Validation of the Gail et al. Model of Breast Cancer Risk Prediction and Implications for Chemoprevention

Beverly Rockhill, Donna Spiegelman, Celia Byrne, David J. Hunter, Graham A. Colditz

Background: **Women and their clinicians are increasingly encouraged to use risk estimates derived from statistical models, primarily that of Gail et al., to aid decision making regarding potential prevention options for breast cancer, including chemoprevention with tamoxifen.** ***Methods:*** **We evaluated both the goodness of fit of the Gail et al. model 2 that predicts the risk of developing invasive breast cancer specifically and its discriminatory accuracy at the individual level in the Nurses' Health Study. We began with a cohort of 82 109 white women aged 45–71 years in 1992 and applied the model of Gail et al. to these women over a 5-year follow-up period to estimate a 5-year risk of invasive breast cancer. All statistical tests were two-sided.** ***Results:*** **The model fit well in the total sample (ratio of expected [E] to observed [O] numbers of cases = 0.94; 95% confidence interval [CI] = 0.89 to 0.99). Underprediction was slightly greater for younger women (<60 years), but in most age and risk factor strata, E/O ratios were close to 1.0. The model fit equally well (E/O ratio = 0.93; 95% CI = 0.87 to 0.99) in a subset of women reporting recent screening (i.e., within 1 year before the baseline); among women with an estimated 5-year risk of developing invasive breast cancer of 1.67% or greater, the E/O ratio was 1.04 (95% CI = 0.96 to 1.12). The concordance statistic, which indicates discriminatory accuracy, for the Gail et al. model 2 when used to estimate 5-year risk was 0.58 (95% CI = 0.56 to 0.60). Only 3.3% of the 1354 cases of breast cancer observed in the cohort arose among women who fell into age–risk strata expected to have statistically significant net health benefits from prophylactic tamoxifen use.** ***Conclusions:*** **The Gail et al. model 2 fit well in this sample in terms of predicting numbers of breast cancer cases in specific risk factor strata but had modest discriminatory accuracy at the individual level. This finding has implications for use of the model in clinical counseling of individual women. [J Natl Cancer Inst 2001;93:358–66]**

There has been growing interest in developing methods to use a woman's risk factor profile to estimate her risk of breast cancer. Women and their clinicians are increasingly encouraged to use risk estimates derived from statistical models, primarily that of Gail et al. *(1),* to aid decision making regarding potential breast cancer prevention options, including chemoprevention with tamoxifen.

The original model of Gail et al. ["model 1" *(1)*], developed in 1989 among a case–control study subsample of regularly screened women participating in the Breast Cancer Detection and Demonstration Project (BCDDP), estimates the absolute risk (probability) that a woman in a program of annual screening will develop invasive or *in situ* breast cancer over a defined age interval. Statisticians modified the original Gail et al. model to predict specifically the risk of developing invasive breast cancer. This modified model, referred to as "model 2" *(2),* was used to determine eligibility for the Breast Cancer Prevention Trial (BCPT) *(3).* The modification of model 1 to model 2 was accomplished by substituting age-specific invasive breast cancer rates for white women from the Surveillance, Epidemiology, and End Results (SEER)[1] Program *(4)* for the breast cancer incidence rates observed in the BCDDP and by use of attributable risk estimates from SEER to obtain the baseline hazard rates *(2).* We focus on this Gail et al. model 2 in our validation analysis.

We consider not only the calibration of the Gail et al. model 2 (i.e., its ability to predict incidence in groups of women, often referred to as goodness of fit) but also its discriminatory accuracy, the ability to separate individuals who will go on to develop different outcomes *(5).* The discriminatory accuracy of statistical models is rarely discussed, yet if a model is indicated for use in the clinical setting to separate individual patients into distinct groups, as is the Gail et al. model 2, such accuracy is relevant. Currently, the U.S. Food and Drug Administration (FDA) guidelines state that women aged 35 years and older with a 5-year risk of breast cancer of 1.67% or greater, as estimated by the Gail et al. model 2, are eligible for prophylactic use of tamoxifen. The manufacturer of Nolvadex (tamoxifen), AstraZeneca Pharmaceuticals LP, Zeneca Inc., Wilmington, DE, now advertises this 1.67% risk cut point as "the line" in breast cancer risk *(6),* implying that this cut point can be used to meaningfully segregate high- and low-risk individuals.

If the average predicted risk for a group of individuals with a certain risk factor profile is 0.10 and the actual proportion of persons in this group who develop disease over the considered time interval is 0.10, the model's predictions are well calibrated, and the model is said to fit well. A discriminating model is one that produces a wide distribution of estimated probabilities and whose estimated probabilities for persons who actually develop disease are consistently larger than the probabilities for individuals who remain disease free. A model that assigns everyone in a

Affiliations of authors: B. Rockhill, C. Byrne, Channing Laboratory, Harvard Medical School and Brigham and Women's Hospital, Boston, MA; D. Spiegelman, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston; D. J. Hunter, G. A. Colditz, Channing Laboratory, Harvard Medical School and Brigham and Women's Hospital, and Department of Epidemiology, Harvard School of Public Health.

Correspondence to: Beverly Rockhill, Ph.D., Channing Laboratory, 181 Longwood Ave., Boston, MA 02115 (e-mail: beverly.rockhill@channing.harvard.edu).

See "Notes" following "References."

population the same estimated probability of disease, say 0.05, is well calibrated if 5% of the population actually develops disease, but the individual estimated probabilities are uninformative because the model has no discriminatory ability *(5)*.

To our knowledge, the discriminatory accuracy of neither the original nor the revised model of Gail et al. has been described. The calibration of the Gail et al. model 1 has been assessed in at least four independent populations *(2,7–9)*. The model was found to overpredict among younger women who were not in a program of regular screening *(7,9)*. The Gail et al. model 1 performed better at younger ages, in terms of agreement between expected and observed numbers of cases, among women who were being screened regularly *(2,8)*.

Costantino et al. *(2)* recently assessed the validity of model 2 with respect to a 5-year risk prediction among white women in the placebo arm of the BCPT who had no history of lobular carcinoma *in situ* and who were receiving annual screening. Overall, the expected (E)/observed (O) ratio was 1.03 (95% confidence interval [CI] = 0.88 to 1.21), indicating that the model was well calibrated in this population.

Here, we evaluate the performance of the Gail et al. model 2 with respect to predicting 5-year risk of invasive breast cancer among white women in the Nurses' Health Study (NHS) cohort. We assess model calibration as well as discriminatory accuracy.

Subjects and Methods

Details of the algorithm and parameters of the Gail et al. models 1 and 2 have been reported previously *(1,2,10)*. Both models 1 and 2 contain the following risk factors: age at menarche (≥14 years, 12–13 years, or <12 years), number of previous breast biopsies (0, 1, or ≥2), presence of atypical hyperplasia in a biopsy specimen (yes or no), age at first live birth (<20 years, 20–24 years, 25–29 years or nulliparous, or ≥30 years), number of first-degree relatives with a history of breast cancer (none, mother or sister, or mother and sister), and age. Both models include interaction terms between number of biopsies and age category (<50 years and ≥50 years) and between age at first birth and number of affected first-degree relatives. As mentioned above, model 2 predicts only invasive breast cancer and employs SEER-derived, age-specific rates of breast cancer rather than rates derived from the BCDDP. Both models allow calculation of the probability of developing breast cancer for a woman of any age between 20 and 80 years and with any pattern of the above risk factors over any specified time interval.

To obtain 5-year risk estimates from the Gail et al. model 2, we employed a FORTRAN code (BCPT.FOR, May 12, 2000) obtained from the National Cancer Institute (NCI) (Gail M, Benichou J, Pee D [Information Management Services, Rockville, MD]: personal communication). This code is the underlying calculating machine, the "Driver Routine," used by the NCI Risk Disk. We conducted extensive checking of the risk estimates obtained from this FORTRAN code for our sample against the risk estimates provided by the interactive Risk Disk for women with identical risk factor profiles and found no discrepancies.

We conducted our validation study in an independent population of women from the NHS. This prospective cohort study began in 1976, when 121 700 married women who were registered nurses aged 30–55 years returned a detailed questionnaire on medical history and lifestyle factors. Subsequent questionnaires have been mailed every 2 years. The population is predominantly (>98%) white, reflecting the ethnic/racial makeup of women who trained as registered nurses in the 1940s, 1950s, and 1960s. Further details on the study, including information on disease confirmation, are published elsewhere *(11,12)*. The protocol for the study was approved by the Human Subjects Research Committee of the Brigham and Women's Hospital (Boston, MA).

We restricted our analyses to the 5-year period from 1992 to 1997 to correspond to the period of time during which the BCPT was conducted. (When the original model of Gail et al. was revised for the BCPT, part of the revision was an update of the incidence rates used to estimate the expected baseline risk of breast cancer that would be consistent with the time period of the trial, from 1992 forward.) There were 104 064 women who responded to the 1992 NHS questionnaire. Of these, 95 743 women either were followed disease free through the 1998 survey or developed breast cancer within 5 years of the 1992 survey. Of these 95 743 women, 85 468 had no history of major disease (cardiovascular disease or cancer, including lobular or ductal carcinoma *in situ*) at baseline. We restricted this sample to white women only who had complete data on the required risk factors in 1992; we were thus left with a starting cohort of 81 209 women aged 46–71 years in 1992. From this pool, 1354 women developed breast cancer within 5 years of the return date of their 1992 questionnaire.

On the basis of our questionnaire data, we were able to classify women with regard to only ever/never history of previous benign biopsy, rather than 0, 1, or greater than or equal to two biopsies as specified in the Gail model. Therefore, all women who reported ever having a biopsy were treated in the model as if they had reported one biopsy. Only a small subset of women in the NHS who reported previous biopsies have been classified with regard to presence of atypical hyperplasia. These women were part of a nested case–control study of benign breast disease and breast cancer. The details of the selection of this subset from the whole cohort, as well as the methods of classification of atypical hyperplasia status, are described elsewhere *(13)*. Of the control women with biopsy specimens, 10.3% were identified as having atypical hyperplasia. Thus, we estimate that approximately 2.1% (20.3% with history of biopsy × 10.3% with atypical hyperplasia) of our baseline cohort for this analysis had atypical hyperplasia. Although we were unable to consider atypical hyperplasia status in the main analyses using the entire cohort, and thus assigned all women with a history of benign breast biopsy to the unknown atypical hyperplasia status category, we conducted secondary analyses on the subset of women (n = 83) known to have atypical hyperplasia to examine their risk of breast cancer as estimated by the Gail et al. model 2. The Gail model treats women who have an unknown atypical hyperplasia status because they have not had a biopsy, or because they don't remember ever having a biopsy, exactly as women with no history of atypical hyperplasia, with respect to risk estimation. However, women who have had at least one biopsy but their atypical hyperplasia status is unknown because they don't know the results of the pathologic assessments are assigned a slightly higher estimated risk than those with no history of atypical hyperplasia.

We compared the expected and observed numbers of breast cancer case patients by age group as well as by risk factor category, including categories of the interaction terms specified in the Gail et al. model 2. We present data on expected and observed numbers of cases, not only for the total sample but also for the subset of the sample who reported in 1992 that they had received a screening mammogram within 1 year before the baseline (approximately two thirds of the baseline sample of 82 109, or 55 301, women reported such a screening mammogram). We also present data on the expected and observed numbers of cases stratified by estimated 5-year risk (<1.67% and ≥1.67%). The analysis on women with an estimated risk greater than or equal to 1.67% (n = 27 225) allows for more ready comparison with the results of Costantino et al. *(2)*, because a 5-year risk of 1.67% or higher was a criterion for entry into the BCPT. The expected number of cases over the 5-year period was calculated by summing the estimated individual risk for each woman predicted by the models, given the covariate values for each woman at the 1992 baseline. The 95% CIs for the E/O ratios were calculated with the use of the Poisson variance for the logarithm of the observed number of cases as follows:

$$95\%\ \text{CIs interval for E/O ratio} = \frac{E}{O}\exp^{\pm 1.96^*\sqrt{\frac{1}{O}}}.$$

We evaluated the discriminatory accuracy of the 5-year risk prediction in several ways. First, we estimated the concordance statistic, an index of predictive discrimination based on the rank correlation between predicted and observed outcomes *(14)*. Potential values of the concordance statistic range from 0.5 to 1.0. The value of the statistic represents the probability that, for a randomly selected pair of individuals, one diseased and one nondiseased, the diseased individual has the higher estimated disease probability. A concordance statistic of 0.5 for a risk model means that the model producing the estimated probabilities performs no better than chance at ranking diseased and nondiseased individuals; 50% of the time the diseased person will have the higher estimated probability, while 50% of the time the nondiseased person will. A concordance statistic of 1.0 means that the model performs perfectly at ranking diseased and nondiseased individuals. The concordance statistic is exactly equivalent to the area under a receiver-operating characteristic curve created by computing sensitivity and specificity, with respect to true disease outcome, at all estimated risk cut points from 0 to 1.0. To produce the concordance statistic, we used a logistic

regression model to regress breast cancer status at the end of the 5-year follow-up on the corresponding estimated risks from the Gail et al. model 2.

Finally, we grouped the estimated 5-year risks into deciles and computed the relative risk for being diagnosed with breast cancer during follow-up, comparing women in the highest decile of estimated risk with those in the lowest. All statistical tests were two-sided.

RESULTS

Table 1 shows the age distribution of the NHS sample in 1992 along with percentages of the sample in the various risk factor strata considered in the Gail et al. model 2. The relative risks from the NHS data are also presented, alongside the relative risks from the BCDDP, which are used in the Gail et al. model 2. In computing these relative risks, we used data from 1986 through 1997, to have a larger number of cases. The differences between the relative risks from the two samples are largest for the cross-classified (age at first birth × number of first-degree relatives) exposure strata pertaining to two or more first-degree relatives with breast cancer. The relative risks in these strata for both of the samples are imprecise; the proportions of women in both samples who fell into these strata were small.

The range of estimated 5-year risks was 0.56%–10.14% for women who remained disease free. Risk estimates covered a somewhat smaller range among the women who developed disease, from 0.54% to 7.51%. The median estimated 5-year risk in women who remained disease free was 1.41%, slightly lower than that of women who developed breast cancer (1.58%).

The range of 5-year estimated risks among the subset of women known to have atypical hyperplasia was smaller than those ranges given above. Among the 83 women who were diagnosed with atypical hyperplasia by 1992 and who also had complete data on the other risk factors in the model, the range of the 5-year estimated risk was 1.08%–6.68%. The median 5-year risk was 1.99%.

Based on the above data and on our previously stated estimate that approximately 10% of all women with a history of biopsy have atypical hyperplasia (amounting to approximately 2.1% of the entire cohort), we can infer that the distribution of estimated 5-year risk in this cohort is little affected by knowledge of atypical hyperplasia status. Our results below thus pertain to the entire cohort, where all of the women who had biopsies were

Table 1. Prevalence of breast cancer risk factors in the Nurses' Health Study (NHS) cohort in 1992 and relative risks from the NHS and the Breast Cancer Detection and Demonstration Project (BCDDP) *(1)* for variables in the Gail et al. *(1)* model 2

	% of women in 1992 in category (n = 82 109)	Relative risk, NHS* (95% CIs)	Relative risk, BCDDP†
Age group, y			
45–49	15.1		
50–54	21.2		
55–59	21.5		
60–64	19.2		
65–69	18.5		
70–74	4.4		
Age at menarche, y			
≥14	19.9	1.00 (referent)	1.00 (referent)
12–13	57.8	1.05 (0.99 to 1.10)	1.10
<12	22.4	1.10 (0.99 to 1.23)	1.21
No. of biopsies, age <50 y			
0	12.3	1.00 (referent)	1.00 (referent)
1	2.8	1.80 (1.60 to 2.06)	1.70
2	—	—	2.88
No. of biopsies, age ≥50 y			
0	67.4	1.00 (referent)	1.00 (referent)
1	17.5	1.62 (1.42 to 1.85)	1.27
2	—	—	1.62
No. of affected first-degree relatives, age at first birth <20 y			
0	0.9	1.00 (referent)	1.00 (referent)
1	0.1	1.59 (1.22 to 2.07)	2.61
≥2	0.0 (n = 2)	2.52 (1.49 to 4.27)	6.80‡
No. of affected first-degree relatives, age at first birth 20–24 y			
0	47.5	1.16 (1.10 to 1.23)	1.24
1	4.8	1.80 (1.53 to 2.12)	2.68
≥2	0.2	2.81 (1.76 to 4.49)	5.78§
No. of affected first-degree relatives, age at first birth 25–29 y or nulliparous			
0	34.8	1.33 (1.18 to 1.49)	1.55
1	3.8	2.04 (1.76 to 2.36)	2.76
≥2	0.2	3.12 (2.46 to 3.95)	4.91
No. of affected first-degree relatives, age at first birth ≥30 y			
0	7.0	1.54 (1.30 to 1.83)	1.93
1	0.8	2.32 (1.84 to 2.93)	2.83
≥2	0.0 (n = 33)	3.48 (2.29 to 5.30)	4.17‖

*Data from 1986–1997 used in estimating relative risks from NHS data.

†The 95% confidence intervals (CIs) are not provided in *(1)*.

‡Relative risk from BCDDP based on eight case patients and no control subjects.

§Relative risk from BCDDP based on 25 case patients and five control subjects.

‖Relative risk from BCDDP based on 19 case patients and six control subjects.

treated as "atypical hyperplasia status unknown" as mentioned previously.

Table 2 presents data on observed and expected numbers of cases in the full sample. The overall ratio of expected to observed number of cases during 5 years of follow-up was 0.94 (95% CI = 0.89 to 0.99). There was little variation by age group, although there was a slightly greater underprediction in the younger age groups (<60 years). The CIs pertaining to the E/O ratios in the age groups were all fairly narrow, and all but one included the value of 1.0. The E/O ratios that deviated most from 1.0 pertained to the strata with small numbers of cases, and thus the ratios were quite imprecise. For instance, there were 12 cases that occurred in the 5-year period to women with a first birth before age 20 years and with no affected first-degree relatives; 6.62 cases were expected, for an E/O ratio of 0.55 (95% CI = 0.31 to 0.97). Only one case was observed among women with an age at first birth of 20–24 years and two or more affected first-degree relatives, while 8.23 were expected, leading to the largest E/O ratio in Table 2, 8.23 (95% CI = 1.16 to 58.40).

Table 3 presents the same data as in Table 2, now restricted to the subset of women who reported in 1992 that they had had a screening mammogram in the previous year. In this subset, the overall E/O ratio was 0.93 (95% CI = 0.87 to 0.99), very close to that found in the full sample. The stratum-specific E/O ratios were also very similar between this subset and the full sample. Again, the E/O ratios that deviated most from 1.0 were those pertaining to the strata with small numbers of cases. We observed no cases in the stratum of women with an age at first birth of less than 20 years and two or more first-degree relatives.

Tables 4 and 5 present data on expected and observed numbers of cases for the total sample, stratified on an estimated 5-year risk. Table 4 contains data on women with an estimated risk of less than 1.67%. The underprediction in this sample was somewhat greater in the total sample; the overall E/O ratio was 0.86 (95% CI = 0.80 to 0.92). There was no consistent pattern in the E/O ratios by age. Again, the most extreme ratios occurred in the sparse risk factor strata; in many strata, the E/O ratio could not be estimated because there were no observed and/or no expected cases. In this low-risk subsample, not surprisingly, there were few women with a family history of breast cancer. Table 5 contains data on women with an estimated 5-year risk of 1.67% or greater. In this subsample, there was a slight overprediction. The overall E/O ratio in this subsample was 1.04 (95% CI = 0.96 to 1.12), the same E/O ratio that was observed in the validation study by Costantino et al. *(2)* among women of similar high-risk status. Once again, in our sample, there was no consistent pattern in the ratios by age, and the highest E/O ratios were found in the sparsest strata.

Table 2. Ratios of expected (E) and observed (O) numbers of breast cancer cases in the Nurses' Health Study (NHS) based on 5-year (1992 to 1997) risk prediction of Gail et al. *(1)* model 2 in a total sample (n = 82 109)

	Observed No. of breast cancer cases in NHS over a 5-y follow-up (n = 1354)	Expected No. of cases over a 5-y period	E/O ratio	95% confidence interval
Overall	1354	1273.42	0.94	0.89 to 0.99
Age group, y				
45–49	142	128.76	0.91	0.77 to 1.07
50–54	235	208.73	0.89	0.78 to 1.01
55–59	295	261.08	0.89	0.79 to 0.99
60–64	291	284.01	0.98	0.87 to 1.09
65–69	313	311.06	0.99	0.89 to 1.11
70–74	78	79.78	1.02	0.82 to 1.28
Age at menarche, y				
⩾14	290	240.84	0.83	0.74 to 0.93
12–13	741	730.52	0.99	0.92 to 1.06
<12	323	302.06	0.94	0.84 to 1.04
No. of biopsies, age <50 y				
0	104	95.94	0.92	0.76 to 1.12
1	38	32.82	0.86	0.63 to 1.19
No. of biopsies, age ⩾50, y				
0	884	853.23	0.97	0.90 to 1.03
1	328	291.43	0.89	0.80 to 0.99
No. of affected first-degree relatives, age at first birth <20 y				
0	12.0	6.62	0.55	0.31 to 0.97
1	2.0	1.77	0.89	0.22 to 3.55
⩾2	0	0.17	—	—
No. of affected first-degree relatives, age at first birth 20–24 y				
0	558.0	466.55	0.84	0.77 to 0.91
1	94.0	106.47	1.13	0.93 to 1.39
⩾2	1.0	8.23	8.23	1.16 to 58.40
No. of affected first-degree relatives, age at first birth 25–29 y or nulliparous				
0	469.0	445.37	0.95	0.84 to 1.04
1	84.0	90.90	1.08	0.87 to 1.34
⩾2	5.0	7.08	1.42	0.59 to 3.40
No. of affected first-degree relatives, age at first birth ⩾30 y				
0	108.0	117.15	1.08	0.90 to 1.31
1	20.0	21.46	1.07	0.69 to 1.66
⩾2	1.0	1.65	1.65	0.23 to 11.68

Table 3. Ratios of expected (E) and observed (O) number of breast cancer cases in the Nurses' Health Study (NHS) based on 5-year (1992 to 1997) risk prediction of Gail et al. *(1)* model 2 in a recently screened (within 1 year before baseline) sample (n = 55 301)

	No. of breast cancer cases observed in NHS over a 5-y follow-up (n = 941)	Expected No. of cases over a 5-y period	E/O ratio	95% confidence interval
Overall	941	875.32	0.93	0.87 to 0.99
Age group, y				
45–49	97	87.04	0.90	0.74 to 1.09
50–54	158	144.70	0.92	0.78 to 1.07
55–59	201	183.99	0.92	0.80 to 1.05
60–64	212	195.45	0.92	0.80 to 1.05
65–69	222	214.39	0.97	0.85 to 1.10
70–74	51	49.75	0.98	0.74 to 1.28
Age at menarche, y				
⩾14	203	164.11	0.81	0.70 to 0.93
12–13	524	505.96	0.97	0.89 to 1.05
<12	214	205.26	0.96	0.84 to 1.10
No. of biopsies, age <50 y				
0	71	64.34	0.91	0.72 to 1.14
1	26	22.70	0.87	0.59 to 1.28
No. of biopsies, age ⩾50 y				
0	609	575.21	0.94	0.87 to 1.02
1	235	213.07	0.91	0.80 to 1.03
No. of affected first-degree relatives, age at first birth <20 y				
0	8	4.30	0.54	0.27 to 1.07
1	2	1.37	0.69	0.17 to 2.74
⩾2	0	0.07	—	—
No. of affected first-degree relatives, age at first birth 20–24 y				
0	374	317.09	0.85	0.77 to 0.94
1	69	79.78	1.16	0.91 to 1.46
⩾2	1	5.92	5.92	0.83 to 42.0
No. of affected first-degree relatives, age at first birth 25–29 y or nulliparous				
0	327	300.75	0.92	0.83 to 1.03
1	62	68.97	1.11	0.87 to 1.43
⩾2	5	5.39	1.08	0.45 to 2.59
No. of affected first-degree relatives, age at first birth ⩾30 y				
0	75	74.98	1.00	0.80 to 1.25
1	17	15.29	0.90	0.56 to 1.44
⩾2	1	1.42	1.42	0.20 to 10.08

Table 6 presents data on the discriminatory accuracy of the Gail et al. model 2 in our cohort. The first column pertains to the total sample. The concordance statistic for the Gail et al. model 2 when used to estimate 5-year risk in the full sample was 0.58 (95% CI = 0.56 to 0.60), indicating that the model performs statistically significantly better than chance (0.5) at discriminating at the individual level between women who will develop disease over a 5-year period and those who will not. This concordance statistic of 0.58 means that there was a 58% probability that a randomly chosen woman in the cohort who developed breast cancer during 5 years of follow-up had a higher estimated risk than a woman who remained disease free. We divided the women into deciles of estimated 5-year risk and calculated the relative risk for actual development of breast cancer over the 5-year period, comparing those women in the top decile of estimated risk with those in the bottom decile. This relative risk was 2.83 (95% CI = 2.19 to 3.65). We computed sensitivity and specificity of the medical indicator cut point of 1.67% 5-year risk, with respect to development of breast cancer. In this cohort, 44% of the 1354 women who developed breast cancer between 1992 and 1997 had a 5-year estimated risk of 1.67% or greater. Of the women who remained free of breast cancer, 66% had an estimated 5-year risk lower than this cut point.

The second column of Table 6 pertains to the subsample of women who reported a screening mammogram in the year before 1992. The concordance statistic, relative risk comparing top to bottom decile of estimated risk, and sensitivity and specificity were all very similar to those found in the full sample.

The recent analysis by Gail et al. *(15)* provides, to our knowledge, the only information currently available on expected risks and benefits associated with chemoprevention of breast cancer with tamoxifen. According to this analysis, among white women with a uterus, the age–risk groups that are estimated to experience a significant ($P<.10$) net gain from chemoprevention with tamoxifen are women 35–49 years of age (regardless of an estimated 5-year risk) and women aged 50–59 years with an estimated 5-year risk greater than or equal to 6.0%. Among white women without a uterus, the following groups are expected to benefit: all women aged 35–49 years (regardless of 5-year risk), women aged 50–59 years with an estimated 5-year risk of 3.0% or greater, and women aged 60–69 years with an estimated 5-year risk of 5.5% or greater. We had data on the women's hysterectomy status (33.7% of the women in our sample in 1992 had no uterus) and thus were able to calculate the proportion of women in this sample who belonged to the above significant net-benefit groups as well as the proportion of the 1354 breast cancer cases that arose from these groups. Only 2.3% of the

Table 4. Ratios of expected (E) and observed (O) number of breast cancer cases in the Nurses' Health Study (NHS) based on 5-year (1992 to 1997) risk prediction of Gail et al. *(1)* model 2 in women with a risk of <1.67% (n = 54 884)

	No. of breast cancer cases observed in NHS over a 5-y follow-up (n = 753)	Expected No. of cases over a 5-y period	E/O ratio	95% confidence interval
Overall	753	650.72	0.86	0.80 to 0.92
Age group, y				
45–49	125	110.97	0.89	0.75 to 1.06
50–54	198	170.35	0.86	0.75 to 0.99
55–59	212	176.98	0.83	0.73 to 0.96
60–64	135	118.51	0.88	0.74 to 1.04
65–69	72	65.29	0.91	0.72 to 1.14
70–74	11	8.62	0.78	0.43 to 1.41
Age at menarche, y				
⩾14	175	129.08	0.74	0.63 to 0.86
12–13	410	379.44	0.93	0.84 to 1.02
<12	168	142.20	0.85	0.73 to 0.98
No. of biopsies, age <50 y				
0	94	86.05	0.92	0.75 to 1.12
1	31	24.92	0.80	0.57 to 1.14
No. of biopsies, age ⩾50 y				
0	519	463.34	0.89	0.82 to 0.97
1	109	76.40	0.70	0.58 to 0.85
No. of affected first-degree relatives, age at first birth <20 y				
0	12	6.53	0.54	0.31 to 0.96
1	0	0.06	—	—
⩾2	0	0	—	—
No. of affected first-degree relatives, age at first birth 20–24 y				
0	499	418.85	0.84	0.77 to 0.92
1	3	1.84	0.61	0.20 to 1.89
⩾2	0	0	—	—
No. of affected first-degree relatives, age at first birth 25–29 y or nulliparous				
0	221	205.03	0.93	0.81 to 1.06
1	1	0.61	0.61	0.09 to 4.36
⩾2	0	0	—	—
No. of affected first-degree relatives, age at first birth ⩾30 y				
0	17	17.76	1.04	0.65 to 1.68
1	0	0.03	—	0.00 to 1.04
>2	0	0	—	—

women in this sample fell into the above significant net-benefit groups, 1.3% into net-benefit groups pertaining to women with a uterus and 1.0% into net-benefit groups pertaining to women without a uterus. A slightly higher proportion of cases in the total sample, 3.3% of the 1354 cases, arose from the significant net-benefit groups (1.8% of the total cases arose from the net-benefit groups pertaining to women with a uterus and 1.5% from the groups pertaining to women without a uterus).

These findings imply that, if only women in our sample who fell into the net-benefit groups had been given chemoprevention with tamoxifen (2.3% of the sample), approximately 1.6%–1.7% of breast cancer cases would have been prevented over a 5-year period [3.3% multiplied by the relative risk of 0.5 observed in the BCPT *(3)*].

Discussion

In this article, we examined both goodness of fit (as reflected in E/O ratios) and the discriminatory accuracy of the Gail et al. model 2 in the NHS cohort.

While the discriminatory accuracy of models of disease risk is rarely examined, it is key to the use of epidemiologic risk equations for clinical risk prediction among individuals. Traditional goodness-of-fit tests do not assess this accuracy. To our knowledge, this is the first analysis of the performance of this model in a general, i.e., not exclusively high-risk, population of U.S. women.

With regard to goodness of fit, we found that the Gail et al. model 2 fit well in all age groups and in virtually all risk factor strata. Ratios of expected to observed numbers of cases deviated strongly from 1.0 only in those strata with very small numbers of cases; thus, these extreme E/O ratios were very unstable. The E/O ratio of 0.94 in the total sample indicated very modest underprediction. The E/O ratio of 1.03 among women with an estimated 5-year risk of 1.67% or greater was identical to that observed by Costantino et al. *(2)* among women with equivalent risk, supporting the belief that results from our sample of nurses are generalizable. The goodness of fit of the model among the subsample of screened women in our cohort was nearly the same as that observed in the total cohort (0.93).

What might account for the modest degree of model underprediction in these data? Because we lacked data on atypical hyperplasia status for the women who reported a history of breast biopsy and information on numbers of biopsies beyond at least one, a small subset of women in our cohort would have been assigned an artificially low 5-year risk estimate from the model. Consequently, the expected numbers of cases would have been artificially low. The estimated relative risks used in

Table 5. Ratios of expected (E) and observed (O) number of breast cancer cases in the Nurses' Health Study (NHS) based on 5-year (1992 to 1997) risk prediction of Gail et al. *(1)* model 2 in women with a risk of ⩾1.67% (n = 27 225)

	No. of breast cancer cases observed in NHS over a 5-y follow-up (n = 601)	Expected No. of cases over a 5-y period	E/O ratio	95% confidence interval
Overall	601	622.70	1.04	0.96 to 1.12
Age group, y				
45–49	17	17.78	1.05	0.65 to 1.68
50–54	37	38.38	1.04	0.75 to 1.43
55–59	83	84.10	1.01	0.82 to 1.26
60–64	156	165.49	1.06	0.91 to 1.24
65–69	241	245.78	1.02	0.90 to 1.16
70–74	67	71.17	1.06	0.84 to 1.35
Age at menarche, y				
⩾14	115	111.76	0.97	0.81 to 1.17
12–13	331	351.08	1.06	0.95 to 1.18
<12	155	159.86	1.03	0.88 to 1.21
No. of biopsies, age <50 y				
0	10	9.89	0.99	0.53 to 1.84
1	7	7.89	1.13	0.54 to 2.37
No. of biopsies, age ⩾50 y				
0	365	389.89	1.07	0.96 to 1.18
1	219	215.03	0.98	0.86 to 1.12
No. of affected first-degree relatives, age at first birth <20 y				
0	0	0.09	—	0.01 to 2.75
1	2	1.71	0.86	0.21 to 3.42
⩾2	0	0.17	—	—
No. of affected first-degree relatives, age at first birth 20–24 y				
0	59	47.70	0.81	0.63 to 1.04
1	91	104.63	1.15	0.94 to 1.41
⩾2	1	8.23	8.23	1.16 to 58.40
No. of affected first-degree relatives, age at first birth 25–29 y or nulliparous				
0	248	240.35	0.97	0.86 to 1.10
1	83	90.29	1.09	0.88 to 1.34
⩾2	5	7.08	1.42	0.59 to 3.40
No. of affected first-degree relatives, age at first birth ⩾30 y				
0	91	99.40	1.09	0.89 to 1.34
1	20	21.43	1.07	0.69 to 1.66
⩾2	1	1.65	1.65	0.23 to 11.68

the Gail et al. model 2 tended to be higher than those observed in the NHS (Table 1); the direction of this discrepancy would have tended to lead to overprediction by the model, not underprediction. However, discrepancies between the two sets of relative risks were modest in the strata containing the large majority of our sample; where discrepancies between relative risks were large, there were few women and few cases in the strata. Thus, it is unlikely that differences in relative risks had a large effect on our overall findings.

Goodness of fit of the Gail et al. model 2 in the recently screened (within 1 year before baseline) subsample of our cohort did not differ from that observed in the total sample: Overall, we observed very modest underprediction (E/O ratio = 0.93), with the CIs of most stratum-specific E/O ratios including 1.0. In their validation of the original model of Gail et al. in the NHS cohort, Spiegelman et al. *(9)* noted that differences in mammographic screening rates may have explained part of the overprediction of the model at younger ages because this model used incidence rates from the regularly screened BCDDP sample, and women in the NHS cohort were likely being screened at lower rates than the BCDDP population. By advancing the time of diagnosis of progressive cases of breast cancer, screening tends to increase rates at low ages, with a compensatory decrease at older ages *(16,17)*. However, in our analysis, the younger age groups tended to show the greatest degree of underprediction, albeit to only a modest degree still. Furthermore, any differences in screening between the BCDDP and the NHS are less relevant with respect to model 2 compared with the original model of Gail et al. Model 2 employs SEER incidence rates, not BCDDP incidence rates. An assertion that that model 2 is appropriate for use only among women undergoing regular screening implies that the magnitudes of the relative risks from the BCDDP were determined by the annual screening of this population. Clearly, relative risks are not invariant across study populations; magnitudes will vary, depending on relative distributions of other risk factors or covariates, for instance. Published analyses from the BCDDP have relied on the assumption that the relative risks obtained in this study are generalizable outside the select group of women volunteering for participation in a large screening study *(18)*.

It seems likely that, with regard to the BCDDP, screening frequency is more closely related to incidence rates than to relative risks for the established risk factors and that the substitution of SEER incidence rates for incidence rates derived from the BCDDP has made model 2 more appropriate than model 1 for widespread use in populations that are not necessarily being screened annually. Clearly, the women who comprise the SEER incidence rates could not be called "a regularly screened popu-

Table 6. Measures of discriminatory accuracy of the Gail et al. *(1)* model 2 in the total sample in the Nurses' Health Study and in a sample of women who reported screening within 1 year before 1992

	Total sample (n = 82 109; 1354 cases)	Recently screened sample* (n = 55 301; 941 cases)
Concordance statistic (95% confidence interval)	0.58 (0.56 to 0.60)	0.59 (0.57 to 0.61)
Relative risk, highest decile of estimated risk compared with lowest decile (95% confidence interval)	2.83 (2.19 to 3.65)	2.89 (2.13 to 3.93)
Sensitivity† at the Food and Drug Administration guideline cut point (5-y risk = 1.67%)	0.44	0.46
Specificity‡ at the Food and Drug Administration guideline cut point (5-y risk = 1.67%)	0.66	0.66

*Sample restricted to women who, in 1992, reported a screening mammogram within the past 1 year (n = 55 301).

†Sensitivity = proportion of all women who developed breast cancer over a 5-year follow-up who had an estimated 5-year risk of ⩾1.67%.

‡Specificity = proportion of all women who remained free of breast cancer over a 5-year follow-up who had an estimated 5-year risk of <1.67%.

lation." Although we do not have information on whether nurses in our cohort are being screened annually, it is likely that this group of health professionals is getting screened at rates higher than those in the general population of women. This may help explain, in part, the slightly greater underprediction at younger ages.

Our results regarding discriminatory accuracy have implications for the use of the Gail et al. model 2 in clinical counseling. Most women who developed breast cancer over the 5-year follow-up had low estimated risks, and the distributions of estimated risks for cases and noncases were nearly indistinguishable. The concordance statistic for the model was better than a pure-chance 0.5 but was still relatively low. The modest discriminatory accuracy of the Gail et al. model 2 is not unexpected nor is it an attribute of this model alone. Given the low relative risks associated with most established breast cancer risk factors, it is unlikely that any breast cancer risk prediction model will have high discriminatory accuracy. It is possible that models incorporating additional predictive variables, such as plasma estrogen levels *(19,20),* mammographic density *(21),* or more complex information on family history of breast and ovarian cancers *(22,23),* may perform somewhat better at individual discrimination.

In clinical counseling using the Gail et al. model 2, many women who are going to develop breast cancer will not be advised to consider chemoprevention with tamoxifen, while, concomitantly, a very large number of women who will remain disease free may be advised to take such action based on the FDA eligibility guidelines. The potential public health consequences of widespread tamoxifen use are implied in the detailed analysis by Gail et al. *(15),* which shows that tamoxifen chemoprevention will cause net public health losses precisely in the large segments of the U.S. female population from which the bulk of breast cancer cases arise, that is, women aged 50–79 years with intact uteri and with estimated breast cancer risk not very different from age-specific average risks. In the NHS sample analyzed here, only 3.3% of breast cancer cases that occurred over a 5-year period arose from women in age–risk strata estimated to have statistically significant health gains from tamoxifen use *(15).* The remaining 96%–97% of the cases arose from women in age–risk strata not expected to have such significant gains (some of these strata are likely to experience significant net health losses, although Gail et al. do not attach statistical significance estimates to the strata with negative health indices). Obviously, our estimate of 3.3% is dependent on the age structure of our sample; if we had had a sample of younger women (e.g., <50 years of age), the estimate of the proportion of cases arising from significant net-benefit groups would have been larger. This 3.3% is not a parameter estimate but rather a sample-dependent quantity, which, in this analysis of a general population of women ranging in age from 45 to 71 years, is surprisingly low.

The tamoxifen risk/benefit analysis of Gail et al. *(15)* is, to date, the only such analysis available on this important topic and, consequently, we have used the findings here. However, Gail et al. point out the limitations of their analysis and the need for further research and modeling in this area. It is unlikely that further research will change two key messages, however, both articulated by Rose *(24).* Rose noted that a preventive strategy designed to meaningfully reduce the population burden of disease will have to be widespread among the population; he also noted that a preventive strategy based on pharmacologic agents will, by necessity, be limited in its use and thus will not have a large impact on population disease burden, since the bulk of disease cases do not arise from the high-risk tail of the risk distribution but rather from the mass of the population with estimated risk right around average. The analysis by Gail et al. *(15)* illustrates the problem of considering chemoprevention of breast cancer with tamoxifen as a meaningful prevention strategy. If such a strategy is limited appropriately to women in groups estimated to experience net benefits, the impact on total disease burden will be small.

The Gail et al. model 2 of breast cancer risk is the first cancer risk prediction model to be widely disseminated to both the public and health-care professionals. To remain relevant for use in the clinical setting, the SEER incidence rates used in the model will need to be updated over time (currently, 1987 rates are used for white women). In addition, better models, with greater discriminatory accuracy, need to be developed if such models are to be used in clinical decision making. Finally, an additional issue regarding use of the model in the clinical setting involves risk communication. The concept of risk is a difficult one to explain to the lay public, many of whom are concerned with answers to dichotomous personal questions such as, "Will I get this disease in the future?" "Should I take this action in order to prevent possible disease?" Women in the United States are particularly anxious about breast cancer, they greatly overestimate their risk of this disease both in the short term and over their lifetime *(25),* and they have a strongly disproportionate fear of this disease relative to other adverse health events *(25).* Estimated 5-year risks of breast cancer from the Gail et al. model 2 will be low, in an absolute sense, for virtually all women. Even a 50-year-old woman with a high 5-year risk of 4% still has a 96% chance of remaining free of breast cancer over the 5-year period, and this should be clearly communicated.

Careful counseling is needed to ensure that such a woman can understand why her risk is high enough that chemoprevention should be considered and yet, at the same time, understand that she is not likely to get breast cancer in the near future, regardless of her decision about chemoprevention. In the clinical setting, where the Gail et al. model 2 is proposed for use as an aid in decision making, most women will be concerned with the question, "Will this drug prevent a case of breast cancer for me, or will it cause something else?" A statistical model cannot address this question at the individual level well.

REFERENCES

(1) Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989; 81:1879–86.

(2) Costantino J, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.

(3) Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 1998;90:1371–88.

(4) National Cancer Institute. SEER Cancer Statistics Review, 1973–1997. http://seer.cancer.gov/publications/CSR 1973–1997.

(5) Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 1988;80:1198–202.

(6) AstraZeneca advertisement. Journal of Women's Health and Gender-Based Medicine. 2000;9: opening pages.

(7) Gail M, Benichou J. Assessing the risk of breast cancer in individuals. In: DeVita VT Jr, Hellman S, Rosenberg SA, editors. Cancer prevention. Philadelphia (PA): Lippincott; 1992. p. 1–15.

(8) Bondy ML, Lustbader ED, Halabi S, Ross E, Vogel VG. Validation of a breast cancer risk assessment model in women with a positive family history. J Natl Cancer Inst 1994;86:620–5.

(9) Spiegelman D, Colditz GA, Hunter D, Hertzmark E. Validation of the Gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst 1994;86:600–7.

(10) Anderson S, Ahnn S, Duff K. NSABP Breast Cancer Prevention Trial risk assessment program. Version 2. NSABP Biostatistical Center Technical Report; 1992.

(11) Colditz GA, Martin P, Stampfer MJ, Willett WC, Sampson L, Rosner B, et al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am J Epidemiol 1986; 123:894–900.

(12) Colditz GA. The Nurses' Health Study: a cohort of US women followed since 1976. J Am Med Women's Assoc 1995;50:40–4.

(13) Jacobs TW, Byrne C, Colditz G, Connolly JL, Schnitt S. Radial scars in benign breast-biopsy specimens and the risk of breast cancer. N Engl J Med 1999;340:430–6.

(14) Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

(15) Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, et al. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 1999;91:1829–46.

(16) Morrison A. Screening in chronic disease. New York (NY): Oxford University Press; 1992.

(17) Gail MH, Benichou J. Validation studies on a model of breast cancer risk. J Natl Cancer Inst 1994;86:573–5.

(18) Brinton LA, Williams RR, Hoover RN, Stegens NL, Feinleib M, Fraumeni JF Jr. Breast cancer risk factors among screening program participants. J Natl Cancer Inst 1979;62:37–44.

(19) Hankinson SE, Willett WC, Manson JE, Colditz GA, Hunter DJ, Spiegelman D, et al. Plasma sex steroid hormone levels and risk of breast cancer in postmenopausal women. J Natl Cancer Inst 1998;90:1292–9.

(20) Toniolo PG, Levitz M, Zeleniuch-Jacquotte A, Banerjee S, Koenig KL, Shore RE, et al. A prospective study of endogenous estrogens and breast cancer in postmenopausal women. J Natl Cancer Inst 1995;87:190–7.

(21) Byrne C. Studying mammographic density: implications for understanding breast cancer. J Natl Cancer Inst 1997;89:531–3.

(22) Claus E, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. Breast Cancer Res Treat 1993;28:115–20.

(23) Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction.

(24) Rose G. The strategy of preventive medicine. New York (NY): Oxford University Press; 1992.

(25) Black WC, Nease RF Jr, Tosteson A. Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. J Natl Cancer Inst 1995;87:720–31.

NOTES

[1]*Editor's note:* SEER is a set of geographically defined, population-based, central cancer registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Registry data are submitted electronically without personal identifiers to the NCI on a biannual basis, and the NCI makes the data available to the public for scientific research.

Manuscript received March 15, 2000; revised December 29, 2000; accepted January 3, 2001.