

# Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually

Mitchell H. Gail,\* Louise A. Brinton, David P. Byar, Donald K. Corle, Sylvan B. Green, Catherine Schairer, John J. Mulvihill

To assist in medical counseling, we present a method to estimate the chance that a woman with given age and risk factors will develop breast cancer over a specified interval. The risk factors used were age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast cancer. A model of relative risks for various combinations of these factors was developed from case-control data from the Breast Cancer Detection Demonstration Project (BCDDP). The model allowed for the fact that relative risks associated with previous breast biopsies were smaller for women aged 50 or more than for younger women. Thus, the proportional hazards assumption was relaxed to allow separate proportional hazards models for those under age 50 and for those of age 50 or more. The baseline age-specific hazard rate, which is the rate for a patient without identified risk factors, is computed as the product of the observed age-specific composite hazard rate times the quantity 1 minus the attributable risk. We calculated individualized breast cancer probabilities from information on relative risks and the baseline hazard rate. These calculations take competing risks and the interval of risk into account. Our data were derived from women who participated in the BCDDP and who tended to return for periodic examinations. For this reason, the risk projections given are probably most reliable for counseling women who plan to be examined about once a year. [J Natl Cancer Inst 81:1879-1886, 1989]

Increasing public awareness of breast cancer risk factors, such as having a relative with breast cancer, has created a demand for informed counseling of patients at elevated risk (1). A woman's decision to embark on a program of intensive surveillance with mammography, or even to undergo prophylactic mastectomy (2), depends on her awareness of the medical options, on personal preferences, and, very importantly, on an individualized estimate of the probability of her developing breast cancer in a defined period. Such an estimate is also useful for designing prevention trials in high-risk subsets of the population and in targeting screening and prevention efforts.

This article provides such individualized estimates of the probability of developing breast cancer for white females

who are willing to have annual examinations, such as participants in the Breast Cancer Detection Demonstration Project (BCDDP) (3). Previous efforts to estimate absolute breast cancer risk focused on different populations, namely, relatives of known breast cancer patients (4,5).

If one had followed large numbers of patients with each possible combination of risk factors, one could estimate the absolute probability of developing breast cancer separately for each risk factor combination from standard life-table methods. However, because of the many possible risk factor combinations, this direct approach is not optimal, even for large data sets, such as those provided by screening and following 284,780 women in the BCDDP (3,6-10). Instead we use a model that assumes proportional hazards over restricted age ranges of less than 50 years and 50 years or more, and we subdivide the problem of estimating the probability of developing breast cancer into three problems: (a) determining which risk factors are important and estimating the relative risk for a subject with a given constellation of risk factors at a given age compared to a subject without the identified risk factors; (b) estimating the baseline age-specific breast cancer hazard rate, which is the rate for a woman without identified risk factors; and (c) projecting the long-term probability of developing breast cancer on the basis of a consideration of competing risks and the results

Received May 18, 1989; revised September 6, 1989; accepted September 14, 1989.

M. H. Gail (Biostatistics Branch), L. A. Brinton, C. Schairer (Environmental Epidemiology Branch), D. P. Byar, D. K. Corle, S. B. Green (Biometry Branch), J. J. Mulvihill (Clinical Epidemiology Branch), Division of Cancer Etiology, National Cancer Institute, Bethesda, MD.

We thank Dr. Larry H. Baker and Kay V. Wagner for providing helpful comments on the original data sources; Professor Duncan Thomas for suggestions leading to table 5; Dr. Richard Costlow, Christine Carter, and others at the National Cancer Institute for making available original data tapes from the Breast Cancer Detection Demonstration Project; Tom Riley of Information Management Services for calculating the age-specific composite rates in table 3; Dr. George Rubin of the Centers for Disease Control for making available data from the Cancer and Steroid Hormone Study; Lynn Ries for providing rates from the Surveillance, Epidemiology, and End Results Program; reviewers for suggesting analyses on atypical hyperplasia; and Jennifer Donaldson for typing the manuscript.

\*Correspondence to: Dr. Mitchell H. Gail, Executive Plaza North, Rm. 403, National Institutes of Health, Rockville, MD 20892.

on relative risk and baseline hazard. We follow this division in the Methods section. Tables for estimating individualized absolute risks are discussed in the Results section.

## Methods

### Selection of Risk Factors and Estimation of Relative Risks for Risk Factor Combinations

Risk factors were determined, and relative risks were estimated from a case-control study conducted among participants in the BCDDP (3,6-10). Cases consisted of women whose breast cancer was incident (not prevalent at first screening) between 1973 and 1980 at 28 participating centers (9). The present analysis was restricted to white females and included women with in situ carcinoma as well as women with invasive carcinoma. For each case, a matched control was selected from women who did not receive a recommendation for biopsy during the same screening period. The matching variables were age at entry into the screening program in 5-year age groups, race, center at which the participant was screened, calendar time of entry into the screening program within 6 months, and length of participation in the screening program (controls had at least as many years of screening as the cases). Risk factor information was collected during home interview from 80% of eligible cases and 83% of controls. Whites made up 91% of the case-control study population.

The present analysis was based on 2,852 white cases and 3,146 white controls for whom all information was known for the variables in the final model. This study represents an extension and enlargement (9) of the case-control study described in (6-8). Among these 5,998 subjects, 4,496 formed matched pairs. Because we wished to base our analysis on the larger total number of subjects, we used an unstratified, unconditional logistic regression rather than a pair-matched analysis. In fact, both methods of analysis were used in initial exploratory work, and the answers obtained from comparable models were very similar. The results below include age as a main effect, but age is simply dichotomized as AGE-CAT = 0 for age less than 50 years and AGE-CAT = 1 for age 50 years or more. Finer categorizations of age had no substantial effect on estimates of relative risk for other variables.

To select variables for inclusion in our risk model, we assessed a variety of factors within this study population. A number of medications were unrelated to breast cancer risk, including oral contraceptives (11), thyroid supplements (12), rauwolfia preparations (13), and diazepam medications (14). Cigarette smoking (15) and consumption of methylxanthines (16) were not associated with elevations in risk. Modest alcohol consumption (17) and long-term menopausal estrogen therapy (9) did appear to increase risk, but the number of appropriately exposed women was limited. Height was associated with an increased risk, although those in the highest quartile were only at a 40% higher risk than those in the lowest quartile (18). Analyses focusing on familial (7), reproductive (19), menstrual (20), and medical history (8) variables determined that the major predictors of risk in this popula-

tion were a family history of breast cancer in a first-degree relative, a late age at first childbirth, early menarche, and multiple previous benign breast biopsies; these four factors were included in our model. These risk factors were also consistently identified in other studies.

The coding for the four risk factors—age at menarche (AGEMEN), number of previous breast biopsies (NBIOPS), age at first live birth (AGEFLB), and number of first-degree relatives (mother or sisters) with breast cancer (NUMREL)—is given in table 1, together with the distributions among cases and controls. Information on bilaterality and on ages of onset of breast cancer among relatives did not improve on the model containing NUMREL.

The relative risks in table 1 were obtained from an unconditional logistic regression (21) that included main effects AGEMEN, NBIOPS, AGEFLB, and NUMREL as well as AGE-CAT and interactions between AGE-CAT and NBIOPS and between AGEFLB and NUMREL. The former interaction was also found in our unreported reanalysis of data from the Cancer and Steroid Hormone Study (22). Our data revealed no other important interactions between age and other risk factors or among the other risk factors. The unconditional logistic regression model, from which table 1 derives, expresses the log of the odds ratio for disease as

$$\begin{aligned} & -0.74948 + 0.09401(\text{AGEMEN}) + 0.52926(\text{NBIOPS}) \\ & + 0.21863(\text{AGEFLB}) + 0.95830(\text{NUMREL}) \\ & + 0.01081(\text{AGE-CAT}) - 0.28804(\text{NBIOPS} \\ & \times \text{AGE-CAT}) - 0.19081(\text{AGEFLB} \times \text{NUMREL}). \end{aligned} \quad [1]$$

In table 1, the coefficients from this equation have been converted into relative risk factors so that the relative risk can be obtained by multiplication. For example, the relative risk for a woman aged 54 (AGE-CAT = 1) with age at menarche of 15 (AGEMEN = 0), one previous breast biopsy (NBIOPS = 1), no live births (AGEFLB = 2), and one mother or sister with breast cancer (NUMREL = 1), compared to another individual of the same age but with no other risk factors, is  $1.000 \times 1.273 \times 2.756 = 3.508$ .

The covariances for the estimated coefficients in equation 1 are presented in table 2.

An important implication of equation 1 is that the proportional-hazards assumption (23) holds separately for the age ranges less than 50 years and 50 years or more, even though proportional hazards are not assumed to hold over all ages.

### Estimating the Baseline Age-Specific Breast Cancer Hazard Rate

Suppose we have obtained estimates of the age-specific breast cancer incidence rate  $h_1^*(t)$  from a composite population, such as all subjects in the BCDDP. The subscript 1 in  $h_1^*(t)$  denotes breast cancer, whereas a subscript 2 will refer to competing causes of death. The variable  $t$  refers to age, and the asterisk indicates that this is a composite hazard. We need to estimate  $h_1(t)$ , the baseline hazard from a subject without known risk factors.

Suppose there are  $I$  risk groups. In table 1,  $I = 3 \times 2 \times 3 \times 4 \times 3 = 216$ . Let  $r_i(t)$  be the relative risk of the  $i$ th risk group compared to the baseline group, and let  $r_1(t) = 1.0$ .

Table 1. Risk model and associated relative risks\*

| Risk factor (code No.)   | Associated<br>relative risk | No. of cases<br>(n = 2,852) | No. of controls<br>(n = 3,146) |       |
|--------------------------|-----------------------------|-----------------------------|--------------------------------|-------|
| AGEMEN (yr)              |                             |                             |                                |       |
| ≥14 (0)                  | 1.000                       | 790                         | 926                            |       |
| 12-13 (1)                | 1.099                       | 1,554                       | 1,735                          |       |
| <12 (2)                  | 1.207                       | 508                         | 485                            |       |
| NBIOPS                   |                             |                             |                                |       |
| Age < 50 yr              |                             |                             |                                |       |
| 0 (0)                    | 1.000                       | 635                         | 794                            |       |
| 1 (1)                    | 1.698                       | 113                         | 93                             |       |
| ≥2 (2)                   | 2.882                       | 66                          | 24                             |       |
| Age ≥50 yr               |                             |                             |                                |       |
| 0 (0)                    | 1.000                       | 1,551                       | 1,817                          |       |
| 1 (1)                    | 1.273                       | 312                         | 300                            |       |
| ≥2 (2)                   | 1.620                       | 175                         | 118                            |       |
| AGEFLB (yr)              | NUMREL                      |                             |                                |       |
| <20 (0)                  | 0 (0)                       | 1.000                       | 167                            | 285   |
|                          | 1 (1)                       | 2.607                       | 44                             | 40    |
|                          | ≥2 (2)                      | 6.798                       | 8                              | 0     |
| 20-24 (1)                | 0 (0)                       | 1.244                       | 708                            | 1,042 |
|                          | 1 (1)                       | 2.681                       | 208                            | 123   |
|                          | ≥2 (2)                      | 5.775                       | 25                             | 5     |
| 25-29 or nulliparous (2) | 0 (0)                       | 1.548                       | 986                            | 1,106 |
|                          | 1 (1)                       | 2.756                       | 247                            | 178   |
|                          | ≥2 (2)                      | 4.907                       | 46                             | 20    |
| ≥30 (3)                  | 0 (0)                       | 1.927                       | 307                            | 291   |
|                          | 1 (1)                       | 2.834                       | 87                             | 50    |
|                          | ≥2 (2)                      | 4.169                       | 19                             | 6     |

\*Relative risk compared to an individual of the same age without any risk factors is estimated by locating the person's associated relative risk for AGEMEN, NBIOPS, and the combination AGEFLB and NUMREL and multiplying these three numbers together.

Table 2. Parameter estimates and covariances\* for equation 1

|                    | NBIOPS  | AGEMEN         | AGEFLB  | NUMREL  | AGECAT  | NBIOPS<br>× AGECAT | AGEFLB<br>× NUMREL |
|--------------------|---------|----------------|---------|---------|---------|--------------------|--------------------|
| Parameter estimate | 0.52926 | 0.09401        | 0.21863 | 0.95830 | 0.01081 | -0.28804           | -0.19081           |
| Covariance         | 0.956   | 0.001<br>0.161 | -0.005  | -0.015  | 0.201   | -0.956             | 0.012              |
|                    |         |                | 0.000   | -0.010  | 0.023   | 0.002              | 0.005              |
|                    |         |                | 0.128   | 0.174   | -0.032  | 0.002              | -0.106             |
|                    |         |                |         | 1.907   | -0.010  | 0.020              | -0.908             |
|                    |         |                |         |         | 0.418   | -0.280             | 0.002              |
|                    |         |                |         |         |         | 1.244              | -0.015             |
|                    |         |                |         |         |         |                    | 0.534              |

\*All covariances are  $10^{-2}$  times the numbers shown. For example, the standard error of the parameter estimate 0.52926 for NBIOPS is  $(0.00956)^{1/2} = 0.098$ .

Suppose that a proportion  $P_i(t)$  of women of age  $t$  are in risk group  $i$ . Then

$$h_1^*(t) = \sum_{i=1}^I P_i(t) h_i(t) r_i(t). \quad [2]$$

Let  $\hat{p}_i(t)$  be the observed proportion of cases of age  $t$  who are in risk group  $i$ . Then  $\hat{p}_i(t)$  estimates

$$\rho_i(t) = P_i(t) h_i(t) r_i(t) / h_1^*(t). \quad [3]$$

Dividing both sides of equation 3 by  $r_i(t)$ , multiplying by  $h_1^*(t)$ , and summing on  $i$ , we obtain

$$h_1(t) = h_1^*(t) \sum_{i=1}^I \{\rho_i(t) / r_i(t)\} \equiv h_1^*(t) F(t). \quad [4]$$

Thus, an estimate  $\hat{h}_1(t)$  is obtained by substitution of the estimate  $\hat{h}_1^*(t)$  from the general population, the estimates  $\hat{p}_i(t)$  from the cases, and the estimates  $\hat{r}_i(t)$  from the case-control study into equation 4. The factor  $F(t)$  equals 1 minus the population attributable risk fraction for women of age  $t$  (10).

Although the age-specific composite incidence rate  $h_1^*(t)$  increases rapidly with age, the conversion factor  $F(t)$  changes little with age. The estimate of  $F(t)$  was 0.5229 for the 814 cases 31-49 years of age, whereas it was 0.5264 for the 2,038 cases 50-81 years of age. These estimates were obtained from the covariate compositions of the BCDDP cases and from the risk model in equation 1 and tables 1 and 2.

To apply these results, we estimated  $h_1^*(t)$  in table 3 from original data tapes on all 284,780 BCDDP participants with entry or follow-up forms. Of these subjects, 73 were inel-

**Table 3.** Age-specific composite incidence rates  $h^*(t)$  for white females in years 2 and 3 following the initial BCDDP screening†

| Age (yr) | No. of BCDDP cases | Rate  |              | Relative risk |
|----------|--------------------|-------|--------------|---------------|
|          |                    | BCDDP | SEER Program |               |
| 20-24    | 0                  | 2.7‡  | 1.3          | 2.10‡         |
| 25-29    | 0                  | 16.8‡ | 8.0          | 2.10‡         |
| 30-34    | 0                  | 60.3‡ | 28.8         | 2.10‡         |
| 35-39    | 53                 | 114.6 | 54.7         | 2.10          |
| 40-44    | 162                | 203.7 | 109.2        | 1.87          |
| 45-49    | 249                | 280.8 | 173.3        | 1.62          |
| 50-54    | 289                | 320.9 | 198.8        | 1.61          |
| 55-59    | 223                | 293.8 | 221.5        | 1.33          |
| 60-64    | 190                | 369.4 | 278.3        | 1.33          |
| 65-69    | 121                | 356.1 | 315.3        | 1.13          |
| 70-74    | 54                 | 307.8 | 331.3        | 0.93          |
| 75-79    | 11                 | 301.3 | 364.0        | 0.83          |

†Incidence rates are the No. of cases (including in situ carcinoma) per  $10^5$  person-years of follow-up. Relative risks = BCDDP rates divided by rates for white females in 1979, as determined by the SEER Program.

‡Too few BCDDP cases were available to obtain a direct estimate of the rate. The method of extrapolation is described in the Methods section.

igible because their cancers were known to have occurred before the initial screening. We confined our analyses to the 243,221 white females for whom we had information on birth date, age at initial screening, and, for cases, age at which cancer developed. To obtain counts of breast cancers, we analyzed a file (24) containing data on persons with histologically confirmed breast cancer. We concentrated our analyses on the 1,354 white females who developed cancer in years 2 and 3 following the initial screening and for whom we had adequate data on date of birth, date of initial screening, and date of cancer diagnosis for calculation of rates. Ascertainment of breast cancer was quite complete for the first 3 years following the initial screening and included cases detected outside the program as well as cases found as a result of the screening (Baker LH: personal communication, 1983). We computed breast cancer rates that were age-specific in 1-year intervals and that were specific for years 1 (including the initial screening), 2, and 3 of the program. Because the rates in years 2 and 3 following the initial screening were similar and neither was consistently larger than the other, the data were combined for these 2 years. These rates were typically less than half the rate for year 1, which included prevalent cases and would be inappropriate for use in table 3. Corrections to account for missing data (~5% of those enrolled) affected both numerator and denominator to a similar degree and produced negligible changes in computed rates. Therefore, we relied solely on the 243,221 white females without missing data.

The rates in table 3 are grouped into 5-year age intervals. We extended estimates into the age range 20-34 of sparse BCDDP cases by multiplying rates from eight population-based cancer registries in the Surveillance, Epidemiology, and End Results (SEER) Program by the relative risk of BCDDP to SEER rates, 2.10, observed for ages 35-39.

## Projecting Long-Term Probabilities of Developing Breast Cancer

The projections that we make are based on the assumption that the age-specific hazard  $h_2(t)$  of dying of causes other than breast cancer is the same for all subjects. This hazard is estimated from 1979 mortality rates for all causes except breast cancer (25). Under this assumption, standard methods in competing risk analysis (26-28) may be used to calculate the probability that a woman who is age  $a$  and who has age-dependent relative risk  $r(t)$  will develop breast cancer by age  $a + \tau$ . The desired probability is

$$P(a, \tau, r(t)) = \int_a^{a+\tau} h_1(t)r(t) \exp\left\{-\int_a^t h_1(u)r(u)du\right\} \{S_2(t)/S_2(a)\}dt, \quad [5]$$

where  $S_2(t) = \exp\{-\int_0^t h_2(u)du\}$  is the probability of surviving competing risks up to age  $t$ . As discussed in (26-28), we do not require an assumption of independence of competing risks for these calculations, which involve only observable cause-specific hazard rates. Equation 5 is sufficiently general to allow relative risks  $r(t)$  to depend on age, as required by equation 1. Thus, no assumption of proportional hazards (23) is made. Dupont (29) recently presented graphs based on his independent derivation of equation 5. These graphs allow one to convert from relative risks to absolute risks, assuming proportional hazards. However, Dupont did not discuss risk factors or the case of nonproportional hazards.

We evaluated equation 5 by dividing the age scale into 13 intervals,  $[0, \tau_1), [\tau_1, \tau_2), \dots, [\tau_{12}, \tau_{13})$ , with widths  $\Delta_1, \Delta_2, \dots, \Delta_{13}$ . The baseline hazard  $h_1(t)$  is assumed to be 0 in the first interval  $[0, \tau_1)$ , and the remaining 12 intervals are defined as in table 3. If  $h_1(t)$ ,  $h_2(t)$ , and  $r(t)$  are assumed to be constant in interval  $j$ , with respective values  $h_{1j}$ ,  $h_{2j}$ , and  $r_j$ , and if we consider only values of  $a$  and  $a + \tau$  in the set  $(0, \tau_1, \dots, \tau_{13})$ , then equation 5 becomes

$$P(a, \tau, r) = \sum_j \{h_{1j}r_j/(h_{1j}r_j + h_{2j})\} \{S_1(\tau_j - 1)/S_1(a)\} \times \{S_2(\tau_j - 1)/S_2(a)\} [1 - \exp\{-\Delta_j(h_{1j}r_j + h_{2j})\}], \quad [6]$$

where the smallest value of  $j$  in the summation satisfies  $\tau_j - 1 = a$  and the largest value of  $j$  satisfies  $\tau_j = a + \tau$ . Quantities like  $S_1(\tau_j - 1)$  in equation 6 are obtained recursively from formulas like  $S_1(\tau_j) = S_1(\tau_j - 1) \exp(-h_{1j}\Delta_j)$ .

## Results

### BCDDP Composite and Baseline Rates

The composite age-specific rates  $h^*(t)$  are consistently higher in the BCDDP population than in the SEER population, except for those women over age 70 (table 3). This difference can be explained partly in terms of self-selection and partly because screening shifts the age at diagnosis to earlier ages. Indeed 140 (10.3%) of the 1,354 detected cancers were in situ, and some of these might never have been detected in routine follow-up. About 4.5% of breast cancers reported to the SEER Program are classified as in situ.

## Projection of Individualized Probabilities of Breast Cancer

Table 4 contains estimates of the probabilities of developing breast cancer over specified follow-up periods,  $\tau$ , for subjects with given age,  $a$ , and relative risks. If both  $a$  and  $a + \tau$  are less than or equal to 50, or if both  $a$  and  $a + \tau$  are 50 or more, only an "initial relative risk," namely, the relative risk at age  $a$ , need be specified. However, if  $a < 50$  and  $a + \tau > 50$ , one must specify the relative risk at age  $a$  (the initial relative risk) and the relative risk at age 50 (the "later relative risk").

As an example, suppose a 40-year-old woman with menarche at age 12, with one previous biopsy for benign breast disease, with no children, and without a history of breast cancer in her mother or sisters asks what her chances are of developing breast cancer within 30 years. From table 1, we estimate the initial relative risk at age 40 of such a person to be  $1.099 \times 1.698 \times 1.548 = 2.89$ . At age 50, the final relative risk drops to  $1.099 \times 1.273 \times 1.548 = 2.17$ , if we assume that no additional risk factors have developed. From table 4, the chance of developing disease is 8.6% for a person with initial and final relative risks of 2.0, 11.9% for initial relative risk of 5.0 and final relative risk of 2.0, 17.0% for initial relative risk of 2.0 and final relative risk of 5.0, and 20.0% for

initial and final relative risks of 5.0. A rough guess near the point (initial relative risk, final relative risk) = (2.0, 2.0) is, therefore, somewhat greater than 8.6%. A more refined estimate is obtained from bivariate linear interpolation starting from the relative risk point (2.0, 2.0), which is nearest to the two relative risks (2.89, 2.17) in our example. The interpolated estimate is then  $8.6 + (2.89 - 2.0)(11.9 - 8.6)/(5.0 - 2.0) + (2.17 - 2.0)(17 - 8.6)/(5.0 - 2.0) = 10.1\%$ . For a follow-up of 10 years, only the initial relative risk of 2.89 need be specified. From linear interpolation in table 4, the estimated risk would be  $2.5 + (2.89 - 2.0)(6.1 - 2.5)/(5.0 - 2.0) = 3.6\%$ . Likewise, had the woman presented at age  $a = 50$  with the same history of risk factors, one would need only to specify her initial relative risk, 2.17. Over 10 years of follow-up, her risk would be estimated as  $3.1 + (2.17 - 2.0)(7.6 - 3.1)/(5.0 - 2.0) = 3.4\%$ .

## Uncertainty in Projections

There are three major sources of uncertainty in these calculations. The first source is the extent to which the BCDDP population is representative of other women who might come to a physician for counseling. We consider this issue of the generalizability of results from the BCDDP in the Discussion section.

Table 4. Projected probability (%) of developing breast cancer within 10, 20, or 30 years of follow-up

| Initial age (yr) | Years of follow-up | Later relative risk* | Initial relative risk*† |      |      |      |      |      |
|------------------|--------------------|----------------------|-------------------------|------|------|------|------|------|
|                  |                    |                      | 1.0                     | 2.0  | 5.0  | 10.0 | 20.0 | 30.0 |
| 20               | 10                 |                      | 0.0                     | 0.1  | 0.2  | 0.5  | 1.0  | 1.4  |
|                  | 20                 |                      | 0.5                     | 1.0  | 2.5  | 4.9  | 9.5  | 14.0 |
|                  | 30                 |                      | 1.7                     | 3.4  | 8.3  | 15.9 | 29.3 | 40.5 |
| 30               | 10                 |                      | 0.5                     | 0.9  | 2.3  | 4.4  | 8.7  | 12.8 |
|                  | 20                 |                      | 1.7                     | 3.3  | 8.1  | 15.6 | 28.8 | 39.9 |
|                  | 30                 | 1.0                  | 3.2                     | 4.8  | 9.5  | 16.9 | 29.9 | 40.8 |
|                  |                    | 2.0                  | 4.7                     | 6.3  | 10.9 | 18.2 | 30.9 | 41.7 |
|                  |                    | 5.0                  | 8.9                     | 10.4 | 14.9 | 21.8 | 34.0 | 44.3 |
|                  |                    | 10.0                 | 15.6                    | 17.1 | 21.2 | 27.6 | 38.8 | 48.3 |
|                  |                    | 20.0                 | 27.6                    | 28.8 | 32.3 | 37.8 | 47.4 | 55.5 |
|                  |                    | 30.0                 | 37.7                    | 38.7 | 41.8 | 46.4 | 54.7 | 61.7 |
| 40               | 10                 |                      | 1.2                     | 2.5  | 6.1  | 11.8 | 22.2 | 31.3 |
|                  | 20                 | 1.0                  | 2.8                     | 4.0  | 7.5  | 13.1 | 23.4 | 32.4 |
|                  |                    | 2.0                  | 4.3                     | 5.5  | 8.9  | 14.5 | 24.5 | 33.4 |
|                  |                    | 5.0                  | 8.6                     | 9.7  | 13.1 | 18.3 | 28.0 | 36.4 |
|                  |                    | 10.0                 | 15.4                    | 16.4 | 19.5 | 24.4 | 33.3 | 41.1 |
|                  |                    | 20.0                 | 27.4                    | 28.4 | 30.9 | 35.2 | 42.7 | 49.5 |
|                  |                    | 30.0                 | 37.7                    | 38.5 | 40.7 | 44.3 | 50.8 | 56.6 |
|                  | 30                 | 1.0                  | 4.4                     | 5.6  | 9.1  | 14.6 | 24.6 | 33.5 |
|                  |                    | 2.0                  | 7.4                     | 8.6  | 11.9 | 17.3 | 27.0 | 35.6 |
|                  |                    | 5.0                  | 15.9                    | 17.0 | 20.0 | 24.9 | 33.7 | 41.5 |
|                  |                    | 10.0                 | 28.3                    | 29.2 | 31.8 | 35.9 | 43.4 | 50.0 |
|                  |                    | 20.0                 | 47.5                    | 48.1 | 50.0 | 53.1 | 58.5 | 63.4 |
|                  |                    | 30.0                 | 61.2                    | 61.6 | 63.1 | 65.3 | 69.3 | 72.8 |
| 50               | 10                 |                      | 1.6                     | 3.1  | 7.6  | 14.6 | 27.1 | 37.7 |
|                  | 20                 |                      | 3.2                     | 6.4  | 15.1 | 27.9 | 47.8 | 61.9 |
|                  | 30                 |                      | 4.4                     | 8.5  | 19.9 | 35.5 | 57.8 | 71.7 |
| 60               | 10                 |                      | 1.8                     | 3.6  | 8.6  | 16.5 | 30.1 | 41.5 |
|                  | 20                 |                      | 3.0                     | 5.9  | 14.0 | 25.9 | 44.6 | 58.2 |
| 70               | 10                 |                      | 1.4                     | 2.7  | 6.7  | 12.9 | 24.1 | 33.7 |

\*The initial relative risk corresponds to the initial age  $a$ . If the initial age is  $< 50$  and if the initial age plus the follow-up specified is  $> 50$ , then a later relative risk at age 50 should also be specified. If the initial age is  $\geq 50$ , only the initial relative risk is required. If the initial age is  $< 50$  and if the initial age plus the years of follow-up does not exceed 50, then only an initial relative risk is required.

†Values in columns are projected probabilities expressed in percents.

The second source of uncertainty is that equation 1 may be misspecified by the inclusion of inappropriate terms, for example. We arrived at equation 1 by carefully studying known risk factors and their interactions and by allowing the relative risks associated with particular factors to vary with age (no proportional hazards assumption). We balanced the desire to fit the available BCDDP data well with the realization that elaborate models suited to one data set are often unreliable (30). It is likely, however, that other data analysts would arrive at somewhat different models. To the extent that additional information could define a truly better model, our procedures are subject to systematic error.

A third source of uncertainty arises from random error, even if the form of the risk equation 1 is correctly specified. Random fluctuations affect our estimates of baseline hazard,  $h_1(t)$ , and our estimates of relative risk coefficients (tables 1 and 2 and equation 1). Preliminary calculations suggest that the latter is much more important. Therefore, we studied the effects of random fluctuations in relative risks by the following simulation procedure. We generated 1,000 random samples of sets of log relative odds coefficients,  $\beta$ , from a multivariate normal distribution with means and covariances as in table 2. For a given type of individual, such as subject 1 or subject 2 in table 5, each of these sets of  $\beta$  coefficients is used to estimate the probability of developing breast cancer through equation 6. Estimated 95% confidence limits on the projected probability of developing breast cancer are the 25th- and 975th-order statistics from this sample, respectively. This procedure for obtaining confidence intervals is called a "parametric bootstrap" (31), and table 5 gives the results for two hypothetical subjects. These confidence intervals do not take into account the smaller uncertainty associated with estimating the baseline hazard.

Subject 1 in table 5 corresponds to a woman with NBIOPS = 1, AGEMEN = 1, AGEFLB = 2, and NUMREL = 0 both before and at age 50. We discussed this woman previously. We estimated that her absolute risks ranged from 1.3% to 10.1% (table 5), depending on the time interval. Note that the ratio of half the width of the confidence interval to the original risk estimate is in the range 0.18–0.27, with slightly more uncertainty above than below the original value. For subject 2, who has a greater risk primarily because she has two affected relatives (table 5), the absolute risks are much higher, and the ratio of half the width of the confidence interval to the original estimate ranges from 0.44 to 0.64,

which is large. Moreover, the confidence intervals (table 5) are asymmetrically situated about the original estimate and include a wider range of higher risks than lower risks.

## Discussion

The projected probabilities in table 4 are intended to provide an aid to decision-making and, despite their considerable uncertainty, may serve as a useful complement to clinical experience and intuition. Major improvements could be made if we better understood risk factors for breast cancer. Some progress in this direction has been reported by Dupont and Page (32–34) and others, who have identified atypical hyperplasia (AH) in biopsy specimens as an important risk factor. We could not incorporate AH in our model because we did not have pathologic assessments of specimens obtained from biopsies performed before 1973. The risk factors that we selected and the relative risks associated with these factors are broadly consistent with those of other population-based studies (35–37).

We now suggest a simple adjustment for using table 1 for a woman with AH. We used data from 16,692 subjects with biopsy-proven benign breast disease [see table 1 of (24)]. These women were found to have benign breast disease at screening during the BCDDP from 1973 to 1979 and were subsequently followed through February 1986 for detection of breast cancer. We analyzed data from 465 women who developed breast cancer more than 6 months after the first biopsy-proven diagnosis of benign breast disease and from 15,849 women who did not. The remaining  $16,692 - 465 = 15,849 = 378$  women were excluded because we lacked data for them on one or more of the following factors included in a logistic model of the odds of developing breast cancer: AGEFLB, NBIOPS, NUMREL, AGEMEN, AGE (0 if  $<50$ , 1 if  $\geq 50$ ), and AH (1 if at least one biopsy in 1973–1979 yielded a diagnosis of AH and 0 otherwise). These factors were coded as in table 1 but were evaluated from a questionnaire in 1979. For example, NBIOPS refers to the total number of previous biopsies as of 1979. The resulting estimated relative risk for AH was 1.96 with 95% confidence interval (1.51, 2.54). As noted in (24), this relative risk is somewhat smaller than that observed in earlier studies (32–34), probably because less stringent criteria for AH were employed in the later study. The theory of attributable risk calculations (38) indicates that alternative, mutually exclusive, and

Table 5. Estimates of probability (%) of developing breast cancer, with 95% confidence intervals

| Age interval<br>(yr) | Subject 1*  |                     | Subject 2†  |                     |
|----------------------|-------------|---------------------|-------------|---------------------|
|                      | Probability | Confidence interval | Probability | Confidence interval |
| 30–40                | 1.3         | 1.0, 1.7            | 6.1         | 3.5, 11.3           |
| 30–60                | 7.9         | 6.6, 9.8            | 32.7        | 19.9, 51.7          |
| 40–50                | 3.6         | 2.8, 4.7            | 16.0        | 9.4, 28.2           |
| 40–70                | 10.1        | 8.4, 12.2           | 39.8        | 25.1, 60.5          |
| 50–60                | 3.4         | 2.8, 4.1            | 15.2        | 9.0, 26.4           |
| 50–80                | 9.2         | 7.7, 11.0           | 36.8        | 23.1, 56.6          |

\*Subject 1 had values NBIOPS = 1, AGEMEN = 1, AGEFLB = 2, and NUMREL = 0 before and at age 50.

†Subject 2 had values NBIOPS = 1, AGEMEN = 2, AGEFLB = 0, and NUMREL = 2 before and at age 50.

exhaustive partitionings of the "exposed" category will yield the same attributable risk, provided (a) that the "unexposed" category is unaltered and (b) that the risk models fit well. Because the presence of AH can be regarded as a further subdivision of the exposure category defined by NBIOPS in table 1, inclusion of AH in model 1 would not alter appreciably our estimates of the baseline age-specific breast cancer hazard rates from equation 4. Therefore, for women without previous breast biopsies, one could use tables 1 and 4 directly. For women with previous breast biopsies, the relative risks in table 1 correspond to a mixture of relative risks for those with and without AH. Using the 7.8% prevalence rate of AH among 16,692 women with biopsy-proven benign breast disease [table 4 in (24)], we estimate this mixture as  $\pi(0.078 \times 1.96 + 0.922) = 1.075\pi$ , where  $\pi$  is the relative risk in the absence of AH. Thus, for women with AH, one can multiply estimates of relative risk from table 1 by  $1.96/1.075 = 1.82$  before referring to table 4, and for women without AH, one can multiply these estimates by  $1/1.075 = 0.93$ . For practical purposes, one can use tables 1 and 4 directly for women whose breast biopsies reveal no AH.

Ottman et al. (4) estimated the probability of developing breast cancer by age 70 in mothers of breast cancer patients (the index case). Because bilateral disease is uncommon, we compare our results with their results for all index cases with unilateral disease, for which the mothers' cumulative risk is 6.5% [table 1 of (4)] with 95% confidence interval (2.6%, 10.4%). For sisters of such index cases [table 2 of (4)], the cumulative risk to age 70 is 17.4% with 95% confidence interval (1.1%, 33.7%). Note that the confidence intervals are extremely wide, which is typical of life tables based on small numbers. It is difficult to compare our estimates with these figures because Ottman et al. made no attempt to define or use other risk factors associated with the mother or sister of the index case, such as age at first live birth. However, if we assume that a woman with one involved relative had no other risk factors, she would have a relative risk of 2.607 (table 1) and, by linear interpolation in table 4, a cumulative risk from ages 40 to 70 of  $8.6 + (2.607 - 2) \times (11.9 - 8.6)/(5 - 2) + (2.607 - 2)(17 - 8.6)/(5 - 2) = 11.0\%$ , which is quite consistent with the results of Ottman et al. (4). Our figure of 11.0% is also in good agreement with the value 12.3% calculated from table 1 in (39). Similarly, Anderson and Badzioch (5) studied the cumulative risk to age 70 among women who had a sister with breast cancer (index patient) and one other immediate family member with breast cancer (index relative of the index patient). They reported a cumulative risk of 18% with 95% confidence interval (12%, 24%) if the index relative was a mother and 14% with 95% confidence interval (9%, 19%) if the index relative was another sister. Our analysis would lead to a relative risk of 6.798 for a woman with two or more family members with breast cancer and a cumulative risk from ages 40 to 70 of  $20 + (24.9 - 20)(6.798 - 5)/(10 - 5) + (31.8 - 20)(6.798 - 5)/(10 - 5) = 26\%$ , which is somewhat higher but is also subject to considerable variability. Compared with simple life-table methods, the techniques used in our study allow one to assess multiple risk factors; in addition, the projected risks as in table 4 are regular, compared with the erratic

estimates one finds with life tables based on small numbers of events (4,5). One obtains these desirable features as a result of model assumptions that must be examined critically.

The rates in table 3 have never before been calculated or published for the original BCDDP population, although rates for a related cohort with extended follow-up (40) and rates expressed in terms of cancers detected per 1,000 screenings (3) have been published.

The composite rates in table 3 were derived for women in the BCDDP who had passed the initial screening without cancer detection and who were in years 2 or 3 of follow-up. Therefore, the derived probabilities in table 4 are most applicable to a woman who had just had a normal examination, including mammography as in the BCDDP, and who would be willing to have annual examinations for an extended period. For women over age 65, there is little difference between BCDDP rates and SEER rates, which indicates that table 4 may be useful for older women under more sporadic follow-up like members of the SEER population.

In the age range 30-64, women in the general (SEER) population have lower rates than women in the BCDDP cohort (table 3). An approximate calculation of attributable risk with the use of data from the Cancer and Steroid Hormone Study (22,35) suggests that increased prevalence of the risk factors in table 1 in the BCDDP cohort can explain only a relative risk of about 1.17 in table 3. Thus, a considerable portion of the increased risk among young women in the BCDDP is due to the accelerated detection of cancers as a result of frequent screening (41). On the basis of such attributable risk calculations, we believe that if the data in tables 1 and 4 are applied to younger women under sporadic follow-up, the incidence of detected disease will be overestimated by roughly 10%-20%. Nonetheless, a woman at high relative risk who is receiving sporadic follow-up might wish to begin a program of annual examinations and adopt other preventive or surveillance measures on the basis of the results shown in table 4, since the intrinsic rate of progression to malignant disease is presumably no slower among women receiving sporadic follow-up than among those seen annually.

The methods reported in this article may be used to help design prevention trials in high-risk populations, because an important determinant of the required sample size is the absolute risk of developing breast cancer in such a population (42).

## References

- (1) MULVIHILL JJ, SAFYER AW, BENING JK: Prevention in familial breast cancer: Counseling and prophylactic mastectomy. *Prev Med* 11:500-511, 1982
- (2) HOSKINS IA: Genetic counseling for cancer patients and their families. *Oncology* 3:84-98, 1989
- (3) BAKER LH: Breast cancer detection demonstration project: Five-year summary report. *CA* 32:194-225, 1982
- (4) OTTMAN R, KING MC, PIKE MC, ET AL: Practical guide for estimating risk for familial breast cancer. *Lancet* 2:556-558, 1983
- (5) ANDERSON DE, BADZIOCH MD: Risk of familial breast cancer. *Cancer* 56:383-387, 1985
- (6) BRINTON LA, HOOVER RN, SZKLO M, ET AL: Menopausal estrogen use and risk of breast cancer. *Cancer* 47:2517-2522, 1981

- (7) BRINTON LA, HOOVER R, FRAUMENI JF JR: Interaction of familial and hormonal risk factors for breast cancer. *JNCI* 69:817-822, 1982
- (8) BRINTON LA, HOOVER RN, FRAUMENI JF JR: Epidemiology of minimal breast cancer. *JAMA* 249:483-487, 1983
- (9) BRINTON LA, HOOVER RN, FRAUMENI JF JR: Menopausal oestrogens and breast cancer risk: An expanded case-control study. *Br J Cancer* 54:825-832, 1986
- (10) BRUZZI P, GREEN SB, BYAR DP, ET AL: Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol* 122:904-914, 1985
- (11) STANFORD JL, BRINTON LA, HOOVER RN: Oral contraceptives and breast cancer: Results from an expanded case-control study. *Br J Cancer* 60:375-381, 1989
- (12) BRINTON LA, HOFFMAN DA, HOOVER R, ET AL: Relationship of thyroid disease and use of thyroid supplements to breast cancer risk. *J Chronic Dis* 37:877-893, 1984
- (13) STANFORD JL, MARTIN EJ, BRINTON LA, ET AL: Rauwolfia use and breast cancer: A case-control study. *JNCI* 76:817-822, 1986
- (14) KLEINERMAN RA, BRINTON LA, HOOVER R, ET AL: Diazepam use and progression of breast cancer. *Cancer Res* 44:1223-1225, 1984
- (15) BRINTON LA, SCHAIRER C, STANFORD JL, ET AL: Cigarette smoking and breast cancer. *Am J Epidemiol* 123:614-622, 1986
- (16) SCHAIRER C, BRINTON LA, HOOVER RN: Methylxanthines and breast cancer. *Int J Cancer* 40:469-473, 1987
- (17) HARVEY EB, SCHAIRER C, BRINTON LA, ET AL: Alcohol consumption and breast cancer. *JNCI* 78:657-661, 1987
- (18) SWANSON CA, BRINTON LA, TAYLOR PR, ET AL: Body size and breast cancer risk assessed in women participating in the breast cancer detection demonstration project. *Am J Epidemiol*. In press
- (19) BRINTON LA, HOOVER R, FRAUMENI JF JR: Reproductive factors in the etiology of breast cancer. *Br J Cancer* 47:757-762, 1983
- (20) BRINTON LA, SCHAIRER C, HOOVER RN, ET AL: Menstrual factors and risk of breast cancer. *Cancer Invest* 6:245-254, 1988
- (21) BRESLOW NE, DAY NE: *Statistical Methods in Cancer Research. Volume 1—The Analysis of Case-Control Studies*. Lyon: IARC, 1980, pp 5-338
- (22) CENTERS FOR DISEASE CONTROL CANCER AND STEROID HORMONE STUDY: Long-term oral contraceptive use and the risk of breast cancer. *JAMA* 249:1591-1595, 1983
- (23) COX DR: Regression models and life tables. *J R Stat Soc* 34:187-220, 1972
- (24) CARTER CL, CORLE DK, MICOZZI MS, ET AL: A prospective study of the development of breast cancer in 16,692 women with benign breast disease. *Am J Epidemiol* 128:467-477, 1988
- (25) NATIONAL CENTER FOR HEALTH STAT: *Vital Statistics in the United States. Mortality, vol II (Part A)*, 1979. Washington, DC: US Govt Print Off, 1984
- (26) GAIL MH: A review and critique of some models used in competing risk analysis. *Biometrics* 31:209-222, 1975
- (27) PRENTICE RL, KALBFLEISCH JD, PETERSON AV, ET AL: The analysis of failure times in the presence of competing risks. *Biometrics* 34:541-554, 1978
- (28) GAIL MH: Competing risks. In *The Encyclopedia of Statistical Sciences* (Kotz S, Johnson NL, eds), vol 2. New York: Wiley, 1982, pp 75-81
- (29) DUPONT WD: Converting relative risks to absolute risks: A graphical approach. *Stat Med* 8:641-651, 1989
- (30) AKAIKE H: Statistical predictor identification. *Ann Inst Stat Math* 22:203-217, 1970
- (31) EFRON B: The jackknife, the bootstrap and other resampling plans. In *CBMS Monograph 38*. Philadelphia: SIAM, 1982, pp 27-36
- (32) DUPONT WD, PAGE DL: Risk factors for breast cancer in women with proliferative breast disease. *N Engl J Med* 312:146-181, 1985
- (33) DUPONT WD, PAGE DL: Breast cancer risk associated with proliferative disease, age at first birth, and a family history of breast cancer. *Am J Epidemiol* 125:769-779, 1987
- (34) PAGE DL, DUPONT WD, ROGERS LW, ET AL: Atypical hyperplastic lesions of the female breast. *Cancer* 55:2698-2708, 1985
- (35) SATTIN RW, RUBIN GL, WEBSTER LA, ET AL: Family history and the risk of breast cancer. *JAMA* 253:1908-1913, 1985
- (36) SEIDMAN H, STELLMAN SD, MUSHINSKI MH: A different perspective on breast cancer risk factors: Some implications of the nonattributable risk. *CA* 32:301-313, 1982
- (37) LUBIN JH, BURNS PE, BLOT WJ, ET AL: Risk factors for breast cancer in women in northern Alberta, Canada, as related to age at diagnosis. *JNCI* 68:211-217, 1982
- (38) WALTER SD: The estimation and interpretation of attributable risk in health research. *Biometrics* 32:829-849, 1976
- (39) SCHWARTZ AG, KING MC, BELLE SH, ET AL: Risk of breast cancer to relatives of young breast cancer patients. *JNCI* 75:665-668, 1985
- (40) MORRISON AS, BRISSON J, KHALID N: Breast cancer incidence and mortality in the breast cancer detection demonstration project. *J Natl Cancer Inst* 80:1540-1547, 1988
- (41) DAY NE, WALTER SD, TABAR L, ET AL: The sensitivity and lead time of breast cancer screening: A comparison of the results of different studies. In *Screening for Breast Cancer* (Day NE, Miller AB, eds). Toronto: Huber, 1988, pp 105-109
- (42) SELF S, PRENTICE R, IVERSON D, ET AL: Statistical design of the Women's Health Trial. *Controlled Clin Trials* 9:119-136, 1988

**NEW!**

## CHEW OR SNUFF IS REAL BAD STUFF

This free four-color brochure informs adolescents about the adverse health and social effects of using chewing tobacco and snuff. The brochure, with a foldout poster, is available in quantities for schools, health professionals, and community organizations.

To order, write the National Cancer Institute, Building 31, Room 10A24, Bethesda, MD 20892, or call the Cancer Information Service toll-free at 1-800-4-CANCER.

