Supporting Statement B

The Sister Study PHASE 2:

Environmental and Genetic Risk Factors for Breast Cancer

(NIH/NIEHS)

Submitted: 20 August 2009

Principal Investigator and Co-Project Officer:

Dale P Sandler PhD
Chief, Epidemiology Branch
National Institute of Environmental Health Sciences
PO Box 12233
Research Triangle Park NC 27709
Phone: 919-541-4668
Fax: 919-541-2511
Email: sandler@niehs.nih.gov

Co-Principal Investigator:

Clarice Weinberg PhD
Chief, Biostatistics Branch
PO Box 12233
Research Triangle Park NC 27709
Phone: 919-541-4927
Fax: 919-541-4311
Email: weinberg@niehs.nih.gov

Project Officer:

Paula S Juras PhD
Epidemiology Branch
National Institute of Environmental Health Sciences
PO Box 12233
Research Triangle Park NC 27709
Phone: 919-541-4668
Fax: 919-541-2511
Email: juras@niehs.nih.gov

Collections of Information Employing Statistical Methods

B.1. Respondent Universe and Sampling Methods

A total of 214,640 new cases of female breast cancer were expected in the US in 2006 according to SEER estimates. In 1990 there were approximately 140,000 new cases, and in 1980, approximately 100,000. On average, since 1980, there were 150,000 new cases a year for a total of 3,750,000 women diagnosed with breast cancer over those 25 years. Based on data from three large population-based breast cancer studies [personal communications from investigators with the Women's Contraceptive and Reproductive Experiences Study (Bernstein), Carolina Breast Cancer Study (Newman), Long Island Breast Cancer Study Project (Gammon)] we estimated that 2/3 of these breast cancer cases have at least one living sister. Although the estimate of 2/3 with a sister seems high, all three studies were remarkably consistent. Our goal of 50,000 sisters thus represented just 2% of possible sisters. Even if we attracted women whose sisters were diagnosed only since 1990, with more than 15 years of cases (by the time recruitment was completed), we would need to enroll a little over 3% of the available sisters. Thus, enrolling a cohort of 50,000 sisters was feasible from a numbers standpoint. No sampling methods were used; all women in the target population of women, aged 35-74 without breast cancer, who have a sister that has been diagnosed with breast cancer, either living or dead, were eligible.

We used SEER age-specific incidence rates for the years 1993-1997 and an estimate of the population from the 2000 census to estimate the average number of female breast cancer cases by age group per year. Using the age distribution of the expected cases between ages 35-74 (assuming that the sisters with and without cancer would be, on average, the same age), we estimated the expected age distribution of sisters who would enroll in the cohort, assuming that women in each age group were equally likely to enroll (see table).

Age distribution of incident breast cancer cases

Age group	Female population (millions)	Rate per 100,000	Cases	Age distribution	Predicted cohort
35-39	11.4	58.4	6,658	0.04472	2,236
40-44	11.3	116.1	13,119	0.08812	4,406
45-49	10.2	198.5	20,247	0.13600	6,800
50-54	9.0	263.7	23,733	0.15942	7,971
55-59	7.0	305.0	21,350	0.14341	7,170
60-64	5.7	353.6	200,155	0.13538	6,769
65-69	5.1	402.7	20,538	0.13795	6,898
70-74	5.0	461.5	23,075	0.15500	7,750
			148,875	1.00000	50,000

Applying age-specific rates to the predicted number of women in each age group in the cohort, we expected a total of 150,287 cases per year under the assumption of no excess risk. If sisters are truly at 2-fold risk, there will be approximately 300 incident breast cancers per year, or 1,500 over the first 5 years.

Then, applying current age specific incidence rates, we estimated that there will be 150 cases per year diagnosed among members of the cohort if their rates are similar to those in the population as a whole. But, assuming a 2-fold risk for sisters based on studies reported in the literature, we would expect 300 cases per year for a total of 1500 cases after five years of follow-up. This estimate did not take into account the increasing risks as women in the cohort pass through one age/risk group to the next, or the possibility that incidence rates may continue to increase. On the other hand, no allowance was made for the possibility that the sisters who enroll would be disproportionately younger since we monitored recruitment by age and made special efforts to enroll older sisters. While this could have led to fewer cases being diagnosed among a younger cohort, it was also likely that these younger sisters would be at even greater than 2-fold risk by virtue of being the sister of someone diagnosed at an early age. Analysis of data from the first 10,000 participants suggested a higher than expected percentage of women with a sister diagnosed before age 45. Thus the power to detect genetic effects and gene-environment interactions may be even greater than expected.

The study of gene-environment interactions requires large sample sizes. The cohort size will be large enough to test many but not all hypotheses regarding such interactions. In many instances, analyses will require assessing gene status among the full 1,500 cases expected to develop after 5 years of follow-up or waiting even longer as additional cases accrue. The power of the study will depend on the frequency of

the polymorphism and the exposure as well as on the number of cases that accrue. In all cases, power will be greater than in a similarly sized cohort from the general population.

Power will be sufficient for testing most main gene or environment effects of interest, often using smaller subsets of the cases that develop. For example, for alleles that occur in 40% of the population, with a Type I error of 5%, we will have 80% power to detect an odds ratio of 2.0 with approximately 130 cases and an equal number of controls (see table). With 450 cases, we can detect an OR of about 4.0 (80% power, 5% Type I error) for a mutation in a cancer gene that affects 1% of the population.

When studying an interaction between two relatively rare factors, one achieves the best power by weighting the sampling toward people who have the factors under study. Thus, the sampling of sisters provides a benefit, not just by increasing the number of cases to be accrued, but precisely because it over samples for genetic factors.

Approximate number of cases needed to detect odds ratios of 1.5, 2.0 and 3.0 with 80% power. Type I error = 5% and an equal number of cases and controls.

Gene frequency (%)	Odds Ratio		
	1.5	**2.0**	**3.0**
1		2400	800
5	1650	550	200
10	950	400	180
20	550	200	80
30	500	160	75
40	425	130	50

Presumably the most powerful design for studying gene-environment interactions would over-sample people likely to be carrying genetic risk factors (as in the sister design) and would simultaneously over-sample women in high-risk areas where there might be more exposure to some important environmental co-factor. Thus, we concentrated efforts to recruit in areas where women were more likely to have exposure to environmental factors that may relate to risk. It will also be possible to over-sample for rare exposures in choosing controls for the nested case-control studies.

B.2. Procedures for the Collection of Information

This is a non-probability sample and represents a subset of the population whose risk is relatively high. The women who volunteered are more interested, more informed, more concerned, and highly motivated to follow through with study requirements, thus minimizing dropout rates. Follow-up data is collected via telephone interview and self-completed written or web-based forms

The analysis plan includes a nested comparison of sisters who do and do not develop breast cancer during the course of follow-up. Using the questionnaire data and biological and environmental samples, we will assess the separate and combined effects of exposures and genes. Ancillary studies will include exploring the etiology of other diseases (e.g. asthma, uterine fibroids, diabetes, thyroid disease, osteoporosis, rheumatoid arthritis and other autoimmune diseases, neurodegenerative diseases, and other cancers) and studying genetic and environmental effects on prognosis and prevention strategies.

B.3. Methods to Maximize Response Rates and Deal with Nonresponse

Since this is a volunteer cohort of motivated women we expect participation to remain quite high throughout the follow-up. Over 95% of participants have completed annual update forms with a protocol that included only minimal attempts to contact the women. Based on this experience, we expect more than 90% response rates for annual and bi/triennial updates. These response rates are comparable to those achieved in other highly motivated cohorts such as the Nurses Health Study. Such high response rates in the Nurses Health Study and the Black Women's cohort are achieved only after as many as a dozen or more questionnaire mailings to participants.

The CATI interviews are scheduled at the convenience of the participant. Participants are sent a reminder about the appointment for the interview and the importance of completing the other requirements of the study. Non-responders are sent follow-up reminders by mail, and are subsequently contacted by phone to determine whether or not they wish to continue their participation. All study activities and correspondence is available in Spanish.

B.4. Test of Procedures or Methods to be Undertaken

Meetings with breast cancer patients and different groups of sisters of breast cancer patients were held during 1999-2000 to determine acceptability of study and recruitment methods. The overwhelming response was not only that the study was vitally important, but also that women were eager to know more about it, and eager to convey the information to their sisters or others who would be eligible. Their feedback helped us design our screening methods and recruitment strategies. Interestingly many women noted that although it might be too late to help themselves, they would participate in the study in the hope that it would provide information that might prevent breast cancer in their daughters!

All procedures and the questionnaires underwent internal testing prior to implementation. Finally, the information gleaned from each follow-up activity allows further refinement of all study materials and procedures. Forms were shortened and modified to streamline data collection, thus reducing the burden on participants.

B.5. Individuals Consulted on Statistical Aspects and Individuals Collecting and/or Analyzing Data

Dr. Clarice Weinberg (919-541-4927). Chief, Biostatistics Branch, NIEHS, a Co-Investigator on this study, developed the statistical approach for the study in conjunction with Dr. Sandler. Data is collected and managed by SSS, with Ms. Deborah Bittner (919-287-4320) as the Project Director —1009 Slater Road, Suite 120, Durham, NC 27703. Data will be analyzed by Drs. Dale Sandler, Jane Hoppin, Stephanie London, and Jack Taylor, Epidemiology Branch, NIEHS; and Dr. Weinberg.