesion, neighborhood socioeconomic status, and whether the Census tract contains major highways or railroads.

An important application of hierarchical modeling with NCS data will be to examine neighborhood effects on various health and developmental outcomes. The collection of standardized information on neighborhoods and counties will permit analysis of effects of both neighborhood-level and individual-level factors.

A hierarchical model can be fitted using either Monte Carlo methods (Gelfand & Smith, 1990) or some approximate methods such as penalized likelihood, penalized quasi-likelihood (Breslow & Clayton, 1993), and restricted iterative generalized least squares (Goldstein, 1995). Interaction terms can be added in the same way as in linear and nonlinear regression models. There are several software programs available for fitting multilevel models, including MLwiN (www.cmm.bristol.ac.uk/), HLM (www.ssicentral.com/), and WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/).

10.4.5 Evaluating Temporal Effects

10.4.5.1 Overview

The longitudinal design of the NCS has the important advantage of permitting the evaluation of temporal effects. There are several statistical tools available for evaluating temporal effects depending on the research question being asked and the type of data being collected. This section discusses longitudinal data analysis, structural equation modeling with longitudinal data, and growth curve models.

10.4.5.2 Longitudinal Data Analysis

Longitudinal data analysis most often refers to repeated measurements on individuals. In the NCS, data will be collected from individuals at regular time points throughout the period of study participation. Specifically, physical, cognitive, and intellectual growth and functioning will be measured over many years. Because these measurements are taken from the same individuals over time, they are correlated and, thus, require special tools for data analysis.

The primary methods used in longitudinal data analysis are generalized estimating equations (GEE) and mixed effects models (see, for example, Fitzmaurice, Laird, & Ware 2004). Since repeated measures may be considered to be “nested” within individuals, multilevel models can be used to reflect the dependent structure of the observations (see Section 10.4.4).

In mixed effects models, individual effects are modeled explicitly. In essence, a separate regression equation is estimated for each individual based on the values of the dependent variable and the independent variables measured at the different time points. The repeated measurements are assumed to be independent within a given subject after the individual intercept and slope for the subject is taken into account. For example, in modeling dental growth, an intercept and slope might be fit for each individual; variations in projected growth about the regression line for an individual are assumed to be independent. In this example, a mixed model would include random effects for growth variations in each individual and fixed effects for factors such as gender that affect overall growth rates.

GEE is associated with marginal models, so-called because they are based on data that are averaged or accumulated for each time point. That is, the basic model is E($\boldsymbol{Y}_i$) = $X_i'\beta$, where $\boldsymbol{Y}_i$ is a response vector at time i, X_i is an independent variable measured at time i, and β is a fixed regression coefficient. For example, Diggle, Liang, and Zeger (1994) show how marginal models using GEE can be used to compare respiratory infection rates between children with and without vitamin A deficiency using examination data from six medical visits, adjusting for the effects of seasonality and age**.** While marginal models and mixed effect models give similar results for continuous outcomes, GEE is more suited for binary outcomes.

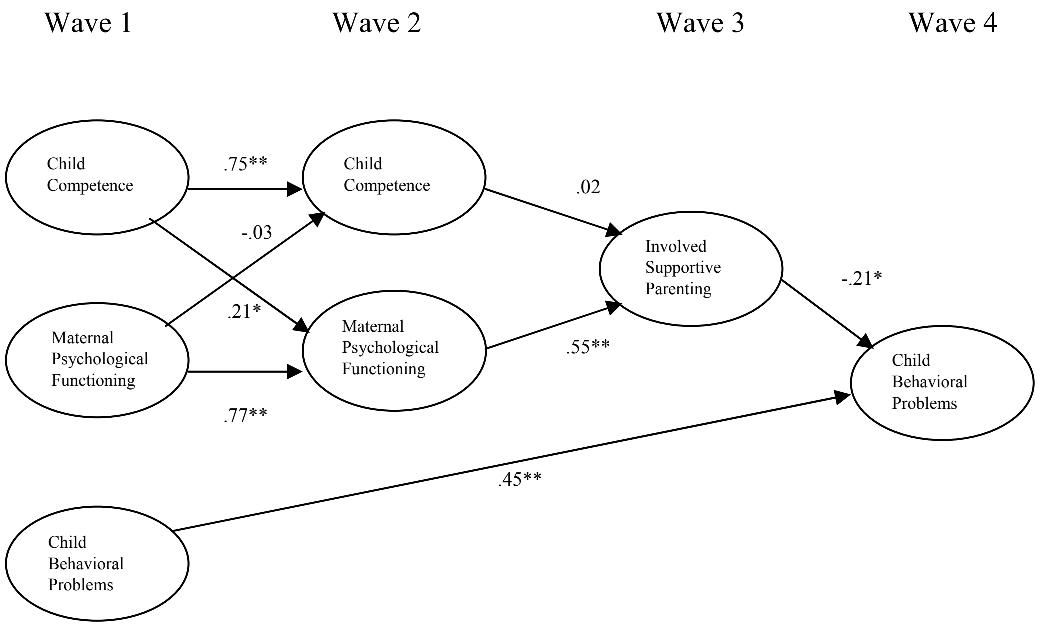
10.4.5.3 Longitudinal Structural Equation Modeling

Structural equation modeling, described in Section 10.4.3, has particular utility for modeling longitudinal data. Data sets such as the NCS that include multiple repeated measures of constructs or indicators are well suited for this analytic strategy. SEM permits modeling of latent variables from multiple indicators and the simultaneous testing of the interrelations among latent variables or individual observed variables. Longitudinal SEM extends this to make multivariate modeling of change over time possible.

When repeated measures over time are represented in SEM models, it is possible to examine predictions of later values of a latent construct while simultaneously accounting for stability of the construct from a previous assessment. Consequently, the predicted outcome represents change in that variable from the previous time point rather than being a static assessment. Because it permits complex multivariate analyses through robust model estimation techniques, SEM provides one of the most statistically elegant methods for investigating stability and change in longitudinal research.

Longitudinal SEM models also permit highly complex, temporally accurate testing of latent variable mediation models, such that the exposures precede the mediator and the outcome follows the mediator temporally, reducing the possibility of reverse-mediation. When previous values of the variables are also accounted for in the model, longitudinal SEM mediation models provide greater support for the hypothesized causal pathways than do models without these multifaceted longitudinal components.

An example of SEM with four waves of data concerning maternal and child adjustment is presented by Brody, Kim, Murry, and Brown (2004). Within these four waves were latent variables at two time points each for maternal psychological functioning, child competence, and child behavioral problems as well as a latent variable for involved supportive parenting (see Figure 10-5 below).


Wave 1
Wave 2
Wave 3
Wave 4
Child Competence
.75**
Child Competence
.02
-.03
Involved Supportive Parenting
-.21*
.21*
Maternal Psychological Functioning
Maternal Psychological Functioning
.55**
Child Behavioral Problems
.77**
.45**
Child Behavioral Problems

* significant path, p < .05
** significant path, p < .01
Source: Brody et al., 2004.

Figure 10-5. Longitudinal SEM Model of Child and Maternal Functioning

The longitudinal nature of the data permits complex inferences about the interrelations of these constructs over time. As can be seen from the figure, Wave 1 child competence predicts changes in maternal psychological functioning between Waves 1 and 2, which subsequently predicts involved supportive parenting at Wave 3. The figure suggests that child competence has effects on later parenting behavior only through its effects on intervening maternal psychological functioning. Involved Supportive Parenting also predicts changes in child behavior problems between Waves 1 and 4. Longitudinal SEM effectively models the complex bidirectionality of influence that parents and children have on each other's functioning over time. The NCS data will permit modeling such as this over the course of many waves of data and for multiple sources of influence.

10.4.5.4 Growth Curve Analysis

The primary goal of growth curve analysis is to describe patterns of change over time and identify predictors that affect these patterns. Growth curves may be modeled using exponential or logistic functions when early growth is rapid. Polynomial growth curves may also be useful. As a familiar example, height has a well understood developmental trajectory where a logistic growth curve fits the first three years of life, followed by a linear growth trend until adolescence when growth again follows a logistic curve (Bock et al., 1973).

Another example of growth curve analysis is given by Cherlin et al. (1998), who describe the effects of parental divorce on the subsequent mental health of their children at ages 7 through 33. This study, which was based on data from the National Child Development Study (Chase-Lansdale et al., 1995), assigned scale scores to emotional problems at a range of ages. Growth in these scales was then modeled as a function of age at time of divorce, gender, economic status, class background, and school achievement. All variables except class background were statistically significant. The significant age at time of divorce variable resulted from higher scores on emotional problems scales for later ages at time of divorce up to age 22.

The latent growth curve model (LGM) is a special case of structural equation modeling and multilevel modeling. LGM treats repeated measures of individual behavior as a function of chronological development. LGM is a type of multilevel model that extends the hierarchical structure to panel data in which individuals are observed across time.

10.4.6 Case-Control Studies

10.4.6.1 Introduction

In case-control studies, a small number of cases (usually persons with a particular disease or other outcome of interest) are compared with a sample of controls, persons without the disease or the outcome being studied. There are several potential situations where this method would be relevant to the NCS. One example would be where blood or other specimens have been stored for many NCS participants but, due to laboratory processing expenses, only a relatively small number of individuals (with a given condition) would actually be selected to have their specimens analyzed. In such a situation, the specimens of an appropriate set of controls (persons without the condition) would also be analyzed and the two sets of data would then be subjected to a case-control analysis. A similar situation might arise if a new hypothesis called for additional measurements to be taken.

A second situation, which is the more traditional application, would arise when individuals with a given outcome are rare. In such a situation, all persons sampled in the NCS with the condition might be included in a comparison with a selected sample of controls to evaluate risk factors, for example.

10.4.6.2 Selection of Controls

The selection of controls is critical to the validity of case-control studies. Quite often, some form of matching is used in this process to control for confounding variables known to influence disease incidence or outcome. The two basic methods of matching are set matching, where each case is matched individually to one or more controls, and frequency matching, where cases and controls are matched in categories (e.g., a group of 15 cases who are male and aged 30 to 39 might be frequency matched to 45 controls who are male and aged 30 to 39). While set matching is a more traditional approach, frequency matching has a number of advantages both operationally and analytically, particularly in large studies.

Because case-control studies typically begin with disease cases that have already occurred, they are subject to significant sources of bias. A key step in the process of ensuring a bias-free case-control study is that cases be representative of all those who develop the disease under investigation. One threat to this process is that cases are often identified as they are diagnosed in a clinical setting, and mild cases or those that result in early mortality will not be diagnosed and are thus missed as cases. This type of bias is called incidence-prevalence bias or survival bias. Case-control studies can also give biased results if the controls are not representative of the population at risk for developing the disease under investigation. To avoid these types of bias, it will be of paramount importance for the NCS to select appropriate and representative cases and controls.

10.4.6.3 Case-Control Studies in the NCS

Nested case-control studies are those where cases are identified within a well-defined cohort, such as the NCS, where controls are selected from within the same cohort. Nested case-control studies combine some of the advantages of both cohort and case-control designs. Depending on how control subjects are selected, a few sources of bias inherent to the case-control design (e.g., recall of events and chronological differences between case and control identification) can be avoided using a nested case-control investigation. Extensive data representing the time period prior to the disease diagnosis will be available for cases, and corresponding data will be available for controls. Another weakness inherent to case-control studies that will be prevented by selecting cases and controls from the NCS cohort is the lack of extensive records before disease diagnosis. Data on many key exposures not typically available in case-control studies, such as dietary patterns, medication use, and environmental exposures, will be available to NCS researchers to determine pre-morbid risk factors more accurately. The NCS is also expected to utilize the matched case-control study design, where individual cases are matched to one or more controls based on similar demographic characteristics suspected of confounding the relationship between the exposure and outcome association under investigation.

10.4.6.4 Analysis of Nested Case-Control Studies

When analyzing nested case-control studies, the use of standard survey weights will generally lead to unacceptably large variability in estimates. By design, controls are matched to cases in terms of key confounding variables, and the distribution of these variables in the cases is usually very different from the distribution in the general population. This feature gives rise to large variation in sampling weights, which, in turn, leads to large standard errors. There are several ways to avoid this problem, two of which we describe below.

One approach is to perform an unweighted analysis. This type of analysis of case-control studies provides estimates of relative risk but not absolute risk. Thus, for example, estimates of regression coefficients would be useful but not of the intercept. Alternatively, if weights are to be used, a possible approach is to weight the sample to the case distribution, thus compensating for any disproportionate sampling of cases or nonresponse. See Scott (2006) for a discussion of these issues.

In case-control studies that used set matching, the multivariate analysis technique most frequently used is conditional logistic regression. This estimation takes into account the pairing or matching of cases and controls with respect to the variables that determined the matching. The interpretation of the coefficients in conditional logistic regression is the same as in ordinary logistic regression except that these coefficients are to be considered "adjusted" not only for the variables included in the model but also for the matching variables.

For the statistical analysis of case-control studies using frequency matching, a more efficient strategy is to use ordinary logistic regression and include the matching variables in the model. Similarly, nested case-control studies can be analyzed in the same way as matched case-control studies, where cases and controls are matched by length of follow-up. As a result, the multivariate analysis technique most often employed is conditional logistic regression, in which the conditional variable is length of follow-up. This type of conditional logistic regression model is similar to the Cox proportional hazards regression model.

10.5 Analysis of Genomic Data

The combination of a longitudinal follow-up of the NCS cohort, its large size, and its comprehensive collection of environmental exposures will provide a rich source of data with which to investigate the contribution of genetic variation to complex diseases such as autism, obesity, and asthma as well as the impact of gene and environment interactions on neurodevelopment, health, and behavior outcomes. For either genome-wide analysis (GWA) or candidate-gene approaches, the outcome can be discrete (e.g., having autism or not), continuous (e.g., quantitative measurements of depression) or censored survival data (e.g., time to onset of type 1 diabetes).

With respect to complex phenotypes that will be studied in the NCS, it is expected that multiple genes as well as gene-environment and gene-gene interactions play a key role. The large sample sizes available will greatly facilitate not only the identification of individual genes, but also the ability to identify gene-gene and gene-environment interactions. It is nevertheless clear that the ultimate success of such GWAs will depend largely on the development of highly complex and innovative analytic strategies. Methods that can efficiently account for the genome-wide linkage disequilibrium patterns and control for genome-wide error rates using false discovery rate procedures are required. Novel statistical methods that can identify gene-gene and gene-environment interactions and methods that can incorporate known biological knowledge, such as networks and pathways, in searching for complex disease genes are also greatly needed. The analysis of genomic data is a field of much active research. Analysis of genotype effects and multilocus genotype-by-genotype interactions (e.g., epistasis) as well as gene-environment interactions can be cast in a regression framework for different types of outcomes where the predictor variables include SNPs, environmental exposures, interactions among SNPs, and SNP-by-environment interactions. Due to the problem of high-dimensionality, standard regression analysis methods cannot be applied directly when many genes are involved because they produce highly variable estimates. Methods developed for analyzing high-dimensional data, such as microarray gene expression, massively parallel signature sequencing (MPSS), and evolutionary trees of haplotypes, may also be utilized. New analytic methods can be expected to emerge in the future, and researchers analyzing the genomic data in the NCS will need to apply the best methods available in every phase of the process. Some specifics of these methods are described below.

10.5.1 Haplotype Analysis

Recent advances in high-throughput technologies and the decrease in genotyping costs have made genome-wide association analysis a feasible tool in the search for genetic contributors of complex traits, including many complex diseases. As the NCS evolves and technologies mature, it is possible that genome-wide genetic profiling for each participant of the study will be available to enable possible genome-wide searches for genetic variants as well as the interactions among genes and between genes and environmental risk factors. One challenge of such data is their very high dimensionality. One solution to this problem is to fully utilize the information from tagging SNPs, haplotypes, and haplotype blocks derived from the HapMap project. Haplotypes are a set of closely linked genetic markers present on one chromosome which tend to be inherited together and which can be utilized as the unit of analyses in order to examine their effects and interactions with the environmental exposures. Haplotype analyses are potentially more powerful in identifying genes predisposing to certain health outcomes. For example, a test for association between the common haplotypes in haplotype blocks and specific outcomes can be conducted. Such association analysis can be performed using sliding windows of a small number of overlapping SNPs. Alternatively, newly developed methods such as Logic regression, FlexTree, and threshold gradient descent procedures can also be applied for considering haplotypes in multiple regions and for identifying haplotype-by-haplotype interactions and haplotype and environmental exposure interactions.

One difficulty with haplotype analysis is that haplotypes are often not observed but must be estimated from the genotyping data. This can be accomplished in a regression analysis and missing data context. The expectation-maximization (EM) algorithm can be developed for estimating the model parameters (Lin, 2004). Such EM-based estimation as well as inference procedures have been developed for binary outcomes in case-control designs and for survival outcomes in prospective cohort designs. Following the nonparametric maximum likelihood approach in Scheike and Juul (2004) for estimation of the Cox model under nested case-control sampling, methods for analyzing age of onset data and for estimating the haplotype effects could be developed in a framework of censored data regression and missing data under the case-control and nested case-cohort samplings. Specifically, we can treat the haplotype phases and the haplotypes of those who are not genotyped as missing data and use the EM algorithm and the nonparametric maximum likelihood approach to estimate the haplotype relative risk and the baseline hazard function (Chen & Li, 2005, in preparation).

10.5.2 Population Stratification

Population stratification is an important issue to consider when studying gene-trait associations using unrelated subjects, since the observed association could be spurious without appropriate adjustment for underlying population strata (Cardon & Palmer, 2003). In our analyses, we will often consider African Americans separately from Caucasians. However, population substrata could still confound gene-trait associations within African Americans and Caucasians, particularly in African Americans (Cardon & Palmer, 2003). Therefore, it will be important to adjust for population stratification in genetic association analyses. If candidate gene studies are conducted, we will select ancestry informative markers in both ethnic groups and use the STRUCTURE (Pritchard & Rosenberg 1999) approach to infer the degree of population stratification as represented by the proportion of ancestry of each individual in the study. We will test the hypothesis of two or more strata (i.e., ethnic subpopulations) within each ethnic group, and the STRUCTURE program will attempt to classify individuals as belonging to one population or another. If there is evidence of population stratification, then the multivariate analyses described previously will be repeated with adjustment for population stratum membership (e.g., using stratified logistic regression analysis). For genome-wide association studies, a recent publication by Price et al. (2006) proposed a method that enables detection and correction of population stratification on a genome-wide scale using the idea of principal components analysis. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations. We will use this method and the EIGENSTRAT software provided by the authors for genome-wide association analysis.

10.5.3 Gene-By-Gene/Gene-By-Environment Interactions

Gene-environment interactions are measured by the effects of clinical/environmental exposures on the disease risk among individuals with different genotypes. Gene classification schemes can be added to the final model for clinical risk factors, and then systematic tests for interactions between gene classification and risk factors can be conducted. Logistic regression models can be utilized to explore three types of interactions. For notation, we define:

R = clinical risk factor(s);

E = exposure(s);

G1 = genotype/haplotype/single nucleotide polymorphism (SNP) 1;

G2 = genotype/haplotype /SNP 2.

Then,

Model 1: Clinical risk factor and gene interaction

$$\text{Logit}[\Pr(\text{outcome})] = b_0 + b_1 R + b_2 G1 + b_3 (R * G1)$$

Model 2: Environmental exposure and gene interaction

$$\text{Logit}[\Pr(\text{outcome})] = b_0 + b_1 E + b_2 G1 + b_3 (E * G1)$$

Model 3: Gene–Gene interaction

$$\text{Logit}[\Pr(\text{outcome})] = b_0 + b_1 G1 + b_2 G2 + b_3 (G1 * G2)$$

In the first model, the outcome variable, say a diagnosis of undescended testes (UDT) would be an indicator of status (0 = absent or 1 = present). A previously selected exposure of interest (e.g., phthalates) would be categorized as absent or present (E = 0 or 1 depending on whether the subject was exposed during gestation) or as a continuous variable (quantitative measure of phthalate exposure from biologic specimens or from an estimated latent variable), and a candidate genotype/haplotype (e.g., HOXA9) would be characterized as nonsusceptible/susceptible (G1 or G2 = 0, 1).

Estimation of the odds ratio for clinical risk factors among the susceptible genotype (G1 = 1) group is then $\exp(b_1 + b_3)$ and for the nonsusceptible group (G1 = 0) is $\exp(b_1)$. A measure of the strength of the interaction can be evaluated by the odds ratio (OR) for susceptible and nonsusceptible, which is expressed in this model as $\exp(b_3)$. A score test for the statistical significance of the interaction OR = 1 is then a test of $b_3 = 0$. The same approach would be taken to model phthalate exposure and gene interactions (model 2), and gene-gene interactions (model 3) (Hwang et al., 1995; Yang & Khoury, 1997; Andrieu & Goldstein, 1998) such as INSL3 and GREAT, both thought to control development of the gubernaculum.

Completion of the statistical analyses described here would allow not only for the assessment of associations between allelic variants in candidate genes but also for interactions between clinical factors and in utero environmental factors such as exposure to phthalates on the risk of the outcome of interest (e.g., nonsyndromic UDT).

Another area of study is the relationship between the environmental exposures and the patterns of somatic mutations in genes. To account for potential dependency of the mutation patterns along the genome, the generalized estimating equation approach for analyzing correlated binary data can be applied to identify how environmental exposures can potentially induce somatic mutations in cells. Clustering analysis methods can also be applied to cluster the mutation patterns based on multivariate binary data and to relate the mutation clusters to environmental exposures.

10.5.4 Regression Tree Approaches to the Analysis of Interactions Between Genes

Because interactions play a key role in the analysis of genomic data, including data involving SNPs, there are multiple approaches to this type of analysis. Since the number of nucleotides involved in susceptibility for a complex traits and conditions is often quite large, special methods are required for analyzing the even larger number of possible interactions of SNPs within and between genes. As an example, the risk of type 1 diabetes can be related to the interaction of multiple SNPs rather than to single variation sites. Analytic approaches such as the adaptive spline and tree-based methods such as MARS and CART (Friedman, 1991; Breiman et al., 1984) can be used to generate interpretable interaction rules among the SNPs. Realizing the limitations of these methods—for example, MARS is efficient on data that has interactions in at most a few variables, and CART only generates rules in disjunctive normal form—the recently developed adaptive regression method, logic regression, may be applied in order to construct predictors as Boolean combinations of the SNPs using simulated annealing.

Such Boolean combinations of the SNPs may not be detected by a standard regression tree as implemented in CART. In order to study and assess gene-by-gene interaction and gene-by-exposure interactions on the risk of developing certain outcomes, the recently developed tree-based method FlexTree (Huang et al., 2004) may be employed. FlexTree is an extension of the binary tree-structured approach such as CART and is particularly applicable to study gene-by-gene and gene-by-environment interactions. The methods work well for both the model where many genes are involved in the predisposition of certain outcomes and the model where only a small list of aberrant genotypes is predisposing.

The Bayesian variable selection approach introduces a latent binary vector to index all possible subsets of variables (George & McCulloch, 1993). A prior distribution is specified for this latent vector and the variable selection is performed based on the posterior model probabilities. When the number p of covariates is large, deriving the posterior probabilities of all 2^p possible models is computationally prohibitive. This can be handled via Markov Chain Monte Carlo (MCMC) stochastic search techniques, which are used to explore the space of variable subsets and search for promising models. At each MCMC iteration, a new candidate model is visited and retained based on its posterior probability relative to the previously visited model. This method is well suited for the analysis of high-dimensional data where the sample size is substantially smaller than the number of covariates $(n << p)$. Another advantage of the Bayesian approach is that it allows the uncertainty inherent in the model selection process to be incorporated in the inference mechanism. This is accomplished via model averaging where the estimation of parameters and the prediction of future outcomes are computed by averaging over a range of likely models.

10.5.4.1 Multifactor Dimensionality Reduction for the Analysis of Interactions Between Genes and Between Genes and the Environment

Multifactor dimensionality reduction is a method designed specifically for investigating multiple gene-gene and gene-environment interactions. In studies using related family members, the programs are currently only applicable for gene-gene but not gene-environment interactions. However, most of the children in the NCS will be unrelated and therefore this method will also be ideal for the investigation of gene-environment interactions. In logistic regression models, the number of possible interaction terms grows exponentially as each additional main effect is added (Ritchie et al., 2003). To reduce the dimensionality in data where interactions are the primary focus, one divides the data into training and test sets and then forms all possible permutations of the chosen interaction terms. The interactions become, in a sense, the "main effects" under investigation (Ritchie et al, 2003). An example would be two genes, each with a dominant and recessive homozygote and a heterozygote, making nine possible combinations. Each cell class is labeled as high or low risk according to a predefined threshold. The MDR model that has the fewest misclassified individuals is selected, and the process is then repeated using 10-fold cross-validation to evaluate predictive ability and reduce spurious significances. Since multiple gene-gene and gene-environment interactions are the rule rather than the exception in many complex health conditions and behaviors being studied in the NCS, this process will allow for more realistic modeling of multiple interactions.

10.5.4.2 Complementary Modeling Approaches

Interactions confirmed with any of the above-mentioned approaches can also be incorporated with other causative factors into path analytic models to identify multiple cross-sectional and/or longitudinal causal pathways (see more descriptive discussions in Section 10.4). An example would be the NCS hypothesis that gene-environment interactions between ozone exposure and polymorphisims of TLR-4 and/or TNF-α play a causal role in asthma onset. This could be tested with one of the methods described above and the significant interactions inserted into a path analysis model with other factors related to the outcome but not to these interactions (e.g., prenatal factors such as low birth rate).

10.5.5 Gene Expression Data-Microarray Analysis

In addition to these established methods, others developed for analyzing high-dimensional data such as microarray gene expression data are worth exploring. Gene microarrays are a method employed for examining the expression of as many as hundreds or thousands of genes in a single tissue (Jarvis, 2006). Although the NCS will not have tissue specimens, it will have whole blood. As the Study evolves, it may be possible to collect the gene expression profiles in whole blood samples over time in order to examine how those gene expression changes detectable in blood are related to development of various health outcomes or to identify potential biomarkers for diseases and investigate how gene expression is affected by environmental exposures. Such data make it possible to learn how expression of different genes and, hence, their coded proteins, interact to provide insight into biochemical pathways and causal mechanisms on a genomic level.

Of particular interest with respect to analysis of these data are the threshold gradient methods (Friedman & Popescu, 2004; Gui & Li, 2005) and the Bayesian variable selection methods (Sha et al., 2004) for identifying important SNPs, environmental exposures, and their interactions for the risk of developing certain health-related outcomes. Bayesian variable selection methods have been developed and used successfully for the analysis of DNA microarray data in the context of multigroup classification (Sha et al., 2004) and clustering (Tadesse, Sha, & Vannucci, 2005). In the former case, the goal is to identify subsets of genes that characterize the different classes and to predict the outcomes for future samples based on their expression profiles. In the latter, the groups from which the observations arose are not known and the goal is to uncover the cluster structure of the observations and identify the discriminating variables. However, methods for analyzing longitudinal gene expression data are still relatively limited (Guo et al., 2003; Tai & Speed, 2006). We propose that the NCS develop new methods in the framework of functional data analysis and empirical Bayes analysis and treat the gene expression profiles over time as curves or functional data. Preliminary analysis of methods based on functional data analysis indicate such methods result in more sensitive procedures for identifying genes that show different expression patterns over time (Hong & Li, 2006; Leng & Mueller, 2006). We propose that the NCS generalize many commonly used multivariate analysis methods such as canonical correlation and correspondence analysis to the functional data for exploratory analysis and for graphically displaying the data. We also propose that the NCS develop regression analysis methods with functional data as predictors to account for gene expression dynamics over time. Functional data analysis provides a natural framework for accounting for gene expression levels measured over time and can potentially consider the dynamic nature of gene expression over time. These methods will be developed and made available to interested researchers on the NCS website.

10.5.6 Family Data

For population-based genetic association studies of complex traits, one of the potential areas of confounding involves the latent population substructures in the study population, which can result in the observation of spurious associations if such substructures exist and are not appropriately accounted for. Family-based study designs such as parent-child trios provide an alternative design for genetic association analysis of complex traits (Spielman & Ewen, 1996). For population-based case-control or cohort designs, genomic controls by typing a set of ancestry informative markers can be employed for adjusting for such population substructures in regression analysis (see more on this above).

Although deviations from the Hardy-Weinberg equilibrium (e.g., existence of migration) can identify systematic errors in genotyping, it is important to carry out other checks. For case-control studies of genetic association, under particular models for genotyping error there is no increase in type I errors of tests for genotype-disease associations (Gordon et al., 2002). (See Section 10.2.3 for a discussion of type I and II errors.) However, if general tests which ignore genotyping error are invalid, one solution is to integrate a realistic error model into association analysis for SNPs. A number of models for measurement error have been proposed and are described by Gordon et al. (2002) along with descriptions of how to appropriately test for association. Similarly, measurement errors in environmental exposures will be accounted for in the context of regression analysis with measurement errors (Ruppert et al., 2003, Chapter 15). Details of dealing with measurement error are provided in Section 10.3.

Genomic imprinting can be loosely defined as the gamete-of-origin dependent modification of phenotype. In other words, the phenotype elicited from a locus is differentially modified by the sex of the parent contributing that particular allele. This process results in a reversible gamete-of-origin specific marking of the genome that ultimately produces a functional difference between the genetic information contributed by each parent. In humans, the term genomic imprinting is usually described as mono-allelic gene expression or the inactivation of either the maternal or paternal allele of a particular locus. One important issue in the context of the NCS is to differentiate between fetal and maternal genotypic effects, which can be tested using the transmission test for linkage disequilibrium (Mitchell, 1997). While both parent-child triads and grandparent-grandchild designs can be used for testing maternal or parent of origin effect, the grandparent-grandchild design may in some situations provide higher power than the parent-child design (Weinberg et al., 1998; Wilcox et al., 1998). However, given the size of the NCS cohort, power should not be a problem (see Section 10.2.3, Power for Subgroups). In the framework of the log linear models as developed in Weinberg et al. (1998), one can also incorporate the environmental covariates into analysis of parents of origin and maternal-mediated genetic effects.

The parents-affected child trio design provides an alternative family-based design for studying associations between candidate genes and the risk of developing diseases or for studying gene-by-environment interactions. Designs based on the genotyping of affected individuals and their parents allow the detection of markers in linkage disequilibrium with disease genes (Spielman & Ewen, 1996). The main advantage of such a design as compared to population non-family cohort design is that it is free from the issue of spurious association caused by potential underlying population substructures. Such designs can be used for confirming the associations found in the standard population-based designs (non-family). Environmental covariates and gene-exposure interactions can be easily taken into account in the analysis (Li & Fan, 2000; Shih & Whittemore, 2002).

10.5.7 Multifactor Dimensionality Reduction and Issues of Multiple Comparisons

One important consideration with respect to genome-wide association studies is the issue of multiple comparisons. This issue usually arises in the context of hypothesis testing; however, even in exploratory studies without a formal hypothesis, there is generally an implicit hypothesis that a given discovered effect is zero.

When conducting multiple hypothesis tests, the type I error rate of 0.05 will hold for each individual test, but the overall probability of making at least one type I error is greatly increased. The most common procedure for protecting against this is to require a stringent "family-wise" error rate adjusted for the number of tests being conducted.

Although the family-wise error rate procedure is popular and performs well in genome-wide linkage analysis, it is too stringent for evaluating multiple loci, which may result in very low power. The false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) provides a new notion of global error for multiple testing procedures. The idea of FDR is to use the expected proportion of false rejections of the null hypothesis among the total number of rejections as the measure of global error. Such a procedure leads to a global cutoff value that is adaptive to the data set (Sabatti et al., 2003). The FDR procedure will identify a lower cutoff level than the universal Bonferroni cutoff if a higher percentage of the null hypotheses tested are truly false. Such a procedure is most effective for the identification of loci with secondary effects. On the other hand, if all the null hypotheses are true (none of the analyzed markers is associated with the disease), controlling FDR is equivalent to controlling family-wise error rates. Although the original procedure by Benjamini and Hochberg was developed for independent tests and *p*-values, recent studies and extensions have indicated the procedure also works well when the tests are not independent, as might be expected in genome-wide association tests (Sabatti et al., 2003; Fernando et al., 2004).

Because health effects are commonly measured by multiple outcomes, the main tools we propose to apply or develop for the data to be collected are the regression models for multiple outcomes (Sammel et al., 1997; Sammel et al., 1999; Geys et al., 1999). We also propose to apply state-of-the-art models such as errors-in-variable models, missing-data methods, smoothing and methods for correlated data, such as longitudinal and spatial data analysis, to assess the health effects of dose, concentration, and duration of exposure. Semiparametric regression and generalized estimation equation methods can be developed for modeling the data and estimating the parameters in order to make fewer assumptions on the underlying models.

Another important application of multifactor dimensionality reduction is the investigation of gene-by-gene interactions. In logistic regression models, the number of possible interaction terms grows exponentially as each additional main effect is added (Ritchie et al., 2003). To reduce the dimensionality in data where interactions are the primary focus, one divides the data into training and test sets and then forms all possible permutations of the interaction terms. So for two genes, each with a dominant and recessive homozygote and a heterozygote, there are nine possible combinations. These combinations or interactions are treated as the "main effects" (Ritchie et al., 2003). Each cell class is labeled as high or low risk according to a predefined threshold. The MDR model that has the fewest misclassified individuals is selected and the process is repeated using 10-fold cross-validation to evaluate predictive ability and reduce spurious significances.

10.5.8 Twin Studies

The NCS cohort is estimated to eventually contain 3,000 twins. These twins will provide a unique opportunity to examine gene-by-environment interactions, especially with respect to complex diseases. Whereas monozygotic twins have identical genotypes, dizygotic twins have the same genetic variation as non-twin siblings. When monozygotic twins have one affected and one nonaffected twin, the ability to investigate gene-by-environment interactions with respect to specific SNPs, haplotypes and environmental factors will greatly facilitate the understanding of causal pathways.