Research article

Open Access

Study design: Evaluating gene–environment interactions in the etiology of breast cancer – the WECARE study

Jonine L Bernstein[1], Bryan Langholz[2], Robert W Haile[2], Leslie Bernstein[2], Duncan C Thomas[2], Marilyn Stovall[3], Kathleen E Malone[4], Charles F Lynch[5], Jørgen H Olsen[6], Hoda Anton-Culver[7], Roy E Shore[8], John D Boice Jr[9], Gertrud S Berkowitz[1], Richard A Gatti[10], Susan L Teitelbaum[1], Susan A Smith[3], Barry S Rosenstein[11], Anne-Lise Børresen-Dale[12], Patrick Concannon[13] and W Douglas Thompson[14]

[1]Department of Community and Preventive Medicine, Mount Sinai School of Medicine, New York, NY, USA
[2]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA
[3]Department of Radiation Physics, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
[4]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[5]Department of Epidemiology, University of Iowa, Iowa City, IA, USA
[6]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark
[7]Department of Medicine, University of California, Irvine, Irvine, CA, USA
[8]Nelson Institute of Environmental Medicine, New York University Medical Center, New York, NY, USA
[9]International Epidemiology Institute, Rockville, MD, USA, and Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[10]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[11]Department of Radiation Oncology, Mount Sinai School of Medicine, New York, NY, USA
[12]Department of Genetics, Norwegian Radium Hospital, Oslo, Norway
[13]Molecular Genetics Program, Benaroya Research Institute at Virginia Mason, Seattle, WA, USA
[14]Department of Applied Medical Sciences, University of Southern Maine, Portland, ME, USA

Corresponding author: Jonine L Bernstein (e-mail: jonine.bernstein@mssm.edu)

Received: 28 Jul 2003 Revisions requested: 26 Sep 2003 Revisions received: 15 Jan 2004 Accepted: 30 Jan 2004 Published: 9 Mar 2004

Breast Cancer Res 2004, **6**:R199-R214 (DOI 10.1186/bcr771)

Abstract

Introduction: Deficiencies in cellular responses to DNA damage can predispose to cancer. Ionizing radiation can cause cluster damage and double-strand breaks (DSBs) that pose problems for cellular repair processes. Three genes (*ATM*, *BRCA1*, and *BRCA2*) encode products that are essential for the normal cellular response to DSBs, but predispose to breast cancer when mutated.

Design: To examine the joint roles of radiation exposure and genetic susceptibility in the etiology of breast cancer, we designed a case-control study nested within five population-based cancer registries. We hypothesized that a woman carrying a mutant allele in one of these genes is more susceptible to radiation-induced breast cancer than is a non-carrier. In our study, 700 women with asynchronous bilateral breast cancer were individually matched to 1400 controls with unilateral breast cancer on date and age at diagnosis of the first breast cancer, race, and registry region, and counter-matched on radiation therapy. Each triplet comprised two women who received radiation therapy and one woman who did not. Radiation absorbed dose to the contralateral breast after initial treatment was estimated with a comprehensive dose reconstruction approach that included experimental measurements in anthropomorphic and water phantoms applying patient treatment parameters. Blood samples were collected from all participants for genetic analyses.

Conclusions: Our study design improves the potential for detecting gene–environment interactions for diseases when both gene mutations and the environmental exposures of interest are rare in the general population. This is particularly applicable to the study of bilateral breast cancer because both radiation dose and genetic susceptibility have important etiologic roles, possibly by interactive mechanisms. By using counter-matching, we optimized the informativeness of the collected dosimetry data by increasing the variability of radiation dose within the case–control sets and enhanced our ability to detect radiation–genotype interactions.

Keywords: bilateral breast cancer, breast cancer susceptibility genes, counter-matching, gene–environment interactions, radiation dosimetry, study design

CB = contralateral breast; CM = counter-matched; DSBs = double-strand breaks; NCC = nested case control; PN = predictive negative; PP = predictive positive; RR = relative risk; RRT = registry radiation treatment indicator; RT = radiation therapy; WECARE = Women's Environmental Cancer and Radiation Epidemiology.

Introduction

The WECARE (for Women's Environmental Cancer and Radiation Epidemiology) Study is a multi-center, population-based case-control study of breast cancer designed to examine the interaction of gene carrier status and radiation exposure in the etiology of breast cancer. We are currently focusing on three major breast cancer susceptibility genes: *ATM*, *BRCA1*, and *BRCA2*. Our underlying hypothesis is that a woman who carries a mutant variant of one of these three genes is more susceptible to radiation-induced breast cancer than a woman who is not a carrier. In addition, the repository of specimens and data that are currently being assembled make it possible to evaluate at a later date a broad range of other genes that lie in related cellular response pathways. This paper addresses the theoretical and practical issues we faced in designing this study to improve cost efficiency and feasibility. Specifically, our study power was enhanced by selecting a genetically enriched study population predisposed to breast cancer, by using a counter-matched design, and by developing a dose reconstruction method to reduce the error in estimating individual radiation exposure derived from imperfect and old medical records. These considerations, coupled with our current understanding of radiogenic breast cancer and genetic susceptibility, provide the basis for the multi-disciplinary WECARE Study, encompassing epidemiologic risk factor information, genetic analyses of large complex genes, and estimation of radiation-absorbed dose to breast tissue.

Background on DNA damage response genes, ionizing radiation, and breast cancer susceptibility

So far, all of the genes known to be associated with increased susceptibility to breast cancer function within DNA damage response pathways. ATM, the product of the gene mutated in the autosomal recessive disorder ataxia–telangiectasia, has a critical function in signaling the presence of DNA double-strand breaks (DSBs) that are induced by ionizing radiation, radiomimetic chemicals, and developmental DNA rearrangement events [1]. Upon activation by ionizing radiation or other DSB-inducing agents, ATM phosphorylates many downstream targets that control pathways whose activation can result in DNA repair, cell cycle checkpoint control, and apoptosis [2]. These targets include the products of several genes implicated in breast cancer susceptibility such as *BRCA1* [3,4] and *CHEK2* [5–7]. There has been evidence from epidemiologic studies of ataxia–telangiectasia families [8,9], mutation screening of *ATM* in breast cancer cohorts [10–12], and animal models [13] that carrier status for at least a subset of *ATM* mutations is also associated with an increased risk for breast cancer. However, the mechanism mediating this increased risk and the potential involvement of radiation exposure have yet to be elucidated.

Ionizing radiation can cause DNA damage that may, in sufficient dosage, result in mutations in oncogenes and tumor suppressor genes [14,15]. It is a known breast carcinogen, especially when exposure occurs before age 40 years when excess risks range from 5.5 to 10.7 cases per 10^4 woman-years/Gy [16,17]. Studies with long-term follow-up, examining different radiation doses and types of exposure, have shown that the risk for developing radiogenic breast cancer is similar for single and fractionated exposures, provided that they represent equal total radiation doses [18–20]. The increased risk is directly proportional to increasing radiation dose, inversely related to age at irradiation, and remains elevated throughout life [21]. For example, among women exposed to radiation between the ages of 10 and 39 years, increased relative risks (RRs) are first seen about 10–15 years after exposure and do not diminish for 35–50 years [21,22]. Very high therapeutic doses received by young women during the treatment of Hodgkin's disease have recently been shown to increase breast cancer risk in a dose-dependent manner [23]. In our study, we focus on the radiation dose received to the contralateral breast (CB) during breast treatment, which ranges from about 1.0 to 7.1 Gy [24–27]. Because many irradiated cells in the CB survive, radiation-induced DNA damage may increase the probability that a tumor may develop in this breast. Findings from the most comprehensive study to date showed that exposure to such radiation to the CB from radiation therapy (RT) increased the risk of developing second primary breast cancer among young women (RR = 1.59, 95% confidence interval 1.07–2.36) and among those who survived at least 10 years after RT (RR = 1.85, 95% confidence interval 1.15 to 2.97) [27]. The goal of the WECARE Study is to examine whether subgroups of women with breast cancer are at heightened risk for developing a second primary breast cancer after RT because of a genetic susceptibility, namely a mutation in *ATM*, *BRCA1*, or *BRCA2*.

Second primary breast cancer – a useful context in which to study gene–environment and gene–gene interactions

For susceptibility genes, such as *BRCA1* and *BRCA2*, the low prevalence of mutation carriers and of selected environmental risk factors limits the informativeness of studies conducted with a random sample obtained from the general population (for example, traditional case-control designs). The WECARE Study was based on the premise that by restricting consideration to women with a first primary breast cancer and then studying the determinants of developing a second primary breast cancer, the power to detect main effects of relatively rare genetic mutations and their interactions with environmental factors is considerably enhanced. This enhancement results because any genetic abnormality that is important in the etiology of breast cancer would be

considerably more prevalent in women who have had a first breast cancer than in the general population [28,29] and genetic factors have a greater role in early-onset cancers and multiple primary cancers. Consequently, the etiologic role for a genetic factor, including its possible interactive effect with environmental factors and other genes, should be more evident [29,30]. Further, studies of other tumor suppressor genes have suggested that persons with mutations have a heightened risk after RT exposure than those with wild-type alleles [31–35]. Lastly, the high prevalence of substantial and quantifiable exposure to ionizing radiation among women with a history of first primary breast cancer makes this group particularly worthy of study for understanding the etiology of radiogenic cancers, something that is virtually impossible with a conventional case-control study design for which the population prevalence of the exposure is so low. Demonstration of an interaction between a gene, or genes, such as *BRCA1* and/or *BRCA2*, and an environmental exposure, such as radiation, will constitute an important scientific and practical finding that will help to elucidate the mechanism through which a genetic characteristic can affect the risk for breast cancer. In turn, this information will contribute to informed decisions on the prudent use of RT in subgroups of the female population.

Design

Sample and rationale

In the WECARE Study of gene–environment interactions of second primary breast cancer, 700 women with asynchronous bilateral breast cancer serve as cases, and 1400 women with unilateral breast cancer who were diagnosed with their cancer at the time that the cases were diagnosed with their first primary, and at the same age, serve as controls. Eligible cases and controls must have survived at least 1 year after their initial diagnosis of breast cancer. All study subjects were identified, recruited, and interviewed through five population-based cancer registries, four in the USA (three Surveillance, Epidemiology and End Results registries, namely Iowa, Los Angeles County [CA], and Seattle [WA], and the Cancer Surveillance Program of Orange County [CA]) and one in Denmark (the Danish Cancer Registry). These registries provided access to an ethnically, racially, and geographically diverse group of women representative of breast cancer cases from the reporting areas.

Women were eligible as cases if they met the following criteria: (1) they were diagnosed between 1 January 1985 and 31 December 2000 with a first primary invasive breast cancer that did not spread beyond the regional lymph nodes at diagnosis and a second primary *in situ* or invasive contralateral breast cancer diagnosed at least 1 year after the first breast cancer diagnosis; (2) they resided in the same study reporting area for both diagnoses; (3) they had no previous or intervening cancer diagnosis; (4) they were under age 55 years at the time of diagnosis of the first primary breast cancer; and (5) they were alive at the time of contact, able to provide informed consent to complete the interview and could give a blood sample.

Our inclusion criteria were designed to create a high-quality database and biorepository of young women with breast cancer to address our specific aims and to facilitate future gene–gene and gene–environment studies. We included both incident and prevalent cases to improve our statistical power and to minimize the number of institutions required. Only young women with breast cancer were included because they were more likely to be genetically predisposed to breast cancer and to have had a greater susceptibility to radiation-induced cancer; effects of radiation on breast cancer decrease with increasing age at exposure and are low after age 45 years. Although ideally we would have therefore only included women diagnosed with first primary breast cancer at a very early age, the logistics of accruing 700 such young women with bilateral breast cancer were not feasible; this restriction would have required additional data collection sites and would also have precluded our competing interest to make the study results applicable to a broader population. We selected 1985 as the earliest date for diagnosis of the first primary to provide an adequate range of follow-up while still ensuring that most treatment records would be available. Our rationale for only including women who were living was the high quality and quantity of DNA that would be available from the blood (as opposed to that from tumor tissue only, if available at all, for deceased women) for the laboratory work as well as the high quality of the self-reported risk factor information (as opposed to the use of surrogate reports). To the extent that the genes of interest are related to survival, this design, by including both living and prevalent cases of breast cancer (both 'cases' and 'controls') in our studies, might have incurred some survival bias. However, because the subjects were drawn from existing cancer registries, with data (for example, age and stage at diagnosis, radiation therapy, race, and histology) on all cancer cases regardless of whether or not they were WECARE Study participants, we will have some ability to look at the potential impact of survival bias on our conclusions. Even in the absence of information on genotype, compared with most case-control studies for which data are available only on those that participate, the availability of these data on everyone is an important advantage. Because the registries included systematic follow-up and comprehensive information on cancers only for residents residing in their reporting areas, we imposed residency requirements for the time at diagnosis of the first and second primaries. We excluded women with a history of previous or intervening cancer diagnosis (other than breast cancer) because the

treatment that they may have received for the previous or intervening cancer adds analytic complexity and may jeopardize comparability between cases and controls. The first primary breast cancer must have been invasive; however, the second primary can have been an *in situ* carcinoma, because women with a history of breast cancer are more likely to be closely monitored for a second primary cancer. We excluded women with disease beyond the regional lymph nodes at diagnosis because any prevalent long-term survivors would be unique in terms of their survival, making it difficult to find suitably matched controls; further, this restriction reduced the unlikely chance that a contralateral breast cancer was a metastasis from the first primary. In conducting this study, we relied on the Surveillance, Epidemiology and End Results definition of second primary breast cancer and did not initiate an independent review of the pathology or medical records before assigning case status.

Controls in the WECARE Study had unilateral breast cancer and were representative of the population at risk for developing a second primary breast cancer; eligible controls were individually matched to cases on year of birth (5-year strata), year of diagnosis (4-year strata), registry region, and race. They must have survived without any subsequent diagnosis of cancer during the interval that elapsed between the matched case's first and second breast cancer diagnoses. At the end of this interval, the control must have had an intact CB and must have been living in the registry region. Although theoretically each case could also have served as a control for intervals less than that which elapsed between her first and second diagnoses, we chose to exclude cases from the control pool because this simplified the process of identifying and recruiting cases and controls and introduced little selection bias when the pool was large so that the selected cases represented only a small percentage of the pool [36]. The registry data on breast cancer diagnoses occurring between 1985 and 2001 were reviewed to identify all women who seemed to meet the eligibility criteria for either case or control status, and the 700 cases and 1400 controls needed to conduct the WECARE Study were recruited from this group of women.

Recruitment efforts began with information contained in registry files for cases in both the US and in Denmark. In the US, if a woman had moved, in order to trace, verify vital status, and update contact information, we linked to publicly available data from other external sources, including national change of address files and mortality files. Because some of the cases were diagnosed with their first primary nearly 15 years ago, tracing these women was challenging and time consuming. In Denmark, these tracing efforts were minimized due to their centralized registry requirements and readily available contact information.

In the US, eligibility was verified by the data collection centers through administration of a brief screening questionnaire during the initial phone contact (after physician permission and the receipt of a letter of introduction to the study). In Denmark, medical records are maintained centrally by the Danish Cancer Society, which made it possible to determine eligibility through an initial review of the medical records followed by a letter of introduction. In the US no such centralized source of medical records exists and abstracting in the various treatment centers could not be undertaken until informed consent had been obtained, which sometimes resulted in substantial additional work to trace and then contact women who were later determined to be ineligible to participate.

Counter-matching and selection of cases and controls

In this study, matching was used to control for confounding due to temporal variables and race, and a technique known as 'counter-matching' was used to improve the statistical efficiency of the design further [37]. Here we describe elements of the counter-matching as implemented in the WECARE Study. A quantitative treatment of its effect on statistical power is given later in the text and a discussion of counter-matching, including details about the implementation, is provided in the Appendix. Because a binary indicator of whether or not a woman received RT as treatment for her first breast cancer was available from the cancer registries at the time of selection of controls, we incorporated counter-matching on the registry radiation treatment indicator (RRT) (according to the registry records). In our counter-matched design, each case and two matched controls form a triplet, in which two members of each triplet are RRT+ (RRT exposed) and one member is RRT– (RRT unexposed).

Practically speaking, beyond the usual process of producing individually matched recruitment lists, only one additional step was required for counter-matching, namely stratification by RRT status of the subpopulation of controls that falls into the same individual matching stratum of a particular case: (1) if the case being matched was RRT+, a set of potential controls was randomly selected from each RRT stratum and one control was recruited from each set; or (2) if the case being matched was RRT–, a set of potential controls was randomly selected from the RRT+ stratum only and two controls were recruited from this single list. This counter-matched design resulted in higher variability in radiation dose than random sampling and enhanced the efficiency of the study for estimating the dose-response effect in exposed women as well as gene–environment interactions involving exposure to radiation. In particular, the probability that all members of a counter-matched set have the same true RT status is much smaller than for random sampling.

Obviously, the tumor-registry-based RRT information available at the time of selection of cases and controls was not as detailed or as accurate as the information that is available through the extensive data collection and dosimetry that were integral parts of the study. However, for the purposes of counter-matching it is sufficient to have only a surrogate for the information on exposure that will actually be incorporated into the analysis. Gains in efficiency are thus still realized, although these gains will not be so large as they would be if error-free information, or true RT exposure information, were available at the time of selection of the controls [37,38].

Collection of data from interview and medical records and blood collection and processing

All women were interviewed with the same pretested, scripted telephone questionnaire that emphasizes the ascertainment of events occurring during the 'at risk' period (time between first and second primary diagnosis, or its equivalent), including known and suspected risk factors for breast cancer that also antedated the 'at risk' period (such as detailed family, reproductive, and medical history), health care use, and treatment. A study phlebotomist was sent to each woman's home to draw the blood sample. These blood samples were in turn sent for DNA extraction and processing, lymphocyte cryopreservation, genotyping, and, ultimately, storage in the WECARE Study biorepository. Initially, we considered collecting buccal cells in the event of a blood refusal, but because the DNA yield would not be sufficient for the mutational screening of all three genes, this approach was not implemented. Medical records, pathology reports, and hospital charts were used to collect detailed treatment (chemotherapy, hormonal therapy, RT) and tumor characteristics (including location in the breast, stage at diagnosis, estrogen and progesterone receptor status, and histology).

A great deal of care has been taken to maintain strict confidentiality of all information; only the data collection site that contacts an individual knows her identity. Personal identifiers have never been included in any electronic or hardcopy transmission of data; archived data and genetic information have been identified only by study identification number. These identification numbers were randomly generated for the WECARE Study and incorporated no identifying information, including that related to the case/control status of the individual, triplet membership, center, or date of enrollment. Each data collection center received Institutional Review Board approval to enroll women in the study and to screen their blood samples for *BRCA* status as well as unknown genetic mutations. Protocols and consent procedures were recently modified for compliance with guidelines established by the Health Insurance Portability and Accountability Act. This new legislation may have a serious negative impact on future research that uses the WECARE repository. For example, we face having to anonymize all data, which would preclude continued follow-up for survival as an endpoint.

Quantification of radiation dose

The aim of radiation dosimetry was to produce individual dose estimates to the CB that are suitable for analysis. Absorbed dose to the CB is a combination of radiation directly from the primary beam, scatter off the collimators and filters, and leakage through the head of the machine. The CB dose depends on several factors, including field configuration, tumor dose, use of wedge filter or beam blocking, and radiation energy. For most treatments the range of dose across the CB varies by a factor of 10 or more from the medial to lateral aspect.

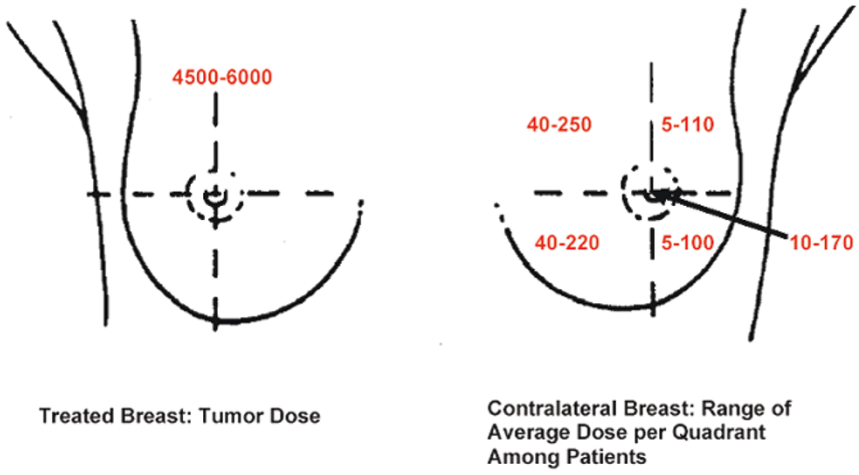
Radiation dose estimates were based on all radiation therapy during the 'at risk' period, including therapy for primary breast cancer, metastases, recurrences, and benign conditions. For each patient, we collected all available information on RT and the location of the breast tumor to provide specific dosimetry for the risk estimates. Sources of RT information included the following: complete basic RT record, summary RT notes, medical record notes and histories; operative reports (for brachytherapy); and physician correspondence. A score was assigned to each patient to indicate the level of information received (1, complete record; 2, partial record; 3, notes or summary only; 4, registry/interview data only).

Individual dose estimates were derived from measurement in tissue-equivalent phantoms. We molded Aquaplast [39] shells of both breasts using women in treatment positions. Shells were made for women with small, medium, and large breasts to test the influence of patient size. The shells were then filled with layers of tissue equivalent material (Superflab), which held thermoluminescent dosimeters. A total of 300–400 dosimeters were used in the CB. The response of the dosimeters was calibrated using an ionization chamber with a calibration traceable to the National Institute of Standards and Technology.

The most common treatment technique was medial and lateral tangential fields with or without a boost field to the tumor bed. Dose to the CB was sometimes modified by use of wedge filters (15, 30, 45, or 60 degrees). For example, a 30 degree wedge results in a dose to the CB about 25% higher than that with a 15 degree wedge. Some patients were also treated with peripheral lymphatic fields, including supraclavicular, axillary, and internal mammary. Figure 1 presents the dose (cGy) to the CB observed among the initial 360 study subjects.

Staff at MDACC were blinded as to the case/control status of subjects; therefore, dosimetry data was reported for all quadrants and the areola for all patients. For each

Figure 1



Range of quadrant doses (cGy) expected in the WECARE Study. Doses were calculated from phantom-derived measurements. Range limits were chosen to correspond to the techniques being used among the WECARE subjects that would result in the lowest and highest doses to each quadrant and the nipple area of the contralateral breast.

case/control set, the coordinating center selected for analysis the region of the breast that contained the CB tumor in the case.

Approximately 7% of the patient records received to date were incomplete in some aspect and 2% of patients had all documentation missing. Missing information was imputed from similar patients treatment at the same hospital. A score reflected which patients had imputed doses. Uncertainties in the CB doses resulted from unknown treatment information that could result in errors larger than 20%, although most uncertainties produced errors of 10% or less.

Mutation carrier status determination

For all three of the current target genes (*ATM*, *BRCA1*, and *BRCA2*), we used the same two-staged approach for determining carrier status. First, all coding regions and flanking intronic regions were screened for mutations or polymorphic variants by denaturing high-performance liquid chromatography (DHPLC), a mutation-screening technique that resolves homoduplexes and heteroduplexes of DNA formed by variant and wild-type sequences by high-performance liquid chromatography. This technique purports a high degree of sensitivity and ease of execution due to the lack of any need for significant sample manipulation after amplification. In the second stage, all variant DHPLC results were followed up by direct nucleotide sequencing of the appropriate amplicons. Because of the size of the WECARE study population and the complexity of the genes under study, five laboratories performed the screening using the same fixed set of protocols. To ensure consistency in screening between and within laboratories, we established a laboratory quality control plan that includes the following: (1) blinded screening of an initial set of 19 samples, including both samples with known mutations and controls by all laboratories; (2) initial screening of the same randomly selected 46 WECARE Study samples by all laboratories; (3) rescreening by one laboratory of a randomly selected 10% sample of all cases screened at each of the participating laboratories; and (4) blinded rescreening of a random 10% sample of each laboratory's own sample by that same laboratory. A more thorough consideration of the issues involved in designing quality control for a distributed screening approach as it applies to the WECARE Study is available [40].

Statistical considerations in using the counter-matched design

Several investigations have shown that counter-matching on a correlate of exposure is more efficient than simple random sampling for assessing exposure dose–response and gene–exposure interactions over a wide range of situations [37,41–43].

To assess the feasibility of possible case-control designs, using an incident series of second primary breast cancer cases as cases and incident first primary breast cancer cases as controls, we calculated the statistical power needed for detecting interaction terms of interest for various levels of allele frequency. To make these power calculations, we performed a limited simulation study using the paradigm of risk set sampling: for each case, two controls would be selected from those at risk at the time of the case's second (contralateral) breast cancer diagnosis who matched the case on all factors (for example, race, age at diagnosis, and registry region). Each

Table 1

Statistical power (percentages) to detect a log-linear trend from various designs for 700 risk sets

				Counter-matching		
		Random sampling		CM 1:2		CM 2:1
2 Gy rate ratio	All[a]	NCC 3	NCC 4	PN = 100%, PP = 100%	PN = 90%, PP = 90%	PN = 100%, PP = 100%
1.4	79	61	63	70	67	63
1.5	87	69	78	84	80	78
1.6	97	86	93	95	92	92
1.7	100	95	97	99	96	97

CM 1:2, counter-matched design in which one member of the triplet is registry radiation treatment negative (RRT–) and two are registry radiation treatment positive (RRT+); CM 2:1, counter-matched design in which two members of the triplet are RRT– and one is RRT+; NCC 3, nested case-control study in which each case is randomly matched to two controls; NCC 4, nested case-control study in which each case is randomly matched to three controls; PN, predictive negative value; PP, predictive positive value. [a]Full cohort power where radiation dose is available for all controls in the risk set and used in the analysis; this is the upper limit on any case-control analysis.

risk set consisted of 100 subjects with covariates generated as follows. RRT status was generated as a Bernoulli trial with 40% probability of RRT+; this approximates the observed percentage. True RT was then (randomly) determined according to the given predictive positive (PP) and predictive negative (PN) probabilities. These were taken to be either PP = 100%, PN = 100% (perfect agreement) or PP = 90%, PN = 90% (low agreement), a range of accuracy in the RRT treatment information that we believe spans the true accuracy [44]. For subjects actually given RT, a dose was assigned from a χ^2 distribution with four degrees of freedom. This distribution was used because it reasonably approximates the dose distribution shape we expect to see in treated subjects on the basis of previous data.

A case was randomly selected from the 100 risk-set subjects with probability based on their assigned rate ratios according to gene carrier status and radiation dose, and the rate model. Seven hundred risk sets were generated in this way. For the statistical power comparison, controls were then drawn from each risk set in several different ways. Random sampling of two or three controls from the 99 in the risk set yields the standard 1:2 or 1:3 nested case-control study (NCC 3 or NCC 4). The other two methods were to counter-match the two controls by RRT status in two different configurations; with one subject from the 'untreated' and two from the 'treated' (CM 1:2) or vice-versa (CM 2:1). The statistical power estimates were based on 200 simulated data sets for each set of parameters.

Statistical power for dose response

For computing the statistical power for detecting the effect of radiation dose on the rate of second breast cancer in the absence of gene susceptibility, the rate ratio was taken to be log-linear as a function of dose and was parameterized by the rate ratio at the 95th centile of the dose distribution. Preliminary data indicated that this corresponds to 2–3 Gy; conservatively, we call this the 2 Gy rate ratio. The 2 Gy rate ratio was taken to be 1.5 on the basis of the findings of Boice and colleagues [27]. Table 1 shows the estimated statistical power for detecting a radiation dose-response in the absence of gene susceptibility according to the different design scenarios. Included in this table are estimates for 'full cohort statistical power' (noted as 'All' in the table) where radiation dose would be available for all controls in the risk set and used in an analysis. The two columns for CM 1:2 show the power of this design when RRT and true RT status are in 'perfect' and 'low' agreement, respectively. There is some loss of statistical power between perfect and low agreement situations, but this loss is no more than 5%. Comparing the CM 1:2 columns with the randomly sampled controls, it is clear that, even with low RRT–true RT agreement, counter-matching provides much higher power than two (NCC 3) or even three (NCC 4) randomly sampled controls. As seen in the final column, and as would be expected on the basis of the reduced within-set variability in radiation dose for sets sampled in this way (see the discussion in the Appendix), CM 2:1 has lower statistical power than the CM 1:2 design.

Statistical power for detecting gene susceptibility

Although it is reasonable to define 'gene susceptibility' as a differential risk associated with an exposure in gene mutation carriers compared with non-carriers, not enough is known about the biological mechanisms between ionizing radiation and mutations in DNA repair genes to predict any particular mathematical form for the statistical interaction that we might observe [45]. In particular, there is some ambiguity (and controversy) about how to define an interactive effect between two factors that are

Table 2

Statistical power for detecting a gene by radiation dose interaction from various study designs

	Proportion of gene carriers = 0.5%				Proportion of gene carriers = 1%			
Study design	PP = 100%, PN = 100%		PP = 90%, PN = 90%		PP = 100%, PN = 100%		PP = 90%, PN = 90%	
Int RR at 2 Gy	70	100	70	100	16	20	16	20
NCC 3	51	55	48	59	56	63	65	71
CM 1:2	66	67	68	73	75	80	74	76
CM 2:1	32	31	43	45	37	47	48	51
	Proportion of gene carriers = 2%				Proportion of gene carriers = 3%			
Study design	PP = 100%, PN = 100%		PP = 90%, PN = 90%		PP = 100%, PN = 100%		PP = 90%, PN = 90%	
Int RR at 2 Gy	8	10	8	10	6	8	6	8
NCC 3	67	84	68	84	73	88	74	89
CM 1:2	78	86	83	86	82	90	80	92
CM 2:1	58	69	71	73	63	79	67	81
	Proportion of gene carriers = 5%				Proportion of gene carriers = 10%			
Study design	PP = 100%, PN = 100%		PP = 90%, PN = 90%		PP = 100%, PN = 100%		PP = 90%, PN = 90%	
Int RR at 2 Gy	5	6	5	6	3	4	3	4
NCC 3	81	88	81	91	71	96	72	86
CM 1:2	88	92	88	94	70	95	74	93
CM 2:1	75	87	81	90	64	89	67	83

The power for detecting a gene by radiation dose interaction varies as a function of the proportion of gene carriers, interaction rate ratio at 2 Gy (Int RR at 2 Gy), and predictive positive (PP) and predictive negative (PN) values for registry-noted radiation treatment for true treatment status. These estimates are based on 700 risk sets on a simulation study with 200 trials for each set of parameters. The proportion of registry-noted radiation treated is 40%, and the rate ratio for 2 Gy exposure in non-carriers is 1.5, which corresponds to the 95th centile of a χ^2, four degrees of freedom distribution. CM 1:2, counter-matched design in which one member of the triplet is unexposed and two are exposed; CM 2:1, counter-matched design in which two members of the triplet are unexposed and one is exposed; NCC 3, nested case-control study in which each case is randomly matched to two controls.

individually associated with disease (that is, that have 'main effects') [46–48]. We note that once we have additional data on genes that lie in the multiple DNA damage response pathways, we will be able to make empirical comparisons between the feasibility of testing associations within causal pathways and associations focused on the effect of single mutant alleles. Here, we consider a situation that would be clearly seen as an indication of biologic gene susceptibility; namely, when there is no main effect of mutation, but the effect of radiation is different in mutation carriers and non-carriers. Specifically, the 2 Gy rate ratio was taken to be 1.5 in mutation non-carriers, carriers were assumed to have no additional risk if unexposed, and mutation by radiation exposure effect was parameterized by the multiplicative interaction at 2 Gy exposure. For example, an interaction rate ratio of 5 would mean that mutation carriers who receive a dose of 2 Gy have $1.5 \times 5 = 7.5$ times the rate of second breast cancer compared with non-radiation-treated subjects. Table 2 shows selected power results for detecting gene mutation–radiation interaction. The prevalence of BRCA1, BRCA2, and ATM mutations in the general population is low, roughly 0.5–1% combined across all genes, but may be as high as 10% among breast cancer cases [12]; thus we present the power estimates for several mutation prevalence levels. Focusing on the CM 1:2 design actually used in the WECARE Study, if the proportion of gene carriers is 1%, 80% power is achieved only for interaction rate ratios of at least 20. The required rate ratio drops quickly with increasing proportion of gene carriers, and 80% statistical power is achieved at an interaction rate ratio of 8 with 3% gene carriers. There is generally little change in power between NCC 3 and CM 1:2 across perfect and low RRT–true RT agreement. Comparing the different designs, CM 1:2 is generally more powerful than simple random sampling NCC 3, and often much more so. In particular, the power advantage is greatest with rare genetic mutations. For comparison, the CM 2:1 design is included and is seen to do poorly, as one might expect on the basis of the

radiation dose results. We note that in the 70–80% statistical power range, an additional control in the standard nested case-control design yields increases of 5–10% power. The CM 1:2 design, as was implemented in the WECARE Study, provides this additional statistical power at no additional up-front cost.

Analysis of radiation dose and gene susceptibility

We are currently preparing data for analysis. Our key exposure variable is estimated radiation dose, which will be treated as a continuous variable in the major analyses, although a categorical approach will also be employed. Analyses will be conducted using, for each case and her matched controls, the estimated absorbed dose to the specific quadrant in which the second primary arose in the case. This approach might be more sensitive to errors in the underlying model for estimating doses, but, if such errors are small, it will be expected to give a more accurate indication of the magnitude of risk for particular levels of absorbed dose. Standard and biologically based approaches to testing interactions (such as [49]), including covariates for the main effects of gene carrier status and radiation dose, together with their product to test for the interaction effect, will be used. Similar approaches could be used to investigate interactions between genes or other factors. A detailed discussion of the likelihood derivation software implementation for analysis, and some examples using the SAS package (SAS Institute, Cary, NC) are given in the Appendix.

Discussion

We designed the WECARE Study to maximize our ability to examine the role of gene–environment and gene–gene interactions in cancer etiology, not only for the three genes that are currently under investigation but also for additional genes that will be investigated in the future. The intensive labor required to locate and enroll participants, to collect the necessary treatment information, and to perform the genotyping in large-scale population-based studies makes the choice of a cost-efficient study design critical. In the WECARE Study, we imposed the following criteria to enhance the utility and increase the statistical power of this resource beyond that obtained by a standard population-based case-control study: (1) we restricted our source population to women with breast cancer to ensure a population enriched in the prevalence of germline mutations which predispose to cancer; (2) we imposed a young age at diagnosis requirement because of the age-dependent effects of radiation to the breast; (3) we quantified the radiation dose received by the CB during therapy because it is sufficient to yield a substantial risk for breast cancer (with doses being several hundred times most diagnostic exposures); (4) we included a large sample size of second primary breast cancers to ensure adequate power for addressing interactions and other comparisons of subgroups; and (5) we used a counter-matched nested case-control design that makes use of the available RRT information to increase efficiency over random sampling.

As in any complex epidemiological study, given limited time and resources, we have had to make compromises on the scope of the study and the necessary number of subjects to recruit, with the amount of environmental risk factor information and biomaterial that we collected. We have tried to be practical about what is necessary for our immediate needs and to balance this with what we would like to have available for future studies. For example, while planning the biorepository we tried to predict some of the future genetic and data analyses that we might be interested in conducting. However, with the rapidly changing field of molecular biology and with the Human Genome Project well under way, our ability to foresee what hypotheses will be relevant to address, even within the next 5 years, is limited. So, to address the unknown genetic mutations that we might be interested in testing and to maximize the utility of the biorepository, we cryopreserved lymphocytes for future transformation into cell lines. For budgetary reasons, this choice meant that we were unable to prepare RNA, which might limit future analyses that we may wish to perform. Also, in theory, we could have increased the precision of our radiation exposure information by including measuring sources other than RT; in particular, we could have included medical exposures before the first breast cancer. However, it seemed unlikely that the relatively small number of women in our study for whom this information would be relevant would justify the considerable cost and effort involved for retrieving information of high enough quality as to be an improvement over the information we gathered through the interview.

To determine whether a radiation-sensitive population of women exists, in our study we will estimate and compare the effects of radiation on the risk for developing contralateral breast cancer in patients who are carriers of certain mutant alleles in comparison with those who are not. RT for treatment of breast cancer can result in substantial exposure to the CB. Boice and colleagues [27] showed that young women, 5 or more years after RT, were most susceptible to radiation-induced second primaries; in this study we therefore restricted the subjects to women under the age of 55 years at diagnosis and included a range of times between primaries to allow for the possibility of an interaction that decreases the expected lag time. Further, by using women exposed to high-dose RT, we increased the likelihood of detecting a gene–radiation interaction, if one exists. An important feature of our study is that the radiation dose is quantifiable, enabling us to examine dose–response relationships.

Although the restriction to a breast cancer patient population limits some of the generalizability of our

findings, the information we gain from this study has implications for our understanding the joint roles of genetic susceptibility and radiation exposure in the etiology of breast cancer in general. The mutation carrier prevalence in the WECARE Study population was expected to be higher than in a traditional case-control study in which the controls are assumed to be cancer free, and so our chances of detecting important associations with specific mutations will be greater. Similarly, exposure to high levels of ionizing radiation, such as that resulting from RT, although rare in the general population, was common in our population as a result of our counter-matched design: two-thirds of the WECARE Study participants will have received RT.

Conclusions

In this paper we have described the key elements of our novel study design combining high-risk women with quantifiable radiation exposure to allow the evaluation of the potential interaction between genetic susceptibility and ionizing radiation with the ultimate goal of determining whether a radiosensitive subpopulation exists. Although we focused on breast cancer and on three genes, our counter-matched study design and research methods are applicable to a broader range of study aims and will help investigators to plan other similar epidemiologic studies.

Appendix: Further discussion of issues related to the counter-matched design

I. Issues related to implementing the counter-matched design

Description of counter-matching as exposure-stratified sampling from matched risk sets

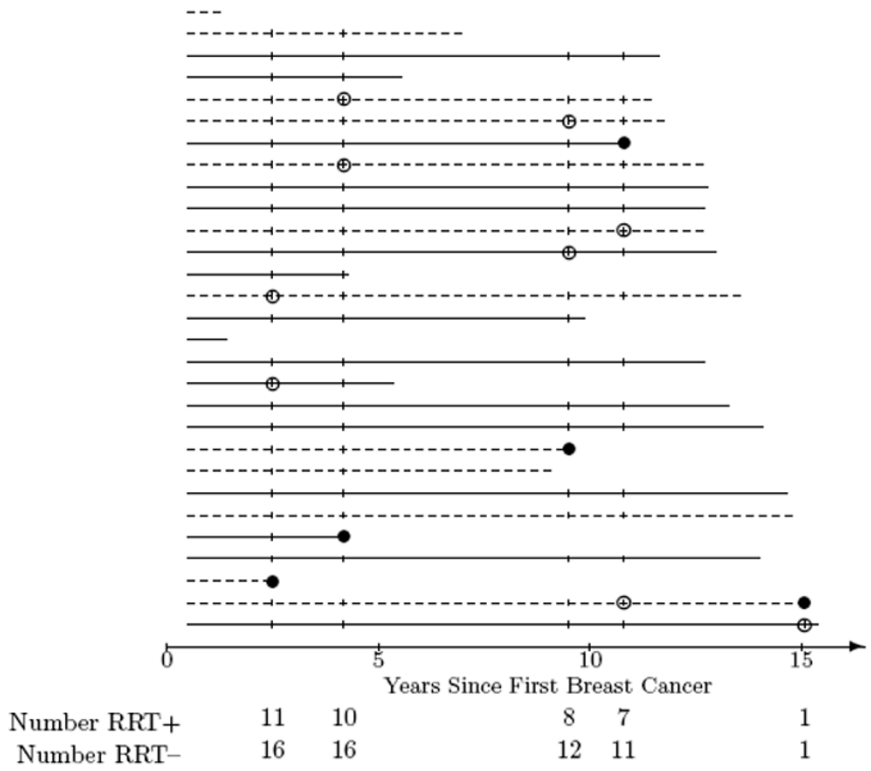
Figure 2 shows a hypothetical and idealized implementation of the CM 1:2 counter-matched sampling for this study, within a matching stratum defined by year of birth (5-year strata), year of diagnosis (4-year strata), registry region, and race/ethnicity. Each line represents a woman's duration on the study, with solid and dashed lines indicating RRT– and RRT+ status, respectively. The filled circle indicates that (and when) a woman developed a contralateral breast cancer. Risk sets are defined as the case and all eligible controls, controls being those women who were observed to have been contralateral breast cancer free at the case's time from first primary. Controls in each risk set in Figure 2 are indicated by short vertical lines. Counter-matched controls (denoted by the open circles) were sampled from those eligible so that the set would have one RRT– and two RRT+ subjects. For example, the first case was RRT+, so one RRT– control was sampled from the 16 RRT– controls in the risk set and one RRT+ is sampled from the 11 RRT+ in the risk set. In the second instance, the case was RRT–, so two controls were sampled from the 10 RRT+ in the risk set, and so forth.

A little intuition and considerations in choosing the counter-matched design

The motivation for using the counter-matched design is based on the fact that in the regression setting, the greater the variability in the covariates, the more precise the regression parameter estimate becomes. CM 1:2 will, on average, yield more radiation dose variability in the sets than randomly sampled controls. This is because counter-matched sets will typically have a subject with zero dose and two that are from the dose distribution, whereas sets with randomly sampled sets will be more likely to have two or even three women who received no radiation treatment (and thus zero dose). By the same reasoning, CM 2:1 would be less efficient than random sampling because CM 2:1 sets would most frequently have two subjects with zero 'true' dose, which will yield less radiation dose variation than random sampling. Thus, the greater variation in the RRT stratified (counter-matched) sets drives the potential gain in information relative to random sampling. In the WECARE Study situation, this intuition is confirmed. The increased variation results in significant cost-efficiencies for assessing radiation dose–response and radiation–genotype interaction, as verified by the simulation studies of power shown in Tables 1 and 2. We note that although counter-matching can increase cost-efficiency, counter-matched sampling on either unrelated or poorly correlated variables will do worse than random sampling. Thus, to realize an efficiency benefit through counter-matching, the sampling stratum variable should be fairly well correlated with the factor of interest. The WECARE Study was designed to maximize the information to assess radiation dose and gene–radiation interactions at the 'expense' of power to investigate other factors (such as main effects of gene; although, as it turns out, the loss is relatively small). We decided that this trade-off was worthwhile because the radiation gene susceptibility question is of primary importance.

In practice, from a list of randomly selected potential controls stratified by RRT status, the investigators at each registry were sent a list of six RRT+ candidate controls if the case was RRT–, and lists of three RRT– and three RRT+ potential controls if the case was RRT+. For each potential control, tracing techniques described in the Design section were used to determine eligibility for the study and, if eligible, the subject was invited to participate in the study. For the first years of the study, potential controls were assessed sequentially, proceeding to the next potential control only when it was determined that the current control was ineligible or declined to participate. In the last year, we decided to process all potential controls simultaneously and recruit the first two subjects who were eligible and agreed to participate. Note that only the RRT specific lists of potential controls distinguish this implementation from that of many standard case-control studies.

Figure 2



Counter-matching (CM) on registry radiation therapy (RRT) status. This figure shows a hypothetical and idealized implementation of the CM 1:2 samples for this study. The solid and dashed lines represent the time on study for RRT-unexposed (RRT–) and RRT-exposed (RRT+) subjects in a matching stratum. Symbols: Filled circle, contralateral breast cancer case; short vertical line, women 'at risk' at the case's time of contralateral breast cancer who may serve as controls; open circle, counter-matched controls sampled from those at risk; number RRT+, number of RRT+ women in the risk set; number RRT–, number of RRT– women in the risk set.

II. Issues pertaining to the analysis of counter-matched data

Likelihood derivation

The appropriate likelihood contribution for CM 1:2 is derived by a standard Bayes formula application as the probability of observing the case given the RRT status and covariates of subjects in the counter-matched set [37,50]. Thus, for the third counter-matched set in Fig. 2, consisting of subjects 9, 18, and 24, the likelihood is given by

$$\lambda_9\, pr(18,24|9)/[\lambda_9\, pr(18,24|9) + \lambda_{18}\, pr(9,24|18) + \lambda_{24}\, pr(9,18|24)],$$

where, for example, λ_9 is the probability that 9 is the case, and $pr(18,24|9)$ is the probability that 18 and 24 are sampled as controls given that 9 is the case. The values for the components are given in Table 3. The λ values are assumed to be from the proportional hazards model and depend on the subjects' covariates that we wish to model, for example $\lambda_9 = \lambda_0\ r(Z_9;\beta)$. The control selection probabilities and weights depend on the number of RRT– and RRT+ subjects in the risk set. Cancellation of common factors yields a 'weighted' conditional logistic likelihood contribution,

$$(8/2)r(Z_9;\beta)/[(8/2)r(Z_9;\beta) + 12r(Z_{18};\beta) + (8/2)r(Z_{24};\beta)], \quad \text{(A1)}$$

in which a subject's rate ratio contribution is multiplied by an RRT 'inverse sampling' weight.

Data analysis

The only difference between an analysis of a counter-matched sample and that of a simple random sample is the inclusion of the weights in the model. This is easy to do with most standard software packages. In particular, consider the commonly used log-linear model, i.e. $r(Z;\beta) = \exp(Z\beta)$. Then, continuing our example, the likelihood contribution (equation A1) for the third counter-matched set equals

$$\exp[Z_9\beta + \log(8/2)]/\{\exp[Z_9\beta + \log(8/2)] + \exp[Z_{18}\beta + \log(12)] + \exp[Z_{24}\beta + \log(8/2)]\}. \quad \text{(A2)}$$

Table 3

Components of the counter-matched (CM) set likelihood calculation for third set illustrated in Fig. 2

Possible CM set configuration					
Case	Controls	*pr*(case)	Case RRT status	*pr*(controls|case)	Weight[a]
9	18, 24	$\lambda r(Z_9;\beta)$	+	1/12 × 1/7	8/2
18	9, 24	$\lambda r(Z_{18};\beta)$	–	2/(8 × 7)	12
24	9, 18	$\lambda r(Z_{24};\beta)$	+	1/7 × 1/12	8/2

The calculation requires the probability of each of the possible case/control combinations, given that subjects 9,18, and 24 are in the counter-matched set. *pr*(case) is the probability that the subject is a case based on the proportional hazards model. For instance, for subject 9, λ is the baseline hazard, $r(Z_9;\beta)$ is the rate ratio based on the covariate value, and β is the rate ratio parameter. *pr*(controls|case) are control selection probabilities. For example, for subject 18, because she is RRT–, *pr*(controls = 9,24|case = 18) is the probability of sampling two RRT+ from the 8 in the risk set. [a]Weights for analysis equal to *pr*(controls|case) × 12 × (8 × 7/2).

Table 4

Hypothetical data from the counter-matched sets depicted in Fig. 2

Set no. (setno)	Subject ID (id)	Case = 1/ control = 0 (cc)	RRT status (rrt)	Number sampled from RRT stratum (m_rrt)	Total in RRT stratum (n_rrt)	True RT status (true_rt)	Radiation dose (rad_dose)	*ATM* mutation carrier status (atm)	Chemotherapy (chemo)
1	12	0	0	1	16	0	0	0	1
1	3	1	1	2	11	1	250	0	1
1	16	0	1	2	11	1	40	0	0
2	5	1	0	1	16	0	0	1	1
2	25	0	1	2	10	1	110	0	0
2	22	0	1	2	10	1	100	0	1
3	18	0	0	1	12	1	220	0	1
3	9	1	1	2	8	1	40	0	1
3	24	0	1	2	8	1	6	1	0
4	23	1	0	1	11	0	0	0	1
4	2	0	1	2	7	1	170	1	0
4	19	0	1	2	7	0	0	0	0
5	1	0	0	1	1	0	0	0	1
5	2	1	1	2	1	1	250	0	1

The structure of the data is as in a standard matched case-control study, but the numbers in each of the sampling strata (n_rrt) are required for analysis. The radiation dose is defined as the absorbed radiation dose (cGy) to quadrant in the contralateral breast where the case's tumor was located. Mutation carrier status is recorded as 0 = non-carrier, 1 = carrier. The status of chemotherapy administration as gathered from medical record review is shown (0 = no chemo, 1 = chemo). ID, identification number; RRT, indicator of radiation therapy as derived from registry records (registry RT; 0 = RRT–; 1 = RRT+); RT, status of radiation therapy administration as gathered from medical record review (True RT; 0 = TRT–, 1 = TRT+).

Thus, a log weight 'covariate' is always included in the model with a coefficient fixed at one. To illustrate the implementation, consider the hypothetical data set shown in Table 4 that includes covariate data and corresponds to the counter-matched sets illustrated in Fig. 2. Note that these data are identical in structure to standard individually matched case-control data with the variable 'setno' identifying case-control sets, 'cc' the case-control indicator, and radiation treatment, genotype, and other covariate data. But also included are the numbers of RRT– and RRT+ subjects that are required for the counter-matching weights. Table 5 shows the SAS code that could be used in an analysis of radiation treatment (yes/no) and of radiation–ATM gene mutation interaction. The DATA step reads in the data and computes the log-weight 'logw' and the interaction term 'rad_atm_int'. The PROC PHREG steps fit conditional logistic regression models. The variable 'setno' is used as both the time and

Table 5

SAS computing code for counter-matched WECARE data (wecare.dat) as illustrated in Table 4

```
data wecare;
        infile 'wecare.dat';
        input id setno cc rrt m_rrt n_rrt true_rt rad_dose atm chemo;
        logw = log(n_rrt/m_rrt);
        rad_atm_int = rad_dose*atm;
run;
* Radiation treatment yes/no, use true RT value;
proc phreg data = wecare;
         model setno*cc(0) = true_rt/offset = logw;
         strata setno;
run;
* Radiation-AT mutation interaction model;
proc phreg data = wecare;
        model setno*cc(0) = rad_dose atm rad_atm_int/offset = logw;
        strata setno;
run;
```

The input variables are as follows: id, subject identification no. (ID); cc, case = 1/control = 0; rrt, Registry radiation therapy (0 = RRT–,1 = RRT+); m_rrt, number sampled from RRT stratum; n_rrt, total in RRT stratum; true_rt, true radiation therapy (0 = TRT–,1 = TRT+); rad_dose, radiation dose; atm, *ATM* mutation (0 = non-carrier,1 = carrier); chemo, chemotherapy (0 = no chemotherapy,1 = chemotherapy).

stratum variable; this is a standard trick to fit case-control data with a Cox regression program. The term cc(0) indicates that cc is the case-control indicator with 0 indicating a control. The 'offset = logw' option in the model statement adds the log of the weights to the log-linear model expression (as in equation A2) and this is included in all analyses to account for the counter-matched sampling design. The first PHREG run is to estimate the rate ratio for radiation therapy (yes/no). Note that the variable used in the analysis is 'true_rt', the radiation therapy status as determined by interview with the subject and review of medical records (rather than 'rrt', the RT status recorded by the cancer registry and used for the counter-matching). The second PHREG run is to fit an interaction model for radiation dose and *ATM* mutation carrier status. The interaction term 'rad_atm_int' captures the relative difference in breast cancer rates associated with radiation dose between *ATM* mutation carriers and non-carriers.

We note that radiation effects are more appropriately modeled as an excess rate ratio, $r(Z;\beta) = 1 + Z\beta$, rather than the log-linear relationship described above. Because SAS allows only the log-linear model, for our actual analysis we will use the package Epicure (Hirosoft International, Inc., Seattle, WA), which accommodates a richer class of rate ratio models (as well as incorporation of the counter-matching weights).

Inaccuracies in RRT

As the 'true_rt' analysis in Table 5 indicates, the RRT numbers used in the sampling weights are determined only by the RRT status and do not change in any way once we determine the 'true' RT status. Incorporation of the RRT weights ensures that the analysis is valid regardless of the accuracy of RRT for true RT status. However, accuracy of RRT for true RT determines the precision with which true RT and gene–RT interactions will be estimated, as can be seen in the power analysis (Tables 1 and 2). As discussed above, in the WECARE Study RRT is highly correlated with true RT and yields a substantial cost-efficiency benefit relative to random sampling.

III. Analytic issues that arise when the data are not perfect

Inaccuracies in the weights

Another issue is that the actual numbers of at-risk RRT– (N_0) and RRT+ (N_1) subjects is less than n_0 and n_1, the RRT–/RRT+ numbers calculated from the WECARE Study registry cohort, based on the number of potential controls. The difference is due to the fact that the registry numbers include women who have died, those who have moved out of the study region, and those who may have refused to participate. On the assumption that a subject in the registry risk set was not in the actual risk set, did not depend on RRT status, and is equal to π, then the expected number of actual subjects is given by $E[N_j] = \pi n_j$. Replacing this expected value in the likelihood contribution (equation A1), π cancels out and we are left with the original weights. This suggests that using the RRT numbers N_j without change yields a valid analysis. Some limited simulation studies indicate that this 'approximation' works well. Further, because we have systematically kept track of the numbers and reasons for all non-eligibility, we will be able to empirically assess this assumption that π does not depend on RRT, and if not, could incorporate RRT (and possibly matching factor) specific π.

Small matching strata

For some WECARE Study cases, only a few potential control subjects, or even no controls for a particular case, were available for recruitment. This was especially true for women diagnosed at a young age. For example, in one registry only 10 potential RRT– controls and 1 potential RRT+ control were available for a Hispanic white RRT– case whose first primary breast cancer was diagnosed between ages 35 and 39 years, between calendar years 1985 and 1988. In this situation we did not have the six RRT+ potential controls needed to implement the counter-matched control selection protocol. In general, choosing

the sampling method based on the risk set characteristics will not result in biased estimation [37]. We therefore decided on the following small-matching-stratum strategy.

1. If there were less than the required number of potential controls in either of the RRT strata, we randomly sampled a contact list of six potential controls, assembled without regard to RRT status from all potential controls and recruited two controls.
2. If fewer than six potential controls were available for the case, all of the potential controls were included on the contact list and, if possible, two were recruited. If only one control was recruited, the set consisted of the case and the single control.
3. If there were no potential controls available or when none could be recruited, the age at diagnosis matching criterion was expanded to include an adjacent age category and controls were counter-matched from this expanded stratum.

In the example provided, where only 10 RRT– and 1 RRT+ controls were available, method 1 applies: we would randomly sample 6 potential controls from the 11 available, without regard to RRT status. This list would be sent to the registry for tracing and the eventual recruitment of two controls. The statistical analysis would be adjusted to reflect these strategies; contributions from subjects in sets with randomly sampled controls would be 'unweighted' in the conditional logistic likelihood.

Incomplete sets and missing covariate data

As is often the situation in case-control studies, there may be some sets at the end of the data collection period that have fewer than two controls for a given case, not because of a lack of potential controls as in the small matching stratum situation but because of tracing or enrollment difficulties. A related problem is that some controls may have partly missing covariate information. For instance, we could have a case for whom radiation dose could not be ascertained, but for whom genotype information is available. A natural approach would be to drop the case from the analysis, but this would result in dropping the entire case-control triplet, a waste of the control information. Although the number of such incomplete sets due to either scenario should be small, we use an approach of Huberman and Langholz [51] to accommodate both problems. This method retains the matching and counter-matching structure and provides a modeling structure to handle the missing information using 'missing indicators'. In the example above, a missing indicator for genotype would be defined for all subjects in the study and defined as 0 if genotype is known and 1 if missing. The genotype variable(s) for subjects with missing genotype would be set to the "baseline" values of those variables. The missing indicator would then be included in all analyses involving genotype. For the triplet in which the case has missing genotype, this method retains the genotype information from controls in the triplet, as well as radiation dose information available on the case. In the analysis, estimated parameters and standard errors related to genotype and radiation are interpreted as if all the data were present while the rate ratio associated with the missing indicator is considered a nuisance parameter and essentially ignored. Other methods are possible, such as complete-case and imputation, but we will explore such options as the need arises.

Competing interests

None declared.

Acknowledgements

This work was supported by the National Institutes of Health, grant numbers U01-CA83178, R01-CA97397, and R01-CA42949. The WECARE Study collaborative group is made up of the following.

Principal Investigator. Dr Jonine L Bernstein (Mount Sinai School of Medicine, New York, NY)

Co-Principal Investigators. Dr W Douglas Thompson, Chair of the Epidemiology and Biostatistics Core (University of Southern Maine, Portland, ME); Dr Robert W Haile (University of Southern California, Los Angeles, CA); Dr Leslie Bernstein, Chair of the Data Collection Core (University of Southern California, Los Angeles, CA); Dr Patrick Concannon, Chair of the Laboratory Core (Benaroya Research Institute at Virginia Mason, Seattle, WA).

Coordinating Center (Mount Sinai School of Medicine). Dr Gertrud S Berkowitz (Epidemiologist); Dr Xiaolin Liang (Informatics Specialist); Dr Susan L Teitelbaum (Project Director); Brooke Levinson (Project Coordinator); Abigail Wallis (Project Coordinator); National Cancer Institute: Dr Daniela Seminara (Program Officer).

Laboratories. Benaroya Research Institute at Virginia Mason: Dr Sharon Teraoka (Laboratory Director), Eric R Olson (Laboratory Manager); University of Southern California: Anh T Diep (Laboratory Director), Dr Nianmin Zhou (Laboratory Manager), Dr Yong Liu (Director of Blood Processing); Norwegian Radium Hospital, Oslo, Norway: Dr Anne-Lise Børresen-Dale (Laboratory Director), Laila Jansen (Laboratory Manager); Mount Sinai School of Medicine: Dr Barry S Rosenstein (Laboratory Director), Dr David P Atencio (Laboratory Manager; University of California at Los Angeles: Dr Richard A Gatti (Consultant).

Data Collection Centers. University of Southern California: Laura Donnelly (Project Manager), Dr Maya Mahue-Giangreco (Project Manager); Danish Cancer Society, Copenhagen, Denmark: Dr Jørgen H Olsen (Director), Dr Lene Mellemkjær (Project Manager); Fred Hutchinson Cancer Research Center: Dr Kathleen E Malone (Director), Noemi Epstein (Project Manager); University of California at Irvine: Dr Hoda Anton-Culver (Director), Joan Largent (Project Manager); University of Iowa: Dr Charles F Lynch (Director), Jeanne DeWall (Project Manager).

Radiation Core. University of Texas, MD Anderson Cancer Center: Dr Marilyn Stovall (Dosimetry Laboratory Director and Chair, Radiation Core), Susan Smith (Quality Assurance Dosimetry Supervisor); New York University: Dr Roy E Shore (Epidemiologist); International Epidemiology Institute and Vanderbilt University: Dr John D Boice Jr (Consultant).

Epidemiology and Biostatistics Core. University of Southern California: Dr Duncan C Thomas, Dr Bryan M Langholz.

External Advisors. Dr Alice Whittemore (Stanford University, Palo Alto, CA); Dr Bruce Ponder (University of Cambridge, Cambridge, UK); Dr William J Schull (University of Texas, Houston, TX).

References

1. Gatti RA, Becker-Catania S, Chun HH, Sun X, Mitui M, Lai CH, Khanlou N, Babaei M, Cheng R, Clark C, Huo Y, Udar NC, Iyer RK: **The pathogenesis of ataxia-telangiectasia. Learning from a Rosetta Stone.** *Clin Rev Allergy Immunol* 2001, **20:**87-108.
2. Kastan MB, Lim DS, Kim ST, Yang D: **ATM – a key determinant of multiple cellular responses to irradiation.** *Acta Oncol* 2001, **40:**686-688.
3. Cortez D, Wang Y, Qin J, Elledge SJ: **Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks.** *Science* 1999, **286:**1162-1166.
4. Gatei M, Scott SP, Filippovitch I, Soronika N, Lavin MF, Weber B, Khanna KK: **Role for ATM in DNA damage-induced phosphorylation of BRCA1.** *Cancer Res* 2000, **60:**3299-3304.
5. Ahn JY, Schwarz JK, Piwnica-Worms H, Canman CE: **Threonine 68 phosphorylation by ataxia telangiectasia mutated is required for efficient activation of Chk2 in response to ionizing radiation.** *Cancer Res* 2000, **60:**5934-5936.
6. Matsuoka S, Rotman G, Ogawa A, Shiloh Y, Tamai K, Elledge SJ: **Ataxia telangiectasia-mutated phosphorylates Chk2 *in vivo* and *in vitro*.** *Proc Natl Acad Sci USA* 2000, **97:**10389-10394.
7. Melchionna R, Chen XB, Blasina A, McGowan CH: **Threonine 68 is required for radiation-induced phosphorylation and activation of Cds1.** *Nat Cell Biol* 2000, **2:**762-765.
8. Swift M, Morrell D, Massey RB, Chase CL: **Incidence of cancer in 161 families affected by ataxia-telangiectasia.** *N Engl J Med* 1991, **325:**1831-1836.
9. Swift M, Reitnauer PJ, Morrell D, Chase CL: **Breast and other cancers in families with ataxia-telangiectasia.** *N Engl J Med* 1987, **316:**1289-1294.
10. Chenevix-Trench G, Spurdle AB, Gatei M, Kelly H, Marsh A, Chen X, Donn K, Cummings M, Nyholt D, Jenkins MA, Scott C, Pupo GM, Dork T, Bendix R, Kirk J, Tucker K, McCredie MR, Hopper JL, Sambrook J, Mann GJ, Khanna KK: **Dominant negative ATM mutations in breast cancer families.** *J Natl Cancer Inst* 2002, **94:**205-215.
11. Dork T, Bendix R, Bremer M, Rades D, Klopper K, Nicke M, Skawran B, Hector A, Yamini P, Steinmann D, Weise S, Stuhrmann M, Karstens JH: **Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients.** *Cancer Res* 2001, **61:**7608-7615.
12. Teraoka SN, Malone KE, Doody DR, Suter NM, Ostrander EA, Daling JR, Concannon P: **Increased frequency of ATM mutations in breast carcinoma patients with early onset disease and positive family history.** *Cancer* 2001, **92:**479-487.
13. Spring K, Ahangari F, Scott SP, Waring P, Purdie DM, Chen PC, Hourigan K, Ramsay J, McKinnon PJ, Swift M, Lavin MF: **Mice heterozygous for mutation in Atm, the gene involved in ataxia-telangiectasia, have heightened susceptibility to cancer.** *Nat Genet* 2002, **32:**185-190.
14. Chakraborty R, Little MP, Sankaranarayanan K: **Cancer predisposition, radiosensitivity and the risk of radiation-induced cancers. IV. Prediction of risks in relatives of cancer-predisposed individuals.** *Radiat Res* 1998, **149:**493-507.
15. Hall EJ: **Radiation, the two-edged sword: cancer risks at high and low doses.** *Cancer J* 2000, **6:**343-350.
16. Horn PL, Thompson WD, Schwartz SM: **Factors associated with the risk of second primary breast cancer: an analysis of data from the Connecticut Tumor Registry.** *J Chronic Dis* 1987, **40:** 1003-1011.
17. John EM, Kelsey JL: **Radiation and other environmental exposures and breast cancer.** *Epidemiol Rev* 1993, **15:**157-162.
18. Shore RE, Hildreth N, Woodard E, Dvoretsky P, Hempelmann L, Pasternack B: **Breast cancer among women given X-ray therapy for acute postpartum mastitis.** *J Natl Cancer Inst* 1986, **77:**689-696.
19. Howe GR, McLaughlin J: **Breast cancer mortality between 1950 and 1987 after exposure to fractionated moderate-dose-rate ionizing radiation in the Canadian fluoroscopy cohort study and a comparison with breast cancer mortality in the atomic bomb survivors study.** *Radiat Res* 1996, **145:**694-707.
20. Little MP, Boice JD Jr: **Comparison of breast cancer incidence in the Massachusetts tuberculosis fluoroscopy cohort and in the Japanese atomic bomb survivors.** *Radiat Res* 1999, **151:** 218-224.
21. Preston DL, Mattsson A, Holmberg E, Shore R, Hildreth NG, Boice JD Jr: **Radiation effects on breast cancer risk: a pooled analysis of eight cohorts.** *Radiat Res* 2002, **158:**220-235.
22. Boice JD Jr, Preston D, Davis FG, Monson RR: **Frequent chest X-ray fluoroscopy and breast cancer incidence among tuberculosis patients in Massachusetts.** *Radiat Res* 1991, **125:**214-222.
23. Travis LB, Hill DA, Dores GM, Gospodarowicz M, van Leeuwen FE, Holowaty E, Glimelius B, Andersson M, Wiklund T, Lynch CF, Van't Veer MB, Glimelius I, Storm H, Pukkala E, Stovall M, Curtis R, Boice JD Jr, Gilbert E: **Breast cancer following radiotherapy and chemotherapy among young women with Hodgkin disease.** *JAMA* 2003, **290:**465-475.
24. Basco VE, Coldman AJ, Elwood JM, Young ME: **Radiation dose and second breast cancer.** *Br J Cancer* 1985, **52:**319-325.
25. Tercilla O, Krasin F, Lawn-Tsao L: **Comparison of contralateral breast doses from 1/2 beam block and isocentric treatment techniques for patients treated with primary breast irradiation with ^{60}CO.** *Int J Radiat Oncol Biol Phys* 1989, **17:**205-210.
26. Fraass BA, Roberson PL, Lichter AS: **Dose to the contralateral breast due to primary breast irradiation.** *Int J Radiat Oncol Biol Phys* 1985, **11:**485-497.
27. Boice JD Jr, Harvey EB, Blettner M, Stovall M, Flannery JT: **Cancer in the contralateral breast after radiotherapy for breast cancer.** *N Engl J Med* 1992, **326:**781-785.
28. Begg CB: **The search for cancer risk factors: when can we stop looking?** *Am J Public Health* 2001, **91:**360-364.
29. Thompson WD: **Methodologic perspectives on the study of multiple primary cancers.** *Yale J Biol Med* 1986, **59:**505-516.
30. Begg CB, Berwick M: **A note on the estimation of relative risks of rare genetic susceptibility markers.** *Cancer Epidemiol Biomarkers Prev* 1997, **6:**99-103.
31. Malkin D, Jolly KW, Barbier N, Look AT, Friend SH, Gebhardt MC, Andersen TI, Borresen AL, Li FP, Garber J: **Germline mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms.** *N Engl J Med* 1992, **326:**1309-1315.
32. Meadows AT, Strong LC, Li FP, D'Angio GJ, Schweisguth O, Freeman AI, Jenkin RD, Morris-Jones P, Nesbit ME: **Bone sarcoma as a second malignant neoplasm in children: influence of radiation and genetic predisposition for the Late Effects Study Group.** *Cancer* 1980, **46:**2603-2606.
33. Rayner CR, Towers JF, Wilson JS: **What is Gorlin's syndrome? The diagnosis and management of the basal cell naevus syndrome, based on a study of thirty-seven patients.** *Br J Plast Surg* 1977, **30:**62-67.
34. Strong LC: **Genetic and environmental interactions.** *Cancer* 1977, **40:**1861-1866.
35. Wong FL, Boice JD Jr, Abramson DH, Tarone RE, Kleinerman RA, Stovall M, Goldman MB, Seddon JM, Tarbell N, Fraumeni JF Jr, Li FP: **Cancer incidence after retinoblastoma. Radiation dose and sarcoma risk.** *JAMA* 1997, **278:**1262-1267.
36. Robins JM, Gail MH, Lubin JH: **More on 'Biased selection of controls for case-control analyses of cohort studies'.** *Biometrics* 1986, **42:**293-299.
37. Langholz B, Goldstein L: **Risk set sampling in epidemiologic cohort studies.** *Statist Science* 1996, **11:**35-53.
38. Langholz B, Goldstein L: **Fitting logistic models using conditional logistic likelihood when there are large strata.** *Comput Sci Statist* 1997, **29:**551-555.
39. **WFR-Aquaplast Corporation** [http://www.wfr-aquaplast.com]
40. Bernstein JL, Teraoka S, Haile RW, Borresen-Dale AL, Rosenstein BS, Gatti RA, Diep AT, Jansen L, Atencio DP, Olsen JH, Bernstein L, Teitelbaum SL, Thompson WD, Concannon P: **Designing and implementing quality control for multi-center screening of mutations in the ATM gene among women with breast cancer.** *Hum Mutat* 2003, **21:**542-550.
41. Andrieu N, Goldstein AM, Thomas DC, Langholz B: **Counter-matching in studies of gene–environment interaction: efficiency and feasibility.** *Am J Epidemiol* 2001, **153:**265-274.
42. Cologne J, Langholz B: **Selecting controls for assessing interaction in nested case-control studies.** *J Epidemiol* 2003, **13:** 193-202.
43. Langholz B, Clayton D: **Sampling strategies in nested case-control studies.** *Environ Health Perspect* 1994, **102 Suppl 8:**47-51.
44. Malin JL, Kahn KL, Adams J, Kwan L, Laouri M, Ganz PA: **Validity of cancer registry data for measuring the quality of breast cancer care.** *J Natl Cancer Inst* 2002, **94:**835-844.

45. Siemiatycki J, Thomas DC: **Biological models and statistical interactions: an example from multistage carcinogenesis.** *Int J Epidemiol* 1981, **10:**383-387.
46. Clayton D, McKeigue PM: **Epidemiological methods for studying genes and environmental factors in complex diseases.** *Lancet* 2001, **358:**1356-1360.
47. Rothman N, Wacholder S, Caporaso NE, Garcia-Closas M, Buetow K, Fraumeni JF Jr: **The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens.** *Biochim Biophys Acta* 2001, **1471**: C1-C10.
48. Brennan P: **Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it?** *Carcinogenesis* 2002, **23:**381-387.
49. Thomas DC. **Temporal effects and interactions in cancer: implications of carcinogenic models**. In *Environmental Epidemiology: Risk Assessment*. Edited by Prentice RL, Whittemore AS. Philadelphia: Society for Industrial and Applied Mathematics; 1982:107-121.
50. Langholz B, Borgan O: **Counter-matching: a stratified nested case-control sampling method.** *Biometrika* 1995, **82:**69-79.
51. Huberman M, Langholz B: **Application of the missing-indicator method in matched case-control studies with incomplete data.** *Am J Epidemiol* 1999, **150:**1340-1345.

Correspondence

Jonine L Bernstein, Department of Community and Preventive Medicine, Mount Sinai School of Medicine, One Gustave L Levy Place, Box 1043, New York, NY 10029-6574, USA. Tel: +1 212 241 8495; fax: +1 212 360 6965; e-mail: jonine.bernstein@mssm.edu